Notes on Elements of Statistical Learning

Kenneth Zhang

June 18, 2026

Contents

Introduction

I will assume that the reader has a fundamental level of exposure to statistics in order to avoid spending excessive time on elementary concepts and notation. These notes are intended to serve as a rigorous but practical companion to *The Elements of Statistical Learning*, with an emphasis on building intuition alongside mathematical understanding.

The primary goal of these notes is not merely to restate the material from the text, but rather to expand upon derivations, clarify underlying assumptions, connect related concepts, and provide additional commentary where certain arguments may feel condensed or abstract on a first reading. Whenever possible, I will attempt to motivate why a method works, what assumptions are required, and how the method should be interpreted in practice.

I will generally prioritize mathematical precision and conceptual clarity over brevity. Some sections may therefore contain derivations or intermediate steps that are omitted in the original text. At the same time, these notes are not intended to be fully self-contained; familiarity with topics such as linear algebra, probability, calculus, and introductory statistical inference will be assumed throughout.

These notes were written primarily for my own learning and reference, though I hope they may also be useful to others studying statistical learning theory, regression, classification, model selection, regularization, and related topics in modern quantitative research.

1 Mathematical Statistics and Decision Theory

1.1 Normal Distribution

I assume that you are very familiar with the properties of the normal distribution so I will avoid explaining properties of it verbosely here. More specifically, we refer to the Normal distribution as a location-scale distribution specified by it's parameters μ and σ^2 where μ affects the location of the mean of the distribution and σ^2 controls the spread (scale) of the distribution's appearance.

Suppose that instead of simply having a normally distribution random variable, let's call it $X \sim \mathcal{N}(\mu, \sigma^2)$, we apply a linear transformation, namely we have $aX + b$. How should we characterize the distribution of such a normal random variable? Evidently, it's characterization is quite clear from it's moment generating function.

$$M_Y(t) = \mathbb{E}[e^{tY}] = \mathbb{E}[e^{t(aX+b)}] = e^{bt}\mathbb{E}[e^{atX}].$$

Note that $\mathbb{E}[e^{atX}]$ is simply $M_X(at)$ by definition. Expanding,

$$e^{bt}M_X(at) = \exp\left\{\frac{1}{2}a^2t^2\sigma^2 + (a\mu + b)t\right\}.$$

Therefore $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

Evidently, we can observe from the mgf that scaling and shifting a normally distributed random variable consequently scales and shifts the mean by a and b respectively, and scales the original variance by a^2.

Suppose now instead that we have i random variables $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ such that $X_1, \ldots, X_n$ are independently and identically distributed. How do we characterized the scaled sum of such X_i's? Again, such a characterization is simple through the moment generating function. Suppose that $Y = \sum_{i=1}^{n} a_i X_i$.

$$M_Y(t) = \prod_{i=1}^{n} M_{a_iX_i}(t) = \prod_{i=1}^{n} M_{X_i}(a_it) = \prod_{i=1}^{n} \exp\left(\frac{1}{2}a_i^2\sigma_i^2t^2 + a_i\mu_it\right) = \exp\left\{\frac{1}{2}\left(\sum_{i=1}^{n} a_i^2\sigma_i^2\right)t^2 + \left(\sum_{i=1}^{n} a_i\mu_i\right)t\right\}.$$

Clearly, we see that the sum of n independently distributed normal random variables is simply a normal random variable parameterized by the sum of their scaled means and squared-summed variances.

Moment generating functions as you may know are particularly useful for characterizing distributions and how they're tails behave. Suppose again that I have n normally distributed independent random variables $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. What is the distribution of the sample mean of such n random variables?

Easy. We again take $Y = \sum_{i=1}^{n} X_i$.

$$M_Y(t) = \{M_{X_1}(t)\}^n = \exp\left(\frac{1}{2}n\sigma^2t^2 + n\mu t\right).$$

Clearly $Y \sim \mathcal{N}(n\mu, n\sigma^2)$. Recall that,

$$\bar{X} = n^{-1}\sum_{i=1}^{n} X_i = n^{-1}Y.$$

Therefore, $\bar{X}$ is simply a random variable scaled by the number of random variables n. So $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$.

1.2 Chi-squared Distribution

A fascinating result is the chi-squared distribution, denoted χ^2_k where k denotes the number of degrees of freedom. You may recall that if $Z \sim \mathcal{N}(0,1)$ then squaring it makes it $Z^2 \sim \chi^2_1$ i.e., a chi-squared distribution with 1-degree of freedom. More generally, if we have k such standard normal random variables, then the sum of such k random variables is simply a χ^2_k random variable,

$$Z_1, Z_2, \ldots, Z_k \sim \mathcal{N}(0,1) \implies Z_1^2 + \cdots + Z_k^2 \sim \chi^2_k.$$

Intuitively, you are counting how many independent standard normal directions you are squaring and adding. Perhaps this is hard to visualize. Suppose for a second that $Z \sim \mathcal{N}(0,1)$. The normal distribution, as you may know, lives on the entire real line, $-\infty < Z < \infty$ and it symmetric about zero i.e., values like $Z = 1$ and $Z = -1$ are equally likely. But now take that random variable and square it thus giving us only non-negative values $0 \leq Z^2 < \infty$. In this case, $Z = 1$ and $Z = -1$ map to the same value so squaring folds the negative half of the normal distribution onto the positive side.

Before you square, the left tail and right tail extend to either side of 0. After squaring, both tails get folded into the positive direction. So, Z^2 is always positive, heavily concentrated near 0 and has a long right tail.

Perhaps it is more intuitive to view by the density has such a shape. Suppose that $Y = Z^2$. Then for $y > 0$ we have that $Z = \sqrt{y}$ or $Z = -\sqrt{y}$. The probability mass at $Y = y$ comes from two normal points $Z = \sqrt{y}$ and $Z = -\sqrt{y}$. Recall the change-of-variables formula,

$$f_Y(y) = f_Z(\sqrt{y}) \left| \frac{d}{dy}\sqrt{y} \right| + f_Z(-\sqrt{y}) \left| \frac{d}{dy}(-\sqrt{y}) \right|.$$

Since the normal density is symmetric then $f_Z(\sqrt{y}) = f_Z(-\sqrt{y})$ and $\left|\frac{d}{dy}\sqrt{y}\right| = \frac{1}{2\sqrt{y}}$. Therefore,

$$f_Y(y) = 2 f_Z(\sqrt{y}) \frac{1}{2\sqrt{y}} = \frac{f_Z(\sqrt{y})}{\sqrt{y}} = \frac{1}{\sqrt{2\pi}} e^{-y/2}.$$

Notice that this precisely the density of a chi-squared random variable with 1 degree of freedom $Y \sim \chi^2_1$. Evidently, the long right tail exists simply because large positive chi-squared values come from normal values far away from zero. For example $Z^2 > 9$ means that $|Z| > 3$ so the right tail χ^2_1 is really the two tails of the normal distribution combined. A higher-dimensional intuition is to take two $Z_1, Z_2 \sim \mathcal{N}(0,1)$ random variables independently. Then $Z_1^2 + Z_2^2 \sim \chi^2_2$. Geometrically, we can observe (Z_1, Z_2) as a random point in the plane. Then $Z_1^2 + Z_2^2$ is the squared distance of that point from the origin.

So a chi-squared random variable measures squared Gaussian distance from the origin. For k dimensions, χ^2_k is the squared length of a k-dimensional standard normal vector $\chi^2_k = \|Z\|^2$.

It is important to note that the random variable must first be standardized. It is not true to say that any normal random variable gives a standard chi-squared distribution.

1.3 t distribution

Suppose that we have two random variables where Z is the standard normal random variable and $Y \sim \chi^2_n$ (potentially the sum of n independent squared standard normal random variables). We also assume that Z and Y are independent. Then $T = \frac{Z}{\sqrt{Y/n}} \sim t_n$ a Student-t distribution with n degrees of freedom.

You may recall to find the pdf of $T \sim t_n$ we can simply use the one-to-one transformation to get the pdf $f_{T,V}(t,v)$ or alternatively get $f_T(t)$ by integrating $\int f_{T,V}(t,v)dv$. We will not do that here as you may have practiced doing such in a mathematical statistics course.

A Student-t random variable is a standard normal random variable divided by a random estimate of its scale. We have that $Z \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2_n$ where $Y = Z_1^2 + \cdots + Z_n^2$ where Z and Y are independent. For independent standard normals $Z_1, \ldots, Z_n$ the quantity $\frac{Y}{n}$ is the average squared size of n standard normal obsevations. So $\sqrt{Y/n}$ acts like a random standard deviation estimate. Therefore, T can be interpreted as normal noise scaled by random scale estimate.

In the case that $n = 1$ and is fixed then $T = \frac{Z}{1} = Z \sim \mathcal{N}(0, 1)$ so the only difference between the standard normal and the Student-t distribution is that the Student-t uses a random denominator. In the case that the denominator is close to 1, T behaves very similarly to a standard normal.

In the case that the denominator is smaller than 1, dividing by it magnifies Z which creates unusually large positive and negative values causing heavier tails than the normal distribution. The degrees of freedom n control how random the denominator is. When n is small, Y/n is a noisy estimate of 1. It can often be noticeably smaller or larger than 1 therefore making $\sqrt{Y/n}$ fluctuate a lot, so the ratio $\frac{Z}{\sqrt{Y/n}}$ has much heavier tails. So for small n, the t_n distribution is more spread out than the standard normal.

In fact, t_1 is the Cauchy distribution. When n is large, by the law of large numbers,

$$\frac{Y}{n} = \frac{Z_1^2 + \cdots + Z_n^2}{n} \approx \mathbb{E}[Z_n^2] = 1.$$

This means that $\sqrt{Y/n} \approx 1$ so $T \approx Z$. As $n \to \infty$ then $t_n \to \mathcal{N}(0, 1)$. This becomes quite significant in future chapters when we construct estimates of sample variance. As the variance esimate becomes more stable, the Student-t becomes normal.

1.4 F distribution

Suppose that $X \sim \chi_n^2$ and $Y \sim \chi_m^2$ where X and Y are independent random variables. Then,

$$F = \frac{X/n}{Y/m} \sim F_{n,m}.$$

Clearly, an F random variable is simply the ratio of two independent average squared standard-normal quantities. Because $X \sim \chi_n^2$ means that $X = Z_1^2 + \cdots + Z_n^2$ so,

$$\frac{X}{n} = \frac{Z_1^2 + \cdots + Z_n^2}{n},$$

which is simply the average squared size of n independent standard normal variables. Similarly,

$$\frac{Y}{m} = \frac{W_1^2 + \cdots + W_m^2}{m}.$$

So F is intuitively the average squared size from group 1 scaled by the average squared size from group 2. In contrast to a Student-t distribution that compares a normal quantity to a random standard deviation estimate, the F distribution instead compares two random variance estimates.

Visually, the F distribution is always non-negative because both X, Y are chi-squared random variables. Unlike the normal and Student-t, the F distribution does not live on the whole real line but rather on $(0, \infty)$. Rather than being symmetric, it is usually right-skewed with a long right tail. The reason is quite trivial. The denominator Y/m is random so if the denominator is unusually small then $F = \frac{X/n}{Y/m}$ can become very large.

The degrees of freedom n and m control how noisy the numerator and denominator are with n controlling the variability of X/n and m controlling the variability of Y/m. Intuitively, the F distribution is centered around the idea of comparing two quantities that should both be around 1. If the two variance-like quantities are similar, then $F \approx 1$. If the numerator variance estimate is much larger than the denominator variance estimate, then $F > 1$, otherwise, if the variance estimate is much smaller than the denominator then $0 < F < 1$.

The most important statistical intuition is that the F distribution appears when comparing variances or explained variation to unexplained variation. Suppose that we are studying $\frac{S_1^2}{S_2^2}$ (i.e., the sample variance estimates of two samples). If both samples from from normal populations with the same variance, then after scaling properly, the ration follows an F distribution. So, a large F statistic means that the explained variation is large relative to the unexplained noise.

Suppose that $T \sim t_m$. What is the distribution of T^2?

Proof. Recall that a T distribution is a standard normal random variable scaled by the average of m chi-squared random variables. Taking the definition and computing we trivially get,

$$T^2 = \frac{Z^2}{(\sqrt{Y/m})^2} = \frac{Z^2}{Y/m} = \frac{Z^2/1}{Y/m} \sim F_{1,m}.$$

Therefore, a squared Student-t distribution is simply a $F_{1,m}$ distribution. □

1.5 Inference and Estimation

Assume throughout that the random variables $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ i.e., they are sampled from independently and identically from a Normal distribution. Recall the classic results from normal-sample distributions,

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \quad \frac{\sqrt{n}(\bar{X}_n - mu)}{\sigma} \sim \mathcal{N}(0,1), \quad \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}.$$

We will use the following section to review some key sample metrics used in inference.

1.5.1 Sample Mean

Recall that the sample mean is defined mathematically as,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Since each X_i is a random variable, then the sample mean is also a random variable. Therefore, it's value depends on the random sample. First, we can compute its expectation,

$$\mathbb{E}[\bar{X}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}\cdot n\mu = \mu.$$

So $\bar{X}_n$ is centered at μ. Recall that since X_i's are independent then $\mathrm{Cov}(X_i, X_j) = 0$ for any $i \neq j$.

$$\mathrm{Var}(\overline{X}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}.$$

Because $\bar{X}_n$ is a linear combination of independent normal random variables, it is also normal. Therefore $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$. Using this result, we can see that the standard deviation of $\bar{X}_n$ is,

$$\sqrt{\mathrm{Var}(\bar{X}_n)} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

If we standardize $\bar{X}_n$, we subtract is mean and divide by its standard deviation,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim \mathcal{N}(0,1).$$

1.5.2 Sample Variance

The sample variance is,

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

The denominator is $n-1$ rather than n because we first estimate μ by $\bar{X}_n$. Once we fix $\bar{X}_n$ the deviations $X_1 - \bar{X}_n, \ldots, X_n - \bar{X}_n$ must satisfy $\sum_{i=1}^{n}(X_i - \bar{X}_n) = 0$. So, there are only $n-1$ freely varying deviations.

Recall from STAT 231 that the total variation around μ is simply the sum of the variation around $\bar{X}_n$ and the variation of $\bar{X}_n$ around μ which gives rise to the following identity,

$$\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}(X_i-\bar{X}_n)^2+n(\bar{X}_n-\mu)^2.$$

Proof. We know that $X_i-\mu=(X_i-\bar{X}_n)+(\bar{X}_n-\mu)$ so squaring both sides gives,

$$\begin{aligned}(X_i-\mu)^2&=[(X_i-\bar{X}_n)+(\bar{X}_n-\mu)]^2\\&=(X_i-\bar{X}_n)^2+2(X_i-\bar{X}_n)(\bar{X}_n-\mu)+(\bar{X}_n-\mu)^2.\end{aligned}$$

Summing over $i=1,\ldots,n$ gives,

$$\sum_{i=1}^{2}(X_i-\mu)^2=\sum_{i=1}^{n}(X_i-\bar{X}_n)^2+2(\bar{X}_n-\mu)\sum_{i=1}^{n}(X_i-\bar{X}_n)+\sum_{i=1}^{n}(\bar{X}_n-\mu)^2.$$

However, notice how,

$$\sum_{i=1}^{n}(X_i-\bar{X}_n)=\sum_{i=1}^{n}X_i-\sum_{i=1}^{n}\bar{X}_n=n\bar{X}_n-n\bar{X}_n=0,\quad\sum_{i=1}^{n}(\bar{X}_n-\mu)^2=n(\bar{X}_n-\mu)^2.$$

Clearly we get that now,

$$\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}(X_i-\bar{X}_n)^2+n(\bar{X}_n-\mu)^2.$$

Dividing by σ^2 we have that,

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2=\frac{\sum_{i=1}^{n}(X_i-\bar{X}_n)^2}{\sigma^2}+\frac{n(\bar{X}_n-\mu)^2}{\sigma^2}.$$

Re-arranging the definition of S_n^2 we get that,

$$\sum_{i=1}^{n}(X_i-\bar{X}_n)^2=(n-1)S_n^2.$$

Therefore,

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2=\frac{(n-1)S_n^2}{\sigma^2}+\left(\frac{\sqrt{n}(\overline{X}_n-\mu)}{\sigma}\right)^2.$$

□

You may recall the theorem,

$$\frac{(n-1)S_n^2}{\sigma^2}\sim\chi^2_{n-1}.$$

Suppose we have data points X_i and standardize each X_i which is gives us Z_i's that are standard normal random variables. Then re-solving for X_i we get $X_i=\mu+\sigma Z_i$. Then,

$$\bar{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i=\frac{1}{n}\sum_{i=1}^{n}(\mu+\sigma Z_i)=\frac{1}{n}\left(n\mu+\sigma\sum_{i=1}^{n}Z_i\right)=\mu+\sigma\bar{Z}_n.$$

Trivially, we then have that $X_i-\bar{X}_n=(\mu+\sigma Z_i)-(\mu+\sigma\bar{Z}_n)=\sigma(Z_i-\bar{Z}_n)$. So $(X_i-\bar{X}_n)^2=\sigma^2(Z_i-\bar{Z}_n)^2$.

Summing the identities from $i = 1, \ldots, n$ we get that,

$$\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \sigma^2 \sum_{i=1}^{n}(Z_i - \bar{Z}_n)^2 \implies \frac{(n-1)S_n^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{\sigma^2} = \sum_{i=1}^{n}(Z_i - \bar{Z}_n)^2.$$

It remains for us to make sense of $\sum_{i=1}^{n}(Z_i - \bar{Z}_n)^2$ i.e., the squared length of the residual vector after removing the average direction. Suppose that $Z = [Z_1 \quad \cdots \quad Z_n]^\top$ is a vector standard normal independent random variables. Since the entries are independent standard normals then $Z \sim \mathcal{N}_n(0, I_n)$. The sample mean direction is the vector $u_1 = \frac{1}{\sqrt{n}}[1 \cdots 1]^\top$. Projecting such Z onto this direction gives $u_1^\top Z = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i = \sqrt{n}\bar{Z}_n$. Extending u_1 to an orthonormal basis of $\mathbb{R}^n$ to get $u_1, u_2, \ldots, u_n$ we can define $W_j = u_j^\top Z$ for $j = 1, \ldots, n$. Because $Z \sim \mathcal{N}_n(0, I_n)$ and the transformation is orthogonal, $W_1, \ldots, W_n \sim \mathcal{N}(0, 1)$.

The first coordinate is $W_1 = u_1^\top Z = \sqrt{n}\bar{Z}_n$ and the remaining coordinates measure variation orthogonal to the mean direction. Note these are residual directions. By preservation of squared length under orthogonal transformations, observe that $\sum_{i=1}^{n} Z_i^2 = \sum_{j=1}^{n} W_j^2$. Similarly, $n\bar{Z}_n^2 = (\sqrt{n}\bar{Z}_n)^2 = W_1^2$.

Using the decomposition identity we see that,

$$\sum_{i=1}^{n}(Z_i - \bar{Z}_n)^2 = \sum_{i=1} Z_i^2 - n\bar{Z}_n^2 = \sum_{j=1} W_j^2 - W_1^2 = \sum_{j=2}^{n} W_j^2.$$

Since $W_2, \ldots, W_n \sim \mathcal{N}(0, 1)$ we get that $\sum_{j=2}^{n} W_j^2 \sim \chi^2_{n-1}$.

Hence we have that,

$$\frac{(n-1)S_n^2}{\sigma^2} = \sum_{i=1}^{n}(Z_i - \bar{Z}_n)^2 \sim \chi^2_{n-1}.$$

The sample variance has $n - 1$ degrees of freedom because one degree of freedom was used to estimate the mean. It is also apparent from this derivation that, from the orthogonal decomposition above, $W_1 = \sqrt{n}\bar{Z}_n$ controls the sample mean because $\bar{X}_n = \mu + \sigma\bar{Z}_n = \mu + \frac{\sigma}{\sqrt{n}}W_1$. Clearly, $\bar{X}_n$ is a function of W_1.

Meanwhile, $\frac{(n-1)S_n^2}{\sigma^2} = \sum_{j=2}^{n} W_j^2$ so S_n^2 is a function of $W_2, \ldots, W_n$ but $W_1, \ldots, W_n$ are independent standard normal random variables. Therefore W_1 is independent of $(W_2, \ldots, W_n)$ and hence any function of W_1 is independent of any function of $(W_2, \ldots, W_n)$. Therefore $\bar{X}_n \perp S_n^2$. Note this is special to the normal distribution.

1.5.3 Unbiasedness of S_n^2 for σ^2

From the result $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$ and the fact that $\mathbb{E}[\chi_k^2] = k$ we have that $\mathbb{E}\left[\frac{(n-1)S_n^2}{\sigma^2}\right] = n - 1$. Then,

$$\mathbb{E}\left[\frac{(n-1)S_n^2}{\sigma^2}\right] = n - 1 \implies \frac{n-1}{\sigma^2}\mathbb{E}[S_n^2] = n - 1.$$

Dividing both sides by $n - 1$ we have that,

$$\frac{1}{\sigma^2}\mathbb{E}[S_n^2] = 1.$$

So therefore $\mathbb{E}[S_n^2] = \sigma^2$ and is therefore an unbiased estimator of the population variance.

Unbiasedness is also apparent through algebra. Recall that,

$$\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X}_n - \mu)^2.$$

Taking expectations,

$$\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right] = \mathbb{E}\left[\sum_{i=1}^{n}(X_i - \mu)^2\right] - \mathbb{E}\left[n(\bar{X}_n - \mu)^2\right].$$

The first term on the right is simply $n\sigma^2$ since $\mathbb{E}[(X_i - \mu)^2] = \sigma^2$. Also since $\mathbb{E}[(\bar{X}_n - \mu)^2] = \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$. Thus we have that $\mathbb{E}[n(\bar{X}_n - \mu)^2] = n \cdot \frac{\sigma^2}{n} = \sigma^2$. Consequently the expectation on the left-hand side is simple just $n\sigma^2 - \sigma^2 = (n-1)\sigma^2$. By the definition of S_n^2, $\mathbb{E}[S_n^2] = \sigma^2$ as we saw previously. So again, $\mathbb{E}[S_n^2] = \sigma^2$.

1.5.4 Student-t from Results of Normality

From above we know that $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim \mathcal{N}(0,1)$ and $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$. These are two independent random variables following from orthogonality of $\bar{X}_n$ and S_n^2. Suppose we let $Z = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim \mathcal{N}(0,1)$ and $Y = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$. Then clearly $Z \perp Y$. By definition of the Student-t distribution,

$$\frac{Z}{\sqrt{Y/(n-1)}} \sim t_{n-1} \iff$$

Substituting and simplifying the denominator we have that,

$$\frac{Z}{\sqrt{Y/(n-1)}} \sim t_{n-1} \implies \frac{\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}}{\sqrt{\frac{1}{n-1}\frac{(n-1)S_n^2}{\sigma^2}}} \sim t_{n-1}$$

$$\implies \frac{\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}}{S_n/\sigma} \sim t_{n-1} \implies \frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \sim t_{n-1}.$$

This is the usual one-sample t-statistic. If σ were known, we would use $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim \mathcal{N}(0,1)$. But if σ were unknown, we replace it with S_n and the extra randomness in the denominator changes the distribution from standard normal to Student-t. From these results, we get the following distributional results.

1.6 One-sample t-statistic

We know from normal sample results that,

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \implies \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \iff \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim \mathcal{N}(0,1).$$

Suppose that we let $Z = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$. Recall the sample variance result $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$. We will let Y represent $\frac{(n-1)S_n^2}{\sigma^2}$. Then $Y \sim \chi^2_{n-1}$. These two results are pretty trivial. From a normal sample we also know that $\bar{X}_n \perp S_n^2$ which means that $Z \perp Y$. Recall the t-statistic defined as $T = \frac{Z}{\sqrt{Y/k}} \sim t_k$ where $Z \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2_k$. Here, $k = n - 1$ which means that $T \sim t_{n-1}$. Substituting definitions for Z and Y we get,

$$\frac{\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}}{\sqrt{\frac{1}{n-1} \cdot \frac{(n-1)S_n^2}{\sigma^2}}} = \frac{\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}}{S_n/\sigma} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \cdot \frac{\sigma}{S_n} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1}.$$

$\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}$ would be a standard normal if σ were know, however, since σ is often unknown, we replace it with the estimate S_n. The uncertainty in the denominator changes the distribution from normal to a Student-t. Under the null hypothesis $H_0 : \mu = \mu_0$ we replace μ by the hypothesized value μ_0 so that we get the one-sample t-test,

$$T = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \sim t_{n-1} \quad \text{under } H_0.$$

We may interpret the component,

$$\frac{S_n}{\sigma} = \sqrt{\frac{S_n^2}{\sigma^2}} = \sqrt{\frac{1}{n-1} \cdot \frac{(n-1)S_n^2}{\sigma^2}} \sim \sqrt{\frac{\chi^2_{n-1}}{n-1}}.$$

From the above, we see that the denominator of the t-statistic is a random estimate of scale. When n is large, $\frac{\chi^2_{n-1}}{n-1} \approx 1$ so $\frac{S_n}{\sigma} \approx 1$. Therefore for large n, $\frac{\bar{X}_n-\mu}{S_n/\sqrt{n}}$ is close to $\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$. This is consistent with $t_{n-1} \to \mathcal{N}(0,1)$ as $n \to \infty$. What about for two samples?

1.7 Two-sample Variance Ratio and F-distribution

Suppose we have two independent normal random samples $X_1, \ldots, X_n \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Assume that the two samples are independent. We define their respective samples means as $\bar{X}_n$ and $\bar{Y}_m$ and respective sample variances as S_1^2 and S_2^2. From the normal sample variance result,

$$\frac{(n-1)S_1^2}{\sigma_1^2} \sim \chi^2_{n-1}, \quad \frac{(m-1)S_2^2}{\sigma_2^2} \sim \chi^2_{m-1}.$$

Since the X and Y samples are independent, any statistic computed from the X sample is independent of any statistic computed from the Y statistic. Hence $\frac{(n-1)S_1^2}{\sigma_1^2} \perp \frac{(n-1)S_2^2}{\sigma_2^2}$. Let U and V be the two statistics respectively, meaning $U \sim \chi^2_{n-1}$ and $V \sim \chi^2_{m-1}$ respectively. Of course, $U \perp V$. Then, by definition of the F distribution,

$$\frac{U/(n-1)}{V/(m-1)} = \frac{\frac{1}{n-1} \cdot \frac{(n-1)S_1^2}{\sigma_1^2}}{\frac{1}{m-1} \cdot \frac{(m-1)S_2^2}{\sigma_2^2}} = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \sim F_{n-1,m-1}.$$

The null hypothesis for comparing two variances is usually $H_0 : \sigma_1^2 = \sigma_2^2$. Under this null hypothesis, $\sigma_1^2 = \sigma_2^2$ so $\sigma_2^2/\sigma_1^2 = 1$ which means that $\frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} = \frac{S_1^2}{S_2^2}$. Hence under $H_0 : \sigma_1^2 = \sigma_2^2$ we have that $\frac{S_1^2}{S_2^2} \sim F_{n-1,m-1}$.

This is the classical two-sample F-test statistic for equality of variances. A more precise way to write $H_0 : \sigma_1^2 = \sigma_2^2$ is $\frac{S_1^2}{S_2^2} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n-1,m-1}$. The null is not really $H_0 = \frac{S_1^2}{S_2^2}$ but rather about the population variances with statistic S_1^2/S_2^2 rather than $H_0 : \sigma_1^2 = \sigma_2^2$.

1.8 Fundamental Statistical Decision Theory

Suppose that $X \in \mathbb{R}^p$ is a random input vector and $Y \in \mathbb{R}$ is a random output. The pair (X, Y) has some joint distribution $\mathbb{P}_{X,Y}$. Then, we define a *prediction rule* as a function $f : \mathbb{R}^p \to \mathbb{R}$ such that after observing $X = x$ we can predict $f(x)$. To decide whether f is good or bad, we need a loss function.

A *loss function* measures how costly it is to predict $f(X)$ when the true response is Y. For regression, the most common choice is squared error loss defined as,

$$L(Y, f(X)) = (Y - f(X))^2.$$

The expected prediction error, also referred to in bayesian statistics as risk, is,

$$\text{EPE}(f) = \mathbb{E}[(Y - f(X))^2].$$

If the joint distributon has a density or probability measure notation, we may write the expected prediction error as,

$$\text{EPE}(f) = \int (y - f(x))^2 \mathbb{P}(dx, dy).$$

This means that we average the squared prediction error over all possible values of X and Y according to their true joint distribution. The goal of decision theory is to find the function f that minimizes the expected loss,

$$f^* = \arg\min_f \mathbb{E}[(Y - f(X))^2],$$

where the function f^* is called the Bayes rule for the chosen loss function. Bayes here does not necessarily imply Bayesian inference but rather means the theoretically optimal decision rule under the true data-generating distribution.

1.8.1 Conditioning and Pointwise Minimization

The key move is to condition on X. By the law of iterated expectation,

$$\mathbb{E}[(Y - f(X))^2] = \mathbb{E}_X[\mathbb{E}[(Y - f(X))^2 \mid X]].$$

So, our expected prediction error is simply,

$$\text{EPE}(f) = \mathbb{E}_X[\mathbb{E}[(Y - f(X))^2 \mid X]].$$

Observe what happens when $X = x$. Once $X = x$ is fixed, the prediction $f(X)$ because the number $f(x)$. Therefore the inner conditional expectation becomes,

$$\mathbb{E}[(Y - f(X))^2 \mid X = x] = \mathbb{E}[(Y - f(x))^2 \mid X = x].$$

If we let $c = f(x)$ then at the point x we only need to minimize $\mathbb{E}[(Y - c)^2 \mid X = x]$ over constants c. therefore the global function optimization problem reduces to the pointwise problem $f^*(x) = \arg\min_c \mathbb{E}[(Y - c)^2 \mid X = x]$. This is the most important conceptual step. Since $f(x)$ only affects the prediction mae at input value x, the optimal function can be chosen separately at each x.

1.8.2 Squared loss and Conditional Mean

We now aim to solve,

$$f^*(x) = \arg\min_c \mathbb{E}[(Y - c)^2 \mid X = x].$$

Suppose we let $m(x) = \mathbb{E}[Y \mid X = x]$ be the conditional mean of Y on $X = x$. Expanding the conditional squared error and using $m(x)$, then differentiating with respect to c, we can compute the critical points.

$$\begin{aligned}
\arg\min_{c \in \mathbb{R}} \mathbb{E}\Big[(Y - c)^2 \mid X = x\Big] &= \arg\min_{c \in \mathbb{R}} \mathbb{E}\Big[Y^2 - 2cY + c^2 \mid X = x\Big] \\
&= \arg\min_{c \in \mathbb{R}} \Big\{\mathbb{E}\Big[Y^2 \mid X = x\Big] - 2c\mathbb{E}[Y \mid X = x] + c^2\Big\} \\
&= \arg\min_{c \in \mathbb{R}} \Big\{\mathbb{E}\Big[Y^2 \mid X = x\Big] - 2cm(x) + c^2\Big\}.
\end{aligned}$$

Solving the first derivative we get,

$$0 = \frac{d}{dc}\Big(\mathbb{E}\Big[Y^2 \mid X = x\Big] - 2cm(x) + c^2\Big) = -2m(x) + 2c \Longrightarrow c = m(x),$$

and the second derivative,

$$\frac{d^2}{dc^2}\Big(\mathbb{E}\Big[Y^2 \mid X = x\Big] - 2cm(x) + c^2\Big) = 2 > 0,$$

thus meaning that this is a critical point of which is a minimum. Therefore we have that $f^*(x) = \mathbb{E}[Y \mid X = x]$. Thus, under squared error loss, the best prediction of Y given $X = x$ is the conditional expectation of Y given $X = x$. This function $f^*(x) = \mathbb{E}[Y \mid X = x]$ which is also referred to as the regression function.

A separate derivation adds and subtracts $m(x)$,

$$\begin{aligned}\mathbb{E}\Big[(Y-c)^2 \mid X=x\Big] &= \mathbb{E}\Big[((Y-m(x))+(m(x)-c))^2 \mid X=x\Big] \\ &= \mathbb{E}\Big[(Y-m(x))^2 \mid X=x\Big] \\ &\quad + 2(m(x)-c)\mathbb{E}[Y-m(x) \mid X=x] \\ &\quad + (m(x)-c)^2 \\ &= \mathbb{E}\Big[(Y-m(x))^2 \mid X=x\Big] + (m(x)-c)^2 \\ &= \mathrm{Var}(Y \mid X=x) + (m(x)-c)^2.\end{aligned}$$

So clearly this is minimized as $c = m(x)$ and therefore we have that,

$$\mathbb{E}[Y-m(x) \mid X=x] = \mathbb{E}[Y \mid X=x] - m(x) = 0.$$

So again, we have that,

$$\arg\min_{c\in\mathbb{R}} \mathbb{E}\Big[(Y-c)^2 \mid X=x\Big] = \arg\min_{c\in\mathbb{R}} (m(x)-c)^2 = m(x).$$

Therefore we conclude that $f^*(x) = \mathbb{E}[Y \mid X=x]$. This decomposition also shows that even the best possible prediction may have non-zero error. The unavoidable component is $\mathrm{Var}(Y \mid X=x)$ which is the noise remaining in Y after knowing $X=x$.

1.8.3 Estimation from Data

The previous solution is theoretical because it assumes that we know the true conditional distribution of Y given $X=x$. In practice we only have training data $\mathcal{T} = \{(x_1,y_1),\ldots,(x_N,y_N)\}$. The target is still $f^*(x) = \mathbb{E}[Y \mid X=x]$ but we must estimate it from the data. A direct idea is to average all observed y_i's with x_i close to x. In the nearest-neighbor form, one might estimate $\hat{f}(x) = \dfrac{1}{k}\displaystyle\sum_{i:x_i\in\mathcal{N}_k(x)} y_i$ where $N_k(x)$ is the set of the k training poits closest to x. This is trying to approximate $\mathbb{E}[Y \mid X=x]$ by a local sample average. First, the expectation is replaced by an empirical average. Second, conditioning exactly on $X=x$ is replaced by conditioning on X being near x because in continuous spaces there is usually no training observation with exactly $x_i = x$. The ideal condition is that as $N\to\infty$, the neighbourhood around x becomes small, while the number of observations inside it becomes large. This is why consistency results often require that $k\to\infty$ and that $k/N\to 0$. The condition $k\to\infty$ ensures that the local average remains stable. The condition that $k/N\to 0$ forces the neighbourhood to shrink so that the local average is genuinely local.

The difficulty is that in high dimensions, neighbourhoods become large very quickly. This is one form of the curse of dimensionality. To capture a fixed fraction of the data in $\mathbb{R}^p$, the radius of the neighborhood must often be much large when p is large. Therefore, local averaging becomes less local, and the estimate can become biased because it averages over points that are far away from x.

1.8.4 Reducing the Decision Problem

Instead of estimating the full conditional expectation $\mathbb{E}[Y \mid X=x]$ non-parametrically, we may restrict the class of functions we are willing to consider. For example, suppose we approximate the regression function by a linear function $f(x) = x^\top\beta$. Now, the decision is no longer an arbitrary function f. The decision is the parameter vector β. The expected prediction error becomes $\mathrm{EPE}(\beta) = \mathbb{E}[(Y-X^\top\beta)^2]$.

We can derive the optimal population least-squares coefficient.

$$\begin{aligned}\mathrm{EPE}(\beta) &= \mathbb{E}\Big[(Y-X^\top\beta)^2\Big] \\ &= \mathbb{E}\Big[Y^2 - 2YX^\top\beta + \beta^\top XX^\top\beta\Big] \\ &= \mathbb{E}[Y^2] - 2\mathbb{E}[YX^\top]\beta + \beta^\top\mathbb{E}[XX^\top]\beta \\ &= \mathbb{E}[Y^2] - 2\beta^\top\mathbb{E}[XY] + \beta^\top\mathbb{E}[XX^\top]\beta.\end{aligned}$$

Taking the gradient w.r.t. β,

$$\begin{aligned}\nabla_\beta \operatorname{EPE}(\beta) &= -2\mathbb{E}[XY] + 2\mathbb{E}[XX^\top]\beta \\ 0 &= -2\mathbb{E}[XY] + 2\mathbb{E}[XX^\top]\beta.\end{aligned}$$

Finally, if $\mathbb{E}[XX^\top]$ is invertible, we get that,

$$\mathbb{E}[XX^\top]\beta = \mathbb{E}[XY] \implies \beta^* = \left(\mathbb{E}[XX^\top]\right)^{-1}\mathbb{E}[XY].$$

This is the population least-squares solution. The important interpretation is that least squares does not condition at each point x. Instead, it uses global structural assumption, that is $f(x) \approx x^\top\beta$, to pool information across many values of X. That pooling can reduce variance dramatically, but introduces bias if the true regression function is not close to linear. The sample least-squares estimator is obtained by replacing population expectations with sample averages. If X is the $N \times p$ design matrix and y is the response vector then $\hat{\beta} = (X^\top X)^{-1}X^\top y$. This is essentially the empirical version of $\beta^* = [\mathbb{E}[XX^\top]]^{-1}\mathbb{E}[XY]$.

1.8.5 Loss functions and Optimality

Squared error loss gives the conditional mean but not universally. The optimal prediction depends on the loss functions. Consider the absolute error loss, $L(Y, f(X)) = |Y - f(X)|$. Minimizing this pointwise gives $f^*(x) = \arg\min_c \mathbb{E}[|Y - c| \mid X = x]$. Suppose that $F_x(c)$ is the conditional distribution function Y given $X = x$ and $F_x(c) = \mathbb{P}(Y \leq c \mid X = x)$. Define $\phi(c) = \mathbb{E}[\mid Y - c \mid\mid X = x]$.

For continuous conditional distributions, we can differentiate,

$$|Y - c| = \begin{cases} c - Y, & Y < c, \\ Y - c, & Y > c \end{cases}.$$

From the above, we get that,

$$\phi(c) = \int_{-\infty}^{c} (c - y)\, dF_X(y) + \int_{c}^{\infty} (y - c)\, dF_X(y).$$

Differentiating this with respect to c, we see that for the first integral, increasing c increases the integrand at rate 1 while for the second integral, increasing c decreases the integrand at rate 1. Therefore,

$$\phi'(c) = \mathbb{P}(Y < c \mid X = x) - \mathbb{P}(Y > c \mid X = x).$$

For continuous distributions, $\mathbb{P}(Y < c \mid X = x) = F_x(c)$ and $\mathbb{P}(Y > c \mid X = x) = 1 - F_x(c)$. So $\phi'(c) = F_x(c) - [1 - F_x(c)] = 2F_x(c) - 1$. Setting this equal to zero gives us $2F_x(c) - 1 = 0$ so evidently we have that $F_x(c) = \frac{1}{2}$. Therefore c is the conditional median. Simply put, the decision theoretic implication is that squared loss gives a conditional mean while absolute loss gives a conditional median.

1.8.6 Classification as Decision Theory

Now suppose the output is categorical rather than real-valued, that is let $G \in \mathcal{G} = \{g_1, \ldots, g_K\}$. A classifier is a function that maps a value in the real-valued feature vector $\mathbb{R}^p$ to a category in the group set $\mathcal{G}$, that is $\hat{G} : \mathbb{R}^p \to \mathcal{G}$. So after observing $X = x$, the classifier chooses a class $\hat{G}(x) \in \mathcal{G}$.

As for any statistical learning model, we require a loss function. For classification, a general loss function can be represented by a matrix, that is $L(g_k, g_l)$ where $L(g_k, g_l)$ is the loss incurred when the true class is g_k and we predict g_l. The expected prediction error of such a classifier is therefore simply just $\operatorname{EPE}(\hat{G}) = \mathbb{E}\left[L(G, \hat{G}(X))\right]$. Conditoning on X, we get that $\operatorname{EPE}(\hat{G}) = \mathbb{E}_X\left[\mathbb{E}\left[L(G, \hat{G}(X)) \mid X\right]\right]$. At a fixed point $X = x$, if we choose the class g, the conditional risk is $R(g \mid x) = \sum_{k=1}^{K} L(g_k, g)\mathbb{P}(G = g_k \mid X = x)$.

Therefore the optimal classifier chooses,

$$\hat{G}^*(x) = \arg\min_{g \in \mathcal{G}} \sum_{k=1}^{K} L(g_k, g) \mathbb{P}(G = g_k \mid X = x).$$

This is the general Bayes classifier under an arbitrary loss matrix.

1.8.7 Zero-one loss and the Bayes Classifier

The most common classification loss is zero-one loss i.e., $L(g_k, g)$ equals 0 if $g = g_k$ and equals 1 if otherwise. Equivalently, this sort of loss can be expressed as a indicator function $\mathbf{1}\{g_k \neq g\}$. For fixed x, suppose we predict class g. The conditional risk is simply,

$$R(g \mid x) = \sum_{k=1}^{K} \mathbf{1}\{g_k \neq g\} \mathbb{P}(G = g_k \mid X = x).$$

The only term ommited from the sum is the probability of the class we predicted correctly. Therefore,

$$R(g \mid x) = \sum_{k : g_k \neq g} \mathbb{P}(G = g_k \mid X = x) = 1 - \mathbb{P}(G = g \mid X = x).$$

Thus,

$$\hat{G}^*(x) = \arg\min_{g \in \mathcal{G}} [1 - \mathbb{P}(G = g \mid X = x)] \iff \hat{G}^*(x) = \arg\max_{g \in \mathcal{G}} \mathbb{P}(G = g \mid X = x).$$

The above follows because minimizing $1 - p$ is the same as maximizing p. So under zero-one loss the optimal classifier assigns x to the most probable conditional class. This is also known as the Bayes classifier. The corresponding minimum possible error rate is called the Bayes error rate or Bayes rate. Under zero-one loss, the conditional probability of making an error at x is simply $1 - \max_{g \in \mathcal{G}} \mathbb{P}(G = g \mid X = x)$.

Averaging over X, the Bayes error rate is,

$$\mathbb{E}_X \left[1 - \max_{g \in \mathcal{G}} \mathbb{P}(G = g \mid X) \right].$$

This is the lowest achievable classification error under the true distribution.

1.8.8 Regression view of Classification

Classification is synonymous with regression using dummy variables. For a K class problem, it is possible for us to define $Y_k = 1\{G = g_k\}$. Then we have that Y_k is an indicator random variable on $G = g_k$.

Computing the conditional expectation,

$$\begin{aligned} \mathbb{E}[Y_k \mid X = x] &= 1 \cdot \mathbb{P}(Y_k = 1 \mid X = x) + 0 \cdot \mathbb{P}(Y_k = 0 \mid X = x) \\ &= \mathbb{P}(Y_k = 1 \mid X = x) \\ &= \mathbb{P}(G = g_k \mid X = x). \end{aligned}$$

Therefore $\mathbb{E}[Y_k \mid X = x] = \mathbb{P}(G = g_k \mid X = x)$. So estimating class probabilities is equivalent to estimating conditional expectations of indicator variables. If we estimate $\hat{p}_k(x) \approx \mathbb{P}(G = g_k \mid X = x)$, then the zero-one Bayes rule is approximated by $\hat{G}(x) = \arg\min_{g_k} \hat{p}_k(x)$. This gives a decision-theoretic justification for probability based classifiers. However, if we use an unrestricted regression method, its fitted values may not be valid probabilities. A valid class probability model should ideally satisfy $0 \leq \hat{p}_k(x) \leq 1$ and $\sum_{k=1}^{K} \hat{p}_k(x) = 1$. Ordinary linear regression does not automatically enforce these constraints.

1.8.9 Bias-Variance Decomposition

Now suppose that we estimate a prediction rul from a random training sample $\mathcal{T}$. Then $\hat{f}_{\mathcal{T}}(x)$ is itself random because it depends on the random training data. Fix a test point x_0 and suppose first that the true response is deterministic with no noise. Then the mean squared error at x_0 is,

$$\text{MSE}(x_0) = \mathbb{E}_{\mathcal{T}}\left[\left(f(x_0) - \hat{f}_{\mathcal{T}}(x_0)\right)^2\right].$$

Let $m_{\mathcal{T}}(x_0) = \mathbb{E}_{\mathcal{T}}[\hat{f}_{\mathcal{T}}(x_0)]$. For cleaner notation, suppose that $\hat{f} = \hat{f}_{\mathcal{T}}(x_0)$ and $f_0 = f(x_0)$ with $m = \mathbb{E}_{\mathcal{T}}[\hat{f}]$.

Then we may write,

$$f_0 - \hat{f} = (f_0 - m) + (m - \hat{f}) \iff (f_0 - \hat{f})^2 = (f_0 - m)^2 + 2(f_0 - m)(m - \hat{f}) + (m - \hat{f})^2.$$

Taking expectation over the training set $\mathcal{T}$,

$$\mathbb{E}_{\mathcal{T}}[(f_0 - \hat{f})^2] = (f_0 - m)^2 + \mathbb{E}_{\mathcal{T}}[(\hat{f} - m)^2].$$

The first term is simply squared bias,

$$\text{Bias}^2(\hat{f}(x_0)) = \left(\mathbb{E}_{\mathcal{T}}[\hat{f}_{\mathcal{T}}(x_0)] - f(x_0)\right)^2,$$

and the second term is variance,

$$\text{Var}_{\mathcal{T}}(\hat{f}_{\mathcal{T}}(x_0)) = \mathbb{E}_{\mathcal{T}}\left[\left(\hat{f}_{\mathcal{T}}(x_0) - \mathbb{E}_{\mathcal{T}}[\hat{f}_{\mathcal{T}}(x_0)]\right)^2\right].$$

Thus our decomposition is simply algebraic,

$$\text{MSE}(x_0) = \text{Var}_{\mathcal{T}}(\hat{f}_{\mathcal{T}}(x_0)) + \text{Bias}^2(\hat{f}_{\mathcal{T}}(x_0)).$$

Suppose that the real data-generating process is $Y = f(X) + \varepsilon$ where $\mathbb{E}[\varepsilon \mid X] = 0$ and $\text{Var}(\varepsilon \mid X = x) = \sigma^2(x)$. At a fixed test point, x_0, the new test response is $Y_0 = f(x_0) + \varepsilon$ and the prediction error is,

$$\text{EPE}(x_0) = \mathbb{E}_{T, Y_0 | x_0}\left[\left(Y_0 - \hat{f}_{\mathcal{T}}(x_0)\right)^2\right].$$

Since we can write $Y_0 = f(x_0) + \varepsilon_0$ then $Y_0 - \hat{f}_{\mathcal{T}}(x_0) = \varepsilon_0 + f(x_0) - \hat{f}_{\mathcal{T}}(x_0)$.

Squaring that formulation we get that,

$$(Y_0 - \hat{f}_{\mathcal{T}}(x_0))^2 = \varepsilon_0^2 + 2\varepsilon_0\left(f(x_0) - \hat{f}_{\mathcal{T}}(x_0)\right) + \left(f(x_0) - \hat{f}_{\mathcal{T}}(x_0)\right)^2.$$

Taking expectation, since we know that $\mathbb{E}[\varepsilon \mid x_0] = 0$ and the test noise is independent of the training sample, the cross-term vanishes so we are left with,

$$\text{EPE}(x_0) = \mathbb{E}[\varepsilon_0^2 \mid x_0] + \mathbb{E}_{\mathcal{T}}\left[\left(f(x_0) - \hat{f}_{\mathcal{T}}(x_0)\right)^2\right].$$

Notice that the second term decomposes into variance plus squared bias. The first term is irreducible noise and remains even if we know the true regression function. The second term measures how sensitive the fitted model is to the training sample and the third term measures the systematic error from using a restricted or imperfect prediction rule.

1.8.10 Least Squares Prediction Error via Decision Theory

Consider the correctly specified linear model $Y = X^\top \beta + \varepsilon$ with $\mathbb{E}[\varepsilon \mid X] = 0$ and $\text{Var}(\varepsilon \mid X) = \sigma^2$, that is, conditional on X, our residuals are distributed with mean 0 and constant variance. Suppose that also the training design matrix is given by $X = [x_1 \cdots x_N]^\top$ and suppose that $y = X\beta + \varepsilon$. The least square estimator is characterized by the normal equation $\hat{\beta} = (X^\top X)^{-1} X^\top y$.

Substituting Y with $X\beta + \varepsilon$ we have that,

$$\begin{aligned}\hat{\beta} &= (X^\top X)^{-1} X^\top (X\beta + \varepsilon) \\ &= (X^\top X)^{-1} X^\top X\beta + (X^\top X)^{-1} X^\top \varepsilon \\ &= \beta + (X^\top X)^{-1} X^\top \varepsilon.\end{aligned}$$

Therefore $\hat{\beta} - \beta = (X^\top X)^{-1} X^\top \varepsilon$. At a test point, the fitted prediction is given by $\hat{y}_0 = x_0^\top \hat{\beta}$.

Substituting this expression for $\hat{\beta}$ gives,

$$\begin{aligned}\hat{y}_0 &= x_0^\top \left[\beta + (X^\top X)^{-1} X^\top \varepsilon\right] \\ &= x_0^\top \beta + x_0^\top (X^\top X)^{-1} X^\top \varepsilon.\end{aligned}$$

Since $\mathbb{E}[\varepsilon \mid X] = 0$, we get that $\mathbb{E}[\hat{y}_0 \mid X] = x_0^\top \beta$. Therefore, the least squares prediction is unbiased at the test point x_0 when the linear model is correctly specified. Computing the conditional variance of $\hat{y}_0$ is trivial. Since we know that $\mathrm{Var}(\varepsilon \mid X) = \sigma^2 I$ we have that,

$$\begin{aligned}\mathrm{Var}(\hat{y}_0 \mid X) &= \mathrm{Var}\left(x_0^\top (X^\top X)^{-1} X^\top \varepsilon \mid X\right) \\ &= x_0^\top (X^\top X)^{-1} X^\top \,\mathrm{Var}(\varepsilon \mid X) X (X^\top X)^{-1} x_0 \\ &= x_0^\top (X^\top X)^{-1} X^\top (\sigma^2 I) X (X^\top X)^{-1} x_0 \\ &= \sigma^2 x_0^\top (X^\top X)^{-1} X^\top X (X^\top X)^{-1} x_0 \\ &= \sigma^2 x_0^\top (X^\top X)^{-1} x_0.\end{aligned}$$

A new test response is $Y_0 = x_0^\top \beta + \varepsilon_0$ so the prediction error is simply $Y_0 - \hat{y}_0 = \varepsilon_0 - x_0^\top (\hat{\beta} - \beta)$. The test noise ε_0 is independent of the training error in $\hat{\beta}$ so it is evident that,

$$\begin{aligned}\mathrm{EPE}(x_0 \mid X) &= \mathbb{E}[(Y_0 - \hat{y}_0)^2 \mid X] \\ &= \mathrm{Var}(\varepsilon_0) + \mathrm{Var}(\hat{y}_0 \mid X) + 0^2 \\ &= \sigma^2 + \sigma^2 x_0^\top (X^\top X)^{-1} x_0\end{aligned}$$

It is evident from the formulation of the expected prediction error for the least squares decision problem that the first term σ^2 represents the irreducible noise and the second term is the estimation variance from learning β from finite data. If the training inputs are random and N is large, then approximately $X^\top X \approx N \,\mathrm{Cov}(X)$ assuming that $\mathbb{E}[X] = 0$. Hence $(X^\top X)^{-1} \approx \frac{1}{N} \mathrm{Cov}(X)^{-1}$ which means,

$$x_0^\top (X^\top X)^{-1} x_0 \approx \frac{1}{N} x_0^\top \,\mathrm{Cov}(X)^{-1} x_0.$$

Averaging over x_0,

$$\begin{aligned}\mathbb{E}_{x_0}\left[x_0^\top \,\mathrm{Cov}(X)^{-1} x_0\right] &= \mathbb{E}_{x_0}\left[\mathrm{tr}\left(x_0^\top \,\mathrm{Cov}(X)^{-1} x_0\right)\right] \\ &= \mathbb{E}_{x_0}\left[\mathrm{tr}\left(\mathrm{Cov}(X)^{-1} x_0 x_0^\top\right)\right] \\ &= \mathrm{tr}\left(\mathrm{Cov}(X)^{-1} \mathbb{E}[x_0 x_0^\top]\right).\end{aligned}$$

If $\mathbb{E}[X] = 0$ then $\mathbb{E}[x_0 x_0^\top] = \mathrm{Cov}(X)$ so,

$$\begin{aligned}\mathbb{E}_{x_0}\left[x_0^\top \,\mathrm{Cov}(X)^{-1} x_0\right] &= \mathrm{tr}\left(\mathrm{Cov}(X)^{-1} \,\mathrm{Cov}(X)\right) \\ &= \mathrm{tr}(I_p) \\ &= p.\end{aligned}$$

This clearly represents the tradeoff that is, if the linear model is correct, there is no bias, but there is still variance from estimating p coefficients using N observations. The variance term grows with model dimension p and shrinks with sample size N approaching ∞.

2 Linear Models for Regression

2.1 Univariate Regression

In this section we will start with the univariate case of linear regression and generalize it with matrices to the multiple predictor regressor case. As in any regression model, the systematic portion is represented by the regression coefficients $\beta_0 + \beta_1 x$ and the random error is represented by ε. In this case, β_0 is the intercept and β_1 is the slope. For any model, we often make a few assumptions regarding the properties of the model as well as the data from which we sample from. From statistical decision theory, a regression function assumes that $\mathbb{E}[Y \mid X]$ is linear in the inputs x_i's and that $\mathbb{E}[Y \mid X]$ is a reasonable linear approximation.

Regardless of the source of the x_i's, the model is to be linear in it's parameters, that is, is able to be written as an equation that is a sum of the coefficients multiplied by variables where unknown parameters (coefficients or weights) appear only to the first power. Each sample from which $\mathbb{E}[Y \mid X]$ is estimated from is to be sampled randomly, or observations are independent across all i. There is no perfect multicollinearity among the regressors, so in the case of multiple regression, the design matrix X will have full column rank.

Conditioning on inputs x_i's, the conditional mean of errors must be zero, that is $\mathbb{E}[\varepsilon_i \mid X] = 0$ or equivalently, $\mathbb{E}[\varepsilon \mid x_{i1}, \dots, x_{ik}] = 0$. We also assume that the errors have constant variance, that is $\text{Var}(\varepsilon_i \mid x_{i1}, \dots, x_{ik}) = \sigma^2$ and are thus normally distributed $\varepsilon_i \mid (x_{i1}, \dots, x_{ik}) \sim \mathcal{N}(0, \sigma^2)$.

We refer to these assumptions as the ***Gauss-Markov Assumptions***.

Theorem 2.1: Gauss-Markov Assumptions

(i) **Linearity in Parameters**: $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$
(ii) **Random Sampling**: Observations $\{(x_i, y_i)\}_{i=1}^n$ are distributed iid
(iii) **No Multicollinearity**: No explanatory variable is an exact linear combination of the others
(iv) **Error Zero Conditional Mean**: $\mathbb{E}[\varepsilon_i \mid (x_{i1}, \dots, x_{ik})] = 0$
(v) **Constant Variance**: $\text{Var}(\varepsilon_i \mid (x_{i1}, \dots, x_{ik})) = \sigma^2$
(vi) **Normally Distributed Errors**: $\varepsilon_i \mid (x_{i1}, \dots, x_{ik}) \sim \mathcal{N}(0, \sigma^2)$

In univariate regression our primary interest is to make sense of the slope parameter. That is, β_1 is the change in the conditional mean of Y when x_1 increases by one unit i.e., $\beta_1 = \mathbb{E}[y \mid x_1 = a + 1] - \mathbb{E}[y \mid x_1 = a]$. This is true if the conditional expectation function is linear meaning that $\mathbb{E}[y \mid X = x] = \beta_0 + \beta_1$ which means simply that,

$$\mathbb{E}[y \mid x = a + 1] - \mathbb{E}[y \mid x = a] = (\beta_0 + \beta_1(a + 1)) - (\beta_0 + \beta_1 a) = \beta_1.$$

2.1.1 Least Squares Estimation

I assume you have exposure to mathematical statistics so I will refrain from going into too much depth. We will, at least for univarate regression, make use of the formulations for sample variance, sample covariance, sample correlation and S_{xx}, S_{xy} throughout these derivations. Given N samples (x_i, y_i) we wish to find closed form equations for (β_0, β_1) such that the residual sum of squares is minimized.

More explicitly, we are minimizing the loss criterion,

$$\text{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_1)^2.$$

Differentiating RSS with respect to β_0 and β_1 gives us two equations, of which we set both to 0. This is similar to gradient descent where we obtain a gradient vector (if working on a plane constructed via column vectors) and move in the direction of the minima of the function. The question of whether such a method will always find a global minimum is something that would be covered in a course on neural network optimization so I will avoid covering any of that here (please refer to my CS 479 notes if you're interested).

Anyways, we now have two derivatives, one with respect to β_0 and the other with respect to β_1. From here, we wish to solve to get equations for β_0 and β_1 in explicit closed forms.

Starting with the derivative wrt β_0 we get,

$$0 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i$$

Bringing $n\beta_0$ to the LHS we simply solve and use the definition of $\bar{y}$ and $\bar{x}$,

$$\beta_0 = \frac{1}{n}\left\{\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i\right\} = \bar{y} - \beta_1 \bar{x}.$$

Therefore, clearly we see that $\beta_0 = \bar{y} - \beta\bar{x}$. Similarly, we solve for β_1 using the derivative wrt β_1.

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i + \beta_1 \bar{x} \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0.$$

Note that $\beta_0 = \bar{y} - \beta_1 \bar{x}$ so the term $\beta_0 \sum_i x_i = \bar{y} \sum_i x_i + \beta_1 \bar{x} \sum_i x_i$.

$$\beta_1 \sum_i x_i^2 - \beta_1 \bar{x} \sum_i x_i = \sum_i x_i y_i - \bar{y} \sum_i x_i.$$

Solving for β_1 we get,

$$\beta_1 \left(\sum_i x_i^2 - \bar{x} \sum_i x_i\right) = \sum_i x_i y_i - \bar{y} \sum_i x_i \iff \beta_1 = \frac{\sum_i x_i y_i - n\bar{y}\bar{x}}{\sum_i x_i^2 - n\bar{x}^2} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}.$$

Recall the definition of S_{xy} and S_{yy} and notice that the numerator and denominator correspond respectively.

$$\beta_1 = \frac{S_{xy}}{S_{xx}}.$$

Notice something interesting here? Recall that the correlation coefficient as $r_{xy} = \frac{s_{xy}}{s_x s_y}$. If we re-arrange for s_{xy} we see that $s_{xy} = r_{xy} s_x s_y$. Substituting this into the slope formula we have that,

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{r_{xy} s_x s_y}{s_x^2} = r_{xy} \cdot \frac{s_y}{s_x}.$$

From this we can see that the simple OLS slope coefficient is simply the correlation between x and y multiplied by the ratio of their sample standard deviations. We also know that sample correlation can be written as $r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$ so $S_{xy} = r_{xy}\sqrt{S_{xx} S_{yy}}$, so plugging in $\hat{\beta}_1 = S_{xy}/S_{xx}$ gives,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{r_{xy}\sqrt{S_{xx} S_{yy}}}{S_{xx}} = r_{xy} = \frac{\sqrt{S_{yy}}}{S_{xx}} = r_{xy} \cdot \sqrt{\frac{S_{yy}}{S_{xx}}}.$$

The population version of this formulation is analogous. If $\mathbb{E}[Y \mid X = x] = \beta_0 + \beta_1 x$ then $\beta_1 = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$. Since we also know that $\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$ then we have that $\text{Cov}(X, Y) = \rho_{XY} \sigma_X \sigma_Y$. Therefore, it is trivial to write β_1 as $\rho_{XY} \frac{\sigma_Y}{\sigma_X}$. Correlation is clearly unitless while the slope has units of units of y per unit of x. Therefore, to convert correlation into a slope, you multiply by the scale ratio s_y/s_x. If x and y are highly positively correlated, then $\hat{\beta}_1$ is positive, otherwise, if they are negatively correlated then $\hat{\beta}_1$ is negative. If they are uncorrelated then $\hat{\beta}_1 = 0$. It is also important to note that the size of the slope also depends on the relative scales of x and y.

Theorem 2.2: Ordinary Least Squares Estimators

Suppose that we observe $(x_i, y_i)_{i=1}^n$ from the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

The least squares estimators minimize the residual sum of squares

$$\text{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

The unique minimizers are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}},$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

2.1.2 Properties of $\hat{\beta}_0$ and $\hat{\beta}_1$

Now that we have assumptions for our model as well as closed form solutions to determinining their coefficients, what are the properties of our least squares coefficients? To characterize them properly, we will assess their biasedness (unbiasedness) by evaluating their expectations and variance formulas.

We claim that the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimates.

Suppose that we have a linear regression model written as $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with $\mathbb{E}[\varepsilon] = 0$, $\text{Var}(\varepsilon_i) = \sigma^2$ and $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$. Equivalently, conditional on the observed regressors $x_1, \ldots, x_N$ we know that $\mathbb{E}[y_i] = \beta_0 + \beta_1 x_i$, $\text{Var}(y_i) = \sigma^2$ and $\text{Cov}(y_i, y_j) = 0$ for $i \neq j$.

Recall the formulation for $\hat{\beta}_1$,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{S_{xx}} = \frac{\sum_{i=1}^N (x_i - \bar{x}) y_i - \bar{y} \sum_{i=1}^N (x_i - \bar{x})}{S_{xx}}.$$

Recall that $\sum_i (x_i - \bar{x}) = \sum_i x_i - N\bar{x} = N\bar{x} - N\bar{x} = 0$. Therefore we can write,

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i (x_i - \bar{x}) y_i.$$

Substituting this into the formulation for $\hat{\beta}_1$ we have that,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^N (x_i - \bar{x}) y_i}{S_{xx}} = \sum_{i=1}^N \frac{x_i - \bar{x}}{S_{xx}} y_i.$$

To simplify our notation for the same of algebraic convenience, from the formula reconstruction for $\hat{\beta}_1$ above, we will let $c = \frac{x_i - \bar{x}}{S_{xx}}$ (i.e., the term multiplying the y_i in the formula above). Formulas for the expectations and variances of $\hat{\beta}_1$ and $\hat{\beta}_0$ do not contain c_i after simplifying so this is purely for easier computation.

We will start with the first primary claim which is that $\mathbb{E}[\hat{\beta}_0] = \beta_0$ and $\mathbb{E}[\hat{\beta}_1] = \beta_1$.

By linearity of expectation, the results follow,

$$\mathbb{E}[\hat{\beta}_1] = \mathbb{E}\left(\sum_i c_i y_i\right) = \sum_i c_i \mathbb{E}[y_i] = \sum_i c_i (\beta_0 + \beta_1 x_i) = \beta_0 \sum_i c_i + \beta_1 \sum_i c_i x_i = \beta_1,$$

and the same for the OLS intercept term,

$$\mathbb{E}[\hat{\beta}_0] = \mathbb{E}[\bar{y} - \hat{\beta}_1\bar{x}] = \mathbb{E}[\bar{y}] - \bar{x}\mathbb{E}[\hat{\beta}_1] = \mathbb{E}\left(\frac{1}{N}\sum_i y_i\right) - \bar{x}\beta_1 = \frac{1}{N}\sum_i \mathbb{E}[y_i] - \bar{x}\beta_1 = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0.$$

Therefore, we have shown that $\hat{\beta}_1$ and $\hat{\beta}_0$ are unbiased linear estimates.

Next, we would like to characterize the uncertainty around the estimates $\hat{\beta}_0, \hat{\beta}_1$. Intuitively, estimating a slope is easier when the x values cover a wide range. If all x_i values are close together, a small amount of veritcal noise can drastically change the fitted slope. If x_i values are spread far apart, the line is more "anchored" so the slope estimate is more stable. Moreover, the intercept is dependent on the slope estimate. That is, if $\hat{\beta}_1$ moves around, then $\hat{\beta}_0$ also moves around, especially when $\bar{x}$ is far from zero. If your data is centered far away from 0, then estimating the intercept requires extrapolating the fitted line back to $x = 0$ which makes the intercept less stable.

By applying properties of variance we simply get,

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\left(\sum_i c_i y_i\right) = \sum_i c_i^2 \mathrm{Var}(y_i) = \sum_i c_i^2\sigma^2 = \sigma^2\sum_i c_i^2 = \frac{\sigma^2}{S_{xx}}.$$

For the slope, we see that the uncertainty of the slope estimate is simply the noise in y scaled by the variation in x. So, if errors ε_i are very noisy meaning that σ^2 is large, then $\hat{\beta}_1$ is more volatile. But if the x_i values are very spread out, meaning S_{xx} is large, then the slope is easier to estimate.

Similarly, we get the variance formulation for the intercept,

$$\mathrm{Var}(\hat{\beta}_0) = \mathrm{Var}(\bar{y} - \hat{\beta}_1\bar{x}) = \mathrm{Var}\left(\frac{1}{N}\sum_i y_i\right) + \bar{x}^2\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\left(\frac{1}{N}\sum_i y_i\right) + \bar{x}^2\frac{\sigma^2}{S_{xx}} = \frac{1}{N^2}\sum_i \mathrm{Var}(y_i) + \bar{x}^2\frac{\sigma^2}{S_{xx}}.$$

We can simplify further to see that,

$$\frac{1}{N^2}(N\sigma^2) + \bar{x}^2\frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{N} + \frac{N\bar{x}^2}{S_{xx}}\sigma^2.$$

Observe that the intercept's sources of uncertainty are derived from two sources. The first is the uncertainty from estimating the average level of y that is σ^2/N which is the same form of variance we encounter when estimating a sample mean. In this case, more data lowers this portion which is why sample size is inversely related to variance. Second, there is uncertainty from estimating the slope, specifically represented by the term $\sigma^2\frac{\bar{x}^2}{S_{xx}}$ which is a direct result of $\bar{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$.

You may ask, what is the covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$? First we must derive two important identities of c_i.

$$\sum_i c_i = \sum_i \frac{x_i - \bar{x}}{S_{xx}} = \frac{1}{S_{xx}}\left(\sum_i x_i - N\bar{x}\right) = \frac{1}{S_{xx}}(N\bar{x} - N\bar{x}) = 0.$$

Similarly, the sum of squared c_i's is also simply just,

$$\sum_i c_i^2 = \sum_i \left(\frac{x_i - \bar{x}}{S_{xx}}\right)^2 = \frac{1}{S_{xx}^2}\sum_i (x_i - \bar{x})^2 = \frac{S_{xx}}{S_{xx}^2} = \frac{1}{S_{xx}}.$$

Therefore, using the definition of $\hat{\beta}_0$, the covariance is simply,

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \mathrm{Cov}(\bar{y} - \bar{x}\hat{\beta}_1, \hat{\beta}_1) = \mathrm{Cov}(\bar{y}, \hat{\beta}_1) - \bar{x}\,\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_1) = \mathrm{Cov}(\bar{y}, \hat{\beta}_1) - \bar{x}\,\mathrm{Var}(\hat{\beta}_1).$$

First we compute the covariance between $\bar{y}$ and the slope coefficient,

$$\begin{aligned}\mathrm{Cov}(\bar{y}, \hat{\beta}_1) &= \mathrm{Cov}\left(\frac{1}{N}\sum_{i=1}^N y_i, \sum_{j=1}^N c_j y_j\right) \\ &= \sum_{i=1}^N\sum_{j=1}^N \frac{1}{N}c_j\,\mathrm{Cov}(y_i, y_j) = \sum_{i=1}^N \frac{1}{N}c_i\,\mathrm{Cov}(y_i, y_i) = \sum_{i=1}^N \frac{1}{N}c_i\sigma^2 = \frac{\sigma^2}{N}\sum_{i=1}^N c_i = 0.\end{aligned}$$

The variance of $\hat{\beta}_1$ is simply,

$$\text{Var}(\hat{\beta}_1) = \text{Var}\left(\sum_{i=1}^{N} c_i y_i\right) = \sum_{i=1}^{N} c_i^2 \,\text{Var}(y_i) = \sigma^2 \sum_{i=1}^{N} c_i^2 = \frac{\sigma^2}{S_{xx}}.$$

Substituting back into the formula for $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$ we get,

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \text{Cov}(\bar{y}, \hat{\beta}_1) - \bar{x}\,\text{Var}(\hat{\beta}_1) = 0 - \bar{x}\frac{\sigma^2}{S_{xx}} = -\frac{\bar{x}\sigma^2}{S_{xx}}.$$

It is clear from the above that the estimated slope and estimated intercept have an inverse relationship namely that they move in opposite directions. If $\bar{x} > 0$ then increasing $\hat{\beta}_1$ forces $\hat{\beta}_0$ downward to keep a fitted line passing through $(\bar{x}, \bar{y})$ when data are centered at a positive $\bar{x}$.

Theorem 2.3: Unbiasedness of the OLS Estimators

Suppose that the simple linear regression model satisfies

$$\mathbb{E}[\varepsilon_i \mid X] = 0.$$

Then the least squares estimators are unbiased:

$$\mathbb{E}[\hat{\beta}_1] = \beta_1, \qquad \mathbb{E}[\hat{\beta}_0] = \beta_0.$$

Theorem 2.4: Variance of the OLS Slope Estimator

Suppose that

$$\text{Var}(\varepsilon_i \mid X) = \sigma^2, \qquad \text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad (i \neq j).$$

Then

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}.$$

Hence the uncertainty of the slope estimate decreases as the spread of the x_i values increases.

2.1.3 Properties of Least Squares Residuals

Recall that we chose $c_i = \frac{x_i - \bar{x}}{S_{xx}}$ since it lets us write the slope estimator as a weighted average of the y_i's. We will need to derive some other properties of c_i. We will derive weighted orthogonality of c_i,

$$\sum_i c_i x_i = \frac{1}{S_{xx}} \sum_i x_i(x_i - \bar{x}) = \frac{1}{S_{xx}} \left(\sum_i x_i^2 - \bar{x} \sum_i x_i\right) = \frac{1}{S_{xx}} \left(\sum_i x_i^2 - N\bar{x}^2\right) = \frac{1}{S_{xx}} \left(\sum_i (x_i - \bar{x})^2\right) = \frac{S_{xx}}{S_{xx}} = 1.$$

The residuals are $r_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$.

Trivially, observe that, the residuals algebraically balance out. The positive residuals and negative residuals cancel exactly so with an intercept in regression, OLS does not systematically overpredict or underpredict on average.

$$0 = \frac{\partial}{\partial \beta_0} \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2 \Big|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = -2 \sum_{i=1}^{N} r_i \implies \sum_{i=1}^{N} r_i = 0.$$

Equivalently, $\sum_i (y_i - \hat{y}_i) = 0$ which implies that $\sum_i y_i = \sum_i \hat{y}_i$. So the average fitted value equals the average observed value $\hat{y} = \bar{y}$. Similarly, this is why the OLS line passes through the sample mean point $(\bar{x}, \bar{y})$. From this, we see that $\sum_i x_i r_i = 0$ implying that the residuals are orthogonal to x.

After fitting the OLS line, there is no remaining linear pattern in the residuals with respect to x. Otherwise, if there still was a systematic relationship between x_i and r_i, then the line could still be tilted more to reduce the squared errors. Therefore, OLS residuals are orthogonal to the variables used to fit the model.

In simple linear regression, the model is built from 1 and x_i because $\hat{y}_i = \hat{\beta}_0 \cdot 1 + \hat{\beta}_1 x_i$. So OLS forces the residuals to be orthogonal to both,

$$\sum_i r_i \cdot 1 = 0, \quad \sum_i r_i x_i = 0.$$

From this, we can also view these as constraints on the residuals of our estimates. If ε_i are known, then $\hat{\sigma}^2 = \frac{1}{N}\sum_i \varepsilon_i^2$ or in other words,

$$\sigma^2 = \mathbb{E}[\varepsilon_i^2] + \text{Var}(\varepsilon_i) = \sigma^2 + 0^2 = \sigma^2.$$

However, from the above, notice how the expectation of the average summed residuals is not equal to σ^2. If we let $S^2 = \frac{1}{N-1}\sum_i r_i^2$ then $\mathbb{E}[S^2] = \sigma^2$. Note here that $N-2$ is the degrees of freedom as we cannot count 2 of the values since we have two constraints on r_i. Two of the r_i's cannot vary freely which is why sample variance for $r_i \sim \mathcal{N}(0, \sigma^2)$ is defined with $N-1$ as a constraint on y_i is $\sum_i (y_i - \bar{y}) = 0$.

Theorem 2.5: Slope as Correlation Times Scale Ratio

The ordinary least squares slope estimator can be written as

$$\hat{\beta}_1 = r_{xy}\frac{s_y}{s_x},$$

where r_{xy} is the sample correlation coefficient and s_x, s_y are the sample standard deviations of x and y respectively. Equivalently, at the population level,

$$\beta_1 = \rho_{XY}\frac{\sigma_Y}{\sigma_X}.$$

2.1.4 Confidence Intervals and Hypothesis Testing

Recall previously when we derived $\mathbb{E}[\hat{\beta}_1]$ and found that it is a unbiased estimate of β_1. Moreover, we found that $\text{Var}(\hat{\beta}_1)$ is $\frac{\sigma^2}{S_{xx}}$ that is our slope estimate is the noise of our y_i's scaled by the variation in our x. Naturally, we can state the distribution of our $\hat{\beta}_1$ as $\mathcal{N}(\beta_1, \sigma^2/S_{xx})$ as a result from the Gauss-Markov assumptions and our derivations previously. This follows because $\hat{\beta}_1$ is a linear combination of the normally distributed errors.

Suppose that we know the true σ^2, then if we want to test the null hypothesis $H_0 : \beta_1 = \beta_1^*$ we can use the test statistic,

$$\frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\sigma^2/S_{xx}}} \sim \mathcal{N}(0,1).$$

However, in practice, we usually do not know the true error variance σ^2. Therefore, we replace it with the unbiased estimator

$$S^2 = \frac{\text{RSS}}{n-2}.$$

This introduces additional sampling uncertainty because S^2 is itself random. Recall that if a standard normal random variable is divided by the square root of an independent chi-squared random variable scaled by its degrees of freedom, then the resulting random variable has a Student-t distribution.

So, therefore our new test statistic is,

$$\frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{S^2/S_{xx}}} \sim t_{n-2},$$

which follows a Student-t distribution with $n-2$ degrees of freedom.

From this, the $100(1-\alpha)\%$ confidence interval for β_1 is therefore,

$$\begin{aligned} 1-\alpha &= \mathbb{P}\left(-t_{n-2,\alpha/2} < \frac{\hat{\beta}_1 - \beta_1}{\widehat{\text{SE}}(\hat{\beta}_1)} < t_{n-2,\alpha/2}\right) \\ &= \mathbb{P}\left(\hat{\beta}_1 - t_{n-1,\alpha/2}\cdot \widehat{\text{SE}}(\hat{\beta}_1) < \beta_1 < \hat{\beta}_1 + t_{n-2,\alpha/2}\widehat{\text{SE}}(\hat{\beta}_1)\right). \end{aligned}$$

In this case we let $\text{SE}(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{S_{xx}}}$ be the true but unknown standard error. Since σ^2 is unknown, we estimate it with $\hat{\text{SE}}(\hat{\beta}_1) = \sqrt{\frac{S^2}{S_{xx}}}$ which is why our t distribution has $n-2$ degrees of freedom.

Therefore our $100(1-\alpha)\%$ confidence interval is,

$$\left[\hat{\beta}_1 - t_{n-2,\alpha/2}\hat{\text{SE}}(\hat{\beta}_1), \hat{\beta}_1 + t_{n-2,\alpha/2}\hat{\text{SE}}(\hat{\beta}_1)\right].$$

Similarly, we can use the test statistic to test $H_0 : \beta_1 = \beta_1^*$ against the alternative hypothesis $H_\alpha : \beta_1 \neq \beta^*$. That is we reject H_0 at the α significance level if $|t| > t_{n-2,\alpha/2}$. Alternative, we could compute the $p = P(|T| \geq |t|)$ value for $T \sim t_{n-2}$ and reject is $p \leq a$. Our null hypothesis can also be written as,

$$H_0 : t = \frac{\hat{\beta}_1 - \hat{\beta}_1^*}{\text{SE}(\hat{\beta}_1)} \sim t_{n-2}.$$

Why do we do this? We computed $\hat{\beta}_1$ from the data, but $\hat{\beta}_1$ is random because the data is randomly sampled. Even if the true slope were exactly β_1^*, sampling variability means that $\hat{\beta}_1$ would not be exactly equal to β_1^*. Thus, by standardizing the difference between $\hat{\beta}_1$ and β_1^*, we measure how many estimated standard errors away is the observed slope from the hypothesized slope which is analogous to the z score.

Why do we reject H_0 if $|t| > t_{n-2,\alpha/2}$? We are basically asking if the observed slope is too many standard errors away from the null value to plausibly occur under the null. The rejection region lies in the two tails because this is essentially a two-sided test. Assuming the null hypothesis is true, the p value therefore tells us what the probability of observing a test statistic at least as extreme as the one we got. Smaller p values mean the observed data is very unusual if H_0 were true so the null hypothesis becomes implausible.

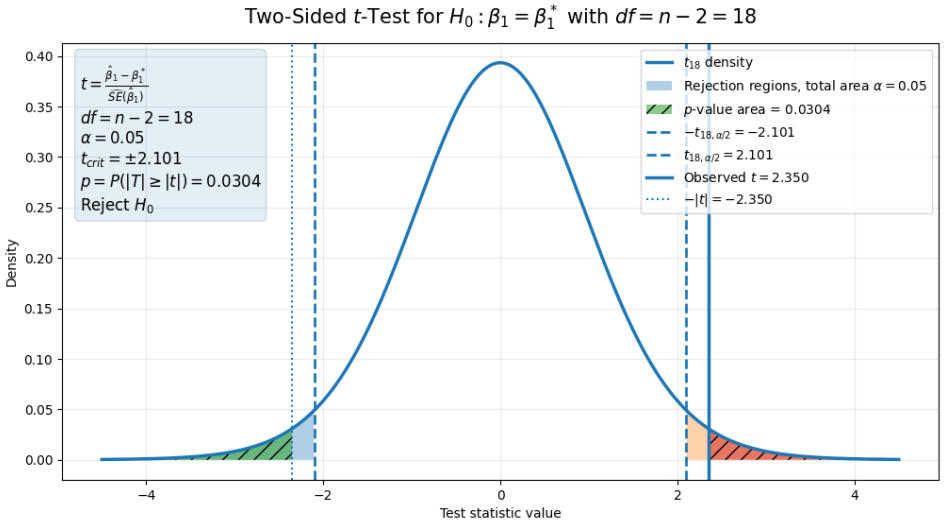

Figure 1: Two-sided t-test rejection regions and observed test statistic.

Theorem 2.6: Confidence Interval for the OLS Slope

A $100(1-\alpha)\%$ confidence interval for β_1 is

$$\left[\hat{\beta}_1 - t_{n-2,\alpha/2}\sqrt{\frac{S^2}{S_{xx}}}, \hat{\beta}_1 + t_{n-2,\alpha/2}\sqrt{\frac{S^2}{S_{xx}}}\right].$$

2.1.5 Inference

For an arbitrary x_0, we refer to $\mu_0 = \beta_0 + \beta_1 x_0$ as the **mean response** and estimate $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$. The predicted mean response at point x_0 is actually a weighted average of the observed response $y_1, \ldots, y_N$. It may appear first that the prediction depends only on the estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$. If you substitute the formulas for the least-squares estimators and simplify, we can write the prediction $\hat{\mu}_0$ as a linear combination of the observed responses.

Suppose we wish to write the prediction as the linear combination,

$$\hat{\mu}_0 = \sum_{i=1}^{N} d_i y_i.$$

d_i in this case would be a discrepancy measure, of which we'll define as,

$$d_i = \frac{1}{N} + \frac{(x_0 - \bar{x})(x_i - \bar{x})}{S_{xx}}$$

which tells us how much influence y_i has on the prediction at point x_0.

The first term gives every observation an equal baseline weight. If this were the only term, the prediction would simply be the sample mean $\bar{y}$. The second term adjust the weights according to the geometric relationship between x_i and the prediction location x_0. The factor $(x_0 - \bar{x})(x_i - \bar{x})$ measures whether x_i lies on the same side of the sample mean as x_0.

If x_i and x_0 are both above $\bar{x}$ or both below $\bar{x}$, then this product is positive so the observation receives more weight in the prediction. If they lie on opposite sides of $\bar{x}$, then the product is negative, so the observation receives less weight. Thus observations whose predictor values are near or directionally aligned with x_0 influence the fitted value more strongly. We aim to show $\hat{\mu}_0 = \sum_i d_i y_i$.

Proof. Suppose that our mean response estimate is defined as $\hat{\mu}_0 = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0$.

The result follows by expanding and simplifying the sum,

$$\begin{aligned}
\hat{\mu}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0 \\
&= \frac{1}{N} \sum_i y_i - \sum_i c_i y_i \bar{x} + \sum_i c_i y_i x_0 \\
&= \sum_i \left(\frac{1}{N} + (x_0 - \bar{x} c_i) \right) y_i \\
&= \sum_i d_i y_i.
\end{aligned}$$

It also follows that $\hat{\mu}_0$ is an unbiased estimate,

$$\mathbb{E}[\hat{\mu}_0] = \mathbb{E}[\hat{\beta}_0 + \hat{\beta}_1 x_0] = \mathbb{E}[\hat{\beta}_0] + x_0 \mathbb{E}[\hat{\beta}_1] = \beta_0 + \beta_1 x_0 = \mu_0.$$

Lastly, we can also derive the mean responses' variance,

$$\begin{aligned}
\text{Var}(\hat{\mu}_0) &= \text{Var}\left(\sum_{i=1}^{N} d_i y_i\right) = \sum_{i=1}^{N} d_i^2 \,\text{Var}(y_i) \\
&= \left[\sum_{i=1}^{N} \frac{1}{N^2} + \sum_{i=1}^{N} \frac{2(x_0 - \bar{x})(x_i - \bar{x})}{N S_{xx}} c_i + \sum_{i=1}^{N} \frac{(x_0 - \bar{x})^2 c_i^2}{S_{xx}}\right] \sigma^2 \\
&= \left[\frac{1}{N} + \sum_{i=1}^{N} \frac{(x_0 - \bar{x})^2 c_i^2}{S_{xx}}\right] \sigma^2 \\
&= \left[\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \sum_{i=1}^{N} c_i^2\right] \sigma^2 \\
&= \left[\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \cdot \frac{1}{S_{xx}}\right] \sigma^2 \\
&= \left[\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right] \sigma^2.
\end{aligned}$$

□

From the results above, we can see that, over repeated samples, the estimate is centered around the true regression mean at x_0. So $\hat{\mu}_0$ may be too high or too low in any one sample but is not systematically too high or systematically too low. Second, the variance formula tells us how uncertain the estimate is, that is, the estimated response is most precise when x_0 is close to $\bar{x}$. If $x_0 = \bar{x}$ then the second term disappears and $\text{Var}(\hat{\mu}_0) = \frac{\sigma^2}{N}$. In that case, this makes estimating the mean response at the center of the observed x values equivalent to estimating the average response. As x_0 moves further away from $\bar{x}$, the term $\frac{(x_0 - \bar{x})^2}{S_{xx}}$ gets larger so the variance of $\hat{\mu}_0$ increases. This means predictions of the mean response become less reliable the farther x_0 is from the center of the data.

Theorem 2.7: Mean Response Estimator

For a fixed predictor value x_0, define the mean response

$$\mu_0 = \beta_0 + \beta_1 x_0,$$

with estimator

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

Then

$$\mathbb{E}[\hat{\mu}_0] = \mu_0,$$

and

$$\text{Var}(\hat{\mu}_0) = \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right] \sigma^2.$$

2.1.6 Predicting Future Values

The goal of regression is wanting the best guess of y given $x = x_p$, that is we want guess y_p.

Thus, our best guess is $\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$.

We claim that $\hat{y}_p$ is an unbiased estimate of y_p.

Proof. Note that claiming $\hat{y}_p$ is unbiased is equivalent to showing $\mathbb{E}[y_p - \hat{y}_p] = 0$.

$$\begin{aligned}
\mathbb{E}[y_p - \hat{y}_p] &= \mathbb{E}[(\beta_0 + \beta_1 x_p + \varepsilon_p) - (\hat{\beta}_0 + \hat{\beta}_1 x_p)] \\
&= \mathbb{E}[(\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_p + \varepsilon_p] \\
&= (\beta_0 - \mathbb{E}[\hat{\beta}_0]) + (\beta_1 - \mathbb{E}[\hat{\beta}_1])x_p + \mathbb{E}[\varepsilon_p] \\
&= (\beta_0 - \beta_0) + (\beta_1 - \beta_1)x_p + 0 \\
&= 0.
\end{aligned}$$

For the variance, I will write an expansion to help simplification,

$$y_p - \hat{y}_p = (\beta_0 + \beta_1 x_p + \varepsilon_p) - (\hat{\beta}_0 - \hat{\beta}_1 x_p) = (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_p + \varepsilon_p = (\hat{\mu}_p - \mu_p) + \varepsilon_p.$$

From this, if we take variances we can obtain,

$$\mathrm{Var}(y_p - \hat{y}_p) = \mathrm{Var}(\hat{\mu}_p - \mu_p) + \mathrm{Var}(\varepsilon_p) = \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\sigma^2 + \sigma^2\right] = \left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right]\sigma^2.$$

Then the test statistic to test $H_0 : y_p = \hat{y}_p$ is,

$$\frac{y_p - \hat{y}_p}{\mathrm{SE}(y_p - \hat{y}_p)} = \frac{y_p - \hat{y}_p}{\sqrt{\left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right] S^2}} = \frac{(y_p - \hat{y}_p)/\sigma}{\sqrt{\left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right] (S^2/\sigma^2)}}.$$

Clearly the numerator is normal with mean 0 and the denominator involves $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$. Hence the ratio of a standard normal and the square root of an independent chi-squared divided by its degrees of freedom yields a t-statistics with a t_{n-1} Student-t distribution.

Trivially, the $100(1-\alpha)\%$ prediction interval for response y_p is simply,

$$\hat{y}_p \pm t_{n-2,\alpha/2} \cdot \mathrm{SE}(y_p - \hat{y}_p).$$

□

Theorem 2.8: Prediction Interval for a Future Response

Suppose that the simple linear regression model is

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

For a future response at x_p, define

$$y_p = \beta_0 + \beta_1 x_p + \varepsilon_p, \qquad \hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p.$$

Then

$$\mathbb{E}[y_p - \hat{y}_p] = 0$$

and

$$\mathrm{Var}(y_p - \hat{y}_p) = \left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right]\sigma^2.$$

Therefore,

$$\frac{y_p - \hat{y}_p}{S\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}} \sim t_{n-2},$$

and a $100(1-\alpha)\%$ prediction interval for the future response y_p is

$$\hat{y}_p \pm t_{n-2,\alpha/2} S\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}.$$

2.1.7 Analysis of Variance

After fitting a model, it is important to assess the fit, usually through some analysis of variance. The distance between a fitted response $\hat{y}_i$ and response y_i is called the *residual error* or sum of squared errors. This is simply the vertical distance between the predicted response and the actual response y_i. The vertical distance between our fitted response $\hat{y}_i$ and mean response $\bar{y}$ is the sum of squared regression or the error sum of squares. This is the variation that the regression line explains relative to the baseline predictor, the sample mean $\bar{y}$.

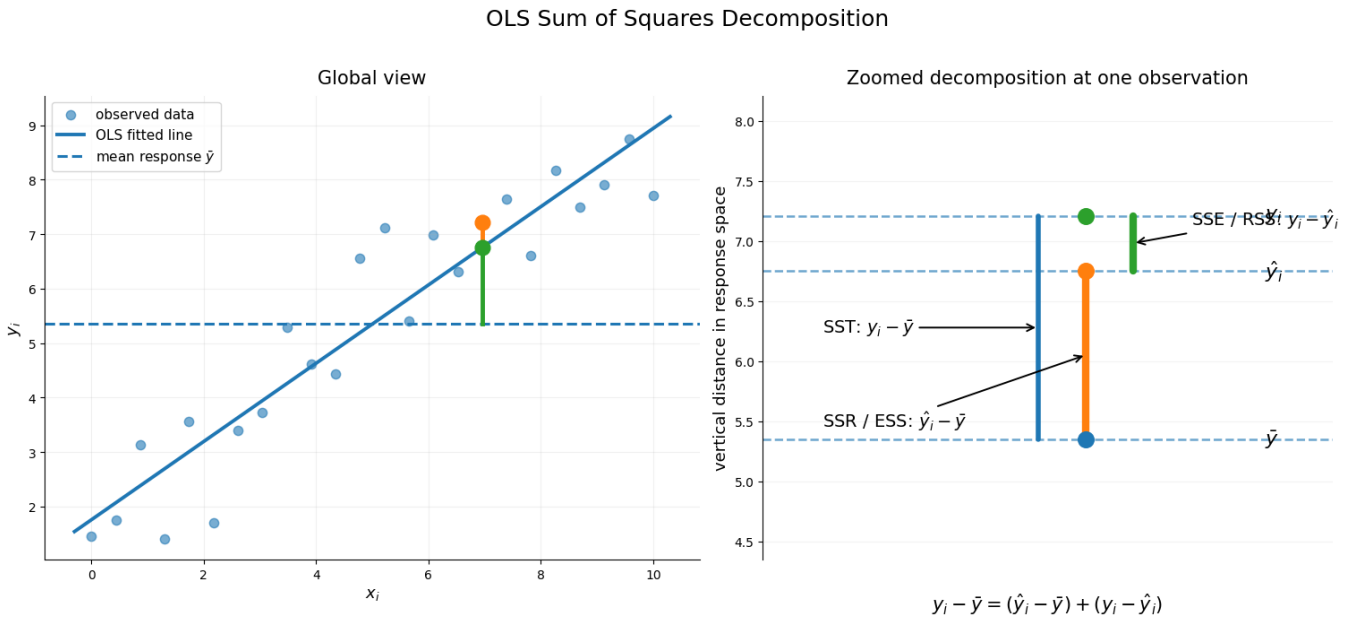

Figure 2: OLS sum of squares decomposition at one observation, showing the total deviation $y_i - \bar{y}$ as the sum of the explained component $\hat{y}_i - \bar{y}$ and the residual component $y_i - \hat{y}_i$.

Thus, the total variability in the response variable can be separated into the variability explained by the regression model and the variability left unexplained by the model. The total sum of squares measures how far the observed responses y_i are from the baseline predictor $\bar{y}$ which is simply predicting every observation using the sample mean.

OLS attempts to construct a regression line that explains as much variation in the response as possible. If the fitted values $\hat{y}_i$ are substantially different from the mean response $\bar{y}$ in the correct direction, then the regression line explains a large portion of the variation in the data leading to a larger regression sum of squares. Conversely, if the observed responses y_i remain far from the fitted values $\hat{y}_i$, then the residual sum of squares is large indicating the model leaves a significant amount of variance unexplained.

Let's first prove that total sum of squares is the sum of the residual sum of squares and explained sum of squares.

Proof. We claim that $\text{TSS} = \text{RSS} + \text{ESS}$.

$$\begin{aligned}
\text{TSS} &= \sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
&= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\
&= \text{RSS} + \text{ESS} + 2\sum_i r_i(\hat{y}_i - \bar{y}) \\
&= \text{RSS} + \text{ESS} + 2\hat{\beta}_1 \cancel{\sum_i r_i x_i} + 2\hat{\beta}_0 \cancel{\sum_i r_i} - 2\bar{y}\cancel{\sum_i r_i} \\
&= \text{RSS} + \text{ESS}.
\end{aligned}$$

Recall that the residuals are $r_i = y_i - \hat{y}_i$ and also in simple linear regression $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Distributing the residuals across the sum and applying OLS orthogonality conditions make the term multiplying $2\hat{\beta}_1$ 0. Therefore, the entire cross-term vanishes and we are left with the sum of the RSS and ESS. □

What are the distributions of the residual sum of squares and explained sum of squares?

Proof. We claim that the explained sum of squares scaled by σ^2 (true unknown variance) is a χ^2_1 random variable.

$$\text{ESS} = \sum_i (\hat{y}_i - \bar{y})^2 = \sum_i (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 = \sum_i \left(\hat{\beta}_1 (x_i - \bar{x})\right)^2 = \hat{\beta}_1^2 \sum_i (x_i - \bar{x})^2 = \hat{\beta}_1^2 S_{xx}.$$

Recall that we know that $\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \sigma^2/S_{xx})$ and that the test statistica $\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim \mathcal{N}(0,1)$. If we are testing the null hypothesis $\beta_1 = 0$ then the test-statistic simplies to,

$$\frac{\hat{\beta}_1 - 0}{\sigma/\sqrt{S_{xx}}} = \frac{\hat{\beta}_1 \sqrt{S_{xx}}}{\sigma} \sim \mathcal{N}(0,1).$$

Recall previously we had that $\text{ESS} = \hat{\beta}_1^2 S_{xx}$. This is exactly,

$$\left(\frac{\hat{\beta}_1 \sqrt{S_{xx}}}{\sigma}\right)^2 = \frac{\hat{\beta}_1^2 S_{xx}}{\sigma^2} = \frac{\text{ESS}}{\sigma^2}.$$

The square of a standard normal random variable is a χ^2_1 random variable. Therefore $\frac{\text{ESS}}{\sigma^2} \sim \chi^2_1$. □

Since we have the distribution of the explained sum of squares, then what is the distribution of the total sum of squares?

Proof. We claim that $\frac{\text{TSS}}{\sigma^2} \sim \chi^2_{n-1}$.

Again, under the null hypothesis $H_0 : \beta_1 = 0$ and for normally distributed responses $y_1, \ldots, y_n \sim \mathcal{N}(\beta_0, \sigma^2)$. We can standardize our responses to $\frac{y_i - \beta_0}{\sigma} \sim \mathcal{N}(0,1)$. Thus by definition of chi-squared random variables,

$$\sum_i \left(\frac{y_i - \beta_0}{\sigma}\right)^2 = \sum_i \frac{(y_i - \beta_0)^2}{\sigma^2} \sim \chi^2_n.$$

We know that $\bar{y} \sim \mathcal{N}(\beta_0, \sigma^2/n)$ so $\frac{\bar{y} - \beta_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$. Therefore the square of $\frac{\bar{y} - \beta_0}{\sigma/\sqrt{n}}$ is simply a χ^2_1 random variable. Following from this, our expansion and simplification of the total sum of squares is simply,

$$\begin{aligned}
\text{SST} &= \sum_{i=1}^N (y_i - \bar{y})^2 = \sum_{i=1}^N \left((y_i - \beta_0) - (\bar{y} - \beta_0)\right)^2 \\
&= \sum_{i=1}^N (y_i - \beta_0)^2 - 2\sum_{i=1}^N (y_i - \beta_0)(\bar{y} - \beta_0) + \sum_{i=1}^N (\bar{y} - \beta_0)^2 \\
&= \sum_{i=1}^N (y_i - \beta_0)^2 - 2(\bar{y} - \beta_0)\sum_{i=1}^N (y_i - \beta_0) + N(\bar{y} - \beta_0)^2 \\
&= \sum_{i=1}^N (y_i - \beta_0)^2 - 2N(\bar{y} - \beta_0)^2 + N(\bar{y} - \beta_0)^2 \\
&= \sum_{i=1}^N (y_i - \beta_0)^2 - N(\bar{y} - \beta_0)^2.
\end{aligned}$$

Scaling each term on the LHS and RHS of the final equation above gives us,

$$\frac{\text{TSS}}{\sigma^2} = \frac{1}{\sigma^2}\sum_i (y_i - \beta_0)^2 - \frac{1}{\sigma^2} N(\bar{y} - \beta_0)^2 \iff \sum_i \frac{(y_i - \beta_0)^2}{\sigma^2} = \frac{\text{TSS}}{\sigma^2} + \left(\frac{y - \beta_0}{\sigma/\sqrt{N}}\right)^2.$$

Clearly the LHS of the final equation above is a χ^2_n random variable since it is the sum of n squared standard normal random variables. The RHS is the sum of a random variable and a χ^2_1 random variable. Since the degrees of freedom on the RHS must add up to the degrees of freedom of the LHS, $\frac{TSS}{\sigma^2}$ must be a χ^2_{n-1} random variable. This is also given by the fact we have a constraint of $\sum_i (y_i - \bar{y}) = 0$. Moreover, since the two random variables on the LHS are independent random variables, we can apply Cochran's Orthogonality Independence Theorem to get the same result. □

Proof. We claim that $\text{RSS}/\sigma^2 \sim \chi^2_{n-2}$ and that RSS is independent of ESS.

Recall that

$$\text{TSS} = \text{RSS} + \text{ESS}.$$

Dividing through by σ^2 gives

$$\frac{\text{TSS}}{\sigma^2} = \frac{\text{RSS}}{\sigma^2} + \frac{\text{ESS}}{\sigma^2}.$$

Now,

$$\frac{\text{TSS}}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \bar{y})^2 \sim \chi^2_{n-1},$$

since centering by $\bar{y}$ removes one degree of freedom.

Further, from the previous derivation,

$$\text{ESS} = \hat{\beta}_1^2 S_{xx}.$$

Since

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right),$$

under the null hypothesis $\beta_1 = 0$ we have

$$\hat{\beta}_1 \sim \mathcal{N}\left(0, \frac{\sigma^2}{S_{xx}}\right).$$

Therefore,

$$\frac{\hat{\beta}_1}{\sigma/\sqrt{S_{xx}}} \sim \mathcal{N}(0,1).$$

Squaring both sides,

$$\frac{\hat{\beta}_1^2}{\sigma^2/S_{xx}} = \frac{\hat{\beta}_1^2 S_{xx}}{\sigma^2} = \frac{\text{ESS}}{\sigma^2} \sim \chi^2_1.$$

Since

$$\frac{\text{TSS}}{\sigma^2} = \frac{\text{RSS}}{\sigma^2} + \frac{\text{ESS}}{\sigma^2},$$

with

$$\frac{\text{TSS}}{\sigma^2} \sim \chi^2_{n-1}, \qquad \frac{\text{ESS}}{\sigma^2} \sim \chi^2_1,$$

and RSS independent of ESS, the additivity property of independent chi-squared random variables implies

$$\frac{\text{RSS}}{\sigma^2} \sim \chi^2_{(n-1)-1} = \chi^2_{n-2}.$$

Therefore,

$$\frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-2},$$

and RSS is independent of ESS. □

2.1.8 F-statistic

From the result above, we know that $\frac{ESS}{\sigma^2}$ and $\frac{RSS}{\sigma^2}$ are chi-squared random variables with 1 and $n-2$ degrees of freedom respectively. Recall from Chapter 1 where the ratio of two chi-squared random variables both scaled by their own degrees of freedom is a random variable with a F distribution, with parameterized specified by the respective degrees of freedom. Applying that logic here, we can derive the F statistic,

$$F = \frac{\frac{ESS}{\sigma^2}/1}{\frac{SSE}{\sigma^2}/n-2} = \frac{ESS}{RSS/n-2} = \frac{MSR}{MSE} \sim F(1, n-2).$$

We can see that the F statistic is simply the ratio between the mean squared regression and mean squared error, or the ratio between the variation explained by regression per regression degree of freedom and the unexplained variation per residual degree of freedom. Distributionally, we can observe that we are sampling from the long right tail of the F distribution, That is, we test $H_0 : \beta_1 = 0$ against the alternative hypothesis $H_\alpha : \beta \neq 0$ and

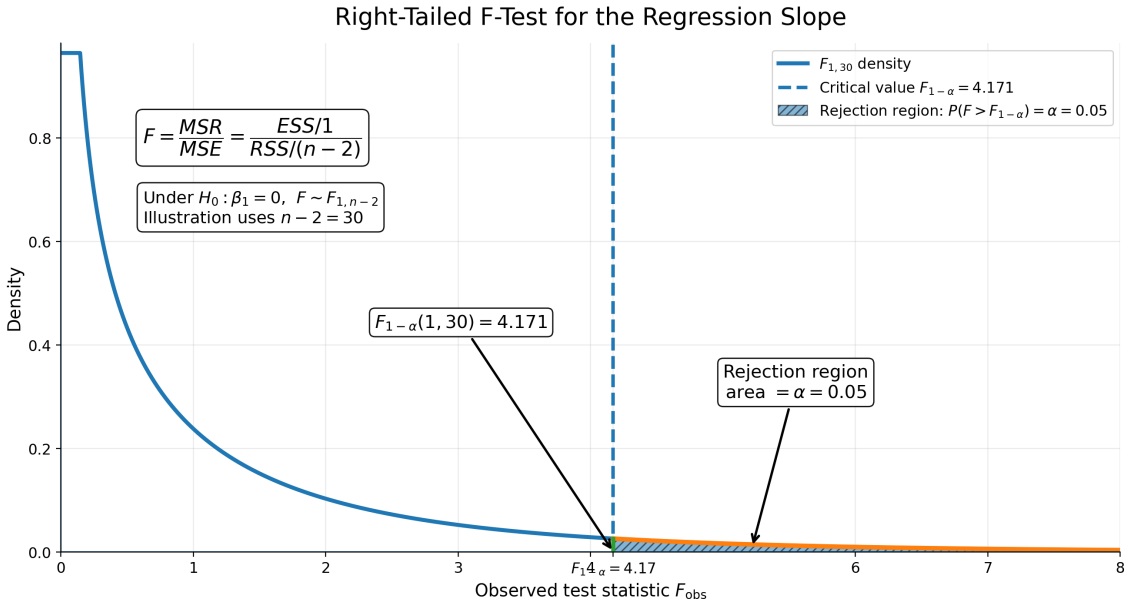

Figure 3: Right-tailed F-test for the regression slope in simple linear regression. Under $H_0 : \beta_1 = 0$, the statistic $F = MSR/MSE = (ESS/1)/(RSS/(n-2))$ follows an $F_{1,n-2}$ distribution. The shaded right tail is the rejection region at significance level α, with critical value $F_{1-\alpha}(1, n-2)$.

reject the null hypothesis at α if $F > F_\alpha(1, n-2)$ which is the upper quantile (right tail) of the F distribution.

How is this related to t-tests? Recall that previously we had used the Student-t test statistic to test $H_0 : \beta_1 = \beta_1^*$, that is $\frac{\hat{\beta}_1 - \beta_1^*}{\text{SE}(\hat{\beta}_1)} \sim t_{n-2}$. In univariate regression, testing $H_0 : \beta_1 = 0$ using the t test is the same as using the F-test. Thus if $X \sim t_{n-2}$ then $X^2 \sim F(1, n-2)$. That is,

$$\frac{\hat{\beta}_1}{\sqrt{\text{Var}(\hat{\beta}_1)}} \sim t_{n-2} \quad \text{under } H_0 : \beta_1 = 0 \iff \left(\frac{\hat{\beta}_1}{\sqrt{\text{Var}(\hat{\beta}_1)}}\right)^2 \sim F_{1,n-2} \iff \frac{MSR}{MSE} = \frac{ESS/1}{RSS/(n-2)} \sim F_{1,n-2}.$$

2.1.9 R^2 Coefficient of Determination

After fitting a regression line, we now have predictors $\hat{y}_i$ that depend on the explanatory variable x_i. Some of the variation in y is now explained by the regression line while the remaining variation is unexplained residual noise. This is the ESS or explained sum of squares as we introduced previously, and the unexplained variation is measured by the residual sum of squares. The natural question to ask now is, what proportion of the total variation in the response variable is explained by the regression model?

We will express ESS as an expression that will make the coefficient of determination more palatable.

$$ESS = \sum_i (\hat{y}_i - \bar{y})^2 = \sum_i (\beta_0 + \hat{\beta}_i x_i - \bar{y})^2 = \sum_i (\hat{\beta}_1 (x_i - \bar{x}))^2 = \hat{\beta}_1^2 \sum_i (x_i - \bar{x})^2 = \hat{\beta}_1 S_{xx}.$$

Then, by the definition of R^2 we simplify to get,

$$R^2 = \frac{ESS}{TSS} = \frac{\hat{\beta}_1^2 S_{xx}}{S_{yy}} = \left(\frac{S_{xx}}{S_{yy}}\right)^2 \frac{S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx} S_{yy}} = r_{xy}^2.$$

We see that in univariate linear regression, the coefficient of determination is exactly the square of the sample correlation between x and y. First, R^2 is fundamentally measuring the strength of the linear relationship between the explanatory variable and the response variable. Since the correlation coefficient measures linear association, squaring it gives the proportion of variation in y explained by the linear relationship with x.

2.2 Multiple Linear Regression

You may notice that a univariate model does not always suffice since we may wish to test the relationship between multiple predictors against a target. Suppose you are a systematic quant who wishes to see if different rolling volatilities and features have a relationship to the forward mid price change of ES futures on a given day. It would be quite computationally inefficient to compute separate univariate regression and do t or F tests on each individual univariate regression separately.

Fortunately, as mathematicians, we have tools such as matrices to deal with multiple features. Suppose that I am that quant and I want to run one (multiple) regression to quantify the change in the response when the predictor X_j increases by one unit when all other predictors are held constant. How might I construct/formalize this problem?

2.2.1 Least Squares Estimation

The tool we are looking for is a feature matrix, where the individual columns of the feature matrix are predictors X_j's corresponding possibly to the features we have engineered for our ES futures mid price prediction problem. To formalize this, suppose that our feature matrix is denoted $X \in \mathbb{R}^{N \times (p+1)}$, that is, a N rows (observations) by $(p+1)$ columns (number of features) matrix, and a response vector $Y \in \mathbb{R}^N$ possibly the forward mid price changes of ES futures. We will also let our coefficient vector be $\beta \in \mathbb{R}^{(p+1)}$ associated with error (noise) vector $\varepsilon \in \mathbb{R}^N$. Then, our linear regression model is,

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}}_{\varepsilon}.$$

Notice how each row of the feature matrix corresponds to the ith sample for $i = 0, 1, \ldots, N$ and each column represents the jth predictor (feature). Each sample (ik) corresponds to a single response (target) y_i in the column vector of responses Y. In machine learning theory, we often refer to our response as our target and predictor(s) as our feature(s). I will use these terms interchangeably in this text.

Similar to unvariate regression, we also must minimize the residual sum of squares to get our OLS estimates. Our criterion remains conceptually the same, but since we are now dealing with matrices and vectors, we must manipulate the formulation of our residual sum of squares.

$$\text{RSS}(\beta) = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2.$$

Since we have let Y be our vector of responses and each x_{ik} corresponds to the ith sample for the k predictor in our feature matrix, the residual sum of squares function can also be written as,

$$\text{RSS}(\beta) = (Y - X\beta)^\top (Y - X\beta).$$

Expanding using basic matrix algebra gives us,

$$Y^\top Y - Y^\top X\beta - \beta^\top X^\top Y + \beta^\top X^\top X\beta = Y^\top Y - 2Y^\top X\beta + \beta^\top X^\top X\beta.$$

Similarly, we must minimize in the direction of the global minimum, so we compute the gradient vector i.e., the derivative with respect to β. Notice that $Y^\top X\beta = \beta^\top X^\top Y$ is a scalar value. From matrix derivative identities, we have

$$\frac{\partial}{\partial Y}(C^\top Y) = C \qquad \text{and} \qquad \frac{\partial}{\partial Y}(Y^\top AY) = (A + A^\top)Y.$$

If A is symmetric, this simplifies to

$$\frac{\partial}{\partial Y}(Y^\top AY) = 2AY.$$

Then, taking the derivative of $\text{RSS}(\beta)$ with respect to the vector β and setting it equal to 0 gives,

$$\frac{d\text{RSS}(\beta)}{d\beta} = -2(X^\top Y) + 2X^\top X\beta = 0.$$

After solving for β, we get the *Normal Equation*,

$$X^\top X\beta = X^\top Y \iff \hat{\beta} = (X^\top X)^{-1}X^\top Y.$$

Note that $X^\top X$ needs to have full column rank (i.e., $\text{rank}(X^\top X) = p + 1$) in order for it to be invertible. If $X^\top X$ is not full column rank, we refer to it as being singular which leads to ill-conditioned betas.

Theorem 2.9: Normal Equation

For a feature matrix $X \in \mathbb{R}^{N\times(p+1)}$, response (target) vector $Y \in \mathbb{R}^N$ and a corresponding vector of coefficients $\beta \in \mathbb{R}^{p+1}$, the Normal Equation for the OLS estimate $\hat{\beta}$ is written as,

$$\hat{\beta} = (X^\top X)^{-1}X^\top y,$$

provided that $X^\top X$ is an invertible matrix.

We claim that any OLS minimizer $\hat{\beta}$ satisfies $X^\top X\hat{\beta} = X^\top Y$. If $X^\top X$ is invertible then the minimizer is unique and $\hat{\beta} = (X^\top X)^{-1}X^\top Y$. Therefore we claim that $X^\top X$ is invertible if and only if $\text{rank}(X) = p + 1$.

Proof. Recall that the OLS objective is given by,

$$S(\beta) = (Y - X\beta)^\top (Y - X\beta).$$

Expanding this we get,

$$\begin{aligned} S(\beta) &= (Y - X\beta)^\top (Y - X\beta) \\ &= Y^\top Y - Y^\top X\beta - (X\beta)^\top Y + (X\beta)^\top X\beta \\ &= Y^\top Y - Y^\top X\beta - \beta^\top X^\top Y + \beta^\top X^\top X\beta. \end{aligned}$$

Since $Y^\top X\beta$ is scalar, it must be equal to $Y^\top X\beta = \beta^\top X^\top Y$ trivially. Therefore $S(\beta) = Y^\top Y - 2\beta^\top X^\top Y + \beta^\top X^\top X\beta$. Differentiating this with respect to β, given that $Y^\top Y$ too is a constant, and setting equal to 0 gives,

$$-2X^\top Y + 2X^\top X\hat{\beta} = 0 \iff X^\top X\hat{\beta} = X^\top Y.$$

Suppose that $X^\top X$ is invertible. Then, we can multiply both sides by $(X^\top X)^{-1}$ giving the normal equation for $\hat{\beta}$. If two vector $\hat{\beta}_1$ and $\hat{\beta}_2$ both solve the normal equation then,

$$X^\top X\hat{\beta}_1 = X^\top Y = X^\top X\hat{\beta}_2.$$

Subtracting shows that $X^\top X(\hat{\beta}_2 - \hat{\beta}_1) = 0$. So it must be the case that $\hat{\beta}_1 = \hat{\beta}_2$. Therefore the OLS minimizer $\hat{\beta}$ is unique. Now we prove that $X^\top X$ is invertible iff $\text{rank}(X) = p + 1$.

$X^\top X$ and X have the same null space. Suppose that $X^\top Xv = 0$. Multiplying on the left by $v^\top$ gives $v^\top X^\top Xv = 0$.

However notice that,

$$v^\top X^\top Xv = (Xv)^\top (Xv) = \|Xv\|^2.$$

Hence, we have that $\|Xv\|^2 = 0$ which implies that $Xv = 0$. So every vector in $\text{Null}(X^\top X)$ is also in $\text{Null}(X)$. Conversely if $Xv = 0$ then multiplying by $X^\top$ gives $X^\top Xv = 0$ So every vector in $\text{Null}(X)$ is also in $\text{Null}(X^\top X)$. Therefore we conclude that $\text{Null}(X^\top X) = \text{Null}(X)$.

Observe that $X^\top X$ is a square $(p+1) \times (p+1)$ matrix. Recall that a square matrix is invertible if and only if its null space is trivial. Thus the statement we aim to prove is $X^\top X$ is invertible if and only if $\text{Null}(X^\top X) = \{0\}$.

Since $\text{Null}(X^\top X) = \text{Null}(X)$, this is equivalent to saying that $\text{Null}(X) = \{0\}$. By Rank-Nullity theorem we have that,

$$\dim(\text{Null}(X)) + \text{rank}(X) = p + 1.$$

Recall that $\text{Null}(X) = \{0\}$ if and only if $\dim(\text{Null}(X)) = 0$ if and only if $\text{rank}(X) = p + 1$, where all three statements are equivalent. Therefore, $X^\top X$ is invertible if and only if $\text{rank}(X) = p + 1$. □

2.2.2 Properties of $\hat{\beta}$

Similar to univariate regression, the assumption of $\hat{\beta}$ still hold. $\hat{\beta}$ is still BLUE, is unbiased, and still possesses constant variance. We will prove these facts now.

Theorem 2.10: Properties of the OLS Estimator

Suppose that the linear model

$$Y = X\beta + \varepsilon$$

satisfies the Gauss–Markov assumptions with

$$\mathbb{E}[\varepsilon] = 0 \quad \text{and} \quad \text{Var}(\varepsilon) = \sigma^2 I.$$

If $X^\top X$ is invertible, then the ordinary least squares estimator

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

satisfies

$$\mathbb{E}[\hat{\beta}] = \beta$$

and

$$\text{Var}(\hat{\beta}) = \sigma^2 (X^\top X)^{-1}.$$

In particular, $\hat{\beta}$ is an unbiased estimator of β with constant variance.

Proof. We claim that $\hat{\beta}$ is unbiased, that is $\mathbb{E}[\hat{\beta}] = \beta$.

$$\begin{aligned}
\mathbb{E}[\hat{\beta}] &= \mathbb{E}[(X^\top X)^{-1} X^\top Y] \\
&= (X^\top X)^{-1} X^\top \mathbb{E}[Y] \\
&= (X^\top X)^{-1} X^\top X\beta \\
&= \beta.
\end{aligned}$$

Similarly, we claim that $\hat{\beta}$ has constant variance,

$$\begin{aligned}
\text{Var}(\hat{\beta}) &= \text{Var}((X^\top X)^{-1} X^\top Y) = (X^\top X)^{-1} \text{Var}(Y) X[(X^\top X)^{-1}] \\
&= (X^\top X)^{-1} X^\top \sigma^2 I X (X^\top X)^{-1} = \sigma^2 (X^\top X)^{-1} X^\top X (X^\top X)^{-1} = \sigma^2 (X^\top X)^{-1}.
\end{aligned}$$

□

Note that the last line for the variance of $\hat{\beta}$ follows because $[(X^\top X)^{-1}]^\top = [(X^\top X)^\top]^{-1} = (X^\top X)^{-1}$. Recall for square matrices we know that $A^{-1}A = I$ implies that $(A^{-1}A)^\top = A^\top (A^{-1})^\top = I$ which means that $(A^\top)^{-1} = (A^{-1})^\top$. There are some interesting properties regarding the matrix that multiplies the response vector Y. More specificaly, the hat matrix is denoted as $H = X(X^\top X)^{-1}X^\top$.

The hat matrix is an orthogonal projection operator onto the column space of X. OLS chooses $\hat{Y}$ to be the point $\mathcal{C}(X)$ is closest to Y in Euclidean distance. Thus $\hat{Y} = \text{proj}_{\mathcal{C}(X)}(Y)$ so is a matrix representation of orthogonal project onto the column space of X.

Proof. We claim that the hat matrix $H = X(X^\top X)^{-1}X^\top$ is idempotent.

$$HH = H^2 = X(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top = X(X^\top X)^{-1}X^\top = H.$$

Similarly, we claim that H is symmetric i.e., that $H = H^\top$. □

2.2.3 Residuals

Recall that the hat matrix (orthogonal projector matrix onto the column space X) is idempotent and symmetric. Then, let $r = Y - \hat{Y} = (I - H)Y$ be our residual vector. It is clear that r is orthogonal to X and Y since $\sum_i r_i = 0$. So we say that $X^\top r = 0$ and $\hat{Y}^\top r = 0$. The expectation $\mathbb{E}[r] = 0$ and $\text{Var}(r) = \sigma^2(I - H)$.

Proof. We claim that the residual vector r defined as $Y - \hat{Y}$ is orthogonal to the feature matrix $X \in \mathbb{R}^{N\times(p+1)}$.

Using the definition of the residuals r we will get,

$$X^\top r = X^\top (Y - \hat{Y}) = X^\top Y - X^\top HY = X^\top Y - X^\top X(X^\top X)^{-1}X^\top Y = X^\top Y - X^\top Y = 0.$$

Clearly $X^\top r$ is the matrix written as,

$$X^\top r = \begin{bmatrix} \sum_i r_i & \cdots & \sum_i X_i r_i \end{bmatrix}^\top = 0.$$

So it is clear that,

$$\hat{Y}^\top r = (\hat{\beta}^\top X^\top) r = \hat{\beta}^\top (X^\top r) = \hat{\beta} \cdot \vec{0} = 0.$$

The computation of the expectation follows directly,

$$\mathbb{E}[r] = \mathbb{E}[(I - H)Y] = (I - H)\mathbb{E}[Y] = (I - H)X\beta = X\beta - X(X^\top X)^{-1}X^\top X\beta = X\beta - X\beta = 0.$$

For variance, recall for a matrix if we want $\text{Var}(AY) = A\,\text{Var}(Y)A^\top$ then,

$$\begin{aligned} \text{Var}(r) &= \text{Var}((I - H)Y) \\ &= (I - H)\,\text{Var}(Y)(I - H)^\top \\ &= (I - H)\sigma^2 I(I - H)^\top \\ &= (I - H)\sigma^2 I(I - H) \\ &= \sigma^2(I - H)(I - H) \\ &= \sigma^2(I - H). \end{aligned}$$

We use the fact that H is symmetric and idempotent. □

Recall that we have $p + 1$ constraints given by the fact that the residuals are to be orthogonal to the design matrix, that is $X^\top_{N\times(p+1)}$. So, given this constraint, we lose $p + 1$ degrees of freedom.

An interesting result from linear algebra regarding residuals can be seen below. If we wish to compute the mean of our residual sum of squares, we can use the trace to sum along the diagonals of the variance of the residuals.

$$\mathbb{E}\left[\sum_i r_i^2\right] = \mathbb{E}[r^\top r] = \mathbb{E}[tr(rr^\top)] = tr(\mathbb{E}[rr^\top]) = tr(\text{Var}(r)).$$

Note that we can write $\text{Var}(r) = \mathbb{E}[(r - \mathbb{E}[r])(r - \mathbb{E}[r])^\top] = 0$. Therefore, by properties of matrices, we can re-order cyclic permutations via properties of trace.

$$\begin{aligned} tr(\text{Var}(r)) &= tr(\sigma^2(I - H)) = \sigma^2[tr(I) - tr(H)] = \sigma^2[n - tr(X(X^\top X)^{-1}X^\top)] \\ &= \sigma^2[n - tr(X^\top X(X^\top X)^{-1})] = \sigma^2[n - (p+1)]. \end{aligned}$$

2.2.4 Sampling Distribution of $\hat{\beta}$ and $\hat{\sigma}^2$

Since we are dealing with matrices now, our response vector Y follows a multivariate normal distribution with mean $X\beta$ and constant variance $\sigma^2 I$. Under normality, our least squares estimates still hold the same properties.

Theorem 2.11: Properties of $\hat{\beta}$

Under normality assumptions, $\hat{\beta}$ has the following properties:

(i) $\hat{\beta} \sim \text{MVN}(\beta, \sigma^2(X^\top X)^{-1})$

(ii) $\hat{\beta}$ and $\hat{\sigma}^2$ are independent.

(iii) $\frac{(N-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi_{N-p-1}$

Proof. We claim that the multiple least squares estimate $\hat{\beta}$ is an unbiased, homoskedastic estimator.

First note that the least squares estimate $\hat{\beta}$ can be written as,

$$\hat{\beta} = (X^\top X)^{-1}X^\top Y = (X^\top X)^{-1}X^\top(X\beta + \varepsilon) = \beta + (X^\top X)^{-1}X^\top\varepsilon.$$

Taking expectations and observing that $(X^\top X)^{-1}X^\top\varepsilon$ is a scalar,

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}[(X^\top X)^{-1}X^\top Y] = \mathbb{E}[(X^\top X)^{-1}X^\top(X\beta + \varepsilon)] = \mathbb{E}[\beta] + \mathbb{E}[(X^\top X)^{-1}X^\top\varepsilon].$$

Note that $\mathbb{E}[\varepsilon] = 0$ so the entire term $\mathbb{E}[(X^\top X)^{-1}X^\top\varepsilon] = 0$. Therefore $\mathbb{E}[\hat{\beta}] = \mathbb{E}[\beta] = \beta$ itself. Therefore, we have shown that the multiple least squares estimate $\hat{\beta}$ is β showing that the OLS estimator $\hat{\beta}$ is indeed unbiased.

We claim that $\hat{\beta}$ has constant variance $\sigma^2(X^\top X)^{-1}$.

$$\begin{aligned} \text{Var}(\hat{\beta}) &= \text{Var}((X^\top X)^{-1}X^\top Y) \\ &= \text{Var}((X^\top X)^{-1}X^\top(X\beta + \varepsilon)) \\ &= \text{Var}((X^\top X)^{-1}X^\top X\beta + (X^\top X)^{-1}X^\top\varepsilon) \\ &= \text{Var}(\beta + (X^\top X)^{-1}X^\top\varepsilon) \\ &= \text{Var}((X^\top X)^{-1}X^\top\varepsilon) \\ &= (X^\top X)^{-1}X^\top \text{Var}(\varepsilon)X[(X^\top X)^{-1}X^\top]^\top \\ &= \sigma^2(X^\top X)^{-1}X^\top X(X^\top X)^{-1} \\ &= \sigma(X^\top X)^{-1}. \end{aligned}$$

Therefore, $\hat{\beta}$ has constant variance defined as $\sigma^2(X^\top X)^{-1}$. □

Proof. We claim that $\hat{\beta}$ and $\hat{\sigma}^2$ are independent.

We can show independence by showing that the two vectors are orthogonal. Recall that $r = (I - H)Y \sim \text{MVN}(0, \sigma^2(I - H))$ and $\hat{\beta} = (X^\top X)^{-1}X^\top Y \sim \text{MVN}(\beta, \sigma^2(X^\top X)^{-}1)$. To show orthogonality, we can show that $\text{Cov}(\hat{\beta}, r)$ is equal to 0. This is because if a random vector is jointly multivariate normal, then they are uncorrelated if and only if they are independent. So in this linear model, we can exploit the Gaussian structure.

$$\begin{aligned}
\text{Cov}(\hat{\beta}, r) &= \text{Cov}((X^\top X)^{-1}X^\top Y, (I - H)Y) \\
&= (X^\top X)^{-1}X^\top \text{Var}(Y)(I - H)^\top \\
&= (X^\top X)^{-1}X^\top \sigma^2 I(I - H) \\
&= \sigma^2(X^\top X)^{-1}X^\top(I - X(X^\top X)^{-1}X^\top) \\
&= \sigma^2(X^\top X)^{-1}X^\top - \sigma^2(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top \\
&= \sigma^2(X^\top X)^{-1}X^\top - \sigma^2(X^\top X)^{-1}X^\top \\
&= \sigma^2 \cdot 0 = 0.
\end{aligned}$$

Clearly we see that the covariance between $\hat{\beta}$ and r is zero, so we claim that they are uncorrelated so they are therefore independent. Therefore $\hat{\beta} \perp r$ meaning that $\hat{\sigma}^2 = \frac{r^\top r}{N-(p+1)}$ is a function of the residuals of our model. Independence of $\hat{\beta}$ and r implies independence of $\hat{\beta}$ and r. □

Proof. We claim that $\frac{(N-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{N-p-1}$.

Recall the definition of $\hat{\sigma}^2$ i.e., it is the RSS divided by it's degrees of freedom.

$$\hat{\sigma}^2 = \frac{r^\top r}{N - p - 1}.$$

Re-arranging we get that,

$$\frac{(N - p - 1)\hat{\sigma}^2}{\sigma^2} = \frac{r^\top r}{\sigma^2} = \left(\frac{r}{\sigma}\right)^\top \left(\frac{r}{\sigma}\right) = r^{*\top} r^*.$$

Recall that our responses are assumed to be jointly multivariate normal $Y \sim \text{MVN}(X\beta, \sigma^2 I)$. Also, we previously wrote $r = (I - H)Y$ so dividing by σ gives the r^* identity above. $r^* = \frac{r}{\sigma} = \frac{(I-H)Y}{\sigma}$. We use the general fact for some $Z \sim \text{MVN}(\mu, \Sigma)$ and A is a matrix, then $AZ \sim \text{MVN}(A\mu, A\Sigma A^\top)$.

Letting $Z = Y$ and $A = \frac{I-H}{\sigma}$ and knowing that $Y \sim \text{MVN}(X\beta, \sigma^2 I)$ clearly,

$$r^* \sim \text{MVN}\left(\frac{(I - H)X\beta}{\sigma}, \frac{(I - H)(\sigma^2 I)(I - H)^\top}{\sigma^2}\right).$$

Observe that $(I - H)X\beta = X\beta - HX\beta$. But since $HX = X$ since H projects onto the column space of X then we have that $HX\beta = X\beta$ meaning that $(I - H)X\beta = 0$. Therefore, $\mathbb{E}[r^*] = 0$.

Similarly, since $H = H^\top$ and we define the covariance as,

$$\frac{(I - H)(\sigma^2 I)(I - H)^\top}{\sigma^2} = (I - H)(I - H)^\top,$$

we know that $(I - H)^\top = I - H$. Similarly since H is idempotent, we can simply $(I - H)^2$ by expanding and collecting like terms which gives us $I - H$. Hence $(I - H)(I - H)^\top = I - H$. Since r^* is simply a transformation of a jointly multivariate normal random variable and we have found that its mean is 0 and variance is $I - H$, then we conclude that,

$$r^* = \frac{(I - H)Y}{\sigma} \sim \text{MVN}(0, I - H).$$

□

2.2.5 Inference for a single β_j

Suppose we have a linear regression model where $Y \sim \text{MVN}(X\beta, \sigma^2 I)$ where X has $p+1$ columns because we estimate p slopes plus one intercept. The OLS estimator therefore satisfies $\hat{\beta} \sim \text{MVN}(\beta, \sigma^2(X^\top X)^{-1})$ describing the entire joint distribution of the whole vector $\hat{\beta}$.

Suppose that we are only interested in one coefficient, let's say $\hat{\beta}_{ii}$. Then, we would look at the ith diagonal element of the covariance matrix. Let v_{ii} denote the ith diagonal element of $(X^\top X)^{-1}$ (recall $X^\top X$ is a square matrix). Then, $\text{Var}(\hat{\beta}_i) = \sigma^2 v_{ii}$. Therefore, for a single $\hat{\beta}_i$ we conclude $\hat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma^2 v_{ii})$.

Equivalently, after standardizing we get that,

$$\frac{\hat{\beta}_i - \beta_i}{\sigma^2 v_{ii}} \sim \mathcal{N}(0,1).$$

We usually do not know σ so we replace it with an unbiased estimator $\hat{\sigma}^2 = \frac{RSS}{N-p-1}$. From the residual sum of squares results, we know that $\frac{(N-p-1)\hat{\sigma}^2}{\sigma^2} = \frac{RSS}{\sigma^2} \sim \chi^2_{N-p-1}$. The degrees of freedom are $N-p-1$ because we started N observations and used up to $p+1$ degrees of freedom estimating the regression coefficients.

Recall that $\hat{\beta}_i$ and $\hat{\sigma}^2$ are independent, so standardizing the coefficient gives a statistic that is independent of the chi-squared random variable, that is,

$$\frac{\hat{\beta}_i - \beta_i}{\sigma^2 v_{ii}} \perp \frac{(N-p-1)\hat{\sigma}^2}{\sigma^2}.$$

What does this setup remind you of? Clearly a Student-t distributed random variable!

Let $Z = \frac{\hat{\beta}_i - \beta_i}{\sigma^2 v_{ii}} \sim \mathcal{N}(0,1)$, a standard normal random variable, and let $Y = \frac{(N-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{N-p-1}$ and $n = N-p-1$ (corresponding degrees of freedom). Dividing Z by the square root of Y scaled by it's degrees of freedom gives us a test statistic with a t_{N-p-1} distribution,

$$\frac{\frac{\hat{\beta}_i - \beta_i}{\sqrt{\sigma^2 v_{ii}}}}{\sqrt{\frac{1}{N-p-1} \cdot \frac{(N-p-1)\hat{\sigma}^2}{\sigma^2}}} = \frac{\frac{\hat{\beta}_i - \beta_i}{\sqrt{\sigma^2 v_{ii}}}}{\frac{\hat{\sigma}}{\sigma}} = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma}^2 v_{ii}}} \sim t_{N-p-1}.$$

The denominator is simply the estimated standard error of $\hat{\beta}_i$ so usually we can write $\text{SE}(\hat{\beta}_i) = \sqrt{\sigma^2 v_{ii}}$. This is precisely a test statistic which tests the null hypothesis about the true coefficient β_i. That is, we test $H_0 : \beta_i = \beta_i^*$ where β_i is the value we are testing against. More commonly, we use $H_0 : \beta_i = 0$ which asks whether the ith predictor has a statistically detectable linear effect on Y after controlling for the other predictors in the regression. We often denote this statistic as t_{obs} since it observes the number of estimated standard errors $\hat{\beta}_i$ is away from the null value β_i^*. If t_{obs} is close to 0, then $\hat{\beta}_i$ is close to the null value so the data are consistent with H_0.

Similarly for a two-sided test as we saw previously,

$$H_0 : \beta_1 = \beta_i^* \quad \text{versus} \quad H_A : \beta_i \neq \beta_i^*.$$

We reject H_0 at the significance level α when $|t_{\text{obs}}| > t_{N-p-1,1-\alpha/2}$ i.e., the $1-\alpha/2$ quantile of the Student-t distribution with $N-p-1$ degrees of freedom. So for a 5% two-sided test, we reject when $|t_{\text{obs}}| > t_{N-p-1,0.975}$. Equivalently, we can use the p value defined as $2\mathbb{P}(t_{N-p-1} \geq |t_{\text{obs}}|)$ then reject H_α at level α if the p-value is less than α. A $100(1-\alpha)\%$ confidence interval for β_1 is therefore defined as,

$$\hat{\beta}_i \pm t_{N-p-1,1-\alpha/2} \cdot \text{SE}(\hat{\beta}_i).$$

That is, we reject H_0 exactly when β_i^* is not inside this confidence interval. For the common test $H_0 : \beta_i = 0$ the test statistic becomes $t_{\text{obs}} = \frac{\hat{\beta}_i}{\text{SE}(\hat{\beta}_i)}$ so if t_{obs} is large in magnitude, then $\hat{\beta}_i$ is many standard errors away from zero meaning there is evidence that $\beta_i \neq 0$.

2.2.6 Inference on $\theta = a^\top \beta$

The general case of the same single-coefficient t test idea for any linear combination of coefficient is the following. Suppose that we have a scalar vector $a = [a_0 \quad \cdots \quad a_p]^\top$ which is a $(p+1)$ dimensional vector of constants. Instead of for a single β_i we want to perform inference on $\theta = a^\top \beta = \sum_i a_i \beta_i$ that is estimate $\hat{\theta} = a^\top \hat{\beta}_i$.

Suppose we have a $(p+1)$ dimensional vector $\in \mathbb{R}^{(p+1)}$ given as $a = [0 \quad 1 \quad -1 \quad 0 \quad \cdots \quad 0]^\top$. Then $\theta = a^\top \beta = \beta_1 - \beta_2$. From this, we can test whether two coefficients are equal. In this example, $\beta_1 = \beta_2$ which is equivalent to writing it as $\beta_1 - \beta_2 = 0$.

Now since $\hat{\beta} \sim \text{MVN}(\beta, \sigma^2 (X^\top X)^{-1})$ any linear combination $\hat{\beta}$ is also normally distributed so $\hat{\theta} = a^\top \hat{\beta}$ is normally distributed. Its expectation is therefore trivial to compute,

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}[a^\top \hat{\beta}] = a^\top \mathbb{E}[\hat{\beta}] = a^\top \beta = \theta.$$

Clearly, $\hat{\theta}$ is an unbiased estimator of θ. The variance,

$$\text{Var}(\hat{\theta}) = \text{Var}(a^\top \hat{\beta}) = a^\top \text{Var}(\hat{\beta}) a = a^\top \sigma^2 (X^\top X)^{-1} a = \sigma^2 a^\top (X^\top X)^{-1} a.$$

Therefore $\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2 a^\top (X^\top X)^{-1} a)$. Standardizing $\hat{\theta}$ gives,

$$\frac{\hat{\theta} - \theta}{\sqrt{\sigma^2 a^\top (X^\top X)^{-1} a}} = \frac{\hat{\theta} - \theta}{\sigma \sqrt{a^\top (X^\top X)^{-1} a}} \sim \mathcal{N}(0, 1).$$

Again, since we often do not know the true variance, we replace σ with an unbiased estimate of the standard deviation,

$$\frac{\hat{\theta} - \theta}{S \sqrt{a^\top (X^\top X)^{-1} a}} \sim t_{N-p-1}.$$

$S = \hat{\sigma} = \sqrt{\frac{\text{RSS}}{N-p-1}}$ where RSS is the residual sum of squares. The estimate S therefore has $N - p - 1$ degrees of freedom since the model estimates $p + 1$ coefficients. For hypothesis testing, we write $H_0 : a^\top \beta = \theta_0$ where θ_0 is the hypothesized value. Then our test statistic is simply,

$$t_{\text{obs}} = \frac{a^\top \hat{\beta} - \theta_0}{S \sqrt{a^\top (X^\top X)^{-1} a}} \sim t_{N-p-1}.$$

For a two sided test with $H_0 : a^\top \beta = \theta_0$ and $H_A : a^\top \beta \neq \beta_0$, we reject at significance level α when $|t_{\text{obs}}| > t_{N-p-1,1-\alpha/2}$. The corresponding confidence interval for $\theta = a^\top \beta$ is given by,

$$a^\top \hat{\beta} \pm t_{N-p-1,1-\alpha/2} \cdot S \sqrt{a^\top (X^\top X)^{-1} a}.$$

2.2.7 Prediction Intervals

We often wish to predict a new response value y_p at a new predictor vector $a_p = (1, x_1, \ldots, x_p)^\top$ including the intercept. The true future response at this predictor value is $y_p = a_p^\top \beta + \varepsilon_p$. Here $a_p^\top$ is the true conditional mean of Y at the predictor value and ε_p is the new random error/noise term for the future observation.

We do not know β so we estimate it using $\hat{\beta}$. Therefore our predicted value is a linear combination of the predictor vector and our OLS estimate, that is, $\hat{y}_p = a_p^\top \hat{\beta}$. Thus, our prediction error is quantified by $y_p - \hat{y}_p$,

$$y_p - \hat{y}_p = (a_p^\top \beta + \varepsilon_p) - a_p^\top \hat{\beta} = \varepsilon_p - a_p^\top (\hat{\beta} - \beta).$$

We aim to characterize the distribution of our prediction $a^\top \hat{\beta}$.

Proof. We claim that for a fixed $a_p^\top \beta$ and given that $a^\top \hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 a^\top (X^\top X)^{-1} a)$ the test statistic,

$$\frac{y_p - \hat{y}_p}{S\sqrt{1 + a_p^\top (X^\top X)^{-1} a_p}} \sim t_{N-p-1}.$$

We start by computing the variance of the prediction error,

$$\begin{aligned}
\text{Var}(y_p - \hat{y}_p) &= \text{Var}\left(\varepsilon_p - a_p^\top(\hat{\beta} - \beta)\right) \\
&= \text{Var}(\varepsilon_p) + \text{Var}(a_p^\top \hat{\beta}) \\
&= \sigma^2 + a_p^\top \text{Var}(\hat{\beta}) a_p \\
&= \sigma^2 + \sigma^2 a_p^\top (X^\top X)^{-1} a_p \\
&= \sigma^2(1 + a_p^\top (X^\top X)^{-1} a_p).
\end{aligned}$$

Standardizing the prediction error now given the variance of the prediction error we get,

$$\frac{y_p - \hat{y}_p}{\sigma\sqrt{1 + a_p^\top (X^\top X)^{-1} a_p}} \sim \mathcal{N}(0, 1).$$

The numerator corresponds to the prediction error and the denominator is the true standard deviation of that prediction error. Since we usually do not know σ^2 we replace it with $S^2 = \frac{RSS}{N-p-1}$ which has $N - p - 1$ degrees of freedom since we estimate $p + 1$ coefficients.

Therefore $S = \sqrt{\frac{RSS}{N-p-1}}$. Replacing σ with S the test statistic becomes,

$$\frac{y_p - \hat{y}_p}{S\sqrt{1 + a_p^\top (X^\top X)^{-1} a_p}} \sim t_{N-p-1}.$$

Therefore a $100(1-\alpha)\%$ prediction interval for the new response y_p is given by,

$$\hat{y}_p \pm t_{N-p-1,1-\alpha/2} \cdot S\sqrt{1 + a_p^\top (X^\top X)^{-1} a_p}.$$

Equivalently we have that $\hat{y}_p \pm t_{N-p-1,1-\alpha/2} \cdot \text{SE}(y_p - \hat{y}_p)$. Notice the extra 1 inside the square root. This comes from the variance of the new observation error ε_p which is what makes the prediction interval wider than a confidence interval for the mean response. □

For the mean response at a_p we are estimating only $\mathbb{E}[Y_p \mid a_p] = a_p^\top \beta$ where the uncertainty is only from estimating β. So the standard error is $S\sqrt{s_p^\top (X^\top X)^{-1} a_p}$. But for predicting a new observation y_p we also need to account for the new random error ε_p which is why we have an extra constant 1 term in the square root.

2.2.8 Gauss-Markov Theorem

We claim that the least squares estimates of the parameters β have the smallest variance among all linear unbiased estimates. This claim although important, highlights an important caveat regarding why unbiased estimates may not always be the wisest. Recall previously we used the least squares estimates to estimate $\hat{\theta} = a^\top \hat{\beta} = a^\top (X^\top X)^{-1} X^\top y$ where $X \in \mathbb{R}^{N \times (p+1)}$ and y is a target vector in $\mathbb{R}^N$. Assuming that the linear model is correct, then we saw that $a^\top \hat{\beta}$ is unbiased which means that,

$$\mathbb{E}[a^\top \hat{\beta}] = \mathbb{E}[a^\top (X^\top X)^{-1} X^\top y] = a^\top (X^\top X)^{-1} X^\top X \beta = a^\top \beta.$$

We will show $\text{Var}(a^\top \hat{\beta})$ is always the lower bound for any other arbitrary estimate $\tilde{\theta} = c^\top y$. Note in this case c is also a scalar vector and y is the same target vector.

Proof. We claim that, for any other estimate $\tilde{\theta} = c^\top y$ for scalar vector $c \in \mathbb{R}^N$ and target vector $y \in \mathbb{R}^N$ that is unbiased for $a^\top \beta$, then $\text{Var}(a^\top \hat{\beta}) \leq \text{Var}(c^\top y)$.

Let $\tilde{\theta} = c^\top Y$ be any other linear estimator of $a^\top \beta$ where $c \in \mathbb{R}^N$. Since $\tilde{\theta}$ is also assumed to be unbiased for $a^\top \beta$ we need $\mathbb{E}[c^\top Y] = a^\top \beta$. Using the definition of $Y = X\beta + \varepsilon$ we get,

$$\begin{aligned}\mathbb{E}[c^\top Y] &= c^\top \mathbb{E}[Y] \\ &= c^\top \mathbb{E}[X\beta + \varepsilon] \\ &= c^\top (X\beta + \mathbb{E}[\varepsilon]) \\ &= c^\top X\beta.\end{aligned}$$

Therefore, for $c^\top Y$ to be unbiased for $a^\top \beta$ we need $c^\top X\beta = a^\top \beta$ for every possible β. Hence we have that $c^\top X = a^\top$ or equivalently $X^\top c = a$. So the condition $X^\top c = a$ is exactly the condition that $c^\top Y$ is unbiased for $a^\top \beta$. Now, we write the OLS estimator $a^\top \hat{\beta}$ as a linear estimator of Y. Since $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ then we have that,

$$a^\top \hat{\beta} = a^\top (X^\top X)^{-1} X^\top Y = [X(X^\top X)^{-1} a]^\top Y.$$

So $a^\top \hat{\beta}$ has the form $a^\top \hat{\beta} = d^\top Y$. Observe that d is the special vector of weights that OLS uses to estimate $a^\top \beta$. You can check as an exercise that $d^\top Y$ is also unbiased that is $X^\top d = a$. Therefore $d^\top Y = a^\top \hat{\beta}$ satisfies the same unbiasedness condition as $c^\top Y$. Since both $c^\top Y$ and $d^\top Y$ are both unbiased, then $X^\top c = a = X^\top d$. Subtracting gives $X6\top(c - d) = 0$. Let $r = c - d$. Then we have that $X^\top r = 0$.

By definition then r is orthogonal to every column in X. d therefore lies in $\text{Col}(X)$ where r lies in $\text{Col}(X)^\perp$ so we have decomposed $c = d + r$ into an orthogonal part in the column space of X and an orthogonal residual portion. So, d and r are orthogonal.

It remains for us to compute the the variances. For the OLS estimator,

$$\text{Var}(a^\top \hat{\beta}) = \text{Var}(d^\top Y) = d^\top \text{Var}(Y) d = d^\top (\sigma^2 I_N) d = \sigma^2 d^\top d.$$

For the arbitrary unbiased estimator $c^\top Y$ we have,

$$\text{Var}(c^\top Y) = c^\top \text{Var}(c) = c^\top (\sigma^2 I_N) c = \sigma^2 c^\top c.$$

However, recall that $c = d + r$,

$$c^\top c = (d + r)^\top (d + r) = d^\top d + d^\top r + r^\top d + r^\top r = d^\top d + r^\top r.$$

Since $r^\top r = \|r\|^2 \geq 0$ we get that $c^\top c \geq d^\top d$. Multiplying both sides of the inequality by $\sigma^2 > 0$ we get that $\sigma^2 c^\top c \geq \sigma^2 d^\top d$. Therefore we have that,

$$\text{Var}(c^\top Y) \geq \text{Var}(d^\top Y).$$

Since $d^\top Y = a^\top \hat{\beta}$ this becomes $\text{Var}(c^\top Y) \geq \text{Var}(a^\top \hat{\beta})$. □

Recall the mean squared error of an estimator $\tilde{\theta}$ in estimating θ,

$$\text{MSE}(\tilde{\theta}) = \mathbb{E}[\tilde{\theta} - \theta]^2 = \text{Var}(\tilde{\theta}) + [\mathbb{E}[\tilde{\theta}] - \theta]^2.$$

The Gauss-Markov Theorem claims that the least squares estimate is the best unbiased linear estimator, that is, it has the smallest mean squared error of all unbiased linear estimators. There very well may be an estimator that trades bias for a larger reduction in variance. As you recall, mean squared error is directly related to prediction accuracy. Suppose we have a prediction of a new response at input x_0, and $Y_0 = f(x_0) + \varepsilon_0$. Then the expected prediction error of estimate $\tilde{f}(x_0) x_0^\top \tilde{\beta}$ can be formulated as,

$$\mathbb{E}[Y_0 - \tilde{f}(x_0)]^2 = \sigma^2 + \mathbb{E}[x_0^\top \tilde{\beta} - f(x_0)]^2 = \sigma^2 + \text{MSE}(\tilde{f}(x_0)).$$

Thus the new σ^2 term is contributed by the variance of the new observation y_0.

2.2.9 Multiple Regression and Orthogonality

In this section I aim not to re-introduce least squares but rather provide an alternative view of the least squares formulation. The least squares problem involves minimizing the residual sum of squares of which can be written as,

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|^2.$$

Expanding this objective gives,

$$(y - x\beta)^\top (y - x\beta) = y^\top y - 2\beta x^\top y + \beta^2 x^\top x.$$

Differentiating with respect to β and setting equal to zero gives,

$$-2x^\top y + 2\beta x^\top x = 0 \implies \beta x^\top x = x^\top y \implies \hat{\beta} = \frac{x^\top y}{x^\top x}.$$

The forms of the numerator and denominator are analogous to the inner product so we can write,

$$\hat{\beta} = \frac{x^\top y}{x^\top x} = \frac{\langle x, y\rangle}{\langle x, x\rangle}.$$

More specifically, it is clear from this form that this is the projection coefficient of y onto x with the fitted vector $\hat{y} = x\hat{\beta}$ and residual vector $r = y - x\hat{\beta}$. The least squares coefficient says that the residual is orthogonal to the predictor. Indeed this follows,

$$\begin{aligned}
\langle x, r\rangle &= \langle x, y - x\hat{\beta}\rangle \\
&= \langle x, y\rangle - \hat{\beta}\langle x, x\rangle \\
&= \langle x, y\rangle - \frac{\langle x, y\rangle}{\langle x, x\rangle} \cdot \langle x, x\rangle = 0.
\end{aligned}$$

So the simple least squares is projection onto the one-dimensional subspace spanned by x. Now suppose that we have multiple predictors $x_1, \ldots, x_p$ and they are mutually orthogonal meaning that $\langle x_j, x_k\rangle = 0$ whenever $j \neq k$. The multiple regression model can be written as $y = \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$ and the least squares fitted value is $\hat{y} = \hat{\beta} x_1 + \cdots + \hat{\beta}_p x_p$. The normal equations also imply that the residual must be orthogonal to every predictor $\langle x_j, y - \hat{y}\rangle = 0$ for every j. Expanding this condition we have,

$$0 = \langle x_j, y - \hat{y}\rangle = \left\langle x_j, y - \sum_{k=1}^{p} \hat{\beta}_k x_k \right\rangle = \langle x_j, y\rangle - \sum_{k=1}^{p} \hat{\beta}_k \langle x_j, x_k\rangle.$$

But the predictors are orthogonal so all cross terms vanish.

Therefore, the terms where $k = j$ remain meaning that,

$$0 = \langle x_j, y\rangle - \hat{\beta}_j \langle x_j, x_j\rangle \implies \hat{\beta}_j = \frac{\langle x_j, y\rangle}{\langle x_j, x_j\rangle}.$$

So, we see that if all the predictors are orthogonal, then each multiple regression coefficient is just the simple univariate regression coefficient of y on that predictor. The problem in real data is that predictors are usually not orthogonal and are often correlated.

So what do we have in our toolbox? Why don't we just orthogonalize them!

Take the simple regression with intercept $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. In vector form, the intercept column is $x_0 = 1 = [1 \quad \cdots \quad 1]^\top$. So our model is just $y = \beta_0 x_0 + \beta_1 x + \varepsilon$. The usual slope estimate is written as $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$. Before using x, remove the part of x that lies in the direction of the intercept column 1. Since the intercept represents constants, projecting x onto 1 gives the constant vector whose entries are the sample mean $\bar{x}$. So the fitted value from regressing x on just an intercept is $\hat{x} = \bar{x}1$. Therefore the residualized version is $z = x - \bar{x}1$.

Notice that z is orthogonal to the intercept vector,

$$\begin{aligned}\langle 1, z\rangle &= 1^\top (x - \bar{x}1) \\ &= \sum_{i=1}^N x_i - \bar{x}\sum_{i=1}^N 1 \\ &= N\bar{x} - \bar{x}N = 0.\end{aligned}$$

So replacing x by $z = x - \bar{x}1$ makes the predictor orthogonal to the intercept.

Then we can write our coefficient on x as,

$$\hat{\beta}_1 = \frac{\langle z, y\rangle}{\langle z, z\rangle} = \frac{\langle x - \bar{x}1, y\rangle}{\langle x - \bar{x}1, x - \bar{x}1\rangle}.$$

Proof. We claim that the orthogonalized form $\hat{\beta}_1 = \frac{\langle x,y\rangle}{\langle z,z\rangle}$ is equivalent to the form $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$.

Observe that,

$$\begin{aligned}\langle z, y\rangle &= \sum_{i=1}^N (x_i - \bar{x})y_i \\ &= \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y} + \bar{y}) \\ &= \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) + \bar{y}\sum_{i=1}^N (x_i - \bar{x}) \\ &= \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}).\end{aligned}$$

So it is clear that $S_{xy} = \langle z, y\rangle$. Similarly for $\langle z, z\rangle$ we have,

$$\langle z, z\rangle = \sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x}) = \sum_{i=1}^N (x_i - \bar{x})^2 = S_{xx}.$$

□

Therefore ordinary simple regression with an intercept is interpreted as removing the part of x explained by the intercept and regress y on the remaining part of x. This is precisely the idea of the Gram-Schmidt orthogonalization procedure. Suppose now that we have predictors $x_0 = 1, x_1, \ldots, x_p$. Our goal is to replace these possibly correlated vectors with orthogonal vectors $z_0, z_1, \ldots, z_p$ that span the same subspace. We start with $z_0 = x_0 = 1$. Then for $j = 1, \ldots, p$ we remove from x_j its projection onto the already-created orthogonal vectors $z_0, \ldots, z_{j-1}$. Since the z's are orthogonal, the projection of x_j onto z_k is simply $\frac{\langle z_k, x_j\rangle}{\langle z_k, z_k\rangle} \cdot z_k$.

Therefore we have that,

$$z_j = x_j - \sum_{k=0}^{j-1} \frac{\langle z_k, x_j\rangle}{\langle z_k, z_k\rangle} \cdot z_k.$$

This is precisely the Gram-Schmidt procedure but we do not normalize the vectors to have length 1. So it is an orthogonal basis but not necessarily an orthonormal basis. Clearly the z's form an orthogonal basis for the same column space as the original predictors. This is significant as least squares is a projection onto the column space. If we change the basis of the column space, the fitted vector $\hat{y}$ does not change. We are simply using a more convenient coordinate system.

Once the predictors have been orthogonalized, our regression problem becomes simpler. Since the z_j's are orthogonal, the least squares fitted vector can be written as,

$$\hat{y} = \sum_{j=0}^{p} \frac{\langle z_j, y \rangle}{\langle z_j, z_j \rangle} \cdot z_j.$$

Note here that $\frac{\langle z_p, y \rangle}{\langle z_p, z_p \rangle}$ is the coefficient on z_p. This is the original multiple regression coefficient $\hat{\beta}_p$ on x_p. Why? When we construct z_p, we have $z_p = x_p$ less some linear combination of $z_0, \ldots, z_{p-1}$. But $z_0, \ldots, z_{p-1}$ are themselves a linear combination of $x_0, \ldots, x_{p-1}$. The coefficient of x_p inside z_p is therefore exactly 1 so when we write $\hat{y}$ in the z basis and convert back to the original x basis, the coefficient multiplying x_p is exactly the coefficient multiplying z_p. hence it is the case that,

$$\hat{\beta}_p = \frac{\langle z_p, y \rangle}{\langle z_p, z_p \rangle}.$$

That is, the coefficient on x_p is obtained by first removing from x_p all parts explained by the previous predictors, and then doing a simple regression of y on the leftover part.

We have yet to assign meaning to "adjusting for" the other variables. Recall that the residual vector z_p is the part of x_p not explained by the earlier predictors. That is $z_p = x_p - \hat{x}_p$ where $\hat{x}_p$ is the fitted value from regressing x_p on $x_0, \ldots, x_{p-1}$. So z_p is the new information in x_p that is not already contained in the other variables. Therefore our inner product form of $\hat{\beta}_p$ is just the simple regression coefficient of y on the part of x_p unexplained by the other predictors i.e., x_j being adjusted for the other variables. This is closely related to the Firsch-Waugh-Lovell theorem.

In modern regression, we would write,

$$\hat{\beta}_j = \frac{x_{j,\perp}^\top, y}{x_{j,\perp}^\top, x_{j,\perp}}$$

where $x_{j,\perp}$ is the residual from regressing x_j on all other predictors. Equivalently, because $x_{j,\perp}$ is orthogonal to all the other predictors, we also have that

$$\hat{\beta}_j = \frac{x_{j,\perp}^\top, y_\perp}{x_{j,\perp}^\top, x_{j,\perp}}$$

where $y_\perp$ is the residual from regressing y on all the other predictors.

In the case that x_p is highly correlated with the previous predictors, most of x_p can therefore be already explained by $x_0, x_1, \ldots, x_{p-1}$. So when we residualize x_p, the leftover vector z_p will be very small. In geometric terms, x_p is almost inside the span of the other predictors, so the component x_p orthogonal to them has very small length.

To see this, we must derive the variance formula. Using the formula we obtained for $\hat{\beta}_p$ previously and the true model $y = X\beta + \varepsilon$ we get that,

$$\hat{\beta}_p = \frac{z_p^\top y}{z_p^\top z_p} = \frac{z_p^\top (X\beta + \varepsilon)}{z_p^\top z_p} = \frac{\beta_p z_p^\top z_p + z_p^\top \varepsilon}{z_p^\top z_p} = \beta_p + \frac{z_p^\top \varepsilon}{z_p^\top z_p}.$$

This follows because z_p is orthogonal to $x_0, \ldots, x_{p-1}$ and because x_p contains z_p plus a linear combination of the earlier predictors. So, the only random part of $\hat{\beta}_p$ is the term $\frac{z_p^\top \varepsilon}{z_p^\top z_p}$. Since $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ we have that $z_p^\top \varepsilon \sim \mathcal{N}(0, \sigma^2 z_p^\top z_p)$. Therefore we have that,

$$\text{Var}(\hat{\beta}_p) = \text{Var}\left(\beta_p + \frac{z_p^\top \varepsilon}{z_p^\top z_p}\right) = \frac{1}{(z_p^\top z_p)^2} \text{Var}(z_p^\top \varepsilon) = \frac{1}{(z_p^\top z_p)^2} \sigma^2 z_p^\top z_p = \frac{\sigma^2}{z_p^\top z_p} = \frac{\sigma^2}{\|z_p\|^2}.$$

If x_p contains a lot of information that is new relative to the other predictors, then z_p is large, so $\|z_p\|^2$ is large and the variance of $\hat{\beta}_p$ is small. Otherwise, if x_p is almost explained by the other predictors, then z_p is small so $\|z_p\|^2$ is small and the variance of $\hat{\beta}_p$ is large. Geometrically, this is the reason why multicollinearity makes regression coefficients unstable. Imagine that x_1 and x_2 span a plane and the response vector y is projected onto that plane to produce the fitted vector $\hat{y}$. If x_1 and x_2 are not orthogonal, it is hard to read off the regression coefficient directly. So we orthogonalize.

Keeping $z_1 = x_1$, then removing from x_2 is projection onto x_1,

$$z_2 = x_2 - \text{proj}_{x_1}(x_2).$$

Now we have that z_2 is orthogonal to x_1. Instead of using the oblique basis x_1, x_2 we use the orthogonal basis z_1, z_2 where the fitted vector $\hat{y}$ is still the projection of y onto the same plane.

In Gram-Schmidt each original column x_j can be written as a linear combination of the orthogonalized vectors up to z_j. More specifically we have that $x_j \in \text{span}\{z_0, \ldots, z_j\}$. So, for each j, we have that $x_j = \gamma_{0j}z_0 + \gamma_{1j}z_1 + \cdots + \gamma_{jj}z_j$. There are no terms involving $z_{j+1} + \cdots + z_p$ since when constructing z_j we only use the previous vectors $z_0, \ldots, z_{j-1}$. So, therefore we write x_j as,

$$x_j = \sum_{i=0}^{j} \gamma_{ij} z_j.$$

If we put this into matrix form, we get $X = Z\Gamma$ where Γ is an upper triangular matrix that is filled with coefficients from the sum γ_{ij}. Note it is upper triangular purely because column x_j only uses $z_0, \ldots, z_j$, that is, the entries below the diagonal are therefore zero. This is analogous to writing the original columns x_j in the new orthogonal basis z_j. The columns of Z must be orthogonal but not necessarily normalized. This means simply that $z_i^\top z_j = 0$ if $i \neq j$. Generally, we had that $z_j^\top z_j = \|z_j\|^2 \neq 1$.

To turn the orthogonal columns into orthonormal columns, we divide each z_j by its length. Suppose we have a diagonal matrix D. Then ZD^{-1} has columns,

$$\frac{z_0}{\|z_0\|}, \frac{z_1}{\|z_1\|}, \ldots, \frac{z_p}{\|z_p\|}.$$

Let $Q = ZD^{-1}$ be such a matrix. Then, Q has orthonormal columns. To check this algebraically is trivial.

$$Q^\top Q = (ZD^{-1})^\top (ZD^{-1}) = D^{-1}Z^\top ZD^{-1}.$$

Because the columns of Z are orthogonal, then we know that $Z^\top Z = D^2$ must hold. Therefore,

$$Q^\top Q = D^{-1}Z^\top ZD^{-1} = D^{-1}D^2D^{-1} = I.$$

Thus, Q has orthonormal columns. Technically, since $Q \in \mathbb{R}^{N\times(p+1)}$, it is usually not a square orthogonal matrix so it is often better to say that Q has orthonormal columns meaning $Q^\top Q = I$.

You may realize that our steps are analogous to deriving a QR decomposition. Starting from $X = Z\Gamma$ we insert the identity matrix in the form $D^{-1}D = I$ so $X = Z\Gamma = ZD^{-1}D\Gamma$. Now define $Q = ZD^{-1}$ as defined earlier and $R = D\Gamma$. Then clearly, $X = QR$ where $Q \in \mathbb{R}^{N\times(p+1)}$ and $R \in \mathbb{R}^{(p+1)\times(p+1)}$ is upper triangular. Multiplying an upper triangular matrix by a diagonal matrix preserves upper triangularity therefore making R upper triangular.

Recall the least squares problem, that is, $\hat{\beta} = \arg\min_\beta \|y - X\beta\|^2$. If $X = QR$ then this becomes, $\hat{\beta} = \arg\min_\beta \|y - QR\beta\|^2$. The usual normal equations are $X^\top X\hat{\beta} = X^\top y$, so substituting $X = QR$ gives,

$$X^\top X\hat{\beta} = X^\top y \iff (QR)^\top (QR)\hat{\beta} = (QR)^\top y \iff R^\top Q^\top QR\hat{\beta} = R^\top Q^\top y.$$

Because $Q^\top Q = I$ then this simplifies to $R^\top R\hat{\beta} = R^\top Q^\top y$. If X has full column rank, then the upper triangular matrix R is invertible. Therefore, we can simply cancel $R^\top$ from both sides so that $R\hat{\beta} = Q^\top y$ and thus, $\hat{\beta} = R^{-1}Q^\top y$.

You may recall from CS 475 or any computation linear algebra course that we do not need to explicitly compute R^{-1} but instead solve the triangular system where $R\hat{\beta} = Q^\top y$. Since R is upper triangular, we can solve this efficiently via back-substitution.

For example if R has upper triangular entries r_{ij} then the system $R\hat{\beta} = Q^\top y$ can be solved starting from the last equation $r_{pp}\hat{\beta}_p = (Q^\top y)_p$. Solving for $\hat{\beta}_p$ then plugging the solution into the previous equation and solving for $\hat{\beta}_{p-1}$ continuing upwards will yield us our solution to the system.

Trivially, the fitted value $\hat{y} = X\hat{\beta}$ is simply,

$$\hat{y} = X\hat{\beta} = QR\hat{\beta} = QR(R^{-1}Q^\top y) = QQ^\top y.$$

Since the columns of Q form an orthonormal basis for the column space of X, the matrix $QQ^\top$ is the orthogonal projection matrix onto the column space of X. So $\hat{y} = QQ^\top y$.

The same idea follows as with the hat matrix $H = X(X^\top X)^{-1}X^\top$. In fact when $X = QR$,

$$\begin{aligned} H &= X(X^\top X)^{-1}X^\top \\ &= QR((QR)^\top(QR))^{-1}(QR)^\top \\ &= QR(R^\top Q^\top QR)^{-1}R^\top Q^\top \\ &= QR(R^\top R)^{-1}R^\top Q^\top. \end{aligned}$$

Since we know that $(R^\top R)^{-1} = R^{-1}(R^\top)^{-1}$ we get,

$$\begin{aligned} H &= QR(R^{-1}(R^\top)^{-1})R^\top Q^\top \\ &= Q[RR^{-1}][(R^\top)^{-1}R^\top]Q^\top \\ &= QIQ^\top \\ &= QQ^\top. \end{aligned}$$

So QR gives the same projection matrix as the usual OLS formula.

2.2.10 Multiple Outputs

Suppose that we are quants and we are predicting several related quantities from the same inputs, that is, we might have that Y_1 is the return of asset 1, Y_2 is the return of asset 2, and Y_3 is the return of asset 3, all predicted using the same input variables $X_1, \ldots, X_p$. The key idea here is that, if every output uses the same design matrix X, then the least squares separates into K ordinary regressions.

Even if the output errors are correlated with each other, the coefficient estimates are still the same as doing each regression separately, provided the covariance matrix is the same for every observation. More concretely, assume that for each output Y_k we can define a separate linear model,

$$Y_k = \beta_{0k} + \sum_{j=1}^{p} X_j\beta_{jk} + \varepsilon_k.$$

So, for output k, the coefficient vector is simply,

$$\beta_k = \begin{bmatrix} \beta_{0k} \\ \beta_{1k} \\ \vdots \\ \beta_{pk} \end{bmatrix}$$

The model for the k-th output is therefore denoted as,

$$Y_k = X\beta_k + \varepsilon_k.$$

Now, suppose we stack all K outputs together so that the response matrix becomes,

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1K} \\ y_{21} & y_{22} & \cdots & y_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{NK} \end{bmatrix} \in \mathbb{R}^{N \times K},$$

where each row in the response matrix above is one observation and each column is one output variable. The design matrix $X \in \mathbb{R}^{N \times (p+1)}$ and the coefficient matrix is defined as $\beta \in \mathbb{R}^{(p+1) \times K}$. The kth columns of β is therefore the coefficient vector for the kth output. Therefore our full model can be written compactly as $Y = X\beta + \varepsilon$. This is exactly the same as ordinary multiple regression except now Y and ε are matrices instead of vectors, and β is now a matrix instead of a vector.

The least squares objective is the sum of squared residuals over all observations and all outputs,

$$\text{RSS}(\beta) = \sum_{k=1}^{K} \sum_{i=1}^{N} (y_{ik} - f_k(x_i))^2.$$

The fitted value matrix is therefore $\hat{Y} = X\beta$ and the residual matrix is therefore $Y - X\beta$. The squared Frobenius norm of this residual matrix is,

$$\|Y - X\beta\|_F^2 = \sum_{i=1}^{N} \sum_{k=1}^{N} (y_{ik} - (X\beta)_{ik})^2,$$

so the least squares objective can be alternatively written as,

$$\text{RSS}(\beta) = \|Y - X\beta\|_F^2.$$

By the trace identity we see that $\|A\|_F^2 = \text{tr}(A^\top A)$ so we can also write the residual sum of squares $\text{RSS}(\beta) = \text{tr}[(Y - X\beta)^\top (Y - X\beta)]$. The trace notation is just a compact way of summing all squared residuals across all outputs. Now we derive the multiple output least squares estimator by expanding the objective,

$$\begin{aligned} RSS(B) &= \text{tr}\Big[(Y - XB)^\top (Y - XB)\Big] \\ &= \text{tr}\Big[(Y^\top - B^\top X^\top)(Y - XB)\Big] \\ &= \text{tr}(Y^\top Y) - \text{tr}(Y^\top XB) - \text{tr}(B^\top X^\top Y) + \text{tr}(B^\top X^\top XB). \end{aligned}$$

Observe that the two middle trace terms equal because they are scalars so $\text{tr}(Y^\top X\beta) = \text{tr}(\beta^\top X^\top Y)$ so the residual sum of squares formulation simplifies to,

$$\text{RSS}(\beta) = \text{tr}(Y^\top Y) - 2\text{tr}(\beta^\top X^\top Y) + \text{tr}(\beta^\top X^\top X\beta).$$

Differentiating the above with respect to β we get,

$$\nabla_\beta \text{RSS}(\beta) = -2X^\top Y + 2X^\top X\beta.$$

Setting the derivative equal to zero we get that,

$$-2X^\top Y + 2X^\top X\beta = 0 \implies X^\top X\beta = X^\top Y.$$

Assuming again that $(X^\top X)$ is invertible we get that,

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y.$$

This is precisely the same formula as ordinary least squares except now Y is a matrix. The reason this works cleanly is that the columns are separable, so we can write $Y = [y_1 \quad y_2 \quad \cdots \quad y_K]^\top$ where each $y_k \in \mathbb{R}^N$ is the vector of responses for output k. Then we have that $\hat{\beta} = (X^\top X)^{-1} X^\top [y_1 \quad y_2 \quad y_3 \quad \cdots \quad y_K]$.

Thus, our OLS estimate for multiple outputs is simply,

$$\hat{\beta} = [(X^\top X)^{-1}X^\top y_1 \quad (X^\top X)^{-1}X^\top y_2 \quad \cdots \quad (X^\top X)^{-1}X^\top y_K].$$

Thus, the kth column is $\hat{\beta}_k = (X^\top X)^{-1}X^\top y_k$. Clearly, estimating β is the same as running K separate OLS regressions, one for each output. Evidently, multiple outputs donot affect on another's least squares estimates. At least for this ordinary least squares criterion, the outputs are just separate columns.

Suppose that the errors across the outputs are now correlated. For one observation i, the error vector ε_i and suppose that $\text{Cov}(\varepsilon_i) = \Sigma$ i.e., the erros in the different outputs may move together. For example, errors in predicting two related assets might be positively correlated. In multivariate Gaussian theory, if we know that $\varepsilon_i \sim \text{MVN}(0, \Sigma)$ then the natural weighted quadratic loss is $\varepsilon_i^\top \Sigma^{-1}\varepsilon_i$. So instead of ordinary RSS we might use,

$$\text{RSS}(\beta; \Sigma) = \sum_{i=1}^{N}(y_i - f(x_i))^\top \Sigma^{-1}(y_i - f(x_i)).$$

Here $y_i \in \mathbb{R}^K$ is the vector of all K responses for observation i and $f(x_i) \in \mathbb{R}^K$ is the vector of fitted values. In matrix form, this weighted objective can be written as,

$$\text{RSS}(\beta; \Sigma) = \text{tr}\left[\Sigma^{-1}(Y - X\beta)^\top (Y - X\beta)\right].$$

Deriving the explicit minimizer we expand,

$$\begin{aligned} RSS(B; \Sigma) &= \text{tr}\Big[\Sigma^{-1}(Y - XB)^\top (Y - XB)\Big] \\ &= \text{tr}(\Sigma^{-1}Y^\top Y) - 2\,\text{tr}(\Sigma^{-1}Y^\top XB) + \text{tr}(\Sigma^{-1}B^\top X^\top XB). \end{aligned}$$

Differentiating with respect to β and setting to zero,

$$\begin{aligned} \nabla_\beta RSS(\beta; \Sigma) &= -2X^\top Y\Sigma^{-1} + 2X^\top X\beta\Sigma^{-1} \\ &= 0. \end{aligned}$$

Then solving for β we get,

$$\begin{aligned} -2X^\top Y\Sigma^{-1} + 2X^\top X\beta\Sigma^{-1} &= 0 \\ X^\top X\beta\Sigma^{-1} &= X^\top Y\Sigma^{-1} \\ X^\top X\beta = X^\top Y \implies \hat{\beta} &= (X^\top X)^{-1}X^\top, \end{aligned}$$

following since Σ^{-1} is invertible so multiplying on the right by Σ gives us our original least squares normal equation. So, even after accounting for constant correlation across the output errors, the least squares coefficient estimator is unchanged. Therefore, if all outputs share the same X and the error covariance is Σ is constant across obsevations then the correlation among output errors does not change the OLS coefficient estimates. Joint inferences, uncertainty estimates, and standard errors can change because of the existence of Σ, but it does not change the point estimate $\hat{\beta}$. If the covariance matrix varies by observation, then our clean separation above breaks. This means that instead of one covariance matrix Σ we might have something like $\text{Cov}(\varepsilon_i) = \Sigma$.

Our weighted objective becomes,

$$\sum_{i=1}^{N}(y_i - f(x_i))^\top \Sigma_i^{-1}(y_i - f(x_i)).$$

The weights now depend on i. Since the covariance structure changes across observations, we can no longer factor out one common Σ^{-1} in a way that cancels from the normal equations. In that case, the problem may no longer decouple into K separate ordinary least squares regressions.

2.3 Cross Validation

We will cover model validation and testing in more depth in a later chapter, but here we'll look at K-fold cross validation as well as leave-one-out cross validation for the purpose of seeing how we estimate model prediction error on new data. The main idea is that if we train and evaluate on the same data, the error is usually too optimistic because the model has already seen those observations.

Naturally, we split the data into training and validation parts.

Recall that we let $H = X(X^\top X)^{-1}X^\top$ be the hat matrix since it puts the hat on y. It is, as you may recall, the orthogonal projection matrix onto the column space of X and is idempotent and symmetric by definition. Therefore, Hy is the orthogonal projection of y onto $\mathrm{Col}(X)$. This is precisely what explains why the coefficients $\hat{\beta}$ can be unstable while the predictions $\hat{y}$ remain stable. If the columns of X are highly correlated, then there can be many very different coefficient vectors that produce almost the same fitted vector $X\beta$.

The coefficients are coordinates of the projection in the basis given by the columns of X. If that basis is nearly linearly dependent, the coordinates can be extremely unstable. But the point being projected onto, namely $\hat{y}$, can still be relatively stable. This is exactly the multi-collinearity phenomenon. The prediction vector is geometrically meaningful since the coefficients depend heavily on the coordinate system used to describe that prediction vector.

2.3.1 K-Fold CV

The purpose of cross-validation (CV) is to estimate prediction error on data not used to fit the model. The training error, $\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$ is usually too optimistic because the same observations are used both to fit the model and to evaluate the model. A flexible model can make the residuals on the training set small without necessarily predicting new data well.

In K-fold CV, the data is divided into K groups. The model is trained on $K-1$ groups and tested on the remaining group. This is repeated until every group has served as the validation group. The validation errors are then averages and this gives an estimate of the test error.

Algorithm 1 K-Fold Cross-Validation

1: Randomly partition the data $\{(x_i, y_i)\}_{i=1}^n$ into K approximately equal-sized folds $\mathcal{I}_1, \ldots, \mathcal{I}_K$.
2: **for** $k = 1, \ldots, K$ **do**
3: Let $\mathcal{I}_k$ be the validation fold.
4: Let $\mathcal{I}_{-k} = \{1, \ldots, n\} \setminus \mathcal{I}_k$ be the training indices.
5: Fit the model $\hat{f}^{(-k)}$ using only the training data $\{(x_i, y_i) : i \in \mathcal{I}_{-k}\}$.
6: Compute the validation error

$$\mathrm{CV}_k = \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \left(y_i - \hat{f}^{(-k)}(x_i)\right)^2.$$

7: **end for**
8: Average the fold errors:

$$\mathrm{CV}_K = \frac{1}{K} \sum_{k=1}^{K} \mathrm{CV}_k.$$

9: **return** CV_K as the estimated test error.

2.3.2 Leave-One-Out CV

Leave-one-out CV or LOOCV is the special case where $K = n$. Each validation set contains one observation. For the i-th observation, we fit the model using all observations except i, then predict the left-out response y_i. We let $\hat{f}^{(-i)}$ denote the fitted regression function obtained after deleting observation i.

The LOO prediction error is therefore,

$$LOOCV = \frac{1}{n}\sum_i \left(y_i - \hat{f}^{(-i)}(x_i)\right)^2.$$

Naively, this requires fitting n different least squares models. But for OLS, there is a remarkable shortcut. We can compute LOOCV exactly from a single model fit. The shortcut formula is simply,

$$y_i - \hat{f}^{(-i)}(x_i) = \frac{y_i - \hat{y}_i}{1 - h_{ii}},$$

where h_{ii} is the ith diagonal entry of the hat matrix. Therefore, LOOCV is defined as,

$$LOOCV = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_{ii}}\right)^2.$$

The formula above basically say that the LOO residual is the ordinary residual inflated by a factor of $1/(1 - h_{ii})$.

Proof. We aim to derive the LOOCV formula. Suppose that $A = X^\top X$ our square matrix and that the full-data OLS estimator is $\hat{\beta} = A^{-1}X^\top y$. Let $x_i^\top$ denote the ith row of X. Equivalently, $x_i \in R^p$ is the column vector of predictors for observation i. If we remove observation i then the new cross product is,

$$X_{-i}^\top X_{-i} = X^\top X - x_i x_i^\top = A - x_i x_i^\top.$$

Similarly, we have that $X_{-i}^\top y_{-i} = X^\top y - x_i y_i$. Therefore the LOO coefficient vector can be written as,

$$\hat{\beta}^{(-i)} = (A - x_i x_i^\top)^{-1}(X^\top y - x_i y_i).$$

The obstacle is the inverse $(A - x_i x_i^\top)^{-1}$.

We will use the Sherman-Morrison formula which says for an invertible matrix A,

$$(A - uu^\top)^{-1} = A^{-1} + \frac{A^{-1}uu^\top A^{-1}}{1 - u^\top A^{-1}u},$$

provided that $1 - u^\top A^{-1}u \neq 0$ and where $u = x_i$ so,

$$(A - x_i x_i^\top)^{-1} = A^{-1} + \frac{A^{-1}x_i x_i^\top A^{-1}}{1 - x_i^\top A^{-1}x_i}.$$

Observe that the hat matrix,

$$h_{ii} = H_{ii} = x_i^\top (X^\top X)^{-1}x_i = x_i^\top A^{-1}x_i.$$

The inverse therefore becomes,

$$(A - x_i x_i^\top)^{-1} = A^{-1} + \frac{A^{-1}x_i x_i^\top A^{-1}}{1 - h_{ii}}.$$

Substituting this identity into the LOO estimator,

$$\begin{aligned}\hat{\beta}^{(-i)} &= \left(A^{-1} + \frac{A^{-1}x_i x_i^\top A^{-1}}{1 - h_{ii}}\right)(X^\top y - x_i y_i) \\ &= A^{-1}X^\top y - A^{-1}x_i y_i + \frac{A^{-1}x_i x_i^\top A^{-1}X^\top y}{1 - h_{ii}} - \frac{A^{-1}x_i x_i^\top A^{-1}x_i y_i}{1 - h_{ii}}.\end{aligned}$$

Since we know that $A^{-1}X^\top y = \hat{\beta}$ and we have that $x_i^\top A^{-1}X^\top y = x_i^\top \hat{\beta} = \hat{y}_i$ and $h_i = x_i^\top A^{-1}x_i = h_i$ we get that the OLS coefficient removed the i observation is given as,

$$\begin{aligned}\hat{\beta}^{(-i)} &= \hat{\beta} - A^{-1}x_i y_i + \frac{A^{-1}x_i \hat{y}_i}{1 - h_i} - \frac{A^{-1}x_i h_i y_i}{1 - h_i} \\ &= \hat{\beta} + A^{-1}x_i\left[-y_i + \frac{\hat{y}_i - h_i y_i}{1 - h_i}\right].\end{aligned}$$

Putting the bracket over a common denominator we can write,

$$\begin{aligned} -y_i + \frac{\hat{y}_i - h_i y_i}{1 - h_i} &= \frac{-(1-h_i)y_i + \hat{y}_i - h_i y_i}{1 - h_i} \\ &= \frac{-y_i + h_i y_i + \hat{y}_i - h_i y_i}{1 - h_i} \\ &= \frac{\hat{y}_i - y_i}{1 - h_i}. \end{aligned}$$

Therefore the OLS coefficient removed the ith is written as,

$$\hat{\beta}^{(-i)} = \hat{\beta} + \frac{A^{-1} x_i (\hat{y}_i - y_i)}{1 - h_i}.$$

The formulation above tells us how much the coefficient vector changes when observation i is removed. We want the LOO fitted value at the left-out point x_i which is $\hat{f}^{(-i)}(x_i) = x_i^\top \hat{\beta}^{(-i)}$. Substituting this formula we get,

$$\begin{aligned} \hat{f}^{(-i)}(x_i) &= x_i^\top \left[\hat{\beta} + \frac{A^{-1} x_i (\hat{y}_i - y_i)}{1 - h_i} \right] \\ &= x_i^\top \hat{\beta} + \frac{x_i^\top A^{-1} x_i (\hat{y}_i - y_i)}{1 - h_i} \\ &= \hat{y}_i + \frac{h_i (\hat{y}_i - y_i)}{1 - h_i} \\ &= \frac{(1-h_i)\hat{y}_i + h_i \hat{y}_i - h_i y_i}{1 - h_i} \\ &= \frac{\hat{y}_i - h_i y_i}{1 - h_i}. \end{aligned}$$

The LOO residual is therefore given as,

$$\begin{aligned} y_i - \hat{f}^{(-i)}(x_i) &= y_i - \frac{\hat{y}_i - h_i y_i}{1 - h_i} \\ &= \frac{(1-h_i)y_i - \hat{y}_i + h_i y_i}{1 - h_i} \\ &= \frac{y_i - \hat{y}_i}{1 - h_i}. \end{aligned}$$

Therefore we have proven the identity,

$$y_i - \hat{f}^{(-i)}(x_i) = \frac{y_i - \hat{y}_i}{1 - h_i}.$$

□

The ordinary residual $e_i = y_i - \hat{y}_i$ is too optimistic here because $\hat{y}_i$ was computed using y_i itself. The diagonal hat value h_i measures how much observation i influences its own fitted value. If h_i is small, then leaving out observation i does not change its prediction very much and $e_i/(1-h_i)$ is close to e_i. If h_i is large, then observation i has pulled the regression fit toward itself so the ordinary residual may be artificially small. Dividing by $1 - h_i$ corrects for the self-influence. High leverage points prove to be very dangerous in such instances. A high-leverage point can have a small residual under full-data fit because it has strongly influences the fitted line or plane. But, when the point is removed, the fitted model may move away from it dramatically producing a large leave-one-out residual. Consider the diagonal entries of the hat matrix. The fitted value is,

$$\hat{y}_i = \sum_j H_{ij} y_j$$

with diagonal entry $H_{ii} = h_{ii}$ is the coefficient multiplying y_i in its own fitted value. Therefore h_{ii} measures the self-dependence of $\hat{y}_i$ on y_i. Since we know that $H = X(X^\top X)^{-1} X^\top X$ the diagonal entry is simply

$h_{ii} = x_i^\top (X^\top X)^{-1} x_i$. This expression shows that leverage depends only on the predictors X and not on the response values y. A point has high leverage because its x_i value is unusual relative to the design matrix and not because its y_i value is large or small. The sum of leverages is the trace of the hat matrix and for OLS we know that $\text{tr}(H) = p$. Naturally, we can define the average leverage as,

$$\frac{1}{n}\sum_i h_{ii} = \frac{p}{n}.$$

This is precisely why there are p degrees of freedom in an OLS fit as the trace of the smoothing matrix measures the total amount of self-influence in the fitted values.

2.3.3 Generalized Cross-Validation

In *Generalized CV*, we replace the individual leverages h_i by their average value. Since the average leverage is given as $\bar{h} = \frac{\text{tr}(H)}{n}$, Generalized CV approximates LOOCV by,

$$\text{GCV} = \frac{1}{n}\sum_i \left(\frac{e_i}{1 - \text{tr}(H)/n}\right)^2.$$

Since the denominator no longer depends on i we see that the above becomes,

$$\begin{aligned}\text{GCV} &= \frac{1}{n} \cdot \frac{\sum_{i=1}^n e_i^2}{(1 - \text{tr}(H)/n)^2} \\ &= \frac{RSS/n}{(1 - \text{tr}(H)/n)^2}.\end{aligned}$$

For OLS we know that $\text{tr}(H) = p$ so,

$$\text{GCV} = \frac{RSS/n}{(1 - p/n)^2}.$$

The numerator RSS/n is the training mean squared error and the denominator inflates it to correct for model flexibility. When p is small relative to n, the factor $1 - p/n$ is close to 1 so the correction is mild. When p is large relative to n, the factor $1 - p/n$ is smaller, so the training error gets inflated more aggressively. The GCV becomes especially important for ridge regression and other linear smoothers. A linear smoother, as you may recall, is any method whose fitted values can be written as $\hat{y} = Ay$ for some matrix A depending on the predictors and tuning parameters but not on y. For OLS, $A = H$ the hat matrix and for ridge regression,

$$\hat{\beta}_{\text{ridge}} = (X^\top X + \lambda I)^{-1} X^\top y,$$

so naturally our fitted ridge values become,

$$\hat{y}_{\text{ridge}} = X\hat{\beta}_{\text{ridge}} = X(X^\top X + \lambda I)^{-1} X^\top y.$$

Thus the ridge has a smoothing matrix defined by it's hat matrix. The fitted values are clearly $\hat{y}_{\text{ridge}} = H_{\text{ridge}} y$. For ridge, H_{ridge} is symmetric but not idempotent. Generally, $H_{\text{ridge}}^2 \neq H_{\text{ridge}}$ so ridge is not an orthogonal projection in the same sense as OLS. Nevertheless, the trace of H_{ridge} still measures effective model complexity i.e., $df(\lambda) = \text{tr}(H_{\text{ridge}})$. In the next section we will see how to derive the degrees of freedom via the trace for ridge but for now,

$$\text{df}(\lambda) = \text{tr}(H_{\text{ridge}}) = \sum_j = \frac{d_j^2}{d_j^2 + \lambda}.$$

The formulation above provides us with a fast way to choose λ. For each candidate λ, we fit ridge, compute $\text{RSS}(\lambda)$, compute $\text{df}(\lambda)$ and pick the λ with the best estimated prediction error.

Algorithm 2 Generalized Cross-Validation for Ridge Regression

Require: Design matrix $X \in \mathbb{R}^{n \times p}$, response vector $y \in \mathbb{R}^n$, candidate tuning parameters Λ
Ensure: Selected tuning parameter $\hat{\lambda}$

1: **for** $\lambda \in \Lambda$ **do**
2: Fit ridge regression:

$$\hat{\beta}_\lambda = (X^\top X + \lambda I)^{-1} X^\top y.$$

3: Compute fitted values:

$$\hat{y}_\lambda = X\hat{\beta}_\lambda.$$

4: Compute residual sum of squares:

$$RSS(\lambda) = \sum_{i=1}^{n} (y_i - \hat{y}_{\lambda,i})^2.$$

5: Compute the ridge smoothing matrix:

$$H_\lambda = X(X^\top X + \lambda I)^{-1} X^\top.$$

6: Compute the effective degrees of freedom:

$$df(\lambda) = \operatorname{tr}(H_\lambda).$$

7: Compute the generalized cross-validation score:

$$\mathrm{GCV}(\lambda) = \frac{RSS(\lambda)/n}{(1 - df(\lambda)/n)^2}.$$

8: **end for**
9: Select

$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda} \mathrm{GCV}(\lambda).$$

10: **return** $\hat{\lambda}$

2.3.4 Tuning Parameters

Suppose a model depends on a tuning parameter λ. For ridge λ controls the strength of shrinkage. For LASSO, λ controls the strength of the $L1$ penalty. For K nearest neighbors, the tuning parameter is K. For LOESS, the tuning parameter controls the size of the local neighbourhood or smoothing span.

For every candidate tuning parameter, we fit the model on training subsets and estimate validation error. We then choose the tuning parameter that minimizes estimated out-of-sample error,

$$\hat{\lambda} = \arg\min_{\lambda} \hat{\mathrm{Err}}(\lambda).$$

Training error almost always favors more flexible models. If a model is allowed to become more complex, it can usually reduce training error. But test error has a bias-variance tradeoff. Too much simplicity gives high bias and too much flexibility gives high variance. Cross-validation can be seen as a way to estimate the sweet spot.

2.4 Shrinkage Methods

Recall that from OLS, we required that the square matrix $X^\top X$ has full column rank. When the columns of X are nearly linearly dependent, then $X^\top X$ becomes ill-conditioned or nearly singular which makes the inverse of $X^\top X$ numerically unstable causing OLS coefficient estimates to become extremely large. As we saw, the variance of $\hat{\beta}$ explodes and small perturbations in our data can cause large swings in coefficients.

Recall the OLS covariance matrix,

$$\text{Var}(\hat{\beta}_{\text{OLS}}) = \sigma^2 (X^\top X)^{-1}.$$

If $X^\top X$ has very small eigenvalues (equivalently very small singular values of X), then its inverse contains very large eigenvalues, inflating coefficient variance. Given that this is inherently a variance and instability problem, in this section we will introduce methods that aim to solve this problem via penalization and shrinkage estimators.

2.4.1 Ridge Regression

Ridge regression modifies the ordinary least squares objective by adding an $L2$ penalty on the slope coefficients,

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

The first term of the ridge normal equation is synonymous to the residual sum of squares criterion used in least squares, that is, how well the model fits the data. The second term penalizes coefficients so that they do not become too large. Clearly, $\lambda \geq 0$ is our tuning parameter that controls the tradeoff. In the case that $\lambda = 0$, ridge is simply just OLS. When λ is large, the penalty is severe so the slope coefficients are forced closer to zero.

Notice how we have *not* penalized the intercept term β_0. The intercept, as you may recall, simply sets the vertical location of the regression surface. Penalizing the intercept would make the result of the regression dependent on arbitrary shifts of the response y. Suppose for some constant c that we replaced every y_i with a shifted response $y_i + c$. Trivially, we would want all fitted values to shift by c but not for the penalty to resist that shift.

Naturally we standardize our predictors and response, after which, we remove the intercept from the algebra. Assume that every column in X is centered and that our responses y are centered as well. Then the ridge objective above can be re-written as,

$$\text{RSS}_\lambda(\beta) = (y - X\beta)^\top (y - X\beta) + \lambda \beta^\top \beta.$$

Note that $X \in \mathbb{R}^{N \times p}, \beta \in \mathbb{R}^p$ and $y \in \mathbb{R}^N$ where N is the number of observations and p is the number of predictors. Deriving the ridge solution by expanding the objective we get,

$$\begin{aligned}
\text{RSS}_\lambda(\beta) &= (y - X\beta)^\top (y - X\beta) + \lambda \beta^\top \beta \\
&= y^\top y - y^\top X\beta - \beta^\top X^\top y + \beta^\top X^\top X\beta + \lambda \beta^\top \beta \\
&= y^\top y - 2\beta^\top X^\top y + \beta^\top X^\top X\beta + \lambda \beta^\top \beta \\
&= y^\top y - 2\beta^\top X^\top y + \beta^\top (X^\top X + \lambda I)\beta.
\end{aligned}$$

Differentiating the last line above wrt β,

$$\nabla_\beta \text{RSS}_\lambda(\beta) = -2X^\top y + 2(X^\top X + \lambda I)\beta.$$

Setting the gradient equal to zero we get,

$$\begin{aligned}
-2X^\top y + 2(X^\top X + \lambda I)\hat{\beta}_{\text{ridge}} &= 0 \\
(X^\top X + \lambda I)\hat{\beta}_{\text{ridge}} &= X^\top y.
\end{aligned}$$

Therefore the closed form normal equation for ridge is,

$$\hat{\beta}_{\text{ridge}} = (X^\top X + \lambda I)^{-1} X^\top y.$$

Comparing this to OLS, we see that ridge simply replace $X^\top X$ with $X^\top X + \lambda I$. This is precisely why ridge stabilizes the inverse. In cases where the squares matrix $X^\top X$ has small eigenvalues, the inverse can explode in those corresponding directions, so adding a penalization term λ to every eigenvalue makes the matrix easier to invert. Geometrically, this can be expressed as a constrained lagrangian problem,

$$\mathcal{L}(\beta, \lambda) = \text{RSS}(\beta) + \lambda \left(\|\beta\|^2 - t \right), \quad \lambda \geq 0.$$

For a fixed active constraint, minimizing the Lagrangian above over β is equivalent to minimizing $\text{RSS}(\beta) + \lambda\|\beta\|^2$ because the term $-\lambda t$ is constant with respect to β. λ and t are two different ways of controlling the same thing. A large λ corresponds to a small allowed coefficient norm t and a small λ corresponds to a large allowed coefficient norm t. Geometrically, OLS tries to find the lowest point of the residual sum of squares surface. Ridge instead does not leave a ball of radius $\sqrt{t}$ in coefficient space. If the unconstrained OLS coefficient vector is too large, ridge then pulls it back toward the origin.

The penalty evidently is dependent on scale, so it is mandatory that we standardize predictors before ridge. Suppose one predictor is measured in dollars and another is measured in cents. The same underlying variable can have coefficients that differ by a factor of 100. Since ridge penalizes β_j^2, it would penalize the coefficient differently depending on the units. Therefore, before ridge, we usually center and scale each predictor so that the columns of X have comparable scale. After centering the intercept is handled separately that is, since the centered model has $\bar{x}_j = 0$ and $\bar{y} = 0$ then the fitted intercept in the original scale is,

$$\hat{\beta}_0 = \bar{y} - \sum_j \bar{x}_j \hat{\beta}_j.$$

If all x_j's are centered, this simply becomes $\hat{\beta}_0 = \bar{y}$ so ridge is usually applied only to the centered slope problem.

Ridge is best understood through SVD. Suppose that our centered design matrix is decomposed into $X = UDV^\top$ where $U^\top U = I$ and $V^\top V = I$ where $D = \text{diag}(d_1, \ldots, d_p)$. Columns of U are orthonormal directions in the column space of X and the columns of V are orthonormal directions in the predictor space. Thus, the singular values d_j measure how much the data vary in those directions.

First, we'll compute $X^\top X$ in SVD form,

$$\begin{aligned} X^\top X &= (UDV^\top)^\top (UDV^\top) \\ &= VD^\top U^\top UDV^\top \\ &= VD^\top DV^\top \\ &= VD^2V^\top. \end{aligned}$$

Since D is diagonal, then $D^\top D = D^2$ so $X^\top X = VD^2V^\top$. This is precisely the eigendecomposition of $X^\top X$ where eigenvectors are the columns v_j of V and the eigenvalues are d_j^2. First, we'll return to the OLS case of the fitted vector using SVD. Assume for the moment that all $d_j > 0$. Then trivially we get,

$$\hat{y}_{\text{OLS}} = X\hat{\beta}_{\text{OLS}} = X(X^\top X)^{-1}X^\top y.$$

Substituting the SVD pieces into the formula for the hat matrix we get,

$$\begin{aligned} X(X^\top X)^{-1}X^\top &= UDV^\top (VD^2V^\top)^{-1}(UDV^\top)^\top \\ &= UDV^\top (VD^{-2}V^\top) VDU^\top \\ &= UD(V^\top V)D^{-2}(V^\top V)DU^\top \\ &= UDD^{-2}DU^\top \\ &= UIU^\top UU^\top. \end{aligned}$$

Therefore $\hat{y}_{\text{OLS}} = UU^\top y$ meaning that OLS projects y onto the column space of X matching what we saw with QR decomposition that is $\hat{y} = QQ^\top y$. The difference is that the matrix Q came from Gram-Schmidt while U comes from SVD. Both are orthonormal bases for the column space of X.

The ridge fitted vector therefore follows as,

$$\hat{y}_{\text{ridge}} = X\hat{\beta}_{\text{ridge}} = X(X^\top X + \lambda I)^{-1}X^\top y.$$

We know that $X^\top X = VD^2V^\top$ and since $VIV^\top = I$,

$$\begin{aligned} X^\top X + \lambda I &= VD^2V^\top + \lambda VV^\top \\ &= V(D^2 + \lambda I)V^\top \\ (X^\top X + \lambda I)^{-1} &= \left[V(D^2 + \lambda I)V^\top\right]^{-1} \\ &= V(D^2 + \lambda I)^{-1}V^\top. \end{aligned}$$

Substituting this into the fitted vector we have,

$$\begin{aligned} \hat{y}^{\text{ridge}} &= X(X^\top X + \lambda I)^{-1}X^\top y \\ &= UDV^\top \left[V(D^2 + \lambda I)^{-1}V^\top\right](UDV^T)^\top y \\ &= UDV^\top V(D^2 + \lambda I)^{-1}V^\top VDU^\top y \\ &= UD(D^2 + \lambda I)^{-1}DU^\top y \\ &= U\left[D(D^2 + \lambda I)^{-1}D\right]U^\top y. \end{aligned}$$

Since D is diagonal we have,

$$D(D^2 + \lambda I)^{-1}D = \text{diag}\left(\frac{d_1^2}{d_1^2 + \lambda}, \ldots, \frac{d_p^2}{d_p^2 + \lambda}\right).$$

Instead of throwing away the directions u_j, ridge keeps them but shrinks their contribution by the factor $\frac{d_j^2}{d_j^2+\lambda}$ which is bounded between 0 and 1. If d_j^2 is large relative to λ then $\frac{d_j^2}{d_j^2+\lambda} \approx 1$ otherwise when d_j^2 is small relative to λ ridge heavily shrinks that direction so the ratio is approximately 0. Therefore, we see that ridge mostly preserves high-variance directions of X and strongly shrinks low-variance directions of X.

Substituting the SVD into the ridge coefficient vector we have,

$$\begin{aligned} \hat{\beta}^{\text{ridge}} &= (X^\top X + \lambda I)^{-1}X^\top y \\ &= \left[V(D^2 + \lambda I)V^\top\right]^{-1}(UDV^\top)^\top y \\ &= V(D^2 + \lambda I)^{-1}V^\top VDU^\top y \\ &= V(D^2 + \lambda I)^{-1}DU^\top y \\ &= \sum_{j=1}^{p} v_j \frac{d_j}{d_j^2 + \lambda} u_j^\top y. \end{aligned}$$

Doing the same for OLS we get,

$$\begin{aligned} \hat{\beta}^{\text{OLS}} &= (X^\top X)^{-1}X^\top y \\ &= (VD^2V^\top)^{-1}(UDV^\top)^\top y \\ &= VD^{-2}V^\top VDU^\top y \\ &= VD^{-1}U^\top y \\ &= \sum_{j=1}^{p} v_j \frac{1}{d_j} u_j^\top y. \end{aligned}$$

Comparing the coefficient direction on v_j, OLS uses $\frac{1}{d_j}u_j^\top y$ while ridge uses $\frac{d_j}{d_j^2+\lambda}u_j^\top y$. Indeed, the ridge coefficient equals the OLS coefficient multiplied by $\frac{d_j^2}{d_j^2+\lambda}$. So, in the principal component basis,

$$\text{Ridge Coefficient} = \text{OLS Coefficient} \times \frac{d_j^2}{d_j^2+\lambda}.$$

We often do not know the real covariance, so we compute the sample covariance matrix using the centered predictors which is $S = \frac{1}{N}X^\top X$. So, using the SVD we have that $X^\top X = VD^2V^\top$. Therefore our sample covariance can be written as,

$$S = \frac{1}{N}VD^2V^\top = V\left(\frac{D^2}{N}\right)V^\top.$$

So, the columns v_j of V are the eigenvectors of the sample covariance matrix. These are the principal component directions. We define the jth principal component score vector as $z_j = Xv_j$. Since we have defined X via its SVD then with $X = UDV^\top$ we have that the jth principal component is,

$$z_j = Xv_j = UDV^\top v_j = UDe_j = d_j u_j.$$

So we have that $z_j = Xv_j = d_j u_j$. Now we derive the variance. Since X is centered and z_j is centered, its sample variance is simply $\text{Var}(z_j) = \frac{1}{N}z_j^\top z_j$. Substituting $z_j = d_j u_j$ from earlier we get that,

$$\begin{aligned}\text{Var}(z_j) &= \frac{1}{N}(d_j u_j)^\top (d_j u_j)\\ &= \frac{1}{N}d_j^2 u_j^\top u_j\\ &= \frac{d_j^2}{N},\end{aligned}$$

because $u_j^\top u_j = 1$ since they are orthonormal. Then the variance of the jth principal component is,

$$\text{Var}(Xv_j) = \frac{d_j^2}{N}.$$

Evidently, large singular values correspond to high-variance principal component directions. Small singular values correspond to low-variance principal component directions. In the case of ridge regression, we shrink low-variance directions the most, that is, if d_j^2 is small, then the data does not vary much in the v_j direction. Moreover, that means that the slope in that direction is hard to estimate so a tiny amount of noise in y can create a large OLS coefficient because OLS divides d_j. Ridge prevents this by replacing $\frac{1}{d_j}$ with $\frac{d_j}{d_j^2+\lambda}$. When d_j is tiny, the factor $\frac{1}{d_j}$ explodes but the ridge factor is small.

Ridge is still a linear smoother because the fitted values are still linear in y. We'll define the ridge hat matrix as,

$$H_\lambda = X(X^\top X + \lambda I)^{-1}X^\top.$$

Then $\hat{y}_{\text{ridge}} = H_\lambda y$. In OLS, our hat matrix was simply $X(X^\top X)^{-1}X^\top$ which has degrees of freedom $\text{tr}(H) = p$ if the intercept has been removed and X has rank p. Ridge generalizes this by defining $\text{df}(\lambda) = \text{tr}(H_\lambda)$. Computing this trace using SVD we have that,

$$H_\lambda = UD(D^2 + \lambda I)^{-1}DU^\top.$$

Defining the diagonal,

$$A_\lambda = D(D^2+\lambda I)^{-1}\text{diag}\left(\frac{d_1^2}{d_1^2+\lambda}, \ldots, \frac{d_p^2}{d_p^2+\lambda}\right).$$

Clearly our ridge hat matrix is just $H_\lambda = UA_\lambda U^\top$ so therefore our degrees of freedom can be written as,

$$\text{df}(\lambda) = \text{tr}(H_\lambda) = \text{tr}(UA_\lambda U^\top) = \text{tr}(A_\lambda I) = \text{tr}(A_\lambda) = \sum_j \frac{d_j^2}{d_j^2+\lambda}.$$

This result simply states that each principal component direction contributes a fractional degree of freedom. So if $\lambda = 0$, then each term is $\frac{d_j^2}{d_j^2} = 1$ and thus $\text{df}(0) = p$ which simply recovers OLS. If $\lambda \to \infty$ then each term becomes $\frac{d_j^2}{d_j^2+\lambda} \to 0$ then $\text{df}(\lambda) \to 0$. Ridge continuously moves from a fully flexible p parameter linear model to an almost completely shrunk model. If an intercept is included and not penalized, then the intercept contributes one additional degree of freedom so the total would be,

$$1 + \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}.$$

Consider the case with correlated predictors. When predictors are highly correlated, $X^\top X$ has at least one very small eigenvalue. Equivalently, X has at least one very small singular value d_j. In OLS, the coefficient in that direction involves division by d_j so,

$$\hat{\beta}_{\text{OLS}} = \sum_{j=1}^{p} v_j \cdot \frac{1}{d_j} u_j^\top y.$$

If d_j is small, the coefficient becomes enormous so the formulation above gives an algebraic form for multicollinearity. The data provides little information about the slope in that direction so the estimated coefficient has high variance. Ridge changes the coefficient to,

$$\hat{\beta}_{\text{ridge}} = \sum_{j=1}^{p} v_j \frac{d_j}{d_j^2 + \lambda} u_j^\top y.$$

Now, the dangerous division by a tiny d_j is removed. If d_j is small then,

$$\frac{d_j}{d_j^2 + \lambda} \approx \frac{d_j}{\lambda},$$

which is evidently small rather than huge. It is important to note that ridge deliberately introduces bias in order to reduce variance. OLS is unbiased under the classical linear model, but it can have enormous variance when predictors are nearly collinear. Ridge shrinks the unstable directions producing a biased estimator with often much smaller prediction error. Suppose the model is $y \mid \beta \sim \mathcal{N}(X\beta, \sigma^2 I)$ and put an independent Gaussian prior on the coefficients $\beta_j \sim \mathcal{N}(0, \tau^2)$ for $j = 1, \ldots, p$. Then the prior density is proportional to $\exp\left(-\frac{1}{2\tau^2}\beta^\top\beta\right)$ and the likelihood is proportional to, $\exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^\top(y - X\beta)\right)$. The posterior density is proportional to likelihood times the prior,

$$\begin{aligned} p(\beta \mid y) &\propto \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^\top(y - X\beta)\right) \exp\left(-\frac{1}{2\tau^2}\beta^\top\beta\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}\left[(y - X\beta)^\top(y - X\beta) + \frac{\sigma^2}{\tau^2}\beta^\top\beta\right]\right). \end{aligned}$$

Maximizing the posterior is equivalent to minimizing the negative log-posterior so we minimize,

$$(y - X\beta)^\top(y - X\beta) + \frac{\sigma^2}{\tau^2}\beta^\top\beta.$$

This is simply ridge with $\lambda = \frac{\sigma^2}{\tau^2}$. The ridge posterior is therefore the posterior mode under a Gaussian prior centered at zero. Since the posterior is Gaussian, the posterior mode and posterior mean coincide. The smaller τ^2 is, the stronger prior belief that coefficients are near zero, and the larger $\lambda = \frac{\sigma^2}{\tau^2}$ becomes.

2.4.2 Matrix Invertibility and Ridge

We aim to show that the ridge regression estimator is always mathematically defined, that is, the regularized Gram matrix $(X^\top X + \lambda I_p)$ is invertible for any strictly positive regularization parameter $\lambda > 0$ regardless of the rank of X.

Proof. Recall that a real symmetric matrix is invertible if and only if it is strictly positive definite, which means in quadratic form with any non-zero vector is strictly greater than zero. Suppose that $a \in \mathbb{R}^p$ is our non-zero vector. Then, the quadratic form associated with the regularized Gram matrix is formulated as,

$$\begin{aligned} a^\top (X^\top X + \lambda I_p) a &= a^\top X^\top X a + a^\top (\lambda I_p) a \\ &= a^\top X^\top X a + \lambda a^\top I_p a \\ &= (Xa)^\top (Xa) + \lambda a^\top a \\ &= \|Xa\|_2^2 + \lambda \|a\|_2^2. \end{aligned}$$

What remains is to show that this quantity expressed in terms of the squared Euclidean ℓ_2 norm is strictly positive i.e., each term of the sum must be non-negative and guaranteed ≥ 0. Observe that the first term $\|Xa\|_2^2$ is a squared Euclidean norm of the vector $Xa \in \mathbb{R}^N$. Because the norm of any real-vector is non-negative, it is guaranteed that $\|Xa\|_2^2 \geq 0$. This term is exactly zero if and only if a lies in the null space of X but is otherwise strictly positive. The second term $\|a\|_2^2$ represents the squared Euclidean norm of an arbitrary vector $a \in \mathbb{R}^p$. Because a is assumed to be non-zero, at least one of its components must be non-zero. Consequently, the sum of its squared components is strictly positive meaning that $\|a\|_2^2 > 0$. The regularization parameter is strictly positive so the product of λ and the strictly positive term $\|a\|_2^2$ is also strictly positive meaning $\lambda \|a\|_2^2 > 0$.

Combining these results, the sum of a non-negative number and a strictly positive number is strictly positive,

$$\|Xa\|_2^2 + \lambda \|a\|_2^2 > 0, \quad \forall a \neq 0.$$

Because the quadratic form is strictly positive for any non-zero vetor a, the matrix $(X^\top X + \lambda I_p)$ is also symmetric and positive definite. Every eigenvalue of $(X^\top X + \lambda I_p)$ is strictly positive, specifically, if λ_i is an eigenvalue of $X^\top X$, then the corresponding eigenvalue of $(X^\top X + \lambda I_p)$ is $\lambda_i + \lambda \geq \lambda > 0$.

Since the determinant of a matrix is equal to the product of its eigenvalues, the determinant of $(X^\top X + \lambda I_p)$ is strictly positive which guarantees that the matrix has full rank and is invertible. □

2.4.3 Coordinate Shrinkage under Ridge

We return to the SVD of ridge regression of the centered design matrix X. We will again let the SVD of the $N \times p$ matrix X assuming that $N \geq p$ be formulated as $X = UDV^\top$ where $U \in \mathbb{R}^{N \times N}$ is an orthogonal matrix whose columns represent left-singular vectors of X spanning the column space of X. The matrix $V \in \mathbb{R}^{p \times p}$ is an orthogonal matrix whose columns represent the right-singular vectors spanning the row space. The matrix $D \in \mathbb{R}^{N \times p}$ is the diagonal matrix containing the singular values of X arranged in descending order $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$. Because U, V are orthogonal matrices, they satisfy the identities $U^\top U = I_N$ and $V^\top V = I_p$. Using this decomposition, the Gram matrix $X^\top X$ is expressed as $X^\top X = V D^2 V^\top$ as we saw before where $D^2 \in \mathbb{R}^{p \times p}$ is the diagonal matrix containing the singular values d_j^2 which are the eigenvalues of $X^\top X$.

As we showed before, assuming that $X^\top X$ is invertible, the OLS fitted values vector can be written using SVD as $\hat{y}_{\text{OLS}} = UU^\top y$ showing that ordinary least squares projects the response vector y directly onto the column space of X using the orthonormal basis U. Writting this projection as sum of outer products,

$$\hat{y}_{\text{OLS}} = \sum_j u_j u_j^\top y,$$

where $u_j \in \mathbb{R}^N$ is the j-th column of U.

To find the corresponding SVD representation for the ridge regression fitted values, the decomposition is substituted into the regularized closed-form expression,

$$\begin{aligned}\hat{y}_{\text{ridge}} &= X\hat{\beta}_{\text{ridge}} = X(X^\top X + \lambda I_p)^{-1}X^\top y \\ &= (UDV^\top)(VD^2V^\top + \lambda VV^\top)^{-1}(VD^\top U^\top)y, \qquad \text{Factor out } V, \\ &= UDV^\top V(D^2 + \lambda I_p)^{-1}V^\top VD^\top U^\top y, \qquad \text{By of Orthogonal Matrix Products}, \\ &= UD(D^2 + \lambda I_p)^{-1}D^\top U^\top y, \qquad V^\top V = I_p.\end{aligned}$$

The product of two diagonal matrices is also diagonal so $D(D^2 + \lambda I_p)^{-1}$ is diagonal. The jth diagonal entry is $\frac{d_j^2}{d_j^2+\lambda}$ so expanding as a summation of the singular vector projections we get,

$$\hat{y}_{\text{ridge}} = \sum_{j=1}^{p} u_j \left(\frac{d_j^2}{d_j^2 + \lambda}\right) u_j^\top y.$$

It is obvious that the ridge regression projects the response vector y onto the orthonormal basis U and scales each coordinate by a shrinkage factor $f_j = \frac{d_j^2}{d_j^2+\lambda}$. Because $\lambda > 0$, the shrinkage factor is strictly less than one meaning that every coordinate is shrunk toward zero. The severity of this shrinkage is proportional to the magnitude of the squared singular value d_j^2 relative to the penalty parameter λ.

Small singular values d_j correspond to directions in the column space of X having small variance. Ridge regression shrinks these directions the most, effectively suppressing the noise-sensitive directions that cause variance inflation.

There are two noteworthy special cases of this shrinkage behaviour.

(i) In the case of orthogonal covariates where $X^\top X = NI_p$, the ridge estimator simplifies to a uniform shrinkage of the OLS estimates, given by $\hat{\beta}_{\text{ridge}} = \frac{N}{N+\lambda}\hat{\beta}_{\text{OLS}}$.

(ii) If two or more variables are identical, their ridge coefficients are mathematically identical so their sum equals what would have been the single ridge coefficient if only one of the variables were included in the model. We refer to this as group-binding behaviour which is characteristic of the quadratic penalty, which distributes the coefficients evenly among identical or highly correlated features.

In wide-data scenarios i.e., when $p >> N$, evaluating the standard estimator requires inverting a massive $p \times p$ matrix which is computationally unfeasible. SVD provides a workaround since X is decomposed into $UDV^\top$ and because the rank of X is at most N, the upper right triangular matrix $R = UD$ is of dimension $N \times p$. The ridge estimator can be re-written to involve only the inversion of an $N \times N$ matrix, reducing the computational operations from $\mathcal{O}(p^3)$ to $\mathcal{O}(pN^2)$.

2.4.4 Dual Formulations and the Woodbury Matrix Identity

In high-dimensional settings, in order to generalize computational efficiency, we use dual forms of the ridge regression estimator, that is, establish equivalence of the primal and dual matrix expressions through direct algebraic substitution. Our target identity is formulated as,

$$(X^\top X + \lambda I_p)^{-1}X^\top = X^\top(XX^\top + \lambda I_N)^{-1}.$$

Both sides of the identity are pre-multiplied by $(X^\top X + \lambda I_p)$ and post-multiplied by $(XX^\top + \lambda I_N)$ so the LHS becomes,

$$(X^\top X + \lambda I_p)\left[(X^\top X + \lambda I_p)^{-1}X^\top\right](XX^\top + \lambda I_N) = X^\top(XX^\top + \lambda I_N) = X^\top XX^\top + \lambda X^\top.$$

The expansion of the RHS also follows as,

$$(X^\top X + \lambda I_p)\left[X^\top(XX^\top + \lambda I_N)^{-1}\right](XX^\top + \lambda I_N) = (X^\top X + \lambda I_p)X^\top = X^\top XX^\top + \lambda X^\top.$$

Because both operations reduce to the identical matrix expression $X^\top X X^\top + \lambda X^\top$, the equivalence is established. Multiplying both sides of the identity on the right by the response vector y yields the dual formulation of the ridge regression estimator,

$$\hat{\beta}_{\text{ridge}} = X^\top (XX^\top + \lambda I_N)^{-1} y.$$

In future sections, we'll see how this dual formulation has implications for kernel-based learning and reproducing kernel Hilbert spaces. By representing the inner products between data points as $K = XX^\top$, the ridge fitted values are computed as $\hat{y} = K(K + \lambda I_n)^{-1} y$ allowing the model to be generalized to infinite-dimensional feature spaces by replacing K with a non-linear kernel function.

2.4.5 Bias and Variance of Ridge

Under the assumption of a true linear model $y = X\beta + \varepsilon$ with mean 0 errors with constant variance, the conditional expectation of the ridge regression estimator is computed as,

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_{\text{ridge}} \mid X] &= \mathbb{E}\left[(X^\top X + \lambda I_p)^{-1} X^\top y \mid X\right] \\
&= (X^\top X + \lambda I_p)^{-1} X^\top \mathbb{E}[y \mid X] \\
&= (X^\top X + \lambda I_p)^{-1} X^\top X \beta.
\end{aligned}$$

For any $\lambda > 0$ this expectation does not equal the true parameter vector β proving that the ridge estimator is biased toward zero. The bias vector is formulated as,

$$\begin{aligned}
\text{Bias}(\hat{\beta}_{\text{ridge}}) &= \mathbb{E}[\hat{\beta}_{\text{ridge}} \mid X] - \beta \\
&= \left[(X^\top X + \lambda I_p)^{-1} X^\top X - I_p\right] \beta \\
&= (X^\top X + \lambda I_p)^{-1} \left[X^\top X - (X^\top X + \lambda I_p)\right] \beta \\
&= -\lambda (X^\top X + \lambda I_p)^{-1} \beta.
\end{aligned}$$

Evidently, the ridge estimator is a continuous function of the regularization parameter λ and the bias vector points in the opposite direction of the true parameter vector β. Moreover, the conditional covariance matrix of the ridge estimator is derived using properties of linear transformations,

$$\begin{aligned}
\text{Var}(\hat{\beta}_{\text{ridge}} \mid X) &= \text{Var}\left((X^\top X + \lambda I_p)^{-1} X^\top y \mid X\right) \\
&= (X^\top X + \lambda I_p)^{-1} X^\top \text{Var}(y \mid X) \cdot X (X^\top X + \lambda I_p)^{-1} \\
&= \sigma^2 (X^\top X + \lambda I_p)^{-1} X^\top X (X^\top X + \lambda I_p)^{-1}.
\end{aligned}$$

By *Loewner Partial Ordering*, we claim that, some coefficient estimate $\hat{\beta}_2$ has smaller variance than another coefficient estimate $\hat{\beta}_1$ if and only if the difference between their covariance matrices is positive definite. Denote the Gram matrix $S := X^\top X$ so the difference between the OLS and ridge covariance matrices is,

$$\begin{aligned}
\Delta_{\text{Var}} &= \text{Var}(\hat{\beta}_{\text{OLS}} \mid X) - \text{Var}(\hat{\beta}_{\text{ridge}} \mid X) \\
&= \sigma^2 S^{-1} - \sigma^2 (S + \lambda I_p)^{-1} S (S + \lambda I_p)^{-1} \\
&= \sigma^2 \left[S^{-1} - (S + \lambda I_p)^{-1} S (S + \lambda I_p)^{-1}\right] \\
&= \sigma^2 \left[S^{-1} - (S + \lambda I_p)^{-1} (S + \lambda I_p - \lambda I_p)(S + \lambda I_p)^{-1}\right] \\
&= \sigma^2 \left[S^{-1} - (S + \lambda I_p)^{-1} + \lambda (S + \lambda I_p)^{-2}\right] \\
&= \sigma^2 \left[S^{-1} - (S + \lambda I_p)^{-1}\right] + \sigma^2 \lambda (S + \lambda I_p)^{-2} \\
&= \sigma^2 \lambda S^{-1} (S + \lambda I_p)^{-1} + \sigma^2 \lambda (S + \lambda I_p)^{-2} \\
&= \sigma^2 \lambda (S + \lambda I_p)^{-1} \left[S^{-1} + (S + \lambda I_p)^{-1}\right].
\end{aligned}$$

Because $S = X^\top X$ is positive definite, its inverse S^{-1} is also positive definite. The sum of the identity matrix $2I_p$ and the positive definite matrix λS^{-1} is strictly positive definite. For any $\lambda > 0$, multiplying by the positive scalar $\lambda\sigma^2$ preserves positive definiteness. Pre- and post- multiplying by symmetric, invertible matrices such as $(S + \lambda I_p)^{-1}$ preserves this property which proves that the difference matrix Δ_{var} is strictly positive definite.

2.4.6 Theobald's Existence Theorem

While providing a reduction in variance, ridge is trivially a biased estimator introduced via its parameter estimates. For ridge to be a viable alternative to OLS, there must exist a range of regularization parameters where the reduction in variance outweighs the increase in bias, resulting in a lower MSE.

We define the MSE matrix of an estimator $\hat{\beta}$ as the expectation of the outer product of its estimation error,

$$\text{MSE}(\hat{\beta}) = \mathbb{E}\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^\top \mid X\right] = \text{Var}(\hat{\beta} \mid X) + \text{Bias}(\hat{\beta})\text{Bias}(\hat{\beta})^\top.$$

For the OLS estimator, which we know is unbiased, the MSE matrix is simply its covariance matrix, $\text{MSE}(\hat{\beta}_{\text{OLS}}) = \sigma^2 S^{-1}$ where $S = X^\top X$. For the ridge estimator, the MSE matrix is formulated as the sum of its covariance matrix and the outer product of its bias vector that is,

$$\text{MSE}(\hat{\beta}_{\text{ridge}}) = \sigma^2(S + \lambda I_p)^{-1} S (S + \lambda I_p)^{-1} + \lambda^2 (S + \lambda I_p)^{-1} \beta\beta^\top (S + \lambda I_p)^{-1}.$$

C.M. Theobald claims that there always exists a strictly positive regularization parameter λ for which the OLS matrix strictly dominates the ridge regression matrix. The difference between these two MSE matrices is simply $\Delta_{\text{MSE}}(\lambda) = \text{MSE}(\hat{\beta}_{\text{OLS}}) - \text{MSE}(\hat{\beta}_{\text{ridge}})$. Using the SVD decomposition and factoring out $(S + \lambda I_p)^{-1}$

$$\begin{aligned}
\Delta_{\text{MSE}}(\lambda) &= \text{MSE}(\hat{\beta}_{\text{OLS}}) - \text{MSE}(\hat{\beta}_{\text{ridge}}) \\
&= (S + \lambda I_p)^{-1} \left[\sigma^2(S + \lambda I_p) S^{-1} (S + \lambda I_p) - \sigma^2 S - \lambda^2 \beta\beta^\top\right] (S + \lambda I_p)^{-1} \\
&= (S + \lambda I_p)^{-1} \left[\sigma^2(S + \lambda I_p) S^{-1} (S + \lambda I_p) - \sigma^2 S - \lambda^2 \beta\beta^\top\right] (S + \lambda I_p)^{-1} \\
&= (S + \lambda I_p)^{-1} \left[\sigma^2(S + 2\lambda I_p + \lambda^2 S^{-1}) - \sigma^2 S - \lambda^2 \beta\beta^\top\right] (S + \lambda I_p)^{-1} \\
&= (S + \lambda I_p)^{-1} \left[2\lambda\sigma^2 I_p + \lambda^2\sigma^2 S^{-1} - \lambda^2 \beta\beta^\top\right] (S + \lambda I_p)^{-1} \\
&= \lambda(S + \lambda I_p)^{-1} \left[\sigma^2(2I_p + \lambda S^{-1}) - \lambda\beta\beta^\top\right] (S + \lambda I_p)^{-1} \\
&= \lambda(S + \lambda I_p)^{-1} T(\lambda) (S + \lambda I_p)^{-1},
\end{aligned}$$

where we let $T = \sigma^2(2I_p + \lambda S^{-1}) - \lambda\beta\beta^\top$. The difference matrix above is strictly positive definite if and only if the inner matrix $T(\lambda)$ is strictly positive definite. Let $a \in \mathbb{R}^p$ be any non-zero vector for $a \neq 0$. The quadratic form associated with $T(\lambda)$ is evaluated as,

$$a^\top T(\lambda) a = a^\top \left[\sigma^2(2I_p + \lambda S^{-1}) - \lambda\beta\beta^\top\right] a = 2\sigma^2\|a\|_2^2 + \lambda\sigma^2 a^\top S^{-1} a - \lambda(a^\top\beta)^2.$$

The conditions under which the quadratic form is strictly positive are well-defined by the Cauchy-Schwarz inequality which is applied to the final term,

$$a^\top T(\lambda) a \geq 2\sigma^2\|a\|_2^2 + \lambda\sigma^2 a^\top S^{-1} a - \lambda\|\beta\|_2^2\|a\|_2^2 = \left(2\sigma^2 - \lambda\|\beta\|_2^2\right)\|a\|_2^2 + \lambda\sigma^2 a^\top S^{-1} a.$$

Again, because $S = X^\top X$ is positive definite, its inverse S^{-1} is also positive definite, meaning the quadratic form $\lambda\sigma^2 a^\top S^{-1} a$ is strictly positive for any $a \neq 0$. Therefore, the entire expression is guaranteed to be strictly positive if the coefficient of the first term is positive, which requires that $2\sigma^2 - \lambda\|\beta\|_2^2 > 0$.

Solving this inequality for λ establishes the bound,

$$0 < \lambda < \frac{2\sigma^2}{\|\beta\|_2^2}.$$

Further optimizations in the quantitative research landscape have yielded formulations to minimize the mean squared error under multicollinearity. Once such formulation is the Modified Unbiased Ridge estimator (MUR) which uses a stochastic constraint to eliminate bias of the traditional ridge estimator while preserving its variance-reduction properties.

The bias, variance and Matrix Mean Squared Error (MMSE) of the MUR estimator is defined as,

$$\begin{aligned}
\text{Bias}(\hat{\beta}_{\text{MUR}}) &= -\lambda(S+\lambda I)^{-1}\beta \\
\text{Var}(\hat{\beta}_{\text{MUR}}) &= \sigma^2 W(S+\lambda I)^{-1}W^\top, \qquad W = I - \lambda(S+\lambda I)^{-1} \\
\text{MMSE}(\hat{\beta}_{\text{MUR}}) &= \sigma^2 W(S+\lambda I)^{-1}W^\top + \lambda^2(S+\lambda I)^{-1}\beta\beta^\top(S+\lambda I)^{-1}.
\end{aligned}$$

2.4.7 The LASSO

Suppose that we have a dataset consisting of N observations and p predictors and let $y \in \mathbb{R}^N$ be our response vector and let $X \in \mathbb{R}^{N\times p}$ denote our design matrix. WLOG, it is assumed that the response vector and predictors have been centered and standardized ensuring that our response has mean of zero and each predictor has a mean of zero and unit ℓ_2-norm which eliminates the need to explicitly penalize the intercept term.

The constraint formulation minimizes RSS to an inequality constraint on the ℓ_1-norm of the coefficient vector,

$$\min_\beta \frac{1}{2}\|y - X\beta\|_2^2 \quad \text{subject to} \quad \|\beta\|_1 \le t.$$

In the formulation above $t \ge 0$ represents a user-specified budget parameter that restricts the allowable magnitude of the coefficients. Alternatively, the penalized formulation adds a separable penalty term proportional to the ℓ_1 norm of the parameters directly to the objective function,

$$\min_\beta \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

$\lambda \ge 0$ is our regularization parameter that governs the trade-off between bias and variance of the coefficient vector. Formally, we use convex analysis and Lagrange multipliers to establish equivalence. The constrained problem can be framed as minimizing the objective function $g(\beta) = \frac{1}{2}\|y - X\beta\|_2^2$ subject to the inequality constraint $h(\beta) = \|\beta\|_1 - t \le 0$. The Lagrangian function for this optimization problem is defined as,

$$\mathcal{L}(\beta, \mu) = \frac{1}{2}\|y - X\beta\|_2^2 + \mu\left(\|\beta\|_1 - t\right)$$

where $\mu \ge 0$ represents the Lagrange multiplier associated with the ℓ_1 constraint. Because the objective function is quadratic and convex, and the constraint set is a closed, convex polytope, the conditions are satisfied for global optimality. We require that there is dual and primal feasibility and complementary slackness with stationarity of the Lagrangian. Suppose $\hat{\beta}_{\text{pen}}$ is the unique minimizer of the penalized objective for a given $\lambda > 0$,

$$\hat{\beta}_{\text{pen}} = \arg\min_\beta \left(\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\right).$$

Setting the budget parameter of the constrained formulation exactly to $t = \|\hat{\beta}_{\text{pen}}\|_1$, any feasible vector β must satisfy $\|\beta\|_1 \le t$ so the global optimality of $\hat{\beta}_{\text{pen}}$ on the unconstrained space implies that for any β, the penalized cost of β is bounded below by the penalized cost of $\hat{\beta}_{\text{pen}}$,

$$\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \ge \frac{1}{2}\|y - X\hat{\beta}_{\text{pen}}\|_2^2 + \lambda\|\hat{\beta}_{\text{pen}}\|_1.$$

Re-arranging the terms of this inequality we get that,

$$\frac{1}{2}\|y - X\beta\|_2^2 \ge \frac{1}{2}\|y - X\hat{\beta}_{\text{pen}}\|_2^2 + \lambda\left(\|\hat{\beta}_{\text{pen}}\| - \|\beta\|_1\right).$$

Since β is feasible in the constrained problem, we know that $\|\beta\|_1 \leq t = \|\hat{\beta}_{\text{pen}}\|_1$ which guarantees that the term $\left(\|\hat{\beta}_{\text{pen}}\| - \|\beta\|_1\right)$ is non-negative. Now, since the regularization parameter λ is strictly posiitve, the second term on the RHS is non-negative which leads to the inequality,

$$\frac{1}{2}\|y - X\beta\|_2^2 \geq \frac{1}{2}\|y - X\hat{\beta}_{\text{pen}}\|_2^2.$$

We conclude that no feasible coefficient vector can achieve a lower residual sum of squares than the penalized solution thereby proving that $\hat{\beta}_{\text{pen}}$ is also a global minimizer of the constrained problem for $t = \|\hat{\beta}_{\text{pen}}\|_1$.

Thenn for the converse, we let $\hat{\beta}_{\text{cons}}$ be the optimal solution to the constrained problem for $t > 0$. If $t \geq \|\hat{\beta}_{\text{OLS}}\|_1$ where $\hat{\beta}_{\text{OLS}}$ is the unconstrained OLS estimator, so the budget constraint is inactive, and the solution corresponds to $\lambda = 0$. However, if the budget parameter is set so that $0 < t < \|\hat{\beta}_{\text{OLS}}\|_1$, the constraint must be active at the optimum, which dictates that $\|\hat{\beta}_{\text{cons}}\|_1 = t$. By the complementary slackness condition with a strictly positive Lagrange multiplier where $\mu = \lambda$ and the stationarity condition of the Lagrangian requires that the zero vector must belong to its subdifferential with respect to β,

$$0 \in \partial_\beta \left(\frac{1}{2}\left\|y - X\hat{\beta}_{\text{cons}}\right\|_2^2\right) + \lambda \partial \left\|\hat{\beta}_{\text{cons}}\right\|_1.$$

The penalized formulation is a sum of convex functions, so the subgradient condition is satisfied and sufficient for $\hat{\beta}_{\text{cons}}$ to be the global minimizer of the penalized objective. This establishes a bi-directional equivalence between the constrained and penalized formulation of the LASSO.

LASSO imposes an ℓ_1 penalty on the coefficients, which replaces the ℓ_2 penalty that ridge imposes. Because the absolute value function $g(x) = |x|$ exhibits a sharp corner at the origin, the ℓ_1 norm penalty is non-differentiable at any point where one or more regression coefficients are exactly equal to zero.

The objective function of the penalized LASSO is formulated as,

$$f(\beta) = \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

This is the sum of a continuously differentiable quadratic function and a non-smooth convex penalty. By applying the additive property of subdifferentials for convex functions, the subdifferential of $f(\beta)$ is the sum of the gradient of the quadratic loss and the subdifferential of the ℓ_1 norm,

$$\partial f(\beta) = -X^\top(y - X\beta) + \lambda \cdot \partial\|\beta\|_1.$$

A coefficient vector $\hat{\beta}$ is a global minimizer of the LASSO objective if and only if the zero vector is contained in its subdifferential which is expressed as $0 \in \partial f(\hat{\beta})$ if and only if $X^\top(y - X\hat{\beta}) = \lambda s$. Decomposing this vector relation into p distinct coordinate-wise optimality conditions, we have that,

$$\partial f(\beta) = \begin{cases} X_j^\top(y - X\hat{\beta}) = \lambda \operatorname{sign}(\hat{\beta}_j) & \text{if } \hat{\beta}_j \neq 0, \\ \left|X_j^\top(y - X\hat{\beta})\right| \leq \lambda & \text{if } \hat{\beta}_j = 0, \end{cases}$$

where $X_j \in \mathbb{R}^N$ represents the jth column of the design matrix X.

2.4.8 Geometry of Sparsity of Feature Selection

We explain the ability of LASSO to perform automatic feature selection by driving coefficients exactly to zero by the geometric interaction between the objective contours and the constraint boundaries. Consider a two-dimensional parameter space containing coefficients β_1 and β_2. The OLS RSS is a quadratic function so its contours of constant error are nested, concentric ellipses centered at the unconstrained OLS estmate $\hat{\beta}_{\text{OLS}}$. Their shape is determined by the empirical covariance of the predictors given by the gram matrix $X^\top X$.

Using the same case, the constraint region defined by the ℓ_1 restriction, $\|\beta\|_1 \leq t$ corresponds to the inequality where $|\beta_1| + |\beta_2| \leq t$. In two-dimensional space, this inequality defines a diamond-shaped region whose vertices are situated on the coordinate axes. In higher-dimensional spaces, this constraint region generalizes so there are more sharp vertices, edges, and low-dimensional faces.

In ridge regression, this region is defined by the quadratic inequality where $\|\beta\|_2^2 \leq t$ which corresponds $\beta_1^2 + \beta_2^2 \leq t$ i.e., a circle in two dimensions and a smooth hyperplane in higher-dimensions which is rotationally invariant and lacks sharp corners or vertices. The constrained LASSO problem aims to find the first point at which this expanding ellipse contour of the RSS touches the constraint boundary. Because the L_1-norm (ball) has sharp vertices on the coordinate axes, it is very probable that the elliptical contour makes initial contact with the constraint boundary at one of these corners. Any contact at a vertex on an axis corresponds to a solution where one of the coordinate values is exactly equal to zero.

As $p \to \infty$ the number of vertices and low-dimensional hyperplanes on the boundary increases, making it more and more likely for the optimal solution to lie on a low-dimensional face where a large subset of the regression coefficients are driven exactly to zero. In contrast, the circular boundary of Ridge regression is smooth and curved, so the elliptical contours will almost certainly touch the circle tangentially at a point that does not lie on any coordinate axis, causing shrunken but non-zero coefficients.

2.4.9 LASSO under Orthogonality

For the case of orthogonality, suppose that the predictors are completely uncorrelated and the design matrix is orthonormal such that $X^\top X = I$. We therefore can derive a closed-form solution for LASSO.

Under orthonormality, the penalized objective function becomes,

$$\begin{aligned}
f(\beta) &= \frac{1}{2}\|y - X\beta\|_2^2 + \lambda \sum_j |\beta_j| \\
&= \frac{1}{2}\left(y^\top y - 2\beta^\top X^\top y + \beta^\top X^\top X\beta\right) + \lambda \sum_j |\beta_j| \\
&= \frac{1}{2} y^\top y - \beta^\top X^\top y + \frac{1}{2}\beta^\top \beta + \lambda \sum_j |\beta_j| \\
&= \frac{1}{2} y^\top y - \beta^\top \hat{\beta}_{\text{OLS}} + \frac{1}{2}\sum_j \beta_j^2 + \lambda \sum_j |\beta_j| \\
&= \frac{1}{2}\sum_j \left(\beta_j - \hat{\beta}_{\text{OLS},j}\right)^2 + \frac{1}{2} y^\top y - \frac{1}{2}\sum_j \hat{\beta}_{\text{OLS},j}^2.
\end{aligned}$$

With respect to our optimization, the terms $\frac{1}{2} y^\top y$ and $\frac{1}{2}\sum_j \hat{\beta}_{\text{OLS},j}^2$ are constant wrt the optimization variable β. Thus, minimizing the overall objective $f(\beta)$ is the same as minimizing the simplified function. Evidently, the simplified objective function is a fully separable function with respect to each individual coordinate β_j which allows us to solve the problem independently for each parameter. Using coordinate-wise subdifferentials of $f_j(\beta_j)$ evaluated and set to zero, we get that,

$$\partial_{\beta_j} f_j(\beta_j) = \beta_j - \hat{\beta}_{\text{OLS},j} + \lambda s_j = 0,$$

so that solving gives $\hat{\beta}_{\text{OLS},j} = \beta_j + \lambda s_j$. $s_j \in \partial|\beta_j|$ is the subgradient of the absolute value function. There are three separate conditions in which we must analyze this optimality equation.

Consider the case that $\beta_j > 0$ and the subgradient is $s_j = 1$. The equaton therefore reduces to,

$$\beta_j + \lambda = \hat{\beta}_{\text{OLS},j} \implies \beta_j = \hat{\beta}_{\text{OLS},j} - \lambda.$$

For the assumption that $\beta_j > 0$ to hold, the unconstrained OLS estimate must satisfy that $\hat{\beta}_{\text{OLS},j} - \lambda > 0$.

Consider the case that $\beta_j < 0$ and the subgradient is $s_j = -1$ then the equation reduces to,

$$\beta_j - \lambda = \hat{\beta}_{\text{OLS},j}.$$

Solving for β_j we see that $\beta_j = \hat{\beta}_{\text{OLS},j} + \lambda$. For the assumption to hold, the unconstrained OLS estimate must satisfy $\hat{\beta}_{\text{OLS},j} + \lambda < 0$ which implies that $\hat{\beta}_{\text{OLS},j} < -\lambda$. Third, if $\beta_j = 0$ then we have that the subgradient is any value $s_j \in [-1, 1]$ and the optimality condition becomes,

$$0 + \lambda s_j = \hat{\beta}_{\text{OLS},j} \implies s_j = \frac{\hat{\beta}_{\text{OLS},j}}{\lambda}.$$

Since the subgradient s_j must lie within the closed interval $[-1, 1]$ this case is valid if and only if the magnitude of the OLS estimates satisfies the inequality,

$$\left| \frac{\hat{\beta}_{\text{OLS},j}}{\lambda} \right| \leq 1 \iff \left| \hat{\beta}_{\text{OLS},j} \right| \leq \lambda.$$

That is, the magnitude of the OLS estimate must be at most λ. Combining these three distinct cases yields the soft-thresholding operator denoted S_λ where $\hat{\beta}_j = S_\lambda(\hat{\beta}_{\text{OLS},j})$ which is simply $\text{sign}(\hat{\beta}_{\text{OLS},j}) \max\left(0, \left|\hat{\beta}_{\text{OLS},j}\right| - \lambda\right)$. It is clear that the soft-thresholding done by LASSO is simply a translation of each coefficient by a constant factor λ and truncating at zero.

2.4.10 Generalization of Ridge and LASSO

Supose that we have the general q-level criterion,

$$\tilde{\beta} = \arg\min_{\beta} \left\{ \sum_i \left(y_i - \beta_0 - \sum_j x_{ij}\beta_j \right)^2 + \lambda \sum_j |\beta_j|^q \right\}, \quad q \geq 0.$$

When $q = 0$, the Bayes criterion above corresponds to variable subset selection as the penalty simply counts the number of non-zero parameters. When $q = 1$, the criterion corresponds to the LASSO (representing an ℓ_1 regularization tuning parameter) while $q = 2$ corresponds to the ℓ_2 penalization term in Ridge. It is also important to note that when $q = 1$, the prior is an independent double exponential or Laplace distribution for each input. Such q is the smallest q such that the the constraint region is convex. It is often not useful to attempt to estimate the most optimal q from data for the extra variance possibly incurred. Value of $q \in (1, 2)$

3 Chapter 3 Exercises

3.1 Exercise 3.1

We aim to show equality between the two identities,

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(N - p_1 - 1)} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}.$$

Proof. First, we examine the numerator of the formula on the LHS, which is, by definition, the formula for an F-statistic. Let RSS_0 denote the residual sum of squares of the reduced model and let the residual sum of squares be RSS_1 for the full model. Then the corresponding predictor count for the full and reduced model are p_1 and p_0, respectively. Trivially, since we remove one coefficient to yield our reduced model, the difference $p_1 - p_0$ must be one since the difference in the number of predictors is one. This simplifies to,

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(N - p_1 - 1)} = \frac{(RSS_0 - RSS_1)/1}{RSS_1/(N - p_1 - 1)} = \frac{RSS_0 - RSS_1}{RSS_1/(N - p_1 - 1)}.$$

Notice the denominator is simply estimated $\hat{\sigma}^2$ using the residual sum of squares of our full model with $N - p_1 - 1$ of freedom so we may re-write the denominator as,

$$F = \frac{RSS_0 - RSS_1}{RSS_1/(N - p_1 - 1)} = \frac{RSS_0 - RSS_1}{\hat{\sigma}^2}.$$

It remains for us to show that the denominator is equivalent to $\hat{\beta}_j^2/v_j$ where v_j is the jth predictor.

WLOG, suppose that we have p predictors $\{x_0, \ldots, x_p\}$ and we wish to analyze the effect of reducing our model i.e., removing a predictor. Suppose we apply Gram-Schmidt onto the p predictors we that we are left with an orthogonal set of vectors $\{z_0, \ldots, z_p\}$. We can say that the z_i's span the x_i's (our original predictors from which the orthogonalized predictors are derived) and that by definition $z_i^\top z_j = 0$ for all $i \neq j$.

Note that z_p is the vector defined as $z_p = x_p - \hat{x}_p$ i.e., the portion of x_p not explained by the other predictors. $\hat{x}_p$ in this case is a fitted value for which we regress x_p on the remaining $x_0, x_1, \ldots, x_{p-1}$ predictors. The orthogonal vector z_p is therefore new information in x_p after adjusting for the other variables.

Our models are therefore written as,

$$\hat{y}_{\text{full}} = \sum_{k=0}^{p} \frac{z_k^\top y}{z_k^\top z_k} z_k, \quad \hat{y}_{\text{reduced}} = \sum_{k=0}^{p-1} \frac{z_k^\top y}{z_k^\top z_k} z_k.$$

Finally, since subtracting the two fitted vectors gives,

$$\hat{y}_{\text{full}} - \hat{y}_{\text{reduced}} = \underbrace{\frac{z_p^\top y}{z_p^\top z_p}}_{=\hat{\beta}_p} \cdot z_p = \hat{\beta}_p \cdot z_p.$$

The difference between the residual sum of squares between the reduced and full model, by properties of the Euclidean Norm, can be re-written as the following:

$$RSS_{\text{reduced}} - RSS_{\text{final}} = \|\hat{y}_{\text{full}} - \hat{y}_{\text{reduced}}\|_2^2 = \|\hat{\beta}_p z_p\|^2 = \hat{\beta}_p^2 \|z_p\|^2 = \hat{\beta}_p^2 z_p^\top z_p$$

It remains to explain the vector product $\hat{\beta}_p^2 z_p^\top z_p$. Recall that if we have a model $Y = X\beta + \varepsilon$ that the coefficient estimates can be formulated using the orthogonal vectors we obtained via Gram-Schmidt. Specifically, we had used $\hat{\beta}_p = \frac{z_p^\top y}{z_p^\top z_p}$ so that the full formulation of the p-th coefficient can be written as,

$$\hat{\beta}_p + \beta_p + \frac{z_p^\top \varepsilon}{z_p^\top z_p}.$$

We derive the variance formula for $\hat{\beta}_p$ to connect the vector product to the z-score notation. Using the same orthogonalized form mentioned before and assumptions under a linear model as well as z_p orthogonality, the variance is as follows:

$$\text{Var}(\hat{\beta}_p) = \text{Var}\left(\frac{z_p^\top \varepsilon}{z_p^\top z_p}\right) = \frac{z_p^\top \text{Var}(\varepsilon) z_p}{(z_p^\top z_p)^2} = \frac{z_p^\top (\sigma^2 I) z_p}{(z_p^\top z_p)^2} = \frac{\sigma^2 z_p^\top z_p}{(z_p^\top z_p)^2} = \frac{\sigma^2}{z_p^\top z_p}.$$

By proposition, recall that v_p is the pth diagonal element of the inverse of the Gram matrix $(X^\top X)$ that is $v_p = \left[(X^\top X)^{-1}\right]_{pp}$. So we see that $\sigma^2 v_p = \frac{\sigma^2}{z_p^\top z_p}$. Naturally, $v_p = \frac{1}{z_p^\top z_p}$ or equivalently $z_p^\top z_p = \frac{1}{v_p}$.

Therefore, the difference in the residual sum of squares between our reduced and full model is,

$$RSS_{\text{reduced}} - RSS_{\text{full}} = \hat{\beta}_p^\top z_p^\top z_p = \hat{\beta}_p^\top \frac{1}{v_p}.$$

Substituting this into the numerator of our simplified F statistic,

$$F = \frac{\hat{\beta}_p^2 / v_p}{\hat{\sigma}^2} = \frac{\hat{\beta}_p^2}{\hat{\sigma}^2 v_p} = \left(\underbrace{\frac{\hat{\beta}_p}{\hat{\sigma}\sqrt{v_p}}}_{:=z_j}\right)^2 = z_j^2,$$

as required. □

3.2 Exercise 3.2

Proof. Refer to my `Jupyter` notebook for the full implementation in `Python`.

Conclusion: I conclude that the simultaneous 95% confidence band is wider than the pointwise 95% confidence band at every value of x. In this simulation, the simultaneous band is about,

$$\frac{\sqrt{4 \cdot F_{4,56}(0.95)}}{t_{0.975,56}} \approx \frac{3.19}{2.00} \approx 1.59,$$

times as wide as the pointwise band. For example, near the center of the data, the pointwise band width is roughly 0.8 while the simultaneous band is roughly 1.25. Near the endpoints, the pointwise width is roughly 2.0 while the simultaneous width is roughly 3.2. Therefore, the simultaneous band is approximately 59% wider across the entire curve. That is, the pointwise band gives 95% confidence for $f(x_0)$ at one fixed value of x_0 while the simultaneous band gives 95% confidence for the entire cubic regression curve at once, so the simultaneous band must be wider to protect against errors over all x-values simultaneously. □

3.3 Exercise 3.3

Proof. For some estimate $\theta = a^\top \beta$ and some other arbitrary linear unbiased estimate of the form $\tilde{\theta} = c^\top Y$, we aim to prove that $\text{Var}(a^\top \hat{\beta}) \leq \text{Var}(c^\top Y)$ for every vector c such that $c^\top Y$ is unbiased for $a^\top \beta$.

Again, we assume linearity in our parameters with $\mathcal{N}(0, \sigma^2 I_N)$ distributed errors ε with a full-rank design matrix X such that $X^\top X$ is invertible. The least squares estimator is thus given as $\hat{\beta} = a^\top (X^\top X)^{-1} X^\top Y$.

We aim to write $a^\top \hat{\beta} = a^\top (X^\top X)^{-1} X^\top Y$ as a linear estimator in Y. Suppose we define $d := X(X^\top X)^{-1} a$ so that $d^\top = [X(X^\top X)^{-1} a]^\top$. Since $X^\top X$ is symmetric, then its inverse is also symmetric. Hence we have that $d^\top a^\top (X^\top X)^{-1} X^\top$. Therefore we see that $d^\top Y = a^\top (X^\top X)^{-1} X^\top Y = a^\top \hat{\beta}$. The least squares estimator of $a^\top \beta$ is itself a linear estimator of Y meaning that $a^\top \hat{\beta} = d^\top Y$.

Suppose we have another linear unbiased estimator of $a^\top \beta$ that is $c^\top Y$. Since it is unbiased, we require that $\mathbb{E}[c^\top Y] = a^\top \beta$ trivially. Using the model $Y = X\beta + \varepsilon$ we have that $\mathbb{E}[Y] = X\beta$ since $\mathbb{E}[\varepsilon] = 0$.

Therefore, $\mathbb{E}[c^\top Y] = c^\top \mathbb{E}[Y] = c^\top X\beta$.

For $c^\top Y$ to be unbiased for $a^\top\beta$ we need $c^\top X\beta = a^\top\beta$ for every possible β. Since this must hold for all β we must have that $c^\top X = a^\top$ or equivalently $X^\top c = a$. It remains for us to check that the least squares weight vector d satisfies the same unbiasedness condition. Since $d = X(X^\top X)^{-1}a$ we get that $X^\top d = X^\top X(X^\top X)^{-1}a = a$ so both c and d satisfy $X^\top c = a$ and $X^\top d = a$. Subtracting the two equations gives,

$$X^\top c - X^\top d = X^\top(c-d) = a - a = 0.$$

Define $r = c - d$. Then $X^\top r = 0$ meaning that r is orthogonal to every column in X or in other words $r \in \mathrm{Col}(X)^\perp$. However, we had defined $d = X(X^\top X)^{-1}a$. Since d is X multiplied by some vector, d is a linear combination of the columns of X. Therefore we have that $d \in \mathrm{Col}(X)$. Clearly, we have decomposed the arbitrary unbiased estimator weight vector c as $c = d + r$ where $d \in \mathrm{Col}(X)$ and $r \in \mathrm{Col}(X)^\perp$. Therefore it must be that $d^\top r = 0$ as well. The vector d is the least squares weight vector and every other unbiased weight vector c differs from d by adding some extra component r orthogonal to the column space of X. That extra component cannot help with unbiasedness because it is invisible to X but it does add variance.

Now we compute the variance of the least squares estimator,

$$\begin{aligned}
\mathrm{Var}(a^\top\hat\beta) &= \mathrm{Var}(d^\top Y) \\
&= d^\top \mathrm{Var}(Y) d \\
&= d^\top(\sigma^2 I_N)d \\
&= \sigma^2 d^\top d.
\end{aligned}$$

The variance of the arbitrary unbiased linear estimator $c^\top Y$ follows,

$$\begin{aligned}
\mathrm{Var}(c^\top Y) &= c^\top \mathrm{Var}(Y) c \\
&= c^\top(\sigma^2 I_N)c \\
&= \sigma^2 c^\top c \\
&= \sigma^2(d+r)^\top(d+r) \\
&= \sigma^2(d^\top d + d^\top r + r^\top d + r^\top r) \\
&= \sigma^2(d^\top d + r^\top r) \\
&= \sigma^2 c^\top c \\
&\geq \sigma^2 d^\top d.
\end{aligned}$$

Thus, since $\sigma^2 c^\top c = \mathrm{Var}(c^\top Y)$ and $\sigma^2 d^\top d = \mathrm{Var}(d^\top Y)$ then we have shown that $\mathrm{Var}(c^\top Y) \geq \mathrm{Var}(a^\top\hat\beta)$.

Now, we can prove the matrix version of the Gauss-Markov result. We showed above that for every fixed vector a, the least squares estimator $a^\top\hat\beta$ has variance no bigger than any other linear unbiased estimator of $a^\top\beta$. We now translate this scalar statement into a matrix inequality.

Again, we assume a linear model with $\mathcal{N}(0, \sigma^2 I_N)$ distributed errors ε and $X \in \mathbb{R}^{N\times(p+1)}$ with full column rank. Given our least squares estimator $\hat\beta$ let $\hat V = \mathrm{Var}(\hat\beta)$ be the variance-covariance matrix of the least squares estimator and let $\tilde V = \mathrm{Var}(\tilde\beta)$ be the variance-covariance matrix of any other linear unbiased estimator $\tilde\beta$ of β. We wish to show that $\hat V \preceq \tilde V$. By definition, we need to prove that $\tilde V - \hat V$ is positive semi-definite.

Since $\tilde\beta$ is assumed to be a linear estimator of β then there exists some matrix $L \in \mathbb{R}^{(p+1)\times N}$ such that $\tilde\beta = LY$. For $\tilde\beta$ to be unbiased we require that $\mathbb{E}[\tilde\beta] = \beta$. But, $\mathbb{E}[\tilde\beta] = \mathbb{E}[LY] = L\mathbb{E}[Y]$. So, it follows that $\mathbb{E}[\tilde\beta] = LX\beta$.

Therefore, unbiasedness requires that $LX\beta = \beta$ for every possible β hence $LX = I_{p+1}$.

We write the least squares estimator in the same linear form and define $(X^\top X)^{-1}X^\top$ and let $\hat\beta = MY$. Notice that M is also unbiased because $MX = (X^\top X)^{-1}X^\top X = I_{p+1}$. Comparing L to the least squares matrix M we can define the difference as $C = L - M$ so then $L = M + C$.

Using the unbiasedness conditions $LX = I$ and $MX = I$ we see that,

$$CV = (L - M)X = LX - MX = I - I = 0.$$

Therefore $CX = 0$ which means that the extra part C added to the least squares estimator is invisible to the signal space generated by X. It does not help preserve unbiasedness and only adds extra noise.

It remains for us to compute the variance-covariance matrix of $\tilde{\beta}$. Since we know that $\tilde{\beta} = LY$ and we have that $\tilde{V} = \text{Var}(\tilde{\beta}) = \text{Var}(LY)$, using $\text{Var}(Y) = \sigma^2 I_N$ we get,

$$\begin{aligned}
\tilde{V} &= \text{Var}(LY) \\
&= L\,\text{Var}(Y)L^\top \\
&= L(\sigma^2 I_N)L^\top \\
&= \sigma^2 LL^\top \\
&= \sigma^2(M + C)(M + C)^\top \\
&= \sigma^2(MM^\top + MC^\top CM^\top CC^\top) \\
&= \sigma^2(MM^\top + CC^\top).
\end{aligned}$$

Since $\hat{\beta} = MY$ then $\hat{V}$ is,

$$\begin{aligned}
\hat{V} &= \text{Var}(\hat{\beta}) \\
&= M\,\text{Var}(Y)M^\top \\
&= M(\sigma^2 I_N)M^\top \\
&= \sigma^2 MM^\top.
\end{aligned}$$

Then $\tilde{V} = \hat{V} + \sigma^2 CC^\top$ which means that $\tilde{V} - \hat{V} = \sigma^2 CC^\top$. Suppose we have some vector $u \in \mathbb{R}^{p+1}$ that is arbitrary. Then,

$$u^\top(\tilde{V} - \hat{V})u = u^\top(\sigma^2 CC^\top)u.$$

Pulling out σ^2 we see that,

$$u^\top(\tilde{V} - \hat{V})u = \sigma^2 u^\top CC^\top u.$$

However by properties of matrices,

$$u^\top CC^\top u = (C^\top u)^\top(C^\top u) = \|C^\top u\|_2^2.$$

Therefore it follows that,

$$u^\top(\tilde{V} - \hat{V})u = \sigma^2\|C^\top u\|_2^2.$$

Since $\sigma^2 > 0$ and squared norms are always non-negative, then $\sigma^2\|C^\top u\|_2^2 > 0$ so $u^\top(\tilde{V} - \hat{V})u \geq 0$ for every vector u. Therefore $\tilde{V} - \hat{V}$ is positive semi-definite. Hence $\hat{V} \preceq \tilde{V}$. Moreover, the least squares is not only the best for every scalar linear functional $a^\top\beta$, its entire covariance matrix is smaller in the positive semi-definite ordering. Equivalently, every direction $u^\top\beta$ has no larger variance under least squares than under any other linear unbiased estimator, as required. □

3.4 Exercise 3.4 (a) and (b)

Proof. Suppose that $X \in \mathbb{R}^{N\times(p+1)}$ has full column rank. Suppose that we also wish to solve the least squares minimization problem formulated as,

$$\hat{\beta} = \arg\min_\beta \|y - X\beta\|_2^2.$$

Notice, we do not need to explicitly compute $X^\top X$ nor do we need to invert it. Rather, we can run Gram-Schmidt once on the columns of X. Running GS-Orthogonalization using Algorithm 3.1, we obtain an orthonormal basis $\{q_0, q_1, \ldots, q_p\}$ for the same column space as X. Putting these columns into a matrix, we obtain matrix Q such that its columns are orthonormal i.e., $Q^\top Q = I$.

At the same time, Gram-Schimdt tells us how to reconstruct each original column x_j from the orthonormal columns $q_0, \ldots, q_j$. Specifically, we can take $x_j = \sum_i r_{ij} q_i$ where $r_{ij} = q_i^\top x_j$ for $i < j$ and get,

$$r_{jj} = \left\| x_j - \sum_{i=1}^{j-1} r_{ij} q_i \right\|_2.$$

Because x_j only uses $q_0, \ldots, q_j$, the coefficients form an upper triangular matrix R. Therefore, the Gram-Schmidt orthogonalization gives the QR decomposition $X = QR$. Substituting this back into the least squares problem,

$$\hat{\beta} = \arg\min_\beta \|y - X\beta\|_2^2 = \arg\min_\beta \|y - QR\beta\|_2^2.$$

Suppose that $\theta = R\beta$. Since X has full column rank, it is invertible, so minimizing over β is equivalent to minimizing over θ. Thus,

$$\min_\beta \|y - QR\beta\|_2^2 = \min_\theta \|y - Q\theta\|_2^2.$$

Now, this is easy because the columns of Q are orthonormal. The least squares projection of y onto $\mathrm{Col}(Q)$ is simply $\hat{y} = QQ^\top y$. If $\hat{y} = Q\hat{\theta}$ then $Q\hat{\theta} = QQ^\top y$. Multiplying both sides by $Q^\top$ we get that $Q^\top Q\hat{\theta} = Q^\top QQ^\top y$. Since $QQ^\top = I$ then this is simply $\hat{\theta} = Q^\top y$. Therefore $R\hat{\beta} = QQ^\top y$.

Since R is upper triangular and invertible we have that,

$$\hat{\beta} = R^{-1} Q^\top y.$$

We need not explicitly compute R^{-1} and instead solve the triangular system $R\hat{\beta} = Q^\top y$ by backsubstitution. Therefore, a single pass of GS produces Q and R. Simultaneously, once each q_j has been constructed, we can also compute $q_j^\top y$ so collecting gives us $Q^\top y$ and $R\hat{\beta}$.

For example, since R is upper triangular, the last coefficient satisfies $r_{pp}\hat{\beta}_p = q_p^\top y$ so $\hat{\beta}_p = \frac{q_p^\top y}{r_{pp}}$. The second last coefficient then satisfies the equation,

$$r_{p-1,p-1}\hat{\beta}_{p-1} + r_{p-1,p}\hat{\beta}_p = q_{p-1}^\top y.$$

Re-arranging the equation above we get an explicit formulation for $\hat{\beta}_p$,

$$\hat{\beta}_p = \frac{q_{p-1}^\top y - r_{p-1,p}\hat{\beta}_p}{r_{p-1,p-1}}.$$

Continuing backward gives every coordinate of $\hat{\beta}$ so in general,

$$\hat{\beta}_j = \frac{q_j^\top y - \sum_{k=j+1}^{p} r_{jk}\hat{\beta}_k}{r_{kk}}, \quad j = p, p-1, \ldots, 0.$$

So the least squares coefficients can be obtained from one Gram-Schmidt pass because that pass simultaneously gives the orthonormal basis Q and the upper triangular coordinate matrix R along with the projected coordinates $Q^\top y$. The final computation is only the triangular solve $R\hat{\beta} = Q^\top y$.

Since Gram-Schmidt constructs an orthonormal coordinate system for the same column space, the fitted vector is unchanged, but the coefficient computation becomes the simple back-substitution problem. □

3.5 Exercise 3.5

Proof. Recall that our original Ridge problem is,

$$\hat{\beta}_{\text{ridge}} = \arg\min_\beta \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

Only the slope coefficients $\beta_1, \ldots, \beta_p$ are penalized and the intercept β_0 is not penalized. We must show that this is equivalent to the centered-predictor problem,

$$\hat{\beta}^c = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left[y_i - \beta_0^c - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)\, \beta_j^c \right]^2 + \lambda \sum_{j=1}^{p} (\beta_j^c)^2 \right\}.$$

Centering the predictors does not change the set of fitted values we can produce and only changes how the intercept is written. Starting with the original linear predictor we can add and subtract $\bar{x}_j$ inside each predictor $x_{ij} = (x_{ij} - \bar{x}_j) + \bar{x}_j$. Substituting this into the linear predictor gives,

$$\begin{aligned}
\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j &= \beta_0 + \sum_j \left[(x_{ij} - \bar{x}_j) + \bar{x}_j \right] \beta_j \\
&= \beta_0 + \sum_j (x_{ij} - \bar{x}_j)\beta_j + \sum_j \bar{x}_j \beta_j \\
&= \left(\beta_0 + \sum_j \bar{x}_j \beta_j \right) + \sum_j (x_{ij} - \bar{x}_j)\beta_j \\
&= \beta_0^c + \sum_j (x_{ij} - \bar{x}_j)\beta_j^c.
\end{aligned}$$

where we have let $\beta_0^c = \beta_0 + \sum_j \bar{x}_j \beta_j$ and $\beta_j^c = \beta_j$ for $j = 1, \ldots, p$.

Therefore the residuals are exactly the same so we can write,

$$y_i - \beta_0 - \sum_j x_{ij}\beta_j = y_i - \beta_0^c - \sum_j^p (x_{ij} - \bar{x}_j)\beta_j^c.$$

Also, since we have defined $\beta_j^c = \beta_j$ for $j \geq 1$, the ridge penalty is exactly the same,

$$\lambda \sum_j \beta_j^2 = \lambda \sum_j (\beta_j^c)^2.$$

Therefore, the full objective value is unchanged under the transformation meaning the two optimization problems are equivalent. The correspondence is therefore $\beta_j^c = \beta_j$ for $j = 1, \ldots, p$ with the intercept $\beta_0^c = \beta_0 + \sum_j \bar{x}_j \beta_j$.

Equivalently, solving for the original intercept yields,

$$\beta_0 = \beta_0^c - \sum_j \bar{x}_j \beta_j^c.$$

Clearly, the slopes are unchanged but the intercept shifts to "absorb" the column means of the predictors. It remains for us to characterize the solution to the centered problem.

Let $x_{ij}^c = x_{ij} - \bar{x}_j$ so that the ridge criterion becomes,

$$L_c(\beta_0^c, \beta_1^c, \ldots, \beta_p^c) = \sum_i \left[y_i - \beta_0^c - \sum_j x_{ij}^c \beta_j^c \right]^2 + \lambda \sum_j (\beta_j^c)^2.$$

For a fixed slope vector $\beta^c_{1:p}$ we minimize wrt β^c_0 and differentiate wrt β^c_0,

$$\begin{aligned}
\frac{\partial L_c}{\partial \beta^c_0} &= -2\sum_i \left[y_i - \beta^c_0 - \sum_j x^c_{ij}\beta^c_j \right] = 0 \\
&= \sum_i y_i - \sum_i \beta^c_0 - \sum_i \sum_j x^c_{ij}\beta^c_j \\
&= \sum y_i - N\beta^c_0 - \sum_j \beta^c_j \sum_i x^c_{ij} \\
&= \sum y_i - N\beta^c_0 - \sum_j \beta^c_j \sum_i (x_{ij} - \bar{x}_j) \\
&= \sum y_i - N\beta^c_0 - \sum_j \beta^c_j \left(\sum_i x_{ij} - \sum_i \bar{x}_j \right) \\
&= \sum y_i - N\beta^c_0 - \sum_j \beta^c_j \left(\sum_i x_{ij} - N\bar{x}_j \right) \\
&= \sum y_i - N\beta^c_0 - \sum_j \beta^c_j \cancel{(N\bar{x}_j - N\bar{x}_j)} \\
&= \sum y_i - N\beta^c_0 \implies \sum_i y_i = N\beta^c_0 \implies \hat{\beta}^c_0 = \bar{y}.
\end{aligned}$$

Conceptually, once the predictors are centered, the unpenalized intercept is just the sample mean of the response. Substituting $\hat{\beta}^c_0 = \bar{y}$ back into the objective the centered problem for $\hat{\beta}^c_{1:p}$ becomes,

$$\hat{\beta}^c_{1:p} = \arg\min_b \left\{ \sum_i \left[(y_i - \bar{y}) - \sum_j (x_{ij} - \bar{x}_j) b_j \right]^2 + \lambda \sum_j \beta_j^2 \right\}.$$

Suppoose we let X_c be the matrix of centered predictors where $x_{ij} - \bar{x}_i$ for $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, p\}$ and the centered responses $y_c = y - \bar{y}\mathbf{1}$. The slope part of the criterion is as follows,

$$\begin{aligned}
L_c(b) &= \|y_c - X_c b\|_2^2 + \lambda \|b\|_2^2 \\
&= (y_c - X_c b)^\top (y_c - X_c b) + \lambda b^\top b \\
&= y_c^\top y_c - 2b^\top X_c^\top y_c + b^\top X_c^\top X_c b + \lambda b^\top b \\
&= y_c^\top y_c - 2b^\top X_c^\top y_c + b^\top (X_c^\top X_c + \lambda I) b.
\end{aligned}$$

Differentiating wrt to b we obtain the gradient,

$$\begin{aligned}
\nabla_b L_c(b) &= -2X_c^\top X_c + 2(X_c^\top X_c + \lambda I)b = 0 \\
&= -2X_c^\top X_c + 2(X_c^\top X_c + \lambda I)b \\
&= X_c^\top y_c + (X_c^\top X_c + \lambda I)b = 0 \\
X_c^\top y_c &= (X_c^\top X_c + \lambda I)b \\
\implies \hat{\beta}^c_{1:p} &= (X_c^\top X_c + \lambda I)^{-1} X_c^\top y_c.
\end{aligned}$$

For $\lambda > 0$, $X_c^\top X_c + \lambda I$ is positive definite and therefore invertible so the centered ridge slope estimate is formulated fully above. Putting the intercept and the slopes together, the solution to the centered problem is $\hat{\beta}^c_0 = \bar{y}$ and $\hat{\beta}^c_{1:p} = (X_c^\top X_c + \lambda I)^{-1} X_c^\top (y - \bar{y}\mathbf{1})$, giving us our result as required. □

3.6 Exercise 3.12

Proof. Assume $X \in \mathbb{R}^{N\times p}$ is centered and $y \in \mathbb{R}^N$ is our centered response. Notice, we do not need to carry the intercept into our algebra. Define the augmented design matrix and augmented response vector respectively as,

$$X_* = \begin{bmatrix} X \\ \sqrt{\lambda} I_p \end{bmatrix} \in \mathbb{R}^{(N+p)\times p}, \qquad y_* = \begin{bmatrix} y \\ 0_p \end{bmatrix} \in \mathbb{R}^{N+p}.$$

0_p is the p-dimensional zero vector. The OLS problem on the augmented data is therefore formulated as,

$$\hat{\beta}_* = \arg\min_{\beta} \|y_* - X_*\beta\|_2^2.$$

Substituting the definition of the augmented design matrix and response vector we get,

$$\begin{aligned} y_* - X_*\beta &= \begin{bmatrix} y \\ 0_p \end{bmatrix} - \begin{bmatrix} X \\ \sqrt{\lambda} I_p \end{bmatrix} \beta \\ &= \begin{bmatrix} y - X\beta \\ 0_p - \sqrt{\lambda}\beta \end{bmatrix} \\ &= \begin{bmatrix} y - X\beta \\ -\sqrt{\lambda}\beta \end{bmatrix}. \end{aligned}$$

Taking squared Euclidean norms, the squared norm of a stacked vector is the sum of squared norms of its blocks,

$$\begin{aligned} \|y_* - X_*\beta\|_2^2 &= \left\| \begin{bmatrix} y - X\beta \\ -\sqrt{\lambda}\beta \end{bmatrix} \right\|_2^2 \\ &= \|y - X\beta\|_2^2 + \|_2^2 - \sqrt{\lambda}\beta\|_2^2 \\ &= \|y - X\beta\|_2^2 + \|_2^2 + \|\sqrt{\lambda}\beta\|_2^2 \\ &= \|y - X\beta\|_2^2 + \|_2^2 + (\sqrt{\lambda}\beta)^\top(\sqrt{\lambda}\beta) \\ &= \|y - X\beta\|_2^2 + \|_2^2 + \lambda\beta^\top\beta \\ &= (y - X\beta)^\top(y - X\beta) + \lambda\beta^\top\beta. \end{aligned}$$

Therefore the minimization problem becomes,

$$\arg\min_{\beta} \|y_* - X_*\beta\|_2^2 = \arg\min_{\beta} \left\{ \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \right\}.$$

So OLS on the augmented data gives exactly the ridge regression estimate. We can also prove this directly through the normal equations. The OLS on the augmented data gives,

$$\begin{aligned} \hat{\beta}_* &= (X_*^\top X_*)^{-1} X_*^\top y_* \\ &= \left(\begin{bmatrix} X \\ \sqrt{\lambda} I_p \end{bmatrix}^\top \begin{bmatrix} X \\ \sqrt{\lambda} I_p \end{bmatrix} \right)^{-1} \begin{bmatrix} X \\ \sqrt{\lambda} I_p \end{bmatrix}^\top y_* \\ &= \left(X^\top X + \underbrace{(\sqrt{\lambda} I_p)(\sqrt{\lambda} I_p)}_{=\lambda I_p} \right)^{-1} \begin{bmatrix} X \\ \sqrt{\lambda} I_p \end{bmatrix}^\top y_* \\ &= \left(X^\top X + \lambda I_p \right)^{-1} \left(X^\top y + \cancel{(\sqrt{\lambda} I_p) \cdot 0_p} \right) \\ &= \left(X^\top X + \lambda I_p \right)^{-1} X^\top y \\ &= \hat{\beta}_{\text{ridge}}. \end{aligned}$$

Therefore we have that $\hat{\beta}_* = \hat{\beta}_{\text{ridge}}$ as required. □

3.7 Exercise 3.16

Proof. Assume that the columns of our design matrix $X \in \mathbb{R}^{N\times p}$ are orthonormal so that $X^\top X = I_p$. Assume that also the variables have already been centered so we can ignore the intercept. The ordinary least squares estimator is $\hat{\beta}_{\text{OLS}} = (X^\top X)^{-1}X^\top y$. Since $X^\top X = I_p$, this becomes $\hat{\beta}_{\text{OLS}} = X^\top y$.

Let $\hat{\beta}_j$ denote the jth ordinary least squares coefficient so that $\hat{\beta}_j = x_j^\top y$.

Starting from the residual sum of squares we get,

$$\begin{aligned}
\text{RSS}(\beta) &= \|y - X\beta\|_2^2 \\
&= (y - X\beta)^\top (y - X\beta) \\
&= y^\top y - 2\beta^\top X^\top y + \beta^\top X^\top X\beta \\
&= y^\top y - 2\beta^\top \hat{\beta} + \beta^\top \beta \\
&= y^\top y - 2\sum_{j=1}^{p} \beta_j \hat{\beta}_j + \sum_{j=1}^{p} \beta_j^2 \\
&= y^\top y + \sum_j \left[(\beta_j - \hat{\beta}_j)^2 - \hat{\beta}_j^2\right] \\
&= \sum_j (\beta_j - \hat{\beta}_j)^2 + \left(y^\top y - \sum_j \hat{\beta}_j^2\right).
\end{aligned}$$

Notice however that the 2nd term in parentheses does not depend on β so every optimization in Table 3.4 reduces to minimizing a coordinate-wise criterion based on $\sum_j (\beta_j - \hat{\beta}_j)^2$. We will use this simplification for the remaining derivations ofr each of the methods.

For *best subset selection* of size M we solve the optimization problem,

$$\hat{\beta}_{\text{subset}} = \arg\min_\beta \|y - X\beta\|_2^2 \qquad \text{subject to} \qquad \|\beta\|_0 \le M.$$

Note that $\|\beta\|_0 \le M$ counts the number of non-zero coefficients. Using trhe orthogonal decomposition above this is equivalent to writing,

$$\hat{\beta}_{\text{subset}} = \arg \min_{\beta:\|\beta\|_0 \le M} \sum_j (\beta_j - \hat{\beta}_j)^2.$$

Consider a single coordinate j so that, if we decide to include coordinate j, then we are allowed to choose β_j freely. The best choice is trivially $\beta_j = \hat{\beta}_j$ because then $(\beta_j - \hat{\beta}_j)^2 = (\hat{\beta}_j - \hat{\beta}_j)^2 = 0$. In the case we decide to exclude coordinate j, then we force $\beta_j = 0$. The contribution to the objective it therefore,

$$(0 - \hat{\beta}_j)^2 = \hat{\beta}_j^2.$$

Excluding the variable j costs $\hat{\beta}_j^2$ while including variable j costs 0. Therefore the best subset of size M should include the M coordinates with the largest values of $\hat{\beta}_j^2$ or equivalently the largest values of $|\hat{\beta}_j|$. Let the ordered absolute values of the OLS coefficients be $|\hat{\beta}_{(1)}| \ge \cdots \ge |\hat{\beta}_{(p)}|$. Then the M-th largest absolute coefficient is $|\hat{\beta}_{(M)}|$ so the coordinate j is kept if $|\hat{\beta}_j| \ge |\hat{\beta}_{(M)}|$ and is set to zero otherwise. Hence we have that,

$$\hat{\beta}_j^{\text{subset}} = \hat{\beta}_j \mathbf{1}\left\{|\hat{\beta}_j| \ge |\hat{\beta}_{(M)}|\right\}.$$

For *Ridge Regression* under the convention matching Table 3.4, we solve the optimization problem,

$$\hat{\beta}_{\text{ridge}} = \arg\min_\beta \left\{\frac{1}{2}\|y - X\beta\|_2^2 + \frac{\lambda}{2}\sum_{j=1}^{p} \beta_j^2\right\}.$$

The factor $\frac{1}{2}$ is only for convention to make the derivative cleaner. Using the orthogonal decomposition we have that,

$$\frac{1}{2}\|y - X\beta\|_2^2 = \frac{1}{2}\sum_{j=1}^{p}(\beta_j - \hat{\beta}_j^2) + \text{constant}.$$

So the ridge objective is equivalent to minimizing,

$$\sum_{j=1}^{p}\left[\frac{1}{2}(\beta_j - \hat{\beta}_j)^2 + \frac{\lambda}{2}\beta_j^2\right].$$

This is a fully separable objective across coordinates so for each j we minimize,

$$\begin{aligned} f_j(\beta_j) &= \frac{1}{2}(\beta_j - \hat{\beta}_j)^2 + \frac{\lambda}{2}\beta_j^2 \\ &= \frac{1}{2}(\beta_j^2 - 2\beta_j\hat{\beta}_j + \hat{\beta}_j^2) + \frac{\lambda}{2}\beta_j^2 \\ &= \frac{1}{2}\beta_j^2 - \beta_j\hat{\beta}_j + \frac{1}{2}\hat{\beta}_j^2 + \frac{\lambda}{2}\beta_j^2 \\ &= \frac{1+\lambda}{2}\beta_j^2 - \hat{\beta}_j\beta_j + \frac{1}{2}\hat{\beta}_j^2. \end{aligned}$$

Differentiating wrt β_j we get,

$$\frac{d}{d\beta_j}f_j(\beta_j) = (1+\lambda)\beta_j - \hat{\beta}_j = 0 \implies (1+\lambda)\beta_j = \hat{\beta}_j \implies \hat{\beta}_j^{\text{ridge}} = \frac{\hat{\beta}_j}{1+\lambda}.$$

Ridge is clearly a constant multiplicative shrinkage rule in the orthonormal case so every coefficient is shrunk toward zero by the same factor $\frac{1}{1+\lambda}$. Notice that this does not set any non-zero coefficient exactly to zero.

If $\hat{\beta}_j \neq 0$ then $\frac{\hat{\beta}_j}{1+\lambda} \neq 0$, hence continuous shrinkage by Ridge. □

3.8 Exercise 3.19

Proof. We wish to show that as the tuning parameter $\lambda \to 0$, then the ridge coefficient $\|\hat{\beta}_{\text{ridge}}\|$ increases, that is, the tuning parameter has a positive relationship with the magnitude of the ridge coefficient.

Suppose that we have centered design matrix X and responses y for the sake of not carrying the intercept. We apply SVD to the centered design matrix such that X can be written in the form $X = UDV^\top$ where U and V are orthonormal matrices such that $U^\top U = I$ and $V^\top V = I$ and $D = \text{diag}(d_1, \ldots, d_r)$ where $d_j > 0$.

The Gram matrix $X^\top X$ in this SVD form becomes,

$$X^\top X = (UDV^\top)^\top(UDV^\top) = VDU^\top UDV^\top = VD^2V^\top.$$

Similarly the $X^\top y$ component becomes,

$$X^\top y = (UDV^\top)^\top y = VDU^\top y.$$

Therefore the SVD of the ridge estimate becomes,

$$\begin{aligned} \hat{\beta}_\lambda^{\text{ridge}} &= (X^\top X + \lambda I)^{-1}X^\top y \\ &= (VD^2V^\top + \lambda I)^{-1}VDU^\top y \\ &= V(D^2 + \lambda I)^{-1}\underbrace{V^\top V}_{=I}DU^\top y \\ &= V(D^2 + \lambda I)^{-1}DU^\top y \\ &= \sum_{j=1}^{r}\frac{d_j}{d_j^2 + \lambda}(u_j^\top y)v_j. \end{aligned}$$

Taking the ℓ_2 norm (Euclidean squared norm) we get,

$$\begin{aligned}\left\|\hat{\beta}_\lambda^{\text{ridge}}\right\|_2^2 &= \sum_{j=1}^{r}\left[\frac{d_j}{d_j^2+\lambda}(u_j^\top y)\right]^2 \\ &= \sum_{j=1}^{r}\frac{d_j^2}{(d_j^2+\lambda)^2}(u_j^\top y)^2.\end{aligned}$$

Differentiating wrt λ we get that,

$$\begin{aligned}\frac{d}{d\lambda}\left\|\hat{\beta}_\lambda^{\text{ridge}}\right\|^2 &= \sum_{j=1}^{r} d_j^2(u_j^\top y)^2\frac{d}{d\lambda}(d_j^2+\lambda)^{-2} \\ &= -2\sum_{j=1}^{r}\frac{d_j^2(u_j^\top y)^2}{(d_j^2+\lambda)^3} \leq 0.\end{aligned}$$

The identity above is trivially less than or equal to zero because each term in the sum is non-negative so the whole derivative is non-positive. So the squared Euclidean norm decreases as $\lambda \to 0$. It is more precise to say non-decreases rather than strictly increases because if $u_j^\top y = 0$ for every direction in the $\text{Col}(X)$, then the ridge estimate is identically zero. But except in such degenerate cases, the norm increases as λ moves toward 0.

For the LASSO, the analogous monotonicity relationship holds naturally for the ℓ_1 norm but not necessarily for the Euclidean norm. Suppose we have $\lambda_1 > \lambda_2 > 0$ and define $\hat{\beta}_1 = \hat{\beta}_{\lambda_1}^{\text{LASSO}}$ and $\hat{\beta}_{\lambda_2}^{\text{LASSO}}$.

Because $\hat{\beta}_1$ minimizes the LASSO objective with the corresponding tuning parameter λ_1

$$\frac{1}{2}\|y - X\hat{\beta}_1\|_2^2 + \lambda_1\|\hat{\beta}_1\|_2 \leq \frac{1}{2}\|y - X\hat{\beta}_2\|_2^2 + \lambda_1\|\hat{\beta}_2\|_1.$$

Because $\hat{\beta}_2$ minimizes the LASSO objective with the corresponding tuning parameter λ_2,

$$\frac{1}{2}\|y - X\hat{\beta}_2\|_2^2 + \lambda_2\|\hat{\beta}_2\|_1 \leq \frac{1}{2}\|y - X\hat{\beta}_1\|_2^2 + \lambda_2\|\hat{\beta}_1\|_1.$$

Summing the two inequalities cancel the residual sum of squares term which leaves us with,

$$\lambda_1\|\hat{\beta}_1\|_1 + \lambda_2\|\hat{\beta}_2\|_1 \leq \lambda_1\|\hat{\beta}_2\|_1 + \lambda_2\|\hat{\beta}_1\|_1 \iff \lambda\|\hat{\beta}_1\|_1 - \lambda_2\|\hat{\beta}_1\|_1 \leq \lambda_1\|\hat{\beta}_2\|_1 - \lambda_2\|\hat{\beta}_2\|_1.$$

Factoring and dividing both sides by $\lambda_1 - \lambda_2$ we have that,

$$(\lambda_1 - \lambda_2)\|\hat{\beta}_1\|_1 \leq (\lambda_1 - \lambda_2)\|\hat{\beta}_2\|_1 \iff \|\hat{\beta}_1\|_2 \leq \|\hat{\beta}_2\|_1.$$

Therefore for $\lambda_1 > \lambda_2$ we see that $\|\hat{\beta}_{\lambda_1}^{\text{LASSO}}\|_1 \leq \|\hat{\beta}_{\lambda_2}^{\text{LASSO}}\|_1$. So, therefore as $\lambda \to 0$ the LASSO estimate increases in its natural penalty norm $\|\cdot\|_1$. This agrees with the constrained LASSO formulation where,

$$\min_\beta \frac{1}{2}\|y - X\beta\|_2^2 \quad \text{subject to} \quad \|\beta\|_2 \leq t,$$

where increasing t allows a larger coefficient budget. It is not guaranteed that for the LASSO, $\|\hat{\beta}_\lambda^{\text{LASSO}}\|_2$ is monotone for the general correlated design. What is guaranteed is monotonicity of the LASSO penalty norm $\|\hat{\beta}_\lambda^{\text{LASSO}}\|_1$. In the special orthogonal case, the LASSO solution is soft-thresholding,

$$\hat{\beta}_j^{\text{LASSO}} = \text{sign}(\hat{\beta}_j^{\text{OLS}})\left(|\hat{\beta}_j^{\text{OLS}}| - \lambda\right)_+,$$

so that every coordinate magnitude increases as $\lambda \to 0$ and so that in that special case, the Euclidean norm also increases. But with correlated predictors, active coefficients can enter, leave, and change direction so the ℓ_2-monotonicity is not the fundamental LASSO property.

For partial least squares, the tuning parameter is not a continuous penalty parameter λ but instead it's the number of PLS steps or components. So, after m steps, the PLS has constructed m score vectors $z_1, \ldots, z_m$ and fits y by projection onto their span. These score vectors are orthogonal in observation space so the fitted vector after m steps can be written as,

$$\hat{y}^{(m)} = \sum_{l=1}^{m} \frac{z_l^\top y}{z_l^\top z_l} \cdot z_l.$$

Because the z_l's are orthogonal we derive it's squared Euclidean norm,

$$\begin{aligned} \left\|\hat{y}^{(m)}\right\|_2^2 &= \sum_{l=1}^{m} \hat{\theta}_l^2 \|z_l\|_2^2 \\ &= \sum_{l=1}^{m+1} \hat{\theta}_l^2 \|z_l\|_2^2. \end{aligned}$$

Simplifying we get the closed form formulation,

$$\left\|\hat{y}^{(m+1)}\right|_2^2 - \left\|\hat{y}^{(m)}\right\|_2^2 = \hat{\theta}_{m+1}^2 \left\|z_{m+1}\right\|_2^2 \geq 0.$$

So, we conclude that PLS has a monotonicity property for the fitted vector norm. Equivalently, the training residual sum of squares decreases as more PLS components are added. But this is not equivalent to saying $\|\hat{\beta}^{\text{PLS, m}}\|_2$ must increase monotonically in the original coefficient coordinates. The PLS components are orthogonal as fitted score vectors in observation space, not necessarily orthogonal as coefficient directions in parameter space. Therefore, adding a new PLS component can re-express the coefficient vector in a way that changes its direction and scale. The algorithm guarantees monotonic improvement in the fitted-value space, not a universal monotone increase of the Euclidean coefficient norm. □

3.9 Exercise 3.29

Proof. We aim to show that, if a variable is copied many times, ridge spreads the coefficient evenly across the copies and effectively penalizes the direction less. Although Ridge handles collinearity stably, its coefficient values depend on how many times a variable is duplicated.

Suppose again that we have centered design X and responses y so there is no intercept in the algebra. Let $x \in \mathbb{R}^N$ denote the single predictor vector. First with one predictor, ridge solves the optimization,

$$\hat{a} = \arg\min_b \left\{ \|y - X\beta\|_2^2 + \lambda b^2 \right\}.$$

The coefficient from this one-variable regression is a. Here, I will derive it explicitly,

$$\begin{aligned} \|y - xb\|_2^2 + \lambda b^2 &= (y - xb)^\top (y - xb) + \lambda b^2 \\ &= y^\top y - 2bx^\top y + b^2 x^\top x + \lambda b^2 \\ &= y^\top y - 2bx^\top y + (x^\top x + \lambda) b^2. \end{aligned}$$

Differentiate the equation above wrt b and set equal to 0,

$$\begin{aligned} \frac{d}{db}\left[y^\top y - 2bx^\top y + (x^\top x + \lambda) b^2\right] &= -2x^\top y + 2(x^\top x + \lambda) b = 0 \\ &= -2x^\top y + 2(x^\top x + \lambda) b = 0 \\ &= -x^\top y + (x^\top x + \lambda) b = 0 \\ &= (x^\top x + \lambda) b = x^\top y. \end{aligned}$$

So the one-variable ridge regression coefficient is,

$$a = \frac{x^\top y}{x^\top x + \lambda}.$$

Now suppose we include an exact copy $x^* = x$ so that the ridge model now has two coefficients, say b_1 and b_2 and the fitted value $xb_1 + xb_2 = x(b_1 + b_2)$. The ridge objective becomes,

$$\min_{b_1,b_2} \left\{ \|y - xb_1 - xb_2\|_2^2 + \lambda(b_1^2 + b_2^2) \right\} = \min_{b_1,b_2} \left\{ \|y - x(b_1 + b_2)\|_2^2 + \lambda(b_1^2 + b_2^2) \right\}.$$

The data-fitting component depends on the sum $\theta = b_1 + b_2$. So, for a fixed total coefficient θ, the residual term $\|y - x\theta\|_2^2$ is fixed. Therefore, for a fixed θ, ridge chooses b_1, b_2 to minimize the penalty $b_1^2 + b_2^2$ subject to $b_1 + b_2 = \theta$. Solving this small constrained problem, since $b_2 = \theta - b_1$, we minimize $b_1^2 + (\theta - b_1)^2$ and expand,

$$b_1^2 + (\theta - b_1)^2 = b_1^2 + \theta^2 - 2\theta b_1 + b_1^2 = 2b_1^2 - 2\theta b_1 + \theta^2.$$

Again, differentiating wrt b_1 we have that,

$$\frac{d}{db_1} \left(2b_1^2 - 2\theta b_1 + \theta^2 \right) = 4b_1 - 2\theta.$$

Setting this equal to zero we get $b_1 = \frac{\theta}{2}$. Since $b_2 = \theta - b_1$ we also get that $b_2 = \theta - \frac{\theta}{2} = \frac{\theta}{2}$. So for any fixed total coefficient θ, the ridge penalty is minimized by splitting the total equally. This proves that the two coefficients must be identical. Now we derive their actual value at equal split,

$$b_1^2 + b_2^2 = \left(\frac{\theta}{2}\right)^2 + \left(\frac{\theta}{2}\right)^2 = \frac{\theta^2}{4} + \frac{\theta^2}{4} = \frac{\theta^2}{2}.$$

Therefore, the two-copy ridge regression problem reduces to the 1-dimension problem,

$$\min_{\theta} \left\{ \|y - x\theta\|_2^2 + \lambda \frac{\theta^2}{2} \right\} \equiv \min_{\theta} \left\{ \|y - x\theta\| + \frac{\lambda}{2}\theta^2 \right\}.$$

Notice that this is simply ridge regression but on x with an effective ridge penalty of $\lambda/2$ on the total coefficient θ. Expanding this formulation and simplifying we get,

$$\begin{aligned} \|y - x\theta\|_2^2 + \frac{\lambda}{2}\theta^2 &= y^\top y - 2\theta x^\top y + \theta^2 x^\top x + \frac{\lambda}{2}\theta^2 \\ &= y^\top y - 2\theta x^\top y + \left(x^\top x + \frac{\lambda}{2} \right) \theta^2. \end{aligned}$$

Differentiating the final formulation above wrt θ and setting equal to zero we get,

$$\begin{aligned} 0 &= -2x^\top y + 2\left(x^\top x + \frac{\lambda}{2} \right) \theta \\ &= -x^\top y + \left(x^\top x + \frac{\lambda}{2} \right) \theta \\ x^\top y &= \left(x^\top x + \frac{\lambda}{2} \right) \theta \\ \hat{\theta} &= \frac{x^\top y}{x^\top x + \frac{\lambda}{2}}. \end{aligned}$$

Since we know that ridge assigns equal weight to both, then $\hat{b}_1 = \hat{b}_2 = \frac{\hat{\theta}}{2}$ so we obtain,

$$\hat{b}_1 = \hat{b}_2 = \frac{1}{2} \cdot \frac{x^\top y}{x^\top x + \frac{\lambda}{2}} = \frac{x^\top y}{2x^\top x + \lambda}.$$

Therefore, after adding exactly one copy, both ridge coefficients are given by the formulation above. Now, to express these in terms of the original one-variable coefficient a, we can re-arrange our formula for a to get $x^\top y = a(x^\top x + \lambda)$ which means that our coefficients are now,

$$\hat{b}_1 = \hat{b}_2 = a \cdot \frac{x^\top x + \lambda}{2x^\top x + \lambda}.$$

Notice that each duplicated coefficient is smaller than the original coefficient in magnitude, assuming that $a \neq 0$, but the sum of the two coefficients is,

$$\hat{b}_1 + \hat{b}_2 = 2 \cdot \frac{x^\top y}{x^\top x + \lambda} = \frac{2x^\top y}{2x^\top x + \lambda}.$$

Compared the original coefficient a, the total coefficient after duplication is larger in magnitude because the effective penalty has become $\frac{\lambda}{2}$. So, duplicating a variable allows ridge to split the coefficient across copies and reduce the penalty paid for the same total effect. Now, we prove the general result for m identical copies.

Suppose in our new run, we include $x_1 = x_2 = \cdots = x_m = x$.

Let their corresponding coefficients be given as $b_1, b_2, \ldots, b_m$ and the fitted value as,

$$xb_1 + xb_2 + \cdots + xb_m = x \sum_{k=1}^{m} b_k.$$

Define the total coefficient to be $\theta = \sum_{k=1}^{m} b_k$ and the ridge objective,

$$\min_{b_1,\ldots,b_m} \left\{ \left\| y - x \sum_{k=1}^{m} b_k \right\|_2^2 + \lambda \sum_{k=1}^{m} b_k^2 \right\}.$$

using θ the residual part can be expressed $\|y - x\theta\|_2^2$. For fixed θ, the residual term is fixed so we need to minimize $\sum_k b_k^2$ subject to the constraint that $\sum_k b_k = \theta$. We can solve this using Cauchy-Schwarz,

$$\begin{aligned}
\theta^2 &= \left(\sum_k b_k \right)^2 \\
&\leq \left(\sum_k 1^2 \right) \left(\sum_k b_k^2 \right) \\
&= m \cdot \sum_k b_k^2.
\end{aligned}$$

Therefore we conclude that, $\theta^2 \leq m \cdot \sum_k b_k^2$ so therefore,

$$\sum_k b_k^2 \geq \frac{\theta^2}{m}.$$

Equality in Cauchy-Schwarz holds exactly when all the b_k's are equal. Since their sum must be θ, equality occurs when all our b_k's are equal to θ/m. Thus, for a fixed total coefficient θ, ridge minimizes the penalty by assiging equal coefficients to all identical copies. At this equal split, we see that,

$$\sum_k b_k^2 = m \cdot \left(\frac{\theta}{m} \right)^2 = m \cdot \frac{\theta^2}{m^2} = \frac{\theta^2}{m}.$$

Now deriving $\hat{\theta}$ given the effective penalty for the optimization problem is $\frac{\lambda}{m}$,

$$\begin{aligned}
\|y - x\theta\|_2^2 + \frac{\lambda}{m}\theta^2 &= y^\top y - 2\theta x^\top y + \theta^2 x^\top x + \frac{\lambda}{m}\theta^2 \\
&= y^\top y - 2\theta x^\top y + \left(x^\top x + \frac{\lambda}{m} \right) \theta^2 \\
&\overset{\partial/\partial\theta}{=} -2x^\top y + 2\left(x^\top x + \frac{\lambda}{m} \right) \theta = 0 \\
&= -x^\top y + \left(x^\top x + \frac{\lambda}{m} \right) \theta.
\end{aligned}$$

So re-arranging we get,

$$x^\top y = \left(x^\top x + \frac{\lambda}{m}\right)\theta$$
$$\hat{\theta} = \frac{x^\top y}{x^\top x + \lambda/m}.$$

Therefore, each individual copy receives the coefficient $\hat{b}_k = \frac{\hat{\theta}}{m}$ so therefore,

$$\hat{b}_k = \frac{1}{m} \cdot \frac{x^\top y}{x^\top x + \lambda/m} = \frac{x^\top y}{mx^\top x + \lambda}.$$

Therefore for $k = 1, \ldots, m$ we get that,

$$\hat{b}_k = \frac{x^\top y}{mx^\top x + \lambda}.$$

meaning that in terms of the original one-copy ridge coefficient we get that the total coefficient across m copies is,

$$\sum_k \hat{b}_k = m \cdot \frac{x^\top y}{mx^\top x + \lambda} = \frac{x^\top y}{x^\top x + \lambda/m}.$$

Therefore, the duplicated variables all receive the same coefficient but the fitted effect behaves like a single-variable ridge regression with penalty λ/m and not λ. As m grows, the effective penalty on the shared direction becomes smaller. This means ridge is stable under exact collinearity in the sense that the solution exists and is unique for $\lambda > 0$ but is not invariant to duplicating predictors. □