STAT 343 Final Review

Rohan Buluswar

December 2025

Introduction

These are my final review notes for STAT 34300, Applied Linear Stat Methods, at the University of Chicago with Prof. Nikolaos Ignatiadis, in the Autumn 2025 quarter.

Statistical Assumptions and Framework, OLS

Notation: We have observations/samples $i = 1, 2, \ldots, n$. Each observation is a vector $X_i \in \mathbb{R}^p$, with p covariates/features. Our observations also come with a response variable $Y_i \in \mathbb{R}$. We can thus construct an $n \times p$ data/design matrix X, with rows equal to $X_1^T, \ldots, X_n^T$. We use $X_{\cdot j}$ to denote the j-th column of X, i.e. the values of the j-th feature for some $1 \leq j \leq p$. Our goal is to find $\beta \in \mathbb{R}^p$ and predict the response variable by $Y_i = X_i^T \beta = \beta^T X_i$.

Definition 0.1 (**Ordinary Least Squares**). *We define $\hat{\beta}$ as*

$$\hat{\beta} \in \text{argmin}\{\frac{1}{n}||Y - X\beta||^2\}$$

$$= \text{argmin}\{\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2\}$$

Proposition 0.1 (**OLS Solution**). *If X is full rank then $\hat{\beta}$ is unique and given by*

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

Note that if $p > n$, the OLS solution will be non-unique.
Assumption-lean Framework: We assume that our observations (X_i, Y_i) are i.i.d. samples from an unknown distribution $\mathbb{P}$. Our task is to use the training data to find a predictor $\hat{m}(\cdot) : \mathbb{R}^p \to \mathbb{R}$. We will be given a new independent test point $(X_f, Y_f) \sim \mathbb{P}$, and need to predict Y_f given X_f as $Y_f = \hat{m}(X_f)$. Our goal is to minimize the expected error of this prediction.

Error: There are two primary kinds of error - the error of the particular model generated from the training data, and the expected error of a model before seeing the training data. We can define the first as

$$\text{Err}(\mathcal{D}_{\text{train}}) = \mathbb{E}_{(X_f, Y_f)\sim\mathbb{P}}[(Y_f - \hat{m}(X_f))^2 | \mathcal{D}_{\text{train}}]$$

We can define the second as

$$\text{Err} = \mathbb{E}_{(X_f, Y_f) \sim \mathbb{P}}[(Y_f - \hat{m}(X_f))^2]$$

Note also that

$$\text{Err} = \mathbb{E}_{\mathcal{D}_{\text{train}} \sim \mathbb{P}^n}[\text{Err}(\mathcal{D}_{\text{train}})]$$

Proposition 0.2 (**CEF**). *The predictor that minimizes the error, $\mathbb{E}[(Y_f - m(X_f)^2]$, is the conditional expectation function:*

$$m^*(x) = \mathbb{E}[Y_i | X_i = x]$$

Our underlying goal is thus to approximate the CEF.

Proposition 0.3 (**BLP**). *We define the best linear predictor, β^*, to be the minimizer of $\mathbb{E}[(Y_f - X_f^T \beta)^2]$. Then it is given by*

$$\beta^* = \mathbb{E}[X_f X_f^T]^{-1} \mathbb{E}[X_f Y_f]$$

The goal of linear regression is thus to approximate the best linear predictor.

Proposition 0.4 (**Error decomposition**). *For any predictor m, we can write*

$$\mathbb{E}[(Y_f - m(X_f))^2] = \mathbb{E}[\mathit{Var}(Y_f | X_f)] + \mathbb{E}[(m^*(X_f) - m(X_f))^2]$$

The first term is called irreducible error, and the second term is the error in approximating the CEF. In particular, if the CEF is linear, then it is the best linear predictor.

Residuals: We may define two different kinds of residuals. The first is the residual of the CEF:

$$\epsilon_f = Y_f - m^*(X_f)$$

Then ϵ_f is orthogonal to all functions of X_f, in the sense that $\mathbb{E}[\epsilon_f h(X_f)] = 0$. The second is the residual of the BLP:

$$\delta_f = Y_f - X_f^T \beta^*$$

Then δ_f is orthogonal to all linear functions of X_f, in the sense that if A is linear, then $\mathbb{E}[\delta_f A(X_f)] = 0$.

Proposition 0.5 (**Error decomposition II**). *If $\beta \in \mathbb{R}^p$ is an arbitrary linear predictor and β^* is the BLP, then*

$$\mathbb{E}[(Y_f - X_f^T \beta)^2] = \mathbb{E}[\mathit{Var}(Y_f | X_f)] + \mathbb{E}[(m^*(X) - X^T \beta^*)^2] + \mathbb{E}[(X^T \beta - X^T \beta^*)^2]$$

The first term is irreducible error, the second term is non-linearity of the system, and the third term is error in approximating the BLP.

Asymptotics, Feature Selection, Regularization

Theorem 0.1 (Convergence of OLS to BLP). *Let $(X_i, Y_i) \sim \mathbb{P}$ and suppose $\mathbb{E}[||X_i||^4] < \infty$, $\mathbb{E}[Y_i^4] < \infty$, and $\mathbb{E}[X_i X_i^T] \succ 0$. Let*

$$\hat{\beta} = \beta^{\hat{OLS}} = (X^T X)^{-1} XY$$

and let β^* be the BLP. Then as $n \to \infty$ (and p is fixed), $\hat{\beta} \xrightarrow{\mathbb{P}} \beta^*$.

Feature selection: We might begin with p features for very large p. However, this will be prone to overfitting and cause issues with interpretation. Thus, we may want to do feature selection. Specifically, we want to pick a subset $S \subseteq [p]$ and then do OLS on only the features from S. There are several methods.

Best subset: For fixed $s < p$, we choose $S \subseteq [p]$ such that $|S| = s$ and OLS on the features in S produces the least empirical error. That is, we minimize

$$\frac{1}{n} \sum (Y_i - X_{i,S} \hat{\beta_S})^2$$

where $X_{i,S}$ is X restricted to relevant features and $\hat{\beta_S}$ is the OLS. In general, this is extremely computationally expensive, as we have to perform OLS $\binom{p}{s}$ times, which is exponential in s.

Forward stepwise: Forward stepwise is a greedy algorithm. We begin with only an intercept, and repeatedly choose a new feature that, in combination with the previous features, decreases empirical error the most. We iterate until we have the desired s features.

Backward stepwise: In backward stepwise, we begin with all possible features, and iteratively remove whichever feature causes the least increases in empirical error, until we have exactly s features.

Ridge: Ridge regression, for some fixed parameter $\lambda > 0$, is the solution to the following problem:

$$\text{argmin} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 + \lambda ||\beta||^2$$

Ridge regression does 'shrinkage', in that it encourages model coefficients to be smaller. However, it does not do feature selection. Moreover, it is well-defined and unique even when $p > n$. This is due to the following result:

Proposition 0.6 (Ridge regression). *The solution to ridge regression is given by*

$$\hat{\beta}(\lambda) = (X^T X + \lambda I)^{-1} X^T Y$$

Note that because $X^T X$ is positive semidefinite, if $\lambda > 0$, $X^T X + \lambda I$ is positive definite. As a result, the ridge regression solution is well-defined.

Ridgeless: Ridgeless regression is defined by

$$\hat{\beta}^{\text{Ridgeless}} = \lim_{\lambda \to 0} \hat{\beta}^{\text{Ridge}}(\lambda)$$

If X has full rank, then the ridgeless regression is exactly equal to OLS. If $p > n$, however, there are many OLS solutions. Ridgeless is then the interpolating predictor with least l_2 norm. This is due to the following formula:

$$\hat{\beta} = \text{argmin}\{||\beta||^2 : \sum(Y_i - X_i^T\beta)^2 = \text{argmin}_\gamma(Y_i - X_i^T\gamma)^2\}$$

LASSO: Lasso is a kind of regularization that accomplishes both shrinkage and feature selection. For $\lambda > 0$, it is defined by

$$\hat{\beta} = \text{argmin}\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \lambda||\beta||_1$$

where

$$||\beta||_1 = \sum_{j=1}^{p}|\beta_j|$$

Unlike Ridge, Lasso is not generally unique. For example, suppose that two features are the same, i.e. $X_{i1} = X_{i2}$ for all i. Let β be the OLS coefficient that would be on X_{i1} if X_{i2} did not exist. In this scenario, Lasso is ambivalent to any split of β into β_1 on X_{i1} and β_2 on X_{i2}, as long as $\beta = \beta_1 + \beta_2$. The same is not true for Ridge, due to the quadratic l_2 regularization.

Linear data reparameterization: Suppose we have a design matrix $X \in \mathbb{R}^{n\times p}$ and an invertible matrix $A \in \mathbb{R}^{p\times p}$. We think of A as a linear reparameterization of the features. Then we can consider a new design matrix, $\tilde{X} = XA$. If we denote the OLS result on X by $\hat{\beta}$ and the OLS on $\tilde{X}$ by $\hat{\gamma}$, then $X\hat{\beta} = \tilde{X}\hat{\gamma}$. Moreover, the class of linear models is the same. That is,

$$\{x \mapsto x^T\beta : \beta \in \mathbb{R}^p\} = \{\tilde{X} \mapsto \tilde{X}^T\gamma : \gamma \in \mathbb{R}^p\}$$

Hence, OLS is completely ambivalent to such reparameterizations. The same is not true for regularized methods like Ridge and Lasso.

Elastic Net: For some $\lambda > 0$ and $\alpha \in [0, 1]$, we define

$$\hat{\beta} = \text{argmin}\sum(Y_i - X_i^T\beta)^2 + \lambda(\alpha||\beta||_1 + (1-\alpha)||\beta||_2)$$

For $\alpha = 1$, this reduces to Lasso, and for $\alpha = 0$, it reduces to Ridge. For $\alpha = (0, 1)$, elastic net achieves feature selection and uniqueness.

Relaxed Lasso: For relaxed lasso, we choose λ and perform standard Lasso to obtain $\hat{\beta}(\lambda)$. We then do OLS on the features with non-zero coefficients in $\hat{\beta}(\lambda)$. As a result, relaxed Lasso accomplishes feature selection with less shrinkage.

Estimating Error, Overfitting

Due to overfitting, the empirical/training error of a predictor is often much better than the true test error. Our next task is thus to estimate the test error of a predictor.

In-sample Prediction Error: Given our input data (X_i, Y_i) for $i = 1, \ldots, n$, and our trained predictor $\hat{m}$, we can define the in-sample prediction error as follows. For each X_i, sample Y_i' from $\mathbb{P}(Y_i|X_i)$ independently from the original Y_i. We then consider

$$\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(Y_i' - \hat{m}(X_i))^2|X]$$

Proposition 0.7 (**Decomposition of in-sample test error**).

$$\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(Y_i' - \hat{m}(X_i))^2|X] = \mathbb{E}[\frac{1}{n}\sum(Y_i - \hat{m}(X_i))^2|X] + \frac{2}{n}\sum \mathrm{Cov}(Y_i, \hat{m}(X_i)|X)$$

The first term in the above sum is the training error, and the second term is called the **optimism**, as it represents the difference between training error and in-sample test error.

Effective degrees of freedom: Suppose that we have homoscedasticity, i.e. there exists $\sigma > 0$ such that $\mathrm{Var}(Y_i|X_i) = \sigma^2$ for all X_i. Then we can rewrite the optimism as

$$(\frac{2\sigma^2}{n})(\frac{1}{\sigma^2}\sum \mathrm{Cov}(Y_i, \hat{m}(X_i)|X))$$

The second term is called the **effective degrees of freedom**, or $\mathrm{edf}(\hat{m})$ and can be roughly interpreted as the number of parameters in the model.

Remark: If $\hat{m}$ is an interpolating predictor, it has $\mathrm{edf}(\hat{m}) = n$. The same is true if $\hat{m}$ is the CEF. To compute edf for a wider class of models, we introduce the following definition.

Linear smoothers: A method of fitting a predictor, $\hat{m}$, is called a linear smoother if the predictions can be written in the form

$$\begin{pmatrix}\hat{m}(X_1)\\ \vdots \\ \hat{m}(X_n)\end{pmatrix} = A(X)\begin{pmatrix}Y_1\\ \vdots \\ Y_n\end{pmatrix}$$

where $A(X) \in \mathbb{R}^{n\times n}$ depends only on the input covariates. For example, OLS is a linear smoother, with $A(X) = X(X^TX)^{-1}X^T$. Second, any interpolating predictor is a linear smoother, with $A(X) = I$. Third, Ridge regression is a linear smoother, with $A(X) = X(X^TX + \lambda I)^{-1}X^T$.

Proposition 0.8. *If $\hat{m}$ is a linear smoother with matrix $A(X)$, its effective degrees of freedom is* $\mathrm{Tr}(A)$.

As a result, when we do OLS with a full-rank design matrix, we get edf $= p$.

Because Lasso is not a linear smoother, it is difficult to calculate the edf. However, if we assume that $Y_i|X_i \sim N(\mathbb{E}[Y_i|X_i], \sigma^2)$, we obtain the following unbiased estimate:

$$\hat{\mathrm{edf}}(\hat{m}) = \sum_{i=1}^{n}\frac{\partial \hat{m}(X_i)}{\partial Y_i}$$

In the case of Lasso, we find that the edf is the expected number of nonzero coordinates.

Now, we would like to find confidence intervals for the test error, i.e.

$$\mathbb{P}(\text{Err}_{\mathcal{D}_{\text{train}}} \in I) \approx 1 - \alpha$$

Confidence intervals: Here is a brief review of how to construct confidence intervals. Suppose we have i.i.d. r.v. U_i with $\text{Var}(U_i) = \tau^2$. We want a CI for $\mu = \mathbb{E}[U_i]$. Let $\hat{\mu} = \overline{U}$, the sample mean, and let

$$\hat{\tau} = \frac{1}{n-1} \sum_{i=1}^{n} (U_i - \overline{U})^2$$

By the Law of Large Numbers, we can show that

$$\hat{\mu} \xrightarrow{\mathbb{P}} \mu, \hat{\tau} \xrightarrow{\mathbb{P}} \tau$$

By the Central Limit Theorem,

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} N(0, \tau^2)$$

Moreover, it is a theorem that

$$\frac{\sqrt{n}((\hat{\mu} - \mu))}{\hat{\tau}} \xrightarrow{d} N(0, 1)$$

Hence, if we define

$$\hat{\text{se}}(\hat{\mu}) = \sqrt{\frac{\hat{\tau}^2}{n}}$$

then we get

$$\frac{\hat{\mu} - \mu}{\hat{\text{se}}(\hat{\mu})} \xrightarrow{d} N(0, 1)$$

Because we care about bounding $|\hat{\mu} - \mu|$, this suffices to create a CI. For a 95% confidence interval, we would thus take $[\hat{\mu} - 2\hat{\text{se}}(\hat{\mu}), \hat{\mu} + 2\hat{\text{se}}(\hat{\mu})]$, for some large n. To return to our setting, we set $U_i = (Y_i' - \hat{m}(X_i))^2$.

Cross-Validation

Confidence intervals for error: In order to estimate the standard error in the above confidence interval, we might use cross-validation. We begin by splitting the data into K subsets. (If $K = n$, this is called Leave-One-Out Cross-Validation, or LOOCV.) Then, for each $j = 1, \ldots, K$, we train $\hat{m}$ on all data blocks except I_j, then evaluate the error on I_j to get Err_j. This gives two different estimates of error:

$$CV = \frac{1}{K} \sum_{j=1}^{K} \text{Err}_j$$

and

$$CV = \frac{1}{n} \sum_{j=1}^{K} \sum_{i \in I_j} (Y_i - \hat{m}_{-I_j}(X_i))^2$$

(These are the same if all blocks have the same size.) Because cross-validation computes errors over different training sets, it is an estimate of the true error, not conditioned on training data. With the above definition, we can construct confidence intervals by $CV \pm 2\hat{\text{se}}(CV)$. That is,

$$CV \pm 2\frac{1}{\sqrt{k}}\frac{1}{\sqrt{k-1}}\sqrt{\sum_{j=1}^{K}(\text{Err}_j - CV)^2}$$

Alternatively, for each $i \in I_j$, we can write

$$s_i = (Y_i - \hat{m}_{-I_j}(X_i))^2$$

and compute the CI

$$CV \pm 2\frac{1}{\sqrt{n}}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(s_i - CV)^2}$$

Model selection: We can also use cross-validation for model selection. After splitting the data into K blocks, we compute the CV error for the model with each λ (or some alternative hyperparameter). We choose the λ associated with minimal CV error, refit it on all training data, and evaluate the error on the test set.

Prediction and Confidence Intervals

We have typically trained a predictor $\hat{m}$ and reported point predictions, $\hat{m}(X_f)$. Now, we would like to report prediction intervals. There are two kinds:

Unconditional prediction intervals: We consider $I(X_f, (X_i, Y_i)_{i=1}^n)$ and require

$$\mathbb{P}[Y_f \in I(X_f)] \geq 1 - \alpha$$

for some tolerance $\alpha \in (0, 1)$.

Conditional prediction intervals: We require that for all X_f,

$$\mathbb{P}[Y_f \in I(X_f)|X_f] \geq 1 - \alpha$$

Note that by taking expectations, conditional coverage implies unconditional coverage. Conditional coverage is thus preferable, but with minimal assumptions, it is impossible to guarantee. We thus focus our attention on unconditional prediction intervals. To accomplish this, we have the following method:

Split-conformal prediction: Suppose we have a training set, a calibration set, and a test point (X_f, Y_f). We train $\hat{m}$ on the training set, and for each X_i $(i = 1, \ldots, k)$ in the calibration set, we compute $s_i = |Y_i - \hat{m}(X_i)|$. Let $S_{(j)}$ be the order statistics of the s_i. Then define

$$\hat{q} = S_{\lceil (k+1)(1-\alpha) \rceil}$$

and create a prediction interval

$$I(X_f) = \hat{m}(X_f) \pm \hat{q}$$

Suppose we want to create asymptotic confidence intervals for the optimal coefficients β_j^*. From previous work, we know that a 95% CI should take the form

$$\hat{\beta}_j \pm 2\hat{\text{se}}(\hat{\beta}_j)$$

Thus, we just need an estimator for the standard error. We have the following result:

Proposition 0.9 (**Sandwich formula**). *We have that*

$$\sqrt{n}(\hat{\beta} - \beta^*) \xrightarrow{d} N(0, \Gamma)$$

where

$$\Gamma = \mathbb{E}[X_i X_i^T]^{-1}\mathbb{E}[X_i X_i^T (Y_i - X_i^T \beta^*)^2]\mathbb{E}[X_i X_i^T]^{-1}$$

In particular, we have

$$\sqrt{n}(\hat{\beta}_j - \beta_j^*) \xrightarrow{d} N(0, \Gamma_{jj})$$

and for any vector c, we have

$$\sqrt{n}(c^T\hat{\beta} - c^T\beta^*) \xrightarrow{d} N(0, c^T\Gamma c)$$

We have the following unbiased estimate for Γ:

Definition 0.2 (**EHW Standard Errors**).

$$\hat{\Gamma} = (\frac{X^T X}{n})^{-1}\frac{1}{n}\sum X_i X_i^T (Y_i - X_i^T\hat{\beta})^2 (\frac{X^T X}{n})^{-1}$$

Moreover, $\hat{\Gamma} \xrightarrow{\mathbb{P}} \Gamma$ as $n \to \infty$. This yields the following asymptotic confidence interval for β_j^*:

$$I = \hat{\beta}_j \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\Gamma}_{jj}}{n}}$$

Remark: Suppose now that we assume linearity and homoskedasticity. That is,

$$\mathbb{E}[Y_f|X_f] = \gamma^T X_f$$

$$\text{Var}(Y_f|X_f) = \sigma^2$$

Then some computation yields

$$\Gamma = \sigma^2\mathbb{E}[X_i X_i^T]^{-1}$$

$$\hat{\Gamma} = \hat{\sigma^2}(\frac{X^T X}{n})^{-1}$$

where

$$\hat{\sigma^2} = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - X_i^T\hat{\beta})^2 \xrightarrow{\mathbb{P}} \sigma^2$$

Bootstrapping

Suppose now that we have i.i.d. data $Z_i = (X_i, Y_i) \sim \mathbb{P}$ and estimate some parameter $\theta(\mathbb{P})$ by $\hat{\theta}(Z_1, \ldots, Z_n)$. (For example, an OLS coefficient.) We want to estimate $\text{Var}(\hat{\theta})$ and use this to create a CI for θ.

If we had access to $\mathbb{P}$, we could sample n data points B times, estimate $\hat{\theta}^1, \ldots, \hat{\theta}^B$, and take

$$\hat{\text{Var}}(\hat{\theta}) = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}^i - \bar{\hat{\theta}})^2$$

However, we do not have access to $\mathbb{P}$, and thus need to estimate it with some $\hat{\mathbb{P}}$. In standard boostrapping, we take $\hat{\mathbb{P}}$ to be the empirical (uniform) distribution on our data Z_i. We then sample from these points (with replacement) to get the $\hat{\theta}^i$.

Confidence Intervals via Bootstrapping

We can use bootstrap to construct confidence intervals for θ in three ways. First, assume that $\theta \approx N(\theta^*, \sigma^2)$ and $\hat{\text{Var}}(\hat{\theta}) \approx \sigma^2$. Then we have the confidence interval

$$\hat{\theta} \pm 2\sqrt{\hat{\text{Var}}(\hat{\theta})}$$

Second, we have the percentile bootstrap. Let $\hat{l}_{\frac{\alpha}{2}}$ be the $\frac{\alpha}{2}$ percentile of $\{\hat{\theta}_1, \ldots, \hat{\theta}_B\}$ and let $\hat{u}_{\frac{\alpha}{2}}$ be the $1 - \frac{\alpha}{2}$ percentile. Then we can simply report

$$[\hat{l}_{\frac{\alpha}{2}}, \hat{u}_{\frac{\alpha}{2}}]$$

Finally, we can use the t-distribution. That is, for each re-sample b, we have

$$\frac{\hat{\theta}_b - \hat{\theta}}{\hat{\text{se}}_b} \overset{d}{\approx} \frac{\hat{\theta} - \theta^*}{\hat{\text{se}}} \implies \hat{t}_b \overset{d}{\approx} \hat{t}$$

If we let $\hat{l}_{\frac{\alpha}{2}}$ and $\hat{u}_{\frac{\alpha}{2}}$ be the quantiles of the $\hat{t}_b$, then we can report

$$[\hat{\theta} - \hat{u}_{\frac{\alpha}{2}}\hat{\text{se}}, \hat{\theta} - \hat{l}_{\frac{\alpha}{2}}\hat{\text{se}}]$$

Bootstrapping as WLS

Instead of resampling, we can think of bootstrap as randomly reassigning weights to the original sample, and then doing WLS (weighted least squares). As long as we assign nonzero probability to all of the original sample points, we can keep our design matrix full-rank. For example, we have Poisson bootstrap:

$$\tilde{w}_1^b, \dots \tilde{w}_n^b \overset{\text{iid}}{\sim} \text{Poiss}(\lambda)$$

$$w_i^b = \frac{\tilde{w}_i^b}{\sum_i \tilde{w}_i^b}$$

This does not solve the prior issue, but the Exponential bootstrap does:

$$\tilde{w}_1^b, \dots \tilde{w}_n^b \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$$

$$w_i^b = \frac{\tilde{w}_i^b}{\sum_i \tilde{w}_i^b}$$

Bootstraps Conditional on X

Now suppose that we fix the covariates and only resample the response variable, i.e. $X_i^b = X_i$ for all i, b.

Residual bootstrap: Assume linearity, so that $\mathbb{E}[Y_i|X_i] = X_i^T\beta^*$. Then define $\hat{\varepsilon}_i = Y_i - X_i^T\hat{\beta}$. We can resample $\hat{\varepsilon}_i^b$ and set $Y_i^b = X_i^T\hat{\beta} + \hat{\varepsilon}_i^b$. If we assume that the ε_i are i.i.d. and independent of X_i, then we can directly sample with replacement from the original residuals.

Wild bootstrap: In wild bootstrap, we set $\hat{\varepsilon}_i^b = \pm\hat{\varepsilon}_i$, each with equal probability, independently for each i. In other words, we multiply by U_i, which is a Rademacher random variable. This does not assume homoscedasticity. We can choose other U_i which are mean-zero and thus construct other similar bootstrap methods.

Causal Inference

Suppose we have some covariates X_i, a 0-1 treatment variable w_i, and a response variable $Y_i \in \mathbb{R}$. The fundamental goal of causal inference is to determine whether the treatment w_i had a causal effect on Y_i. Our framework is as follows: for each i, there are two potential outcomes: $Y_i(0)$ and $Y_i(1)$. We assume $Y_i = Y_i(w_i)$. The assumption that Y_i only depends on its own treatment is called SUTVA: Stable Unit Treatment Value Assumption. In sum, the data takes the form

$$(X_i, w_i, Y_i(0), Y_i(1)) \overset{\text{iid}}{\sim} \mathbb{P}$$

The individual treatment effect is defined by

$$\tau = Y_i(1) - Y_i(0)$$

and the average treatment effect (ATE) is defined by

$$\text{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)]$$

This is the fundamental causal quantity that we seek to estimate. For example, we can use the following estimator

$$\hat{\tau}^{\mathrm{DM}} = \frac{1}{\#\{w_i = 1\}} \sum_{w_i=1} Y_i - \frac{1}{\#\{w_i = 0\}} \sum_{w_i=0} Y_i$$

but unfortunately we have

$$\hat{\tau}^{\mathrm{DM}} \xrightarrow{\mathbb{P}} \mathbb{E}[Y_i(1)|w_i = 1] - \mathbb{E}[Y_i(0)|w_i = 0]$$

which is in general not the ATE, as the distribution of $X_i, Y_i(0)$, and $Y_i(1)$ may depend on w_i. So, we need to fix this problem.

Randomized Control Trial: Suppose we have direct control of w_i and for each i, we set $w_i \sim \mathrm{Ber}(p)$ with $w_i \perp (X_i, Y_i(0), Y_i(1))$. Then

$$\mathbb{E}[Y_i(1)|w_i = 1] = \mathbb{E}[Y_i(1)]$$

and

$$\mathbb{E}[Y_i(0)|w_i = 0] = \mathbb{E}[Y_i(0)]$$

so that

$$\hat{\tau}^{\mathrm{DM}} \xrightarrow{\mathbb{P}} \mathbb{E}[Y_i(1) - Y_i(0)] = \mathrm{ATE}$$

Given this, we can create a CI for the ATE in the form

$$\hat{\tau} \pm 2\hat{\mathrm{se}}(\hat{\tau})$$

To estimate the standard error, we can equivalently think of $\hat{\tau}$ as the OLS coefficient in the regression

$$Y_i = \alpha + \tau w_i + \varepsilon_i$$

and use the EHW standard errors. Alternatively, we can consider the null hypothesis H_0: ATE $= 0$, and compute a p-value. Under H_0,

$$\frac{\hat{\tau}^{\mathrm{DM}}}{\hat{\mathrm{se}}} \xrightarrow{d} N(0, 1)$$

so our p-value has the form

$$p = 2(1 - \Phi(\frac{|\hat{\tau}^{\mathrm{DM}}|}{\hat{\mathrm{se}}}))$$

We can also consider Fisher's sharp null hypothesis:

$$\forall i,\ Y_i(0) = Y_i(1)$$

We can test this null hypothesis with a randomization test. For $b = 1, \ldots, B$ iterations, we permute the values of w_i and recompute $\hat{\tau}^{\mathrm{DM},b}$. Then we have the p-value

$$p = \frac{1 + \sum_{b=1}^{B} \mathbb{I}(|\hat{\tau}^b| \geq |\hat{\tau}|)}{1 + B}$$

and under H_0, $\mathbb{P}(p \leq \alpha) \leq \alpha$. This allows us to control the Type-I error.

Adjusted Difference of Means: We can incorporate information about the covariates X_i as follows. We instead perform the regression

$$Y_i = \alpha + \tau w_i + \gamma^T X_i + \varepsilon_i$$

and denote the OLS coefficient on $w_i$by $\hat{\tau}^{\text{adj}}$. Then under the RCT setting,

$$\hat{\tau}^{\text{adj}} \xrightarrow{\mathbb{P}} \text{ATE}$$

If $p = \frac{1}{2}$ is the probability of assigning treatment, then $\hat{\tau}^{\text{adj}}$ will typically have lower asymptotic variance than $\hat{\tau}^{\text{DM}}$.

We can alternatively consider the following regression:

$$Y_i = \alpha + \tau w_i + \gamma^T X_i + \delta^T w_i (X_i - \overline{X}) + \varepsilon_i$$

and denote the OLS coefficient on w_i by $\hat{\tau}^{\text{Lin}}$. Then under the RCT assumptions,

$$\hat{\tau}^{\text{Lin}} \xrightarrow{\mathbb{P}} \text{ATE}$$

and for all p,

$$\text{AVAR}(\hat{\tau}^{\text{Lin}}) \leq \text{AVAR}(\hat{\tau}^{\text{DM}})$$

Observational Studies: Now, suppose we are simply given a dataset. Then we cannot control the assignment w_i, and these methods are not valid. To do causal inference, we need to make two stronger assumptions:

- Unconfoundedness: $w_i \perp (Y_i(0), Y_i(1)) \,\Big|\, X_i$
- Overlap: $\mathbb{P}(w_i = 1 | X_i) \in (0, 1)$ a.s.

Let us define

$$\mu_1(X_i) = \mathbb{E}[Y_i | X_i, w_i = 1]$$

Then under unconfoundedness, we have

$$\mathbb{E}[Y_i(1)] = \mathbb{E}_{X_i}[\mu_1(X_i)]$$

so that the ATE is given by

$$\text{ATE} = \mathbb{E}[\mu_1(X_i) - \mu_0(X_i)]$$

If we knew these functions, we could estimate the ATE by

$$\hat{\text{ATE}} = \frac{1}{n} \sum_{i=1}^{n} \mu_1(X_i) - \mu_0(X_i)$$

We can estimate the CEF $\mu_1(X_i)$ by any ML method, e.g. a neural network. If we know that

$$||\hat{\mu}_{(w)} - \mu_{(w)}||_\infty \xrightarrow{\mathbb{P}} 0$$

for both $w = 0$ and $w = 1$ then using the above estimator,

$$\hat{\text{ATE}} \xrightarrow{\mathbb{P}} \text{ATE}$$

We can again use Lin's method by running the regression

$$Y_i = \alpha + \tau w_i + \gamma^T X_i + \delta^T w_i (X_i - \overline{X}) + \varepsilon_i$$

$$\implies \hat{\mu}_1(X_i) = \hat{\alpha} + \hat{\tau} + \hat{\gamma}^T X_i + \hat{\delta}^T (X_i - \overline{X})$$

$$\hat{\mu}_0(X_i) = \hat{\alpha} + \hat{\gamma}^T X_i + \hat{\delta}^T (X_i - \overline{X})$$

Under this method, we again have

$$\implies \hat{\text{ATE}} = \hat{\tau}$$

Unlike the RCT setting, this is only consistent for the ATE if the functions $\hat{\mu}_1(X_i)$ and $\hat{\mu}_0(X_i)$ are genuinely linear in 1 and X_i.

Multivariate Normal Distributions

To begin, we define a standard normal univariate distribution $Z \sim N(0, 1)$ by the density function

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2})$$

We then define $Z \sim N(0, I_n)$ by taking

$$Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} N(0, 1)$$

$$Z = (Z_1, \ldots, Z_n)^T$$

Then for $\Sigma \succeq 0$ and $\mu \in \mathbb{R}^n$, let $\Sigma = AA^T$ be the Cholesky decomposition. Then to define $Z \sim N(\mu, \Sigma)$, we set

$$\tilde{Z} \sim N(0, I_n) \ , Z = \mu + A\tilde{Z}$$

Note that if $AA^T = \tilde{A}\tilde{A}^T$ then $A\tilde{Z} + \mu \overset{d}{=} \tilde{A}\tilde{Z} + \mu$, so this is a valid definition. Using this definition, we can check that $\mathbb{E}[Z] = \mu$ and $\text{Var}(Z) = \Sigma$.

Proposition: Let $Z \sim N(\mu, \Sigma)$. Then $AZ + b \sim N(A\mu + b, A\Sigma A^T)$.

Proposition: If $Z \sim N(\mu, \Sigma)$ and $\Sigma \succ 0$ then

$$f_Z(z) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left((z - \mu)^T \Sigma^{-1} (z - \mu)\right)$$

In this case, we can also consider $\Lambda = \Sigma^{-1}$ and parametrize a normal distribution as $N(\mu, \Lambda)$.

Partitioning and Marginalization

Partitioning: Consider $Z \sim N(\mu, \Sigma) \in \mathbb{R}^n$ and write

$$Z = \begin{pmatrix} Z_A \\ Z_B \end{pmatrix} \sim N(\mu, \Sigma) = N\left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix} \right)$$

where $Z_A \in \mathbb{R}^m$, $Z_B \in \mathbb{R}^l$, and $m + l = n$. Then we have

$$Z_A \sim N(\mu_A, \Sigma_A A)$$

Proposition: In the above setting, $Z_A \perp Z_B \iff \Sigma_{AB} = 0$.

Corollary: Let $Z \sim N(\mu, \sigma^2 I_n)$ with $\sigma^2 > 0$. Let A, B satisfy $AB^T = 0$. Then $AZ \perp BZ$.

Conditioninal Distribution: Let Z be partitioned as above and assume that $\Sigma \succ 0$ is invertible, so that $\Lambda = \Sigma^{-1}$ exists. Then

$$Z_A | Z_B \sim N\left(\mu_A - \Lambda_{AA}^{-1}\Lambda_{AB}(Z_B - \mu_B), \Lambda_{AA}\right)$$

or

$$Z_A | Z_B \sim N\left(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(Z_B - \mu_B), \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\right)$$

In the case $m = l = 1$, we have $\Sigma_{AB} = \Sigma_{BA} = \rho\sigma_A\sigma_B$, and this formula reduces to

$$Z_A | Z_B \sim N\left(\mu_A + \rho\frac{\sigma_A}{\sigma_B}(Z_B - \mu_B), \sigma_A^2(1 - \rho^2)\right)$$

Notably, $\mathbb{E}[Z_A | Z_B]$ is linear in 1 and Z_B.

Regression to the mean: The above results can explain the phenomenon of regression to the mean. Suppose we have $X, Y \sim N(\mu, \sigma^2)$ with correlation ρ. If we think of X as a pre-treatment result and Y as a post-treatment result then we have ATE equal to 0. But suppose we only treat patients with $X_i \geq \mu$. Then if

$$\hat{\tau} = \frac{1}{\#\{i : w_i = 1\}} \sum_{w_i = 1} (Y_i - X_i)$$

we will have

$$\hat{\tau} \xrightarrow{\mathbb{P}} \mathbb{E}[Y_i - X_i | X_i \geq \mu] < 0$$

by iterated expectation. That is because for each X_i,

$$\mathbb{E}[Y_i | X_i] = \mu + \rho(X_i - \mu) = (1 - \rho)\mu + \rho X_i < X_i$$

when $X_i > \mu$. Hence, for almost all $X_i \geq \mu$,

$$\mathbb{E}[Y_i - X_i] < 0$$

and the inequality follows. Thus, our estimator is inconsistent.

Regression Discontinuity Design: We can choose $w_i = \mathbb{I}\{X_i \geq c\}$ but only consider $X_i \in [c - h, c + h]$ for small h. Then under mild assumptions, $\hat{\tau}^{\mathrm{DM}} \xrightarrow{\mathbb{P}} \mathbb{E}[Y_i(1) - Y_i(0) | X_i = c]$.

Homoscedastic Normal Linear Models

In general, assuming linearity and homoscedastic normal residuals is useful because it allows for finite-sample results (as opposed to purely asymptotic) and yields better power and Type-I error for small sample size. Specifically, we assume

$$Y_i|X_i \sim N(X_i^T\beta, \sigma^2)$$

independently for some fixed β and σ^2. In matrix notation,

$$Y|X \sim N(X\beta, \sigma^2 I_n)$$

We also assume that $p < n$, so that X is typically full-rank. In general, the following theory is all for a fixed design matrix X, and all expectations are conditioned on X.

Example: If $Z \sim N(\mu, \Sigma)$ is some normal random variable, then $Z_j|Z_{-j}$ satisfies these properties.

Again, our primary goal is to estimate β, which we can do via $\hat{\beta}^{\text{OLS}}$. We can compute that

$$\hat{\beta}|X \sim N(\beta, \sigma^2(X^TX)^{-1})$$

so that the OLS estimate is again consistent. In particular,

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2(X^TX)^{-1}_{jj})$$

Also note that $(X^TX)^{-1} = \frac{1}{n}(\frac{X^TX}{n})^{-1}$ and $\frac{X^TX}{n} \xrightarrow{\mathbb{P}} \mathbb{E}[X_iX_i^T]$ as $n \to \infty$, so this variance is shrinking at rate $\frac{1}{n}$.

Confidence Intervals

If we know σ^2, we have a CI for β_j of the form

$$\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}}\sigma\sqrt{(X^TX)^{-1}_{jj}}$$

In practice we do not know σ, so we estimate it by

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - X_i^T\hat{\beta})^2 = \frac{1}{n-p}\mathbf{RSS}$$

Then we have an asymptotic CI of the form

$$\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}$$

This requires linearity and homoscedasticity, but does not use normality.

Projection

Given our assumption on the distribution of $Y|X$, if we denote the projection matrix by $P_X = X(X^TX)^{-1}X^T$, then

$$P_XY|X \sim N(P_XX\beta, \sigma^2P_X)$$
$$(I - P_X)Y|X \sim N(0, \sigma^2(I - P_X))$$

Moreover,

$$P_XY \perp (I - P_X)Y \Big| X$$

Finally, we can note that $\hat{\beta}$ is a function of P_XY by

$$\hat{\beta} = (X^TX)^{-1}X^TP_XY$$

and $\hat{\sigma}^2$ is a function of $(I - P_X)Y$ by

$$\hat{\sigma}^2 = \frac{1}{n-p}||(I - P_X)Y||^2$$

As a result, we conclude that

$$\hat{\beta} \perp \hat{\sigma}^2 \Big| X$$

χ^2-distribution and t-distribution

Definition: Let $Z_1, \ldots, Z_k \overset{\text{iid}}{\sim} N(0,1)$. Then we define the χ^2_k distribution by

$$\chi^2_k = \sum_{j=1}^{k} Z_j^2$$

Hence, we have

$$\hat{\sigma}^2|X \sim \frac{\sigma^2}{n-p}\chi^2_{n-p}$$

Definition: We define the t-distribution with k degrees of freedom by

$$\frac{N(0,1)}{\sqrt{\frac{\chi^2_k}{k}}}$$

where the numerator and denominator are independent.

According to all of our prior results, we have

$$\frac{\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{(X^TX)^{-1}_{jj}}}}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}} \sim t_{n-p}$$

$$\implies \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}} \sim t_{n-p}$$

This is a finite-sample pivot, and allows us to construct confidence intervals with exact coverage:

$$\hat{\beta}_j \pm t_{n-p,1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}$$

This is consistent with our earlier asymptotic confidence interval, because as $n \to \infty$, $n-p \to \infty$, and $t_{n-p,1-\frac{\alpha}{2}} \to z_{1-\frac{\alpha}{2}}$ as well.

Prediction Intervals

Suppose now we want to predict Y_{n+1} given a future X_{n+1}. Our ideal point estimate would naturally be of the form $X_{n+1}^T\beta$, so we begin by estimating this value. Specifically, we have

$$X_{n+1}^T\hat{\beta}|X, X_{n+1} \sim N(X_{n+1}^T\beta, \sigma^2 X_{n+1}^T(X^TX)^{-1}X_{n+1})$$

$$\implies \frac{X_{n+1}^T\hat{\beta} - X_{n+1}^T\beta}{\hat{\sigma}\sqrt{X_{n+1}^T(X^TX)^{-1}X_{n+1}}} \sim t_{n-p}$$

This would provide a CI for $X_{n+1}^T\beta$, but we would like a CI for Y_i, which still has variance around that point estimate. But we can observe that

$$Y_{n+1} - X_{n+1}^T\beta \sim N(0, \sigma^2)$$

$$X_{n+1}^T\beta - X_{n+1}^T\hat{\beta} \sim N(0, \sigma^2 X_{n+1}^T(X^TX)^{-1}X_{n+1})$$

and

$$Y_{n+1} - X_{n+1}^T \perp X_{n+1}^T\beta - X_{n+1}^T\hat{\beta}$$

$$\implies Y_{n+1} - X_{n+1}^T\hat{\beta} \sim N\left(0, \sigma^2(1 + X_{n+1}^T(X^TX)^{-1}X_{n+1})\right)$$

If we know σ^2, this yields a prediction interval of the form

$$X_{n+1}^T\hat{\beta} \pm 1.96\sigma\sqrt{1 + X_{n+1}^T(X^TX)^{-1}X_{n+1}}$$

and if we do not, it yields a prediction interval of the form

$$X_{n+1}^T\hat{\beta} \pm t_{n-p,1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{1 + X_{n+1}^T(X^TX)^{-1}X_{n+1}}$$

Note that we have found a prediction interval with coverage even when conditioned on X and X_{n+1}, which was not previously possible.

Multiple Testing and Simultaneous Inference

Definition: A p-value is something that (under H_0) satisfies $\mathbb{P}_{H_0}(p \leq \alpha) \leq \alpha$. If equality holds, then $p \sim U([0,1])$.

We use p-values to make decisions about hypotheses as follows: for some fixed α, we reject H_0 if $p \leq \alpha$, and otherwise we fail to reject. Then the Type-I error is $\leq \alpha$. Now, suppose we have a family of null hypotheses, $H_1, \ldots, H_m$. We might ask which nulls are rejected, whether the global null (intersection of all H_i) is rejected, or some other question. Let δ_i be 1 if H_i is rejected and 0 otherwise. Moreover, let

$$\mathcal{H}_0 = \{i : 1 \leq i \leq m, H_i \text{ is true}\}$$

Then the number of Type-I errors is given by

$$V = \sum_{i \in \mathcal{H}_0} \delta_i$$

Note that if each H_i has probability α of being falsely rejected, we still have

$$\mathbb{E}[V] = \sum_{i \in \mathcal{H}_0} \alpha$$

which might be quite large if the global null is true, or if many H_i are true.

Definition: The Family-Wise Error Rate (FEWR) is defined by

$$\mathbb{P}(V \geq 1)$$

or the probability of at least one false discovery. Suppose we want $\mathbb{P}(V \geq 1) \leq \alpha$. Then we can use Bonferroni, and test each H_i at level $\frac{\alpha}{m}$. Then

$$\mathbb{P}(V \geq 1) \leq \mathbb{E}(V) = \sum_{i \in \mathcal{H}_0} \mathbb{P}(p_i \leq \frac{\alpha}{m}) \leq \frac{\#\mathcal{H}_0 \alpha}{m} \leq \alpha$$

Note that we reject the global null if $\exists i$ such that $p_i \leq \frac{\alpha}{m}$ or equivalently if $\min p_i \leq \frac{\alpha}{m}$, or $m \min p_i \leq \alpha$. Problematically, for large m, $\frac{\alpha}{m}$ will be very small, and we have little power. Note that we can also use Bonferroni for joint confidence sets, by taking the product of m $1 - \frac{\alpha}{m}$ confidence sets.

To do better than Bonferroni, we need to change the notion of error. We can consider the False Discovery Rate:

$$\text{FDR} = \frac{\sum_{i \in \mathcal{H}_0} \delta_i}{\sum_{i=1}^{m} \delta_i}$$

Now, we seek to have FDR $\leq \alpha$. We can use the Benjamini-Hochberg procedure. Let $p_{(1)}, \ldots, p_{(m)}$ be the sorted p-values. Find the largest k^* such that $p_{(k)} \leq \frac{k\alpha}{m}$, and reject all H_i with $p_i \leq p_{(k^*)}$. If no such k^* exists, then we reject nothing. **Remark:** this will make at least as many discoveries as Bonferroni, and FDR $\leq$ FWER.

Remark: If the p_i are independent, then BH controls FDR at α.

Confidence Sets for β

Now, we return to the homoscedastic linear model and seek to find a confidence set for β. We can already construct a rectangular set using the Bonferroni method. However, we can also see that if $\hat{\theta} \sim N(\theta, \Gamma) \in \mathbb{R}^p$ then

$$\Gamma^{-\frac{1}{2}}(\hat{\theta} - \theta) \sim N(0, p)$$
$$\implies ||\Gamma^{-\frac{1}{2}}(\hat{\theta} - \theta)||_2^2 \sim \chi_p^2$$

This result yields an ellipsoidal confidence set of the form

$$\mathcal{S} = \{\theta : ||\Gamma^{-\frac{1}{2}}(\hat{\theta} - \theta)||_2^2 \leq \chi_{p,1-\alpha}^2\}$$

Under our linear model,

$$||X(\beta - \hat{\beta})||_2^2 \sim \sigma^2 \chi_p^2$$

Thus, if we knew σ, we would have a confidence set given by

$$\mathcal{S} = \{\beta : \frac{||X(\beta - \hat{\beta})||^2}{\sigma^2} \leq \chi_{p,1-\alpha}^2\}$$

When we don't, we instead have the following pivot:

$$\frac{\frac{||X(\beta-\hat{\beta})||^2}{p}}{\hat{\sigma}^2} = \frac{\frac{||X(\beta-\hat{\beta})||^2}{p\sigma^2}}{\frac{\hat{\sigma}^2}{\sigma^2}} \sim \frac{\frac{\chi_p^2}{p}}{\frac{\chi_{n-p}^2}{n-p}}$$

Definition: F-distribution. Suppose $U \sim \chi_k^2$, $V \sim \chi_l^2$, and $U \perp V$. Then we define

$$\frac{\frac{U}{k}}{\frac{V}{l}} = \frac{\frac{\chi_k^2}{k}}{\frac{\chi_l^2}{l}} \sim F_{k,l}$$

In other words, we conclude that

$$\frac{\frac{||X(\beta-\hat{\beta})||^2}{p}}{\hat{\sigma}^2} \sim F_{p,n-p}$$

Finally, this yields a confidence set of the form

$$\mathcal{S} = \{\beta : \frac{||X(\beta - \hat{\beta})||^2}{p\hat{\sigma}^2} \leq F_{1-\alpha,p,n-p}\}$$

F-tests

Suppose we have some $Z \in \mathbb{R}^{n \times k}$ with $\text{Col}(Z) \subsetneq \text{Col}(X)$. We want to test the null hypothesis H_0 that $X\beta \in \text{Col}(Z)$. Typically, we might think of Z as a restriction to certain columns of X. Then

the null hypothesis is that the optimal prediction depends only on those features, or $\beta_j = 0$ for all other j.

We can fit two models:

$$(A) : Y \sim Z \implies \text{obtain the OLS estimate } \hat{\gamma} \implies \text{RSS}(A) = \sum_{i=1}^{n} (Y_i - Z_i^T \hat{\gamma})^2$$

$$(B) : Y \sim X \implies \text{obtain the OLS estimate } \hat{\beta} \implies \text{RSS}(B) = \sum_{i=1}^{n} (Y_i - X_i^T \hat{\beta})^2$$

Note that we always have $\text{RSS}(B) \leq \text{RSS}(A)$, but the difference will allow us to test H_0. Note that

$$\text{RSS}(A) = ||(I - P_Z)Y||^2$$

$$\text{RSS}(B) = ||(I - P_X)Y||^2$$

Moreover,

$$||(I - P_Z)Y||^2 = ||(I - P_X)Y||^2 + ||(P_X - P_Z)Y||^2$$

$$\implies \text{RSS}(A) - \text{RSS}(B) = ||(P_X - P_Z)Y||^2 \overset{H_0}{\sim} \sigma^2 \chi^2_{p-k}$$

Moreover, we have

$$\text{RSS}(A) - \text{RSS}(B) \perp \text{RSS}(B)$$

$$\text{RSS}(B) \sim \sigma^2 \chi^2_{n-p}$$

$$\implies \frac{\frac{\text{RSS}(A) - \text{RSS}(B)}{p-k}}{\frac{\text{RSS}(B)}{n-p}} \sim F_{p-k,n-p}$$

Thus, we can find an exact p-value for this null hypothesis.