Perhaps, if I am very lucky, the feeble efforts of my lifetime will someday be noticed, and maybe, in some small way, they will be acknowledged as the greatest works of genius ever created by Man.

—Eric Laber

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write!

—Mark Sargent

Chapter 1

Linear regression review

1.1 Introduction

Our primary tool for developing our intuition will the linear regression model. In almost any new predictive problem (regression, classification, or reinforcement learning) a linear model is a good place to start. If you are unable to work out the asymptotic behavior of a statistical procedure when everything is linear, it's unlikely that you'll have much luck in a more general setting. This is not, however, a book about linear models. (For those interested in In a book on linear models, there are many excellent references available: Seber and Lee (2012) and Agresti (2015) are two examples.) Rather, our goal is create a primer for researchers in industry and academia who must create new methods for prediction and decision making that are rigorous and scalable.

1.1.1 Notation basics

We assume that the observed data are $\{\mathbf{Z}_i\}_{i=1}^n$ which comprise an i.i.d. draws from the fixed but unknown distribution P. The distribution P is often called the data-generating model or simply the generative model. Let $\mathbf{Z} \sim P$ denote a generic input-output pair. We assume that $\mathbf{Z} \in \mathbb{R}^{p+1}$, to make things concrete, it may be useful to think of $\mathbf{Z} = (\mathbf{X}, Y)$ where $\mathbf{X} \in \mathbb{R}^p$ is an input and $Y \in \mathbb{R}$ is an output. Throughout these notes we will be using operator notation to represent averages

taken with respect to the empirical or data generating distribution. This notation come with a bit of a learning curve, but it greatly simplifies many of the expressions and asymptotic arguments that we'll be using. Furthermore, this notation is widely used in the statistics ML literature, thus learning to read and use this notation will help to make this literature more accessible.

Let $\mathbb{P}_n$ denote the discrete distribution that puts a point mass of n^{-1} at each of the observed data points $\{\mathbf{Z}_i\}_{i=1}^n$; more formally,

$$\mathbb{P}_n \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{Z}_i},$$

where δ_u is the Dirac delta function. Because $\mathbb{P}_n$ is a discrete distribution, expectations taken with respect to $\mathbb{P}_n$ are sums. For example, if $Z \sim \mathbb{P}_n$ then the mean of Z is

$$\mathbb{E}_{\mathbf{Z}\sim\mathbb{P}_n}\mathbf{Z} = \int \mathbf{z} d\mathbb{P}_n(\mathbf{z}) = \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i = \overline{\mathbf{Z}}_n.$$

In operator notation, the preceding expression is written simply as $\mathbb{P}_n\mathbf{Z}$. More generally, for a function $f : \operatorname{dom}\mathbf{Z} \to \mathbb{R}^q$, we have

$$\mathbb{P}_n f(\mathbf{Z}) = \int f(\mathbf{z}) d\mathbb{P}_n(\mathbf{z}) = n^{-1}\sum_{i=1}^n f(\mathbf{Z}_i).$$

Even more compact is to write $\mathbb{P}_n f$ with the argument $\mathbf{Z}$ being made implicit. Similarly, we can write expectations take with respect to the data-generating distribution using operator notation

$$Pf(\mathbf{Z}) = \int f(\mathbf{z}) dP(\mathbf{z}) = \mathbb{E}_P f(\mathbf{Z}).$$

Note that using operator notation, the expectation is taken only with respect to the generic argument $\mathbf{Z}$. To see why this is important, suppose that we have a class of functions $\mathcal{F} = \{f(\mathbf{z};\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ and that we construct an estimator $\widehat{\boldsymbol{\theta}}_n \in \boldsymbol{\Theta}$ using the observed data $\{\mathbf{Z}_i\}_{i=1}^n$, then $\mathbb{P}_n f(\mathbf{Z};\widehat{\boldsymbol{\theta}}_n) = n^{-1}\sum_{i=1}^n f(\mathbf{Z}_i;\widehat{\boldsymbol{\theta}}_n)$ and similarly $Pf(\mathbf{Z};\widehat{\boldsymbol{\theta}}_n) = \int f(\mathbf{z};\widehat{\boldsymbol{\theta}}_n) dP(\mathbf{z})$. Thus, we are not averaging over the randomness in $\widehat{\boldsymbol{\theta}}_n$; this is important in settings where we are evaluating the performance of a fitted model for example. If we want to average over everything, e.g., the distribution of $\mathbf{Z}$ and that of $\widehat{\boldsymbol{\theta}}_n$ we will use the standard notation $\mathbb{E}f(\mathbf{Z};\widehat{\boldsymbol{\theta}}_n)$.

Exercise 1.1.1. Write the population covariance of $\mathbf{Z}$ and the sample covariance constructed from $\{\mathbf{Z}_i\}_{i=1}^n$ using operator notation.

Answer: The covariance is $\mathbf{\Sigma} = P(\mathbf{Z} - P\mathbf{Z})(\mathbf{Z} - P\mathbf{Z})^\intercal = P\mathbf{Z}\mathbf{Z}^\intercal - (P\mathbf{Z})(P\mathbf{Z})^\intercal$ and the sample covariance is $\widehat{\mathbf{\Sigma}}_n = \mathbb{P}_n(\mathbf{Z} - \mathbb{P}_n\mathbf{Z})(\mathbf{Z} - \mathbb{P}_n\mathbf{Z})^\intercal = \mathbb{P}_n\mathbf{Z}\mathbf{Z}^\intercal - (\mathbb{P}_n\mathbf{Z})(\mathbb{P}_n\mathbf{Z})^\intercal$.

Exercise 1.1.2. Write the strong law of large numbers using operator notation.

Answer: Let $\{\mathbf{Z}_i\}_{i=1}^n$ be i.i.d. with $P||\mathbf{Z}||_1 = P\sum_{j=1}^{p+1}|Z_j| < \infty$ then $(\mathbb{P}_n - P)\mathbf{Z} \to 0$ almost surely.

Exercise 1.1.3. Assume that Z is univariate with a continuous distribution. Write the population and sample median using operator notation.

Answer:. The population median is given by $m^* = \arg\min_m P|Z - m|$ while the sample median is given by $\widehat{m}_n = \arg\min_m \mathbb{P}_n|Z - m|$.

Exercise 1.1.4. Write the central limit theorem using operator notation.

Answer: Assume $\{\mathbf{Z}_i\}_{i=1}^n$ are i.i.d. with finite mean $\boldsymbol{\mu} = P\mathbf{Z}$ and covariance $\mathbf{\Sigma} = P(\mathbf{Z} - P\mathbf{Z})(\mathbf{Z} - P\mathbf{Z})^\intercal$ then $\sqrt{n}(\mathbb{P}_n - P)\mathbf{Z} \rightsquigarrow \text{Normal}(0, \mathbf{\Sigma})$, where '$\rightsquigarrow$' denotes convergence in distribution.

It will also be convenient to use so-called little-oh-pee and big-oh-pee notation when describing the convergence of estimators. Let $\{W_n\}_{n\geq 1}$ be a sequence of random variables and let $\{c_n\}_{n\geq 1}$ be a sequence positive constants. We write $W_n = o_P(c_n)$ if (read: W_n is little-oh-pee of c_n) if W_n/c_n converges to zero in probability, i.e., for any $\epsilon > 0$ we have $P(|W_n| > c_n\epsilon) \to 0$. We write $W_n = O_P(c_n)$ (read: W_n is big-oh-pee of c_n) if W_n/c_n is bounded in probability, i.e., for any $\epsilon > 0$ there exists M so that $P(|W_n| > c_nM) \leq \epsilon$ for all n. For example, if $\{\mathbf{Z}_i\}_{i=1}^n$ is an i.i.d. sample with finite mean then from the (weak) law of large numbers we have $||(\mathbb{P}_n - P)\mathbf{Z}|| = o_P(1)$. When $\{\mathbf{W}_n\}_{n\geq 1}$ is a sequence of random vectors we say $\mathbf{W}_n = o_P(c_n)$ to mean that each component of $\mathbf{W}_n$ is $o_P(c_n)$; similarly, we define $\mathbf{W}_n = O_P(c_n)$ component-wise. Thus, provided appropriate moments exist, we have $\sqrt{n}(\mathbb{P}_n - P)\mathbf{Z} = O_P(1)$.

Exercise 1.1.5. Suppose that $\{W_n\}_{n\geq 1}$ are i.i.d. with mean zero and variance σ^2. Show (without appeal to the law of large numbers) that $\mathbb{P}_nW = o_P(1)$.

Answer: This follows from Markov's inequality. To see this, let $\epsilon > 0$ be arbitrary, then

$$\begin{aligned} P\left(|\frac{1}{n}\sum_{i=1}^{n} W_i| > \epsilon\right) &= P\left\{\left(\sum_{i=1}^{n} W_i\right)^2 > n^2\epsilon^2\right\} \\ &\leq \frac{\mathbb{E}\left\{\left(\sum_{i=1}^{n} W_i\right)^2\right\}}{n^2\epsilon^2} \\ &= \frac{\mathrm{Var}\left(\sum_{i=1}^{n} W_i\right)}{n^2\epsilon^2} \\ &= \frac{\sigma^2}{n\epsilon^2} \to 0, \end{aligned}$$

where we have used the fact that $PW = 0$ so that $\mathrm{Var}(W) = PW^2$ and that $\mathbb{E}W_iW_j = 0$ if $i \neq j$. There are stronger versions of this result (i.e., that rely on weaker conditions) but their proofs are more complicated.

1.1.2 Linear regression review

To begin we consider an i.i.d. regression setting. In later sections, we will consider the case in which the inputs (covariates) are dependent upon past observations as happens in online prediction problems, online decision problems, and adaptive experimental designs. We assume that the observed data are $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ which contain n independent copies of the input-output pair $(\mathbf{X}, Y) \sim P$. We assume $\mathbf{X} \in \mathbb{R}^p$ and $Y \in \mathbb{R}$; the goal is to predict the (yet unseen) output Y upon observing the input $\mathbf{X}$. For example, in building a survival model for breast cancer patients, $\mathbf{X}$ might represent covariates characterizing the patient's health status, e.g., age, tumor size, co-morbid conditions, HER2 status, family history, lifestyle measures, etc., and Y might be their log survival time under standard care. Another example is predicting the revenue of a new restaurant in which case $\mathbf{X}$ might contain restaurant attributes, e.g., food type, price, quality, ratings of other restaurants managed by the same group, square-footage, per-square foot lease price, local demographics, competitor performance, etc., and Y might be average monthly revenue.

To construct a predictive model, we will: (i) posit a class of candidate models, (ii) specify a criterion for comparing models, and (iii) estimate the model in our class the optimize the specified

criterion.[1] Given a map $f : \mathbb{R}^p \to \mathbb{R}$ so that, using f, we predictive output $f(\mathbf{x})$ upon observing $\mathbf{X} = \mathbf{x}$ we might choose to measure the quality of f using its mean squared error, i.e.,

$$\mathrm{MSE}(f) \triangleq P\left\{Y - f(\mathbf{X})\right\}^2,$$

and given a class $\mathcal{F}$ of maps from $\mathbb{R}^p$ into $\mathbb{R}$, define the optimal within this class as $f^{\mathrm{opt}} = \arg\min_{f\in\mathcal{F}} \mathrm{MSE}(f)$.

Exercise 1.1.6. Let $\mathcal{F}$ be the class of all (measurable) maps from $\mathbb{R}^p$ into $\mathbb{R}$, show that $f^{\mathrm{opt}} = \arg\min_{f\in\mathcal{F}} \mathrm{MSE}(f)$ is given by $f^{\mathrm{opt}}(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x})$.
Answer: Let $\mu(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x})$ and let f be arbitrary. Then

$$\begin{aligned}
\mathrm{MSE}(f) &= P\left\{Y - f(\mathbf{X})\right\}^2 \\
&= P\left\{Y - \mu(\mathbf{X}) + \mu(\mathbf{X}) - f(\mathbf{X})\right\}^2 \\
&= P\left\{Y - \mu(\mathbf{X})\right\}^2 + P\left\{\mu(\mathbf{X}) - f(\mathbf{X})\right\}^2,
\end{aligned}$$

where we have used $P\left\{Y - \mu(\mathbf{X})\right\}\left\{\mu(\mathbf{X}) - f(\mathbf{X})\right\} = 0$ (use the law of the iterated expectation). It follows immediately the that MSE is minimized by taking $f^{\mathrm{opt}}(\mathbf{x}) = \mu(\mathbf{x})$ for (almost all) $\mathbf{x} \in \mathbb{R}^p$.

Suppose that we take our class $\mathcal{F}$ to be the set of linear models, $\mathcal{F}_{\mathrm{Lin}} = \{f(\mathbf{x};\boldsymbol{\beta}) = \mathbf{x}^\intercal\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathcal{B}\}$, where $\mathcal{B}$ is a subset of $\mathbb{R}^p$.[2] Note that we do not assume that the conditional mean, $\mu(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x})$, belongs to $\mathcal{F}_{\mathrm{Lin}}$, i.e., we do not assume that the generative model is linear. Rather, we are considering a linear *working model.*. The acknowledgement that one's model may (and probably is) incorrect is important when analyzing the behavior of a predictive model. It is also important the performance measures like the MSE are meaningful even if the model is misspecified. For example, if we fit a linear model and then deploy the model to make predictions, it is meaningful to talk about how that model will perform regardless of whether the linear model is correct. However, as

[1] In some problems it may be natural to consider multiple-objectives when evaluating a model, e.g., mean utility and risk, treatment efficacy and side-effects, patient benefit and implementation cost. We will assume that any and all competing outcomes have been folded into a single composite utility measure.

[2] At times we may need $\mathcal{B}$ to be compact or have other properties. For now, it suffices to thinks of this as all of $\mathbb{R}^p$.

we will see, tests and confidence intervals for parameters indexing a linear model must be couched in terms of projects of the true model which may or may not be meaningful in a given domain context.

The optimal linear model, i.e., the member of $\mathcal{F}_{\text{Lin}}$ that minimizes the MSE, is indexed by parameters

$$\boldsymbol{\beta}^* \quad = \quad \arg\min_{\boldsymbol{\beta}} P\left(Y - \mathbf{X}^\intercal \boldsymbol{\beta}\right)^2,$$

which, after differentiating and setting to zero, can be seen to solve the (population level) normal equations

$$P(Y - \mathbf{X}^\intercal \boldsymbol{\beta}^*)\mathbf{X} = 0 \quad \Rightarrow \quad \boldsymbol{\beta}^* = \left(P\mathbf{X}\mathbf{X}^\intercal\right)^{-1} P\mathbf{X}Y.$$

Thus, within $\mathcal{F}_{\text{Lin}}$ the optimal model is given by $f^{\text{opt}}(\mathbf{x}) = \mathbf{x}^\intercal \boldsymbol{\beta}^*$. From the normal equations we see that the residuals are orthogonal to the f^{opt} in the sense that $P(Y - \mathbf{X}^\intercal \boldsymbol{\beta}^*) f^{\text{opt}}(\mathbf{X}) = P(Y - \mathbf{X}^\intercal \boldsymbol{\beta}^*)\mathbf{X}^\intercal \boldsymbol{\beta}^* = 0$.

Exercise 1.1.7. Define $\epsilon = Y - \mu(\mathbf{X})$ to the be the 'residual.' Show that $P\mathbf{X}\epsilon = 0$, thus showing $Y = \mu(\mathbf{X}) + \epsilon$ where $\mathbb{E}(\epsilon|\mathbf{X}) = 0$.

Answer: We have $\mathbb{E}\mathbf{X}\epsilon = \mathbb{E}\left\{\mathbb{E}(\mathbf{X}\epsilon|\mathbf{X})\right\} = \mathbb{E}\left[\mathbf{X}\mathbb{E}\left\{Y - \mu(\mathbf{X})\middle|\mathbf{X}\right\}\right] = 0$.

Exercise 1.1.8. Show that $f^{\text{opt}}(\mathbf{x}) = \mathbf{x}^\intercal \boldsymbol{\beta}^*$ is the projection of $\mu(\mathbf{x})$ onto $\mathcal{F}_{\text{Lin}}$, i.e., $f^{\text{opt}} = \arg\min_{f \in \mathcal{F}_{\text{Lin}}} P\left\{\mu(\mathbf{X}) - f(\mathbf{X})\right\}^2$.

Answer: Let $\widetilde{f} = \arg\min_{f \in \mathcal{F}_{\text{Lin}}} P\left\{\mu(\mathbf{X}) - f(\mathbf{X})\right\}^2$ be such that $\widetilde{f}(\mathbf{x}) = \mathbf{x}^\intercal \widetilde{\boldsymbol{\beta}}$. Then

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} P\left\{\mu(\mathbf{X}) - \mathbf{X}^\intercal \boldsymbol{\beta}\right\}^2 \\
&= \left(P\mathbf{X}\mathbf{X}^\intercal\right)^{-1} P\mathbf{X}\mu(\mathbf{X}) \\
&= \left(P\mathbf{X}\mathbf{X}^\intercal\right)^{-1} P\mathbf{X}(Y - \epsilon) \\
&= \left(P\mathbf{X}\mathbf{X}^\intercal\right)^{-1} P\mathbf{X}Y + 0 \quad \text{(Using exercise 1.2.2.)} \\
&= \boldsymbol{\beta}^*.
\end{aligned}$$

Thus, $\widetilde{f} = f^{\text{opt}}$. When I was a kid, I used to hear people (which people? don't ask, it disrupts the narrative) talk about linear models as projections of the conditional mean. I didn't know what they meant for a long time. I wanted to spare you from a similar life of quiet desperation.

Our focus for estimation and inference will be on $\boldsymbol{\beta}^*$. Thus, our focus is on the projection of the conditional mean of Y given $\mathbf{X}$. The natural estimator $\widehat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}^*$ is the ordinary least squares estimator

$$\widehat{\boldsymbol{\beta}}_n \quad = \quad \arg\min_{\boldsymbol{\beta}} \mathbb{P}_n \left(Y - \mathbf{X}^\intercal \boldsymbol{\beta}\right)^2,$$

from which it can be seen that $\widehat{\boldsymbol{\beta}}_n$ satisfies the (sample) normal equations

$$\mathbb{P}_n(Y - \mathbf{X}^\intercal \widehat{\boldsymbol{\beta}}_n)\mathbf{X} = 0 \quad \Rightarrow \widehat{\boldsymbol{\beta}}_n = \left(\mathbb{P}_n \mathbf{X}\mathbf{X}^\intercal\right)^{-1} \mathbb{P}_n \mathbf{X} Y. \tag{1.1}$$

In many linear regression texts, the least squares estimator is introduced using matrix notation so that $\widehat{\boldsymbol{\beta}}_n = (\mathbb{X}^\intercal \mathbb{X})^{-1} \mathbb{X}^\intercal \mathbb{Y}$ where $\mathbb{X} \in \mathbb{R}^{n \times p}$ is the design matrix and $\mathbb{Y} \in \mathbb{R}^n$ the vector of outcomes. The expression in (1.1) is often easier to work with and makes plain that $\widehat{\boldsymbol{\beta}}_n$ is a plug-in estimator of $\boldsymbol{\beta}^*$ obtained by replacing P with $\mathbb{P}_n$ in the definition of $\boldsymbol{\beta}^*$. Of course, the expressions are equivalent as can be seen by noting $\mathbb{X}^\intercal \mathbb{X} = \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\intercal = n \mathbb{P}_n \mathbf{X}\mathbf{X}^\intercal$ and $\mathbb{X}^\intercal \mathbb{Y} = \sum_{i=1}^n \mathbf{X}_i Y_i = n \mathbb{P}_n \mathbf{X} Y$.

1.1.3 Basic asymptotic inference for linear models

A primary focus throughout these notes is quantifying uncertainty in a statistical estimator through asymptotic distributional approximations. Two tools that we will use repeatedly are Slutsky's theorem and the continuous mapping theorem. We state these informally here; for a more rigorous statement see, for example, Resnick (2019). Slutsky's theorem says that if a sequence of random variables Z_n converges in distribution to Z, i.e., $Z_n \rightsquigarrow Z$, and C_n is sequence of random variables that converge in probability to a constant c, i.e., $C_n \to_p c$, then $C_n Z_n \rightsquigarrow cZ$ and $Z_n + C_n \rightsquigarrow Z + c$. For our purposes, the continuous mapping theorem states that if we have a sequence of random variables $Z_n \in \mathbb{R}^p$ with $Z_n \rightsquigarrow Z$ and a continuous map from $f : \mathbb{R}^p \to \mathbb{R}^q$, then $f(Z_n) \rightsquigarrow f(Z)$; this result holds is much greater generality (see Kosorok 2008 for a more general and rigorous

statement). We also state, as a reminder, that if $\mathbf{Z} \sim \text{Normal}_p(\mu, \Omega)$ is a p-dimensional Gaussian random vector and $\mathbf{M} \in \mathbb{R}^{q \times p}$ then $\mathbf{MZ} \sim \text{Normal}_q(\mathbf{M}\mu, \mathbf{M}\Omega\mathbf{M}^\intercal)$.

We are now ready to derive the limiting distribution of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)$. We reiterate that the derivations here do not assume that the linear model is correct. Recall that

$$\widehat{\boldsymbol{\beta}}_n = (\mathbb{P}_n \mathbf{X}\mathbf{X}^\intercal)^{-1} \mathbb{P}_n \mathbf{X} Y,$$

so that if we subtract $\boldsymbol{\beta}^*$ from both sides and scale by $\sqrt{n}$ we obtain

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) = (\mathbb{P}_n \mathbf{X}\mathbf{X}^\intercal)^{-1} \sqrt{n}\mathbb{P}_n \mathbf{X}(Y - \mathbf{X}^\intercal \boldsymbol{\beta}^*),$$

where we have used $(\mathbb{P}_n \mathbf{X}\mathbf{X}^\intercal)^{-1} \mathbb{P}_n \mathbf{X}(Y - \mathbf{X}^\intercal\boldsymbol{\beta}^*) = (\mathbb{P}_n \mathbf{X}\mathbf{X}^\intercal)^{-1} \mathbb{P}_n \mathbf{X} Y - (\mathbb{P}_n \mathbf{X}\mathbf{X}^\intercal)^{-1} \mathbb{P}_n \mathbf{X}\mathbf{X}^\intercal\boldsymbol{\beta}^* = (\mathbb{P}_n \mathbf{X}\mathbf{X}^\intercal)^{-1} \mathbb{P}_n \mathbf{X} Y - \boldsymbol{\beta}^*$. The (population-level) normal equations state that $P(Y - \mathbf{X}^\intercal\boldsymbol{\beta}^*)\mathbf{X} = 0$ from which we obtain

$$\begin{aligned}
\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) &= (\mathbb{P}_n \mathbf{X}\mathbf{X}^\intercal)^{-1} \sqrt{n}(\mathbb{P}_n - P)\mathbf{X}(Y - \mathbf{X}^\intercal\boldsymbol{\beta}^*) \\
&= (P\mathbf{X}\mathbf{X}^\intercal)^{-1} \sqrt{n}(\mathbb{P}_n - P)\mathbf{X}(Y - \mathbf{X}^\intercal\boldsymbol{\beta}^*) + o_P(1) \\
&\rightsquigarrow \text{Normal}\left\{0, (P\mathbf{X}\mathbf{X}^\intercal)^{-1} P\mathbf{X}\mathbf{X}^\intercal(Y - \mathbf{X}\boldsymbol{\beta}^*)^2 (P\mathbf{X}\mathbf{X}^\intercal)^{-1}\right\},
\end{aligned}$$

where we have $\mathbb{P}_n \mathbf{X}\mathbf{X}^\intercal \to P\mathbf{X}\mathbf{X}^\intercal$ almost surely by the strong law of large numbers, Slutsky's theorem, and the central limit theorem. Let $\Sigma = (P\mathbf{X}\mathbf{X}^\intercal)^{-1} P\mathbf{X}\mathbf{X}^\intercal(Y - \mathbf{X}^\intercal\boldsymbol{\beta}^*)^2 (P\mathbf{X}\mathbf{X}^\intercal)^{-1}$. The plug-in estimator $\widehat{\Sigma}_n$ of Σ is obtained by replacing P with $\mathbb{P}_n$ and $\boldsymbol{\beta}^*$ with $\widehat{\boldsymbol{\beta}}_n$, i.e.,[3]

$$\widehat{\Sigma}_n = (\mathbb{P}_n \mathbf{X}\mathbf{X}^\intercal)^{-1} \mathbb{P}_n \mathbf{X}\mathbf{X}^\intercal(Y - \mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n)^2 (\mathbb{P}_n \mathbf{X}\mathbf{X}^\intercal)^{-1}.$$

We can use this to construct (asymptotic) confidence intervals for linear combinations of $\boldsymbol{\beta}^*$ as follows. Let $v \in \mathbb{R}^p$ be a fixed vector, then we know that $v^\intercal\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \rightsquigarrow \text{Normal}(0, v^\intercal\Sigma v)$. Thus, $v^\intercal\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)/\sqrt{v^\intercal\widehat{\Sigma}_n v}$ converges in distribution to a standard normal random variable. Let z_u

[3]This the familiar 'sandwich' covariance estimator (Huber, 1967; White, 1980).

denote the $u \times 100$ percentile of a standard normal distribution. Given $\alpha \in (0,1)$ we have

$$\begin{aligned} P\left(z_{\alpha/2} \leq \frac{v^\intercal \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)}{\sqrt{v^\intercal \widehat{\Sigma}_n v}} \leq z_{1-\alpha/2} \right) &= 1 - \alpha + o(1) \\ \iff P\left(v^\intercal \widehat{\boldsymbol{\beta}}_n - \frac{z_{1-\alpha/2}\sqrt{v^\intercal \widehat{\Sigma}_n v}}{\sqrt{n}} \leq v^\intercal \boldsymbol{\beta}^* \leq v^\intercal \widehat{\boldsymbol{\beta}}_n - \frac{z_{\alpha/2}\sqrt{v^\intercal \widehat{\Sigma}_n v}}{\sqrt{n}} \right) &= 1 - \alpha + o(1), \end{aligned}$$

thus a $(1-\alpha) \times 100\%$ asymptotic confidence interval for $v^\intercal \boldsymbol{\beta}^*$ is

$$\left[v^\intercal \widehat{\boldsymbol{\beta}}_n - \frac{z_{1-\alpha/2}\sqrt{v^\intercal \widehat{\Sigma}_n v}}{\sqrt{n}}, v^\intercal \widehat{\boldsymbol{\beta}}_n - \frac{z_{\alpha/2}\sqrt{v^\intercal \widehat{\Sigma}_n v}}{\sqrt{n}} \right].$$

The term *asymptotic* here refers to the fact that the coverage has a $o(1)$ term attached to it, i.e., the nominal coverage of $(1-\alpha) \times 100\%$ is guaranteed as the sample size grows large.

We can also use the limiting distribution of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)$ to construct a confidence region for $\boldsymbol{\beta}^*$ in $\mathbb{R}^p$. In particular, a $(1-\alpha) \times 100\%$ Wald-type confidence region for $\boldsymbol{\beta}^*$ is given by

$$\zeta_{1-\alpha,n} = \left\{ \boldsymbol{\beta} \in \mathbb{R}^p \,:\, n\left(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right)^\intercal \widehat{\Sigma}_n^{-1} (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \leq \chi^2_{1-\alpha,p} \right\}.$$

To establish that this is a valid confidence region, we need to show $P(\boldsymbol{\beta}^* \in \zeta_{1-\alpha,n}) \geq 1 - \alpha + o(1)$. It follows that $\widehat{\Sigma}_n^{-1/2}\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \rightsquigarrow \text{Normal}(0, I_p)$ and recall if $\mathbf{Z} \sim \text{Normal}(0, I_P)$ then $\mathbf{Z}^\intercal \mathbf{Z} = \sum_{j=1}^p Z_j^2$ follows a χ^2-distribution with p degrees of freedom. Thus, we have

$$\begin{aligned} P\left(\boldsymbol{\beta}^* \in \zeta_{1-\alpha,n}\right) &= P\left\{ n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)^\intercal \widehat{\Sigma}_n^{-1} (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \leq \chi^2_{1-\alpha,p} \right\} \\ &= P\left[\left\{ \widehat{\Sigma}_n^{-1/2}\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \right\}^\intercal \left\{ \widehat{\Sigma}_n^{-1/2}\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \right\} \leq \chi^2_{1-\alpha,n} \right] \\ &= 1 - \alpha + o(1), \end{aligned}$$

where we have used $\left\{ \widehat{\Sigma}_n^{-1/2}\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \right\}^\intercal \left\{ \widehat{\Sigma}_n^{-1/2}\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \right\} \rightsquigarrow \chi^2_p$, which follows from Slutsky's theorem and the continuous mapping theorem.

Exercise 1.1.9. Suppose that $f : \mathbb{R}^p \to \mathbb{R}$ is continuously differentiable which satisfies $\nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta}^*) \neq$

0. Use the asymptotic normality of $\widehat{\boldsymbol{\beta}}_n$ to construct a confidence interval for $f(\boldsymbol{\beta}^*)$[4]. What goes wrong if $\nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta}^*) = 0$?

Answer: A natural estimator for $f(\boldsymbol{\beta}^*)$ is $f(\widehat{\boldsymbol{\beta}}_n)$, let's take this estimator and expand it around $f(\boldsymbol{\beta}^*)$ using a Taylor series expansion:

$$f(\widehat{\boldsymbol{\beta}}_n) \quad = \quad f(\boldsymbol{\beta}^*) + \nabla_\beta f(\widetilde{\boldsymbol{\beta}}_n)^\intercal (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*),$$

where $\widetilde{\boldsymbol{\beta}}_n$ is an intermediate point between $\widehat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}^*$. Thus, because $\widehat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}^*$ with probability one and because $\nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta})$ is continuous we have $\nabla_\beta f(\widetilde{\boldsymbol{\beta}}_n) \to_p \nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta}^*)$ so that

$$\begin{aligned} \sqrt{n}\left\{ f(\widehat{\boldsymbol{\beta}}_n) - f(\boldsymbol{\beta}^*) \right\} &= \nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta}^*)^\intercal \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) + o_P(1) \\ &\rightsquigarrow \text{Normal}\left\{0, \nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta}^*)^\intercal \Sigma \nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta}^*)\right\}. \end{aligned}$$

Using a plug-in estimator for the variance we have

$$\left[f(\widehat{\boldsymbol{\beta}}_n) - \frac{z_{1-\alpha/2}\sqrt{\nabla_{\boldsymbol{\beta}} f(\widehat{\boldsymbol{\beta}}_n)^\intercal \widehat{\Sigma}_n \nabla_{\boldsymbol{\beta}} f(\widehat{\boldsymbol{\beta}}_n)}}{\sqrt{n}}, f(\widehat{\boldsymbol{\beta}}_n) - \frac{z_{\alpha/2}\sqrt{\nabla_{\boldsymbol{\beta}} f(\widehat{\boldsymbol{\beta}}_n)^\intercal \widehat{\Sigma}_n \nabla_{\boldsymbol{\beta}} f(\widehat{\boldsymbol{\beta}}_n)}}{\sqrt{n}} \right],$$

is a $(1-\alpha)\times 100\%$ asymptotic confidence interval for $f(\boldsymbol{\beta}^*)$. Note that if the gradient had vanished at $\boldsymbol{\beta}^*$ the above argument would not have gone through and we would have needed to use a second-order Taylor series and scaled by n rather than $\sqrt{n}$.

The asymptotic distribution of $\widehat{\boldsymbol{\beta}}_n$ can be used to construct confidence intervals for complex (even black-box) functions of $\boldsymbol{\beta}^*$. Consider a map $f : \mathbb{R}^p \to \mathbb{R}$ and our goal is to construct a confidence set for $f(\boldsymbol{\beta}^*)$. However, we do impose any structure on f, e.g., it need not be differentiable or even continuous. For example, consider a K-arm clinical trial in which we have encoded treatments using a one-hot encoding so that treatment k is coded as e_k the kth row of the $K \times K$ identify matrix. Thus, for each patient we observe the input-output pair $(\mathbf{X}, Y)$ where $\mathbf{X} \in \{e_1, \ldots e_K\}$ denotes the assigned treatment and $Y \in \mathbb{R}$ is the outcome coded so that higher values are better. The

[4]Here $\nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta}^*)$ is shorthand for $\nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$.

linear model $\mathbb{E}(Y|\mathbf{X}=\mathbf{x}) = \mathbf{x}^\intercal\boldsymbol{\beta}^*$ encodes a separate mean for each treatment, i.e., $\boldsymbol{\beta}^*_k$ is the mean outcome under treatment k. Suppose we are interested in the mean outcome under the optimal treatment, i.e., $f(\boldsymbol{\beta}^*) = \max_j \boldsymbol{\beta}^*_j$. Then f is not differentiable and it can be shown that one cannot use Taylor series (or bootstrap) approximations without modification (see Laber et al., 2014 for additional mathematical details). As another example, in retail assortment selection (i.e., deciding which products should be stocked in a store), f may represent a complex optimization system that takes a coefficient value $\boldsymbol{\beta}$ as input and outputs a scalar, e.g., it may control for substitutability of items, cost, downstream effects, local availability, variety, etc. A confidence interval for $f(\boldsymbol{\beta}^*)$ may help us understand how uncertainty in $\widehat{\boldsymbol{\beta}}_n$ propagates through the system.

Let $\zeta_{1-\alpha,n}$ denote our Wald-confidence interval for $\boldsymbol{\beta}^*$, a so-called projection interval for $f(\boldsymbol{\beta}^*)$ is given by

$$\text{Proj}(\zeta_{1-\alpha,n}, f) = \bigcup_{\boldsymbol{\beta}\in\zeta_{1-\alpha,n}} f(\boldsymbol{\beta}).$$

To see why this is a valid confidence interval for $f(\boldsymbol{\beta}^*)$, note that

$$P\left\{f(\boldsymbol{\beta}^*) \in \text{Proj}(\zeta_{1-\alpha,n}, f)\right\} \quad \geq \quad P\left(\boldsymbol{\beta}^* \in \zeta_{1-\alpha,n}\right) = 1-\alpha+o(1).$$

The projection interval holds for any f that maps $\mathbb{R}^p$ to $\mathbb{R}$. Thus, a natural question to ask is: What is the price of not imposing any structure on the function f? In general, the price is conservatism. Projection intervals can be quite conservative in some settings (note the inequality in derivation rather than equality) especially when the dimension p is larger (see Laber et al., 2014 for examples).

1.1.4 Measures of performance

In most prediction problems, predictive performance is a key concern. However, which measure of performance is most appropriate depends on the underlying motivation for the problem including how a predictive model will be used by stakeholders. Here, we distinguish between three types of predictive performance and discuss if and when they are likely to be of interest. In regression problems, these measures of performance usually coincide in large samples; however, in classification and decision problems, these measures need not coincide even asymptotically (see Chapters X, Y).

Throughout this section we focus on variants of mean-squared error, though analogous results hold for other measures of risk, e.g., mean absolute deviation.

Conditional test error

Perhaps the most intuitive measure of predictive performance is the conditional test error, which measures the performance of fitted model on an independent test population. In the case of a linear model with estimated coefficients $\widehat{\boldsymbol{\beta}}_n$, the conditional test error is given by

$$C(\widehat{\boldsymbol{\beta}}_n) = P\left(Y - \mathbf{X}^\intercal \widehat{\boldsymbol{\beta}}_n\right)^2 = \int \left(y - \mathbf{x}^\intercal \widehat{\boldsymbol{\beta}}_n\right)^2 dP(\mathbf{x}, y),$$

which is the mean-squared error of $\widehat{\boldsymbol{\beta}}_n$ applied to a new (and independent) input-output pair. Because $C(\widehat{\boldsymbol{\beta}}_n)$ is a function of $\widehat{\boldsymbol{\beta}}_n$ it is a random variable. Thus, this is an example of a data-dependent estimand which is a bit unusual in statistical inference (in introductory statistics classes, we are used to constructing confidence sets or conducting tests with fixed but unknown parameters). Thus, it worth reviewing what we mean by a confidence interval for a data-dependent parameter. Let $\widehat{\ell}_n$ and $\widehat{u}_n$ be statistics, i.e., functions of the observed data, then we say that $\left[\widehat{\ell}_n, \widehat{u}_n\right]$ is a valid $(1-\alpha) \times 100\%$ (asymptotic) confidence interval for $C(\widehat{\boldsymbol{\beta}}_n)$ if

$$P\left\{\widehat{\ell}_n \leq C(\widehat{\boldsymbol{\beta}}_n) \leq \widehat{u}_n\right\} \geq 1 - \alpha + o(1),$$

where the probability statement is over the observed data so that all three expressions inside the probability statement are treated as random. In some settings, we may be interested in construct a *conditional* confidence interval for $C(\widehat{\boldsymbol{\beta}}_n)$ so that the endpoints $\widehat{\ell}_n$ and $\widehat{u}_n$ satisfy

$$P\left\{\widehat{\ell}_n \leq C(\widehat{\boldsymbol{\beta}}_n) \leq \widehat{u}_n \middle| \widehat{\boldsymbol{\beta}}_n\right\} \geq 1 - \alpha + o_P(1),$$

in some ways a conditional confidence interval is more natural as it conditions on the fitted model but it can often be more difficult to construct than an unconditional interval.

To construct an (unconditional) confidence interval for the conditional test error we use the

limiting distribution of the estimated conditional test error $\mathbb{P}_n \left(Y - \mathbf{X}^\intercal \widehat{\boldsymbol{\beta}}_n\right)^2$. Write

$$\begin{aligned}\sqrt{n}\left\{\mathbb{P}_n \left(Y - \mathbf{X}^\intercal \widehat{\boldsymbol{\beta}}_n\right)^2 - C(\widehat{\boldsymbol{\beta}}_n)\right\} &= \sqrt{n}\left\{\mathbb{P}_n \left(Y - \mathbf{X}^\intercal \widehat{\boldsymbol{\beta}}_n\right)^2 - P\left(Y - \mathbf{X}^\intercal \widehat{\boldsymbol{\beta}}_n\right)^2\right\} \\ &= \sqrt{n}(\mathbb{P}_n - P)\left(Y - \mathbf{X}^\intercal \widehat{\boldsymbol{\beta}}_n\right)^2,\end{aligned}$$

note that we cannot apply the central limit theorem directly here because the terms $(Y_i - \mathbf{X}_i^\intercal \widehat{\boldsymbol{\beta}}_n)^2$, $i = 1, \ldots, n$ are not independent. We will first have to manipulate the expression to put into a form to which the central limit applies. In derivations like this, if you are unsure of where to start, it is often useful to start adding and subtracting terms that you think parameters are converging to. Here, we will try adding and subtracting $\mathbf{X}^\intercal \boldsymbol{\beta}^*$ to obtain

$$\begin{aligned}\sqrt{n}(\mathbb{P}_n - P)\left(Y - \mathbf{X}^\intercal \boldsymbol{\beta}^* + \mathbf{X}^\intercal \boldsymbol{\beta}^* - \mathbf{X}^\intercal \widehat{\boldsymbol{\beta}}_n\right)^2 &= \sqrt{n}(\mathbb{P}_n - P)\left(Y - \mathbf{X}^\intercal \boldsymbol{\beta}^*\right)^2 \\ &\quad -2\sqrt{n}(\mathbb{P}_n - P)(Y - \mathbf{X}^\intercal \boldsymbol{\beta}^*)\mathbf{X}^\intercal(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \\ &\quad +\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)^\intercal(\mathbb{P}_n - P)\mathbf{X}\mathbf{X}^\intercal(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \\ &= \sqrt{n}(\mathbb{P}_n - P)(Y - \mathbf{X}^\intercal \boldsymbol{\beta}^*)^2 + O_P(1/\sqrt{n}) \\ &\rightsquigarrow \text{Normal}\left[0, P\left\{(Y - \mathbf{X}^\intercal \boldsymbol{\beta}^*)^2 - P(Y - \mathbf{X}^\intercal \boldsymbol{\beta}^*)^2\right\}^2\right],\end{aligned}$$

where we have used the central limit theorem and Slutsky's theorem. Note that have implicitly assumed that all requisite moments exist.

To construct a confidence interval for $C(\widehat{\boldsymbol{\beta}}_n)$ we can use the plug-in estimator of the variance, say $\widehat{\tau}_n^2$, which is given by

$$\widehat{\tau}_n^2 = \mathbb{P}_n \left\{(Y - \mathbf{X}^\intercal \widehat{\boldsymbol{\beta}}_n)^2 - \mathbb{P}_n \left(Y - \mathbf{X}^\intercal \widehat{\boldsymbol{\beta}}_n\right)^2\right\}^2.$$

A $(1-\alpha) \times 100\%$ asymptotic confidence interval for $C(\widehat{\boldsymbol{\beta}}_n)$ is thus $\mathbb{P}_n(Y - \mathbf{X}^\intercal \widehat{\boldsymbol{\beta}}_n)^2 \pm z_{1-\alpha/2}\widehat{\tau}_n/\sqrt{n}$.

Expected test error

The expected test error is given by $A_n = \mathbb{E}C(\widehat{\boldsymbol{\beta}}_n)$ which averages the conditional test error over training data and thereby captures the average performance of a predictive model at a given sample size. Thus, the expected test error measures the *algorithm* used to construct the predictive model rather than a given model. The expected test error is of interest when comparing methods for repeated deployment in a given domain. For example, suppose that each week a company refreshes its models using only the previous week's data (discarding data that's more than a week old might be necessary if the system is highly non-stationary or if there are storage constraints). Suppose we are considering two candidate algorithms, e.g., a linear model and a neural network, we might compare their average performance across datasets of size equal to the size of the weekly data sets.

We could posit an estimator $\mathbb{P}_n(Y - \mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n)^2$ of A_n and derive the limiting distribution of $\sqrt{n}\left\{A_n - \mathbb{P}_n(Y - \mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n)^2\right\}$. However, this does not capture the dependence of A_n on the sample size which may be of interest in this setting; i.e., we first may want to construct and estimator of the sequence $\{A_m\}_{m\geq n}$. As a first step, interchange the order of expectation to obtain

$$\begin{aligned} A_n &= P\mathbb{E}\left(Y - \mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n\right)^2 \\ &\approx P\int (Y-\omega)^2\,\phi\left(\omega; \mathbf{X}^\intercal\boldsymbol{\beta}^*, \frac{\mathbf{X}^\intercal\Sigma\mathbf{X}}{n}\right)d\omega, \end{aligned}$$

where $\phi(\cdot;\mu,\sigma^2)$ denotes the density of a normal random variable with mean μ and variance σ^2. Computing the integral we obtain

$$A_n \approx P\left\{\frac{\mathbf{X}^\intercal\Sigma\mathbf{X}}{n} + (\mathbf{X}^\intercal\boldsymbol{\beta}^*)^2 - 2YX^\intercal\boldsymbol{\beta}^* + Y^2\right\},$$

taking the plug-in estimator of the right-hand-side we have, for any $m \geq n$,

$$\begin{aligned} \widehat{A}_m &= \mathbb{P}_n\left\{\frac{\mathbf{X}^\intercal\widehat{\Sigma}_n\mathbf{X}}{m} + (\mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n)^2 - 2YX^\intercal\widehat{\boldsymbol{\beta}}_n + Y^2\right\} \\ &= \mathbb{P}_n\left\{\frac{\mathbf{X}^\intercal\widehat{\Sigma}_n\mathbf{X}}{m} + (Y - \mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n)^2\right\}, \end{aligned}$$

which can be seen to equal the in-sample error plus an offset that decays to zero like $O(1/m)$. Let m_n be sequence of non-negative integers such that $m_n/n \to c$ as $n \to \infty$ for some finite constant $c > 0$. The limiting distribution of $\sqrt{m}(\widehat{A}_{m_n} - A_{m_n})$ is obtained by writing

$$\begin{aligned}
\sqrt{m_n}\left(\widehat{A}_{m_n} - A_{m_n}\right) &= \frac{1}{m_n}\mathbb{P}_n\mathbf{X}^\intercal\widehat{\Sigma}_n\mathbf{X} + \sqrt{\frac{m_n}{n}}\sqrt{n}(\mathbb{P}_n - P)(Y - \mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n)^2 \\
&\quad + \sqrt{\frac{m_n}{n}}\sqrt{n}\left[P\left(Y - \mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n\right)^2 - \mathbb{E}P\left(Y - \mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n\right)^2\right] \\
&= c\sqrt{n}(\mathbb{P}_n - P)(Y - \mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n)^2 \\
&\quad - c\sqrt{n}\left[2PY\mathbf{X}^\intercal(\widehat{\boldsymbol{\beta}}_n - \mathbb{E}\widehat{\boldsymbol{\beta}}_n) - \widehat{\boldsymbol{\beta}}_n^\intercal P\mathbf{X}\mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n + \mathbb{E}\widehat{\boldsymbol{\beta}}_n^\intercal P\mathbf{X}\mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n\right] + o_P(1) \quad (1.2)
\end{aligned}$$

We have already shown that $\sqrt{n}(\mathbb{P}_n - P)(Y - \mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n)^2 = \sqrt{n}(\mathbb{P}_n - P)(Y - \mathbf{X}^\intercal\boldsymbol{\beta}^*)^2 + o_P(1)$, we now must consider the behavior of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \mathbb{E}\widehat{\boldsymbol{\beta}}_n)$. It follows that $\mathbb{E}\widehat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}^* + o_P(1/\sqrt{n})$ as

$$\sqrt{n}\left(\mathbb{E}\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*\right) = \mathbb{E}\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \to 0,$$

where we have used $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \rightsquigarrow \text{Normal}(0, \Sigma)$. Thus, after some algebra, we can show the term in line (1.2) is $o_P(1)$ so that

$$\sqrt{m_n}\left(\widehat{A}_{m_n} - A_{m_n}\right) = c\sqrt{n}(\mathbb{P}_n - P)(Y - \mathbf{X}^\intercal\boldsymbol{\beta}^*)^2 + o_P(1),$$

from which we can derive an approximate confidence interval for A_{m_n}.

Unconditional test error

The unconditional test error is given by $R = P(Y - \mathbf{X}^\intercal\boldsymbol{\beta}^*)^2$ and captures the best possible performance of a linear model fit using the features $\mathbf{X}$ in a given context. The unconditional error has value in assessing the value of a predictive model in a given context, e.g., one might consider automating predictions that were previously based on expert judgment. The unconditional error is also valuable as a means of assessing whether it is worth investing in the collection of additional covariates, e.g., if R is above some threshold one may wish to expend the set of predictors or consider a more flexible/expressive class of models.

The derivation of a confidence interval for the unconditional test error, parallels the arguments used above. The plug-in estimator of the unconditional test error is $\mathbb{P}_n(Y - \mathbf{X}^\intercal \widehat{\boldsymbol{\beta}}_n)^2$ and we base inference for R on the limiting distribution of

$$\begin{aligned}
\sqrt{n}\left\{\mathbb{P}_n\left(Y - \mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n\right)^2 - R\right\} &= \sqrt{n}\left\{\mathbb{P}_n(Y - \mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n)^2 - P(Y - \mathbf{X}^\intercal\boldsymbol{\beta}^*)^2\right\} \\
&= \sqrt{n}\left\{\mathbb{P}_n(Y - \mathbf{X}^\intercal\boldsymbol{\beta}^* + \mathbf{X}^\intercal\boldsymbol{\beta}^* - \mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n)^2 - P(Y - \mathbf{X}^\intercal\boldsymbol{\beta}^*)^2\right\} \\
&= \sqrt{n}(\mathbb{P}_n - P)(Y - \mathbf{X}^\intercal\boldsymbol{\beta}^*)^2 + o_P(1),
\end{aligned}$$

where we have used the fact that the cross term vanishes, i.e., $\mathbb{P}_n(Y - \mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n)\mathbf{X}^\intercal(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) = 0$, and that $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)^\intercal\mathbb{P}_n\mathbf{X}\mathbf{X}^\intercal(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) = O_P(1/\sqrt{n})$. Thus, from the central limit theorem and Slutsky's theorem we have

$$\sqrt{n}\left\{\mathbb{P}_n(Y - \mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n)^2 R\right\} \rightsquigarrow \text{Normal}(0, \tau^2),$$

where $\tau^2 = P\left\{(Y - \mathbf{X}^\intercal\boldsymbol{\beta}^*)^2 - P(Y - \mathbf{X}^\intercal\boldsymbol{\beta}^*)^2\right\}^2$. A $(1-\alpha)\times 100\%$ confidence interval for R is thus $\mathbb{P}_n(Y - \mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n)^2 \pm Z_{1-\alpha/2}\widehat{\tau}_n/\sqrt{n}$ where $\widehat{\tau}_n = \mathbb{P}_n\left\{(Y - \mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n)^2 - \mathbb{P}_n(Y - \mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n)^2\right\}^2$.

1.1.5 Bonus content: derivation of conditional test error

We derived a marginal confidence interval for the conditional test error $C(\widehat{\boldsymbol{\beta}}_n)$, i.e., the coverage was averaged over the distribution of $C(\widehat{\boldsymbol{\beta}}_n)$ as well as the distribution of the interval itself. However, we motivated the conditional test error as a means of evaluating the performance of the fitted model on a test population. As noted previously, viewing the conditional error in this way, it may be more natural to construct an interval $\left[\widehat{\ell}_n, \widehat{u}_n\right]$ such that coverage holds *conditionally* on the model, i.e., given confidence level $\alpha \times 100\%$ the interval satisfies

$$P\left\{\widehat{\ell}_n \leq C(\widehat{\boldsymbol{\beta}}_n) \leq \widehat{u}_n \middle| \widehat{\boldsymbol{\beta}}_n\right\} \geq 1 - \alpha + o_P(1).$$

A standard approach for constructing conditional confidence intervals in nice problems (i.e., where everything is asymptotically normal) is to derive the joint distribution of both the target statistic

and the conditioning variable, show this joint distribution is normal, and then use the fact that conditional distributions from a joint normal model are also normal. In this setting, we will derive the joint distribution of the conditional test error and the estimated coefficients. Write

$$\begin{aligned}\left\{ \begin{array}{l} \sqrt{n}(\mathbb{P}_n - P)(Y - \mathbf{X}^\intercal \widehat{\boldsymbol{\beta}}_n)^2 \\ \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \end{array} \right\} &= \sqrt{n}(\mathbb{P}_n - P) \left\{ \begin{array}{l} (Y - \mathbf{X}^\intercal \boldsymbol{\beta}^*)^2 \\ (P\mathbf{X}\mathbf{X}^\intercal)^{-1}(Y - \mathbf{X}^\intercal \boldsymbol{\beta}^*)\mathbf{X} \end{array} \right\} + o_P(1) \\ &\rightsquigarrow \text{Normal}\left(0, \Omega\right),\end{aligned}$$

where $\Omega \in \mathbb{R}^{(p+1)\times(p+1)}$ is the covariance matrix of

$$\left\{ \begin{array}{l} (Y - \mathbf{X}^\intercal \boldsymbol{\beta}^*)^2 \\ (P\mathbf{X}\mathbf{X}^\intercal)^{-1}(Y - \mathbf{X}^\intercal \boldsymbol{\beta}^*)\mathbf{X} \end{array} \right\}.$$

Let Ω be partitioned as

$$\Omega = \left(\begin{array}{cc} \Omega_{1,1} & \Omega 1,2 \\ \Omega_{2,1} & \Omega_{2,2} \end{array} \right),$$

where $\Omega_{1,1}$ is the variance of $(Y - \mathbf{X}^\intercal \boldsymbol{\beta}^*)^2$, $\Omega_{1,2}$ is the covariance of $(Y - \mathbf{X}^\intercal \boldsymbol{\beta}^*)^2$ and $(P\mathbf{X}\mathbf{X}^\intercal)^{-1}(Y - \mathbf{X}^\intercal \boldsymbol{\beta}^*)\mathbf{X}$, and $\Omega_{2,2}$ is the covariance of $(P\mathbf{X}\mathbf{X}^\intercal)^{-1}(Y - \mathbf{X}^\intercal \boldsymbol{\beta}^*)$. Then it follows that

$$\left\{ \begin{array}{l} \sqrt{n}(\mathbb{P}_n - P)(Y - \mathbf{X}^\intercal \boldsymbol{\beta}^*)^2 \\ \widehat{\boldsymbol{\beta}}_n \end{array} \right\} \approx \text{Normal} \left\{ \left(\begin{array}{c} 0 \\ \boldsymbol{\beta}^* \end{array} \right), \left(\begin{array}{cc} \Omega_{1,1} & \Omega_{1,2}/\sqrt{n} \\ \Omega_{2,1}/\sqrt{n} & \Omega_{2,2}/n \end{array} \right) \right\}$$

so that

$$\sqrt{n}(\mathbb{P}_n - P)(Y - \mathbf{X}^\intercal \widehat{\boldsymbol{\beta}}_n) | \widehat{\boldsymbol{\beta}}_n \approx \text{Normal}\left\{ \Omega_{1,2}\Omega_{2,2}^{-1}\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*), \Omega_{1,1} - \Omega_{1,2}\Omega_{2,2}^{-1}\Omega_{2,1} \right\}.$$

We cannot use the above expression to construct a conditional confidence interval for $P(Y - \mathbf{X}^\intercal \widehat{\boldsymbol{\beta}}_n)^2$ as it depends on $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)$. However, we can use the idea of projection interval which was introduced in previous sections. If $\boldsymbol{\beta}^*$ were known then a valid $(1-\alpha) \times 100\%$ conditional interval

for the conditional test error is

$$\mathfrak{W}_{1-\alpha,n}\left(\boldsymbol{\beta}^*|\widehat{\boldsymbol{\beta}}_n\right) = \left[\mathbb{P}_n(Y-\mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n)^2 - \frac{z_{1-\alpha/2}\sqrt{\widehat{\Omega}_{1,1,n}-\widehat{\Omega}_{1,2,n}\widehat{\Omega}_{2,2,n}^{-1}\widehat{\Omega}_{2,1,n}} - \widehat{\Omega}_{1,2,n}\widehat{\Omega}_{2,2,n}^{-1}\sqrt{n}(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}^*)}{\sqrt{n}},\right.$$
$$\left.\mathbb{P}_n(Y-\mathbf{X}^\intercal\widehat{\boldsymbol{\beta}}_n)^2 - \frac{z_{\alpha/2}\sqrt{\widehat{\Omega}_{1,1,n}-\widehat{\Omega}_{1,2,n}\widehat{\Omega}_{2,2,n}^{-1}\widehat{\Omega}_{2,1,n}} - \widehat{\Omega}_{1,2,n}\widehat{\Omega}_{2,2,n}^{-1}\sqrt{n}(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}^*)}{\sqrt{n}}\right],$$

let $\zeta_{n,1-\eta}$ denote a $(1-\eta)\times 100\%$ confidence region for $\boldsymbol{\beta}^*$ then a $(1-\alpha-\eta)\times 100\%$ projection conditional confidence interval for the the conditional test error (what an elegant phrase!) is given by

$$\text{Proj}\left\{\zeta_{1-\eta,n}, \mathfrak{W}_{1-\alpha,n}\left(\cdot|\widehat{\boldsymbol{\beta}}_n\right)\right\} = \bigcup_{\boldsymbol{\beta}^*\in\zeta_{1-\eta,n}} \mathfrak{W}_{1-\alpha,n}\left(\boldsymbol{\beta}^*|\widehat{\boldsymbol{\beta}}_n\right),$$

which will generally be larger than our unconditional confidence interval for the conditional test error (this is not too surprising given that a conditional confidence interval is always an unconditional interval by the law of iterated expectation). While this excess width reduces power for comparisons with the learned model it also gives a more robust characterization of the learned model's performance. Note that one can obtain a marginal coverage of, say 90%, by having 70% of confidence intervals conditionally cover 99.9% of the time and 30% of the intervals conditionally cover 67% of the time, which is undesirable.

The world needs another penalized regression method.

—Nobody (circa 2010).[5]

Penalization is like the word 'myself,' it's used too much (often incorrectly) by people wishing to seem more intelligent than they are.

—Joel Vaughan

[5]Relax, I'm 98.5 percent joking.

Chapter 2

Linear regression and regularization

2.1 Introduction

The bias-variance trade-off is a central concept the construction of high-quality predictive models. Intuitively, more flexible (expressive) predictive models have more degrees of freedom which reduces bias but increases variance. Put another way, the more structure we impose on a model, e.g,. through a parsimonious parameterization, smoothness conditions, additivity, etc., the less information (data) we need to fit the model but the greater the risk of misspecification. In the context of prediction, we might define an optimal trade-off between bias and variance in terms of prediction accuracy, e.g., on a conceptual spectrum of models ranging from low-variance and high-bias to high-variance and low-bias we wish to find a point on this spectrum that maximizes prediction accuracy.[1] As we will see, regularization offers a means of operationalizing such a spectrum of models through a *solution path* and facilitating data- and expert-driven selection of a model on this spectrum. We shall also explore how the set of models on this spectrum can be used to inform interactive model-building. Conversely, automated model-building procedures can be used as a reproducible surrogate for human-in-the-loop interactive model-building, e.g., we can study the properties of algorithmic model-selection procedures to gain insights about the operating characteristics of models constructed using exploratory data analyses and iterative refinements based

[1] In some problems the focus may be on discovery or characterization of relationships between the outcome and key predictors. In these cases, predictive accuracy may not be the primary goal, rather the focus may be on false discovery rates, power, or some other inferential task.

on subjective judgements.

2.1.1 The bias-variance trade-off

We assume that the observed data are $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ which comprise n i.i.d. copies of $(\mathbf{X}, Y) \sim P$, where $\mathbf{X} \in \mathbb{R}^p$ are the inputs (predictors) and $Y \in \mathbb{R}$ is the output. Let $\mu(\mathbf{x}) \triangleq \mathbb{E}(Y|\mathbf{X} = \mathbf{x})$ and consider an estimator $\widehat{f}_n$ of μ constructed from the observed data. The mean-squared error associated with $\widehat{f}_n(\mathbf{x})$ thus decomposes as

$$\begin{aligned}
\mathbb{E}\left\{Y - \widehat{f}_n(\mathbf{X})\right\}^2 &= \mathbb{E}\left(\{Y - \mu(\mathbf{X})\} + \left[\mu(\mathbf{X}) - \mathbb{E}\left\{\widehat{f}_n(\mathbf{X})|\mathbf{X}\right\}\right] + \left[\mathbb{E}\left\{\widehat{f}_n(\mathbf{X})|\mathbf{X}\right\} - \widehat{f}_n(\mathbf{X})\right]\right)^2 \\
&= \sigma_Y^2 + \mathbb{E}\left[\mu(\mathbf{X}) - \mathbb{E}\left\{\widehat{f}_n(\mathbf{X})|\mathbf{X}\right\}\right]^2 + \mathbb{E}\left[\mathbb{E}\left\{\widehat{f}_n(\mathbf{X})|\mathbf{X}\right\} - \widehat{f}_n(\mathbf{X})\right]^2 \\
&= \text{Residual Error} + \text{Bias}^2 + \text{Variance}, \qquad (2.1)
\end{aligned}$$

where we have used to law of iterated expectations to show that the cross-terms vanish. The residual error cannot be reduced without changing the data-generating model, e.g., by collecting additional predictors or reducing uncertainty in the system (using more precise instruments to collect the data etc.) Thus, in building a predictive model focus is often on balancing bias and variance. The bias captures systematic deviations between the estimator and the target $\mu(\mathbf{x})$ so that one might expect the bias to decrease with the flexibility of the estimator (see below for an example). Conversely, as the variance captures instability of the estimator, so that one might expect the variance to increase with the flexibility of an estimator.

To develop out intuition about the bias-variance trade-off and model complexity we consider a special case in which we can compute the MSE in closed form. Suppose that $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \sim_{i.i.d.} \text{Normal}_p(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{p\times p}$ is positive definite. Recall that $\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\intercal = n\mathbb{P}_n\mathbf{X}\mathbf{X}^\intercal \sim \text{Wishart}(\Sigma, n)$ and that $(\sum_{i=1}^n \mathbf{X}\mathbf{X}_i^\intercal)^{-1} = n^{-1}(\mathbb{P}_n\mathbf{X}\mathbf{X}^\intercal)^{-1} \sim \text{Inv} - \text{Wishart}(\Sigma^{-1}, n)$. Further assume that the linear model is correctly specified so that $Y = \mathbf{X}^\intercal\boldsymbol{\beta}^* + \epsilon$ and that the error are independent with mean zero and variance $\sigma_\epsilon^2 < \infty$. For any vector $\mathbf{v} \in \mathbb{R}^p$ and subset $\mathcal{S} \subsetneq \{1, \ldots, p\}$ write $\mathbf{v_S} = \{\mathbf{v}_s : s \in \mathcal{S}\}$. Assume that $\beta_j^* = 0$ for all $j \in \mathcal{J} \subseteq \{1, \ldots, p\}$ so that $Y = \mathbf{X}_{\mathcal{J}^c}^\intercal\boldsymbol{\beta}_{\mathcal{J}^c}^* + \epsilon$,

where $\mathcal{J}^c = \{1, \ldots, p\} \setminus \mathcal{J}$. Suppose that $\mathcal{J}$ were *known.* A natural estimator would thus be

$$\widehat{\boldsymbol{\beta}}_{\mathcal{J}^c,n} = \arg\min_{\boldsymbol{\beta}} \mathbb{P}_n \left(Y - \mathbf{X}_{\mathcal{J}^c}^\intercal \boldsymbol{\beta}_{J^c}\right)^2,$$

i.e., by fitting least squares using only the predictors with non-zero coefficients. Consider an alternate non-empty subset $\mathcal{K} \subseteq \{1, \ldots, p\}$ and corresponding least squares estimator $\widehat{\boldsymbol{\beta}}_{\mathcal{K},n} = \arg\min_{\boldsymbol{\beta}_\mathcal{K}} \mathbb{P}_n \left(Y - \mathbf{X}_\mathcal{K}^\intercal \boldsymbol{\beta}_\mathcal{K}\right)^2$. A natural question to ask is if/when $\widehat{\boldsymbol{\beta}}_{\mathcal{J}^c,n}$ will have lower MSE than $\widehat{\boldsymbol{\beta}}_{\mathcal{K},n}$. It may be tempting at first blush to think that this must *always* be the case, but as we will show, that using the 'right model' need not lead to the most accurate predictions especially when the signal is small.

Exercise 2.1.1. Compute the MSE of $\widehat{f}_{\mathcal{J}^c,n}(\mathbf{x}) = \mathbf{x}^\intercal \widehat{\boldsymbol{\beta}}_{\mathcal{J}^c,n}$.

Answer: Recall that

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{\mathcal{J}^c,n} &= \left(\mathbb{P}_n \mathbf{X}_{\mathcal{J}^c} \mathbf{X}_{\mathcal{J}^c}^\intercal\right)^{-1} \mathbb{P}_n \mathbf{X}_{\mathcal{J}^c} Y \\
&= (\mathbb{P}_n \mathbf{X}_{\mathcal{J}^c} \mathbf{X}_{\mathcal{J}^c})^{-1} \mathbb{P}_n \mathbf{X}_{\mathcal{J}^c} (\mathbf{X}_{\mathcal{J}^c}^\intercal \boldsymbol{\beta}_{\mathcal{J}^c}^* + \epsilon) \\
&= \boldsymbol{\beta}_{\mathcal{J}^c}^* + \mathbb{P}_n \mathbf{X}_{\mathcal{J}^c} \mathbf{X}_{\mathcal{J}^c})^{-1} \mathbb{P}_n \mathbf{X}_{\mathcal{J}^c} \epsilon,
\end{aligned}$$

taking expectations of both sides and using the fact that ϵ is independent of $\mathbf{X}$ and has mean zero yields $\mathbb{E}\widehat{\beta}_{\mathcal{J}^c,n} = \boldsymbol{\beta}_{\mathcal{J}^c}^*$ so that $\mathbb{E}\left[\mu(\mathbf{X}) - \mathbb{E}\left\{\widehat{f}_{\mathcal{J}^c,n}(\mathbf{X})|\mathbf{X}\right\}\right]^2 = \mathbb{E}\left(\mathbf{X}^\intercal \boldsymbol{\beta}_{\mathcal{J}^c}^* - \mathbf{X}^\intercal \boldsymbol{\beta}_{\mathcal{J}^c}^*\right)^2 = 0$. Thus, the bias term in our MSE decomposition (2.1) is zero. Let $\widetilde{\mathbf{X}}$ be an independent draw from the distribution of $\mathbf{X}$. The variance term in (2.1) is

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left\{\widehat{f}_{\mathcal{J}^c,n}(\mathbf{X})|\mathbf{X}\right\} - \widehat{f}_{\mathcal{J}^c,n}(\mathbf{X})\right]^2 &= \mathbb{E}\left(\mathbf{X}^\intercal \boldsymbol{\beta}_{\mathcal{J}^c}^* - \mathbf{X}^\intercal \widehat{\boldsymbol{\beta}}_{\mathcal{J}^c,n}\right)^2 \\
&= \mathbb{E}\left\{\widetilde{\mathbf{X}}^\intercal (\mathbb{P}_n \mathbf{X}\mathbf{X}^\intercal)^{-1} (\mathbb{P}_n \mathbf{X}\epsilon)^2 (\mathbb{P}_n \mathbf{X}\mathbf{X}^\intercal)^{-1} \widetilde{\mathbf{X}}\right\}. \quad (2.2)
\end{aligned}$$

Note that

$$(\mathbb{P}_n \mathbf{X}\epsilon)^2 = \frac{1}{n^2} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T \epsilon_i^2 + \frac{1}{n^2} \sum_{i \neq j} \mathbf{X}_i \mathbf{X}_j^\intercal \epsilon_i \epsilon_j$$

so that

$$\mathbb{E}\left\{(\mathbb{P}_n\mathbf{X}\epsilon)^2 \,\middle|\, \mathbf{X}_1,\ldots,\mathbf{X}_n\right\} = \frac{\sigma_\epsilon^2}{n}\mathbb{P}_n\mathbf{X}\mathbf{X}^\intercal.$$

Thus, we can write (2.2) as

$$\begin{aligned}
\mathbb{E}\left\{\widetilde{\mathbf{X}}^\intercal(\mathbb{P}_n\mathbf{X}\mathbf{X}^\intercal)^{-1}(\mathbb{P}_n\mathbf{X}\epsilon)^2(\mathbb{P}_n\mathbf{X}\mathbf{X}^\intercal)^{-1}\widetilde{\mathbf{X}}\right\} &= \mathbb{E}\left[\mathbb{E}\left\{\widetilde{\mathbf{X}}^\intercal(\mathbb{P}_n\mathbf{X}\mathbf{X}^\intercal)^{-1}(\mathbb{P}_n\mathbf{X}\epsilon)^2(\mathbb{P}_n\mathbf{X}\mathbf{X}^\intercal)^{-1}\widetilde{\mathbf{X}} \,\middle|\, \mathbf{X}_1,\ldots,\mathbf{X}_n,\widetilde{\mathbf{X}}\right\}\right] \\
&= \frac{\sigma_\epsilon^2}{n}\mathbb{E}\left\{\widetilde{\mathbf{X}}^\intercal(\mathbb{P}_n\mathbf{X}\mathbf{X}^\intercal)^{-1}\widetilde{\mathbf{X}}\right\} \\
&= \mathrm{tr}\left[\frac{\sigma_\epsilon^2}{n}\mathbb{E}\left\{\widetilde{\mathbf{X}}^\intercal(\mathbb{P}_n\mathbf{X}\mathbf{X}^\intercal)^{-1}\widetilde{\mathbf{X}}\right\}\right] \\
&= \frac{\sigma_\epsilon^2}{n}\mathbb{E}\left[\mathrm{tr}\left\{\widetilde{\mathbf{X}}^\intercal(\mathbb{P}_n\mathbf{X}\mathbf{X}^\intercal)^{-1}\widetilde{\mathbf{X}}\right\}\right] \\
&= \frac{\sigma_\epsilon^2}{n}\mathbb{E}\left[\mathrm{tr}\left\{(\mathbb{P}_n\mathbf{X}\mathbf{X}^\intercal)^{-1}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\intercal\right\}\right],
\end{aligned}$$

where tr denotes the trace and we have used the fact that the trace and expectation commute and that $\mathrm{tr}(UV) = \mathrm{tr}(UV)$. Pulling the trace out of the expectation and using the fact that $\widetilde{\mathbf{X}}$ is independent of $\mathbf{X}_1,\ldots,\mathbf{X}_n$, the last term in the preceding display is thus equal to

$$\begin{aligned}
\frac{\sigma_\epsilon^2}{n}\mathrm{tr}\left\{\mathbb{E}(\mathbb{P}_n\mathbf{X}\mathbf{X}^\intercal)^{-1}\Sigma\right\} &= \frac{\sigma_\epsilon^2}{n}\mathrm{tr}\left[\left\{\frac{n\Sigma^{-1}}{n-p-1}\right\}\Sigma\right] \\
&= \frac{\sigma_\epsilon^2 p}{n-p-1},
\end{aligned}$$

where we have used the fact that $(n\mathbb{P}_n\mathbf{X}\mathbf{X}^\intercal)^{-1}$ has an Inverse-Wishart distribution with mean Σ and n degrees of freedom and that $\mathrm{tr}(I_p) = p$.

Now consider the MSE of $\widehat{\boldsymbol{\beta}}_{\mathcal{K},n}$. It follows that

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{\mathcal{K},n} &= \left(\mathbb{P}_n\mathbf{X}_\mathcal{K}\mathbf{X}_\mathcal{K}^\intercal\right)^{-1}\mathbb{P}_n\mathbf{X}_\mathcal{K}Y \\
&= (\mathbb{P}_n\mathbf{X}_\mathcal{K}\mathbf{X}_\mathcal{K}^\intercal)^{-1}\mathbb{P}_n\mathbf{X}_\mathcal{K}(\mathbf{X}_{\mathcal{J}^c}^\intercal\boldsymbol{\beta}^*_{\mathcal{J}^c} + \epsilon) \\
&= (\mathbb{P}_n\mathbf{X}_\mathcal{K}\mathbf{X}_\mathcal{K}^\intercal)^{-1}\mathbb{P}_n\mathbf{X}_\mathcal{K}\mathbf{X}_{\mathcal{J}^c}^\intercal\boldsymbol{\beta}^*_{\mathcal{J}^c} + (\mathbb{P}_n\mathbf{X}_\mathcal{K}\mathbf{X}_\mathcal{K}^\intercal)^{-1}\mathbb{P}_n\mathbf{X}_\mathcal{K}\epsilon.
\end{aligned}$$

Suppose that $\mathcal{K} \subseteq \mathcal{J}$ and write

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\mathcal{K},n} &= \left(\mathbb{P}_n \mathbf{X}_\mathcal{K} \mathbf{X}_\mathcal{K}^\intercal\right)^{-1} \mathbb{P}_n \mathbf{X}_\mathcal{K} \left(\mathbf{X}_\mathcal{K}^\intercal \boldsymbol{\beta}^*_\mathcal{K} + \mathbf{X}^\intercal_{\mathcal{J}^c \setminus \mathcal{K}} \boldsymbol{\beta}_{\mathcal{J}^c \setminus \mathcal{K}}\right) + (\mathbb{P}_n \mathbf{X}_\mathcal{K} \mathbf{X}_\mathcal{K}^\intercal)^{-1} \mathbb{P}_n \mathbf{X}_\mathcal{K} \epsilon \\ &= \boldsymbol{\beta}^*_\mathcal{K} + \left(\mathbb{P}_n \mathbf{X}_\mathcal{K} \mathbf{X}_\mathcal{K}^\intercal\right)^{-1} \mathbb{P}_n \mathbf{X}_\mathcal{K} \mathbf{X}^\intercal_{\mathcal{J}^c \setminus \mathcal{K}} \boldsymbol{\beta}_{\mathcal{J}^c \setminus \mathcal{K}} + (\mathbb{P}_n \mathbf{X}_\mathcal{K} \mathbf{X}_\mathcal{K}^\intercal)^{-1} \mathbb{P}_n \mathbf{X}_\mathcal{K} \epsilon\end{aligned}$$

For any fixed $\mathbf{x}$ it follows that

$$\mathbb{E} \mathbf{x}_\mathcal{K}^\intercal \widehat{\boldsymbol{\beta}}_{\mathcal{K},n} = \mathbf{x}^\intercal \boldsymbol{\beta}^*_\mathcal{K} + O(||\mathbf{x}^\intercal \beta^*_{\mathcal{J}^c \setminus \mathcal{K}}||),$$

thus, we can choose $\boldsymbol{\beta}^*_{\mathcal{J}^c \setminus \mathcal{K}}$ to be nonzero while making the bias arbitrarily small. (This hardly surprising.) Similarly, for an independent draw $\widetilde{\mathbf{X}}$ the variance term is given by

$$\mathrm{Var}\left\{\widetilde{\mathbf{X}}_\mathcal{K}^\intercal \left(\mathbb{P}_n \mathbf{X}_\mathcal{K} \mathbf{X}_\mathcal{K}^\intercal\right)^{-1} \mathbb{P}_n \mathbf{X}_\mathcal{K} \mathbf{X}^\intercal_{\mathcal{J}^c \setminus \mathcal{K}} \boldsymbol{\beta}^*_{\mathcal{J}^c \setminus \mathcal{K}} + \widetilde{\mathbf{X}}_\mathcal{K}^\intercal \left(\mathbb{P}_n \mathbf{X}_\mathcal{K} \mathbf{X}_\mathcal{K}^\intercal\right)^{-1} \mathbb{P}_n \mathbf{X}_\mathcal{K} \epsilon\right\}.$$

If $||\boldsymbol{\beta}_{\mathcal{J}^c \setminus \mathcal{K}}||$ is small, the second term in variance will dominate. Applying a parallel argument to the computation of the variance in the full-model (above) we see that the variance is equal to

$$\frac{\sigma_\epsilon^2 |\mathcal{K}|}{n - |\mathcal{K}| - 1} + O(||\boldsymbol{\beta}_{\mathcal{J}^c \setminus \mathcal{K}}||^2),$$

where $|\mathcal{K}|$ is the number of elements in $\mathcal{K}$. Thus, as $\boldsymbol{\beta}_{\mathcal{J}^c \setminus \mathcal{K}}$ goes to zero the ratio of the MSEs of the reduced model to the full model behaves like $|\mathcal{K}|(n - p - 1)/p(n - |\mathcal{K}| - 1)$. It is therefore straightforward to construct a generative model in which using the correct model is arbitrarily worse than using a reduced model.

Bibliography

Huber, P. J. (1967). Under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: Weather modification*, Volume 5, pp. 221. Univ of California Press.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, 817–838.