MATH 343 / 643 Homework #2

Professor Adam Kapelner

Friday 18th April, 2025

Problem 1

This problem is about OLS estimation in regression. You can assume that

$$
\begin{aligned}
\boldsymbol{X} &:= [\mathbf{1}_n \mid \boldsymbol{x}_{\cdot 1} \mid \ldots \mid \boldsymbol{x}_{\cdot p}] \quad \text{with column indices } 0, 1, \ldots, p \text{ and row indices } 1, 2, \ldots, n \\
\boldsymbol{H} &:= \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T \\
\boldsymbol{Y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\mathcal{E}} \\
\boldsymbol{B} &:= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y} \\
\hat{\boldsymbol{Y}} &= \boldsymbol{H}\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} \\
\boldsymbol{E} &:= \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}
\end{aligned}
$$

where the entries of $\boldsymbol{X}$ are assumed fixed and known and the entries of $\boldsymbol{\beta}$ are the unknown parameter).

(a) [easy] When we "do inference" for the linear model, what is the parameter vector?

The parameter vector in the linear model is $\boldsymbol{\beta}$, which contains all the unknown regression coefficients. Based on the given notation, it would be:

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$$

Where β_0 is the intercept and $\beta_1, \ldots, \beta_p$ are the coefficients for the predictor variables.

(b) [easy] When we "do inference" for the linear model, what are considered the fixed and known quantities?

The fixed and known quantities in the linear model are:

- The design matrix $\mathbf{X}$ (which contains all observed predictor variables)
- The response vector $\mathbf{y}$ (which is the observed realization of the random variable $\mathbf{Y}$)

(c) [easy] When we "do inference" for the linear model, what are considered the random quantities? And what is the notation for their corresponding realizations?

The random quantities in the linear model are:

- The response vector $\mathbf{Y}$ (with realization $\mathbf{y}$)
- The error vector $\boldsymbol{\varepsilon}$ (with realization $\mathbf{e}$)
- The estimator $\mathbf{B}$ (with realization $\mathbf{b}$)
- The fitted values $\hat{\mathbf{Y}}$ (with realization $\hat{\mathbf{y}}$)
- The residuals $\mathbf{E}$ (with realization $\mathbf{e}$)

(d) [easy] What is the "core assumption" in which the classic linear model inference follows?

The "core assumption" for the classic linear model inference is:

$$\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$$

This means the error terms are multivariate normally distributed with mean vector of zeros, are homoscedastic and are uncorrelated.

(e) [easy] From the core assumption, derive the distribution of $\boldsymbol{B}$.

Starting with the core assumption: $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$

And given that $\mathbf{B} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ and $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Substituting:

$$\mathbf{B} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})$$
$$\mathbf{B} = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}$$

Since $\boldsymbol{\varepsilon}$ follows a normal distribution and $\mathbf{B}$ is a linear transformation of $\boldsymbol{\varepsilon}$, $\mathbf{B}$ also follows a normal distribution.

The mean is:

$$E[\mathbf{B}] = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E[\boldsymbol{\varepsilon}] = \boldsymbol{\beta}$$

The variance is:

$$Var[\mathbf{B}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Var[\boldsymbol{\varepsilon}]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$
$$Var[\mathbf{B}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\sigma^2\mathbf{I}_n)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$
$$Var[\mathbf{B}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

Therefore, the distribution of $\mathbf{B}$ is:

$$\mathbf{B} \sim \mathcal{N}_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

(f) [easy] From this result, derive the distribution of B_j.

Since $\mathbf{B} \sim \mathcal{N}_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$, the individual coefficient B_j follows a univariate normal distribution.

The mean is $E[B_j] = \beta_j$

The variance is $Var[B_j] = \sigma^2[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}$, where $[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}$ is the (j,j)-th element of the matrix $(\mathbf{X}^T\mathbf{X})^{-1}$.

Therefore:

$$B_j \sim \mathcal{N}(\beta_j, \sigma^2[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj})$$

(g) [easy] From this result, derive the distribution of B_j standardized.

To standardize B_j, we subtract its mean and divide by its standard deviation:

$$\frac{B_j - \beta_j}{\sqrt{\sigma^2[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}} = \frac{B_j - \beta_j}{\sigma\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}}$$

Since B_j follows a normal distribution, this standardized version follows a standard normal distribution:

$$\frac{B_j - \beta_j}{\sigma\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}} \sim \mathcal{N}(0,1)$$

(h) [easy] from the core assumption, derive the distribution of $\hat{\boldsymbol{Y}}$.

From the core assumption, derive the distribution of $\hat{\mathbf{Y}}$.

Given:

$$\hat{\mathbf{Y}} = \mathbf{HY} = \mathbf{XB} \tag{1}$$
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2}$$
$$\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}_n, \sigma^2\mathbf{I}_n) \tag{3}$$

Substituting $\mathbf{Y}$ into the expression for $\hat{\mathbf{Y}}$:

$$\hat{\mathbf{Y}} = \mathbf{H}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \tag{4}$$
$$= \mathbf{HX}\boldsymbol{\beta} + \mathbf{H}\boldsymbol{\varepsilon} \tag{5}$$

Since $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the hat matrix (projection matrix), and $\mathbf{HX} = \mathbf{X}$:

$$\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{H}\boldsymbol{\varepsilon} \tag{6}$$

Since $\boldsymbol{\varepsilon}$ follows a normal distribution and $\hat{\mathbf{Y}}$ is a linear transformation of $\boldsymbol{\varepsilon}$, $\hat{\mathbf{Y}}$ also follows a normal distribution.

The mean is:

$$E[\hat{\mathbf{Y}}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{H}E[\boldsymbol{\varepsilon}] \tag{7}$$
$$= \mathbf{X}\boldsymbol{\beta} + \mathbf{H}\mathbf{0}_n \tag{8}$$
$$= \mathbf{X}\boldsymbol{\beta} \tag{9}$$

The variance is:

$$Var[\hat{\mathbf{Y}}] = Var[\mathbf{H}\boldsymbol{\varepsilon}] \tag{10}$$
$$= \mathbf{H}Var[\boldsymbol{\varepsilon}]\mathbf{H}^T \tag{11}$$
$$= \mathbf{H}(\sigma^2\mathbf{I}_n)\mathbf{H}^T \tag{12}$$
$$= \sigma^2\mathbf{H}\mathbf{H}^T \tag{13}$$

Since $\mathbf{H}$ is symmetric and idempotent (i.e., $\mathbf{H}^T = \mathbf{H}$ and $\mathbf{H}^2 = \mathbf{H}$):

$$Var[\hat{\mathbf{Y}}] = \sigma^2\mathbf{H}\mathbf{H} \tag{14}$$
$$= \sigma^2\mathbf{H} \tag{15}$$

Therefore, the distribution of $\hat{\mathbf{Y}}$ is:

$$\hat{\mathbf{Y}} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{H}) \tag{16}$$

(i) [easy] From this result, derive the distribution of $\hat{Y}_i$.

Since $\hat{\mathbf{Y}} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{H})$, an individual fitted value $\hat{Y}_i$ follows a univariate normal distribution.

The mean is $E[\hat{Y}_i] = \mathbf{x}_i\boldsymbol{\beta}$, where $\mathbf{x}_i$ is the i-th row of $\mathbf{X}$.

The variance is $Var[\hat{Y}_i] = \sigma^2 H_{ii}$, where H_{ii} is the i-th diagonal element of the hat matrix $\mathbf{H}$.

Therefore:

$$\hat{Y}_i \sim \mathcal{N}(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2 H_{ii}) \tag{17}$$

(j) From this result, derive the distribution of $\hat{Y}_i$ standardized.

To standardize $\hat{Y}_i$, we subtract its mean and divide by its standard deviation:

$$\frac{\hat{Y}_i - \mathbf{x}_i\boldsymbol{\beta}}{\sqrt{\sigma^2 H_{ii}}} = \frac{\hat{Y}_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma\sqrt{H_{ii}}} \tag{18}$$

Since $\hat{Y}_i$ follows a normal distribution, this standardized version follows a standard normal distribution:

$$\frac{\hat{Y}_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma\sqrt{H_{ii}}} \sim \mathcal{N}(0, 1) \tag{19}$$

(j) [easy] From this result, derive the distribution of $\hat{Y}_i$ standardized.

To standardize $\hat{Y}_i$, we subtract its mean and divide by its standard deviation:

$$\frac{\hat{Y}_i - \mathbf{x}_i\boldsymbol{\beta}}{\sqrt{\sigma^2 H_{ii}}} = \frac{\hat{Y}_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma\sqrt{H_{ii}}} \tag{20}$$

Since $\hat{Y}_i$ follows a normal distribution, this standardized version follows a standard normal distribution:

$$\frac{\hat{Y}_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma\sqrt{H_{ii}}} \sim \mathcal{N}(0,1) \tag{21}$$

(k) [easy] from the core assumption, derive the distribution of $\boldsymbol{E}$.

Given:

$$\mathbf{E} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y} \tag{22}$$
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{23}$$
$$\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}_n, \sigma^2\mathbf{I}_n) \tag{24}$$

Substituting $\mathbf{Y}$ into the expression for $\mathbf{E}$:

$$\mathbf{E} = (\mathbf{I}_n - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \tag{25}$$
$$= (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon} \tag{26}$$

Since $\mathbf{HX} = \mathbf{X}$, we have $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \mathbf{0}$:

$$\mathbf{E} = (\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon} \tag{27}$$

Since $\boldsymbol{\varepsilon}$ follows a normal distribution and $\mathbf{E}$ is a linear transformation of $\boldsymbol{\varepsilon}$, $\mathbf{E}$ also follows a normal distribution.

The mean is:

$$E[\mathbf{E}] = (\mathbf{I}_n - \mathbf{H})E[\boldsymbol{\varepsilon}] \tag{28}$$
$$= (\mathbf{I}_n - \mathbf{H})\mathbf{0}_n \tag{29}$$
$$= \mathbf{0}_n \tag{30}$$

The variance is:

$$Var[\mathbf{E}] = Var[(\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}] \tag{31}$$
$$= (\mathbf{I}_n - \mathbf{H})Var[\boldsymbol{\varepsilon}](\mathbf{I}_n - \mathbf{H})^T \tag{32}$$
$$= (\mathbf{I}_n - \mathbf{H})(\sigma^2\mathbf{I}_n)(\mathbf{I}_n - \mathbf{H})^T \tag{33}$$
$$= \sigma^2(\mathbf{I}_n - \mathbf{H})(\mathbf{I}_n - \mathbf{H})^T \tag{34}$$

Since $(\mathbf{I}_n - \mathbf{H})$ is symmetric and idempotent:

$$Var[\mathbf{E}] = \sigma^2(\mathbf{I}_n - \mathbf{H})(\mathbf{I}_n - \mathbf{H}) \tag{35}$$
$$= \sigma^2(\mathbf{I}_n - \mathbf{H}) \tag{36}$$

Therefore, the distribution of $\mathbf{E}$ is:

$$\mathbf{E} \sim \mathcal{N}_n(\mathbf{0}_n, \sigma^2(\mathbf{I}_n - \mathbf{H})) \tag{37}$$

(l) [easy] From this result, derive the distribution of E_i.

Since $\mathbf{E} \sim \mathcal{N}_n(\mathbf{0}_n, \sigma^2(\mathbf{I}_n - \mathbf{H}))$, an individual residual E_i follows a univariate normal distribution.

The mean is $E[E_i] = 0$.

The variance is $Var[E_i] = \sigma^2(1 - H_{ii})$, where H_{ii} is the i-th diagonal element of the hat matrix $\mathbf{H}$.

Therefore:

$$E_i \sim \mathcal{N}(0, \sigma^2(1 - H_{ii})) \tag{38}$$

(m) [easy] From this result, derive the distribution of E_i standardized.

To standardize E_i, we subtract its mean and divide by its standard deviation:

$$\frac{E_i - 0}{\sqrt{\sigma^2(1 - H_{ii})}} = \frac{E_i}{\sigma\sqrt{1 - H_{ii}}} \tag{39}$$

Since E_i follows a normal distribution, this standardized version follows a standard normal distribution:

$$\frac{E_i}{\sigma\sqrt{1 - H_{ii}}} \sim \mathcal{N}(0, 1) \tag{40}$$

(n) [easy] From the core assumption, show that $\frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top\boldsymbol{\mathcal{E}} \sim \chi^2_n$.

Given the core assumption: $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$

We can standardize this by dividing by σ:

$$\frac{\boldsymbol{\varepsilon}}{\sigma} \sim \mathcal{N}_n(\mathbf{0}_n, \mathbf{I}_n)$$

By definition, if $\mathbf{Z} \sim \mathcal{N}_n(\mathbf{0}_n, \mathbf{I}_n)$, then $\mathbf{Z}^\top\mathbf{Z} \sim \chi^2_n$ (the sum of squares of n independent standard normal random variables follows a chi-square distribution with n degrees of freedom).

In our case, $\mathbf{Z} = \frac{\boldsymbol{\varepsilon}}{\sigma}$, so:

$$\mathbf{Z}^\top \mathbf{Z} = \frac{\boldsymbol{\varepsilon}^\top}{\sigma}\frac{\boldsymbol{\varepsilon}}{\sigma} = \frac{\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}}{\sigma^2} \sim \chi^2_n$$

Therefore:

$$\frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} \sim \chi^2_n$$

(o) [easy] Let $\boldsymbol{B}_1 = \boldsymbol{H}$ and let $\boldsymbol{B}_2 = \boldsymbol{I}_n - \boldsymbol{H}$. Justify the use of Cochran's theorem and then find the distributions of $\frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top \boldsymbol{B}_1 \boldsymbol{\mathcal{E}}$ and $\frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top \boldsymbol{B}_2 \boldsymbol{\mathcal{E}}$.

To apply Cochran's theorem, we need to verify that: 1. $\mathbf{B}_1$ and $\mathbf{B}_2$ are symmetric matrices. 2. $\mathbf{B}_1$ and $\mathbf{B}_2$ are idempotent matrices (i.e., $\mathbf{B}_1^2 = \mathbf{B}_1$ and $\mathbf{B}_2^2 = \mathbf{B}_2$). 3. $\mathbf{B}_1 + \mathbf{B}_2 = \mathbf{I}_n$ (the sum equals the identity matrix).

For $\mathbf{B}_1 = \mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$: - $\mathbf{H}$ is symmetric: $\mathbf{H}^\top = [\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]^\top = \mathbf{X}[(\mathbf{X}^\top\mathbf{X})^{-1}]^\top\mathbf{X}^\top = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{H}$ - $\mathbf{H}$ is idempotent: $\mathbf{H}^2 = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{H}$

For $\mathbf{B}_2 = \mathbf{I}_n - \mathbf{H}$: - $\mathbf{B}_2$ is symmetric: $\mathbf{B}_2^\top = (\mathbf{I}_n - \mathbf{H})^\top = \mathbf{I}_n^\top - \mathbf{H}^\top = \mathbf{I}_n - \mathbf{H} = \mathbf{B}_2$ - $\mathbf{B}_2$ is idempotent: $\mathbf{B}_2^2 = (\mathbf{I}_n - \mathbf{H})(\mathbf{I}_n - \mathbf{H}) = \mathbf{I}_n - \mathbf{H} - \mathbf{H} + \mathbf{H}^2 = \mathbf{I}_n - \mathbf{H} - \mathbf{H} + \mathbf{H} = \mathbf{I}_n - \mathbf{H} = \mathbf{B}_2$

Clearly, $\mathbf{B}_1 + \mathbf{B}_2 = \mathbf{H} + (\mathbf{I}_n - \mathbf{H}) = \mathbf{I}_n$

Also, the rank of $\mathbf{H}$ is $\text{rank}(\mathbf{H}) = \text{rank}(\mathbf{X}) = p+1$, and the rank of $\mathbf{I}_n - \mathbf{H}$ is $n-(p+1)$.

Therefore, by Cochran's theorem:

$$\frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top \mathbf{B}_1 \boldsymbol{\varepsilon} = \frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top \mathbf{H} \boldsymbol{\varepsilon} \sim \chi^2_{p+1}$$

$$\frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top \mathbf{B}_2 \boldsymbol{\varepsilon} = \frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top (\mathbf{I}_n - \mathbf{H}) \boldsymbol{\varepsilon} \sim \chi^2_{n-(p+1)}$$

Moreover, these two chi-square distributions are independent.

(p) [easy] Show that $\frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top \boldsymbol{B}_1 \boldsymbol{\mathcal{E}} = \frac{1}{\sigma^2}\,||\boldsymbol{X}(\boldsymbol{B} - \boldsymbol{\beta})||^2$.

Starting with the left-hand side:

$$\frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top \mathbf{B}_1 \boldsymbol{\varepsilon} = \frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top \mathbf{H} \boldsymbol{\varepsilon}$$

Where $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$

Substituting:

$$\frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top \mathbf{H} \boldsymbol{\varepsilon} = \frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{\varepsilon}$$

Recall that $\mathbf{B} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} = \boldsymbol{\beta} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}$

So $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon} = \mathbf{B} - \boldsymbol{\beta}$

Substituting this into the expression:

$$\frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon} = \frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top \mathbf{X}(\mathbf{B}-\boldsymbol{\beta})$$

Now, $\mathbf{X}(\mathbf{B}-\boldsymbol{\beta})$ represents the difference between the fitted values using the estimated coefficients and the fitted values using the true coefficients.

$$\frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top \mathbf{X}(\mathbf{B}-\boldsymbol{\beta}) = \frac{1}{\sigma^2}[\mathbf{X}(\mathbf{B}-\boldsymbol{\beta})]^\top\boldsymbol{\varepsilon}$$

Since $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{XB} = \hat{\mathbf{Y}}$, we have:

$$\frac{1}{\sigma^2}[\mathbf{X}(\mathbf{B}-\boldsymbol{\beta})]^\top\boldsymbol{\varepsilon} = \frac{1}{\sigma^2}[\mathbf{X}(\mathbf{B}-\boldsymbol{\beta})]^\top[\mathbf{X}(\mathbf{B}-\boldsymbol{\beta})]$$

This is the squared norm:

$$\frac{1}{\sigma^2}[\mathbf{X}(\mathbf{B}-\boldsymbol{\beta})]^\top[\mathbf{X}(\mathbf{B}-\boldsymbol{\beta})] = \frac{1}{\sigma^2}\|\mathbf{X}(\mathbf{B}-\boldsymbol{\beta})\|^2$$

Therefore:

$$\frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top \mathbf{B}_1\boldsymbol{\varepsilon} = \frac{1}{\sigma^2}\|\mathbf{X}(\mathbf{B}-\boldsymbol{\beta})\|^2$$

(q) [harder] Why is the term $||\boldsymbol{X}(\boldsymbol{B}-\boldsymbol{\beta})||^2$ used to measure the model's "estimation error"?

(r) [easy] Show that $\frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top \boldsymbol{B}_2\boldsymbol{\mathcal{E}} = \frac{1}{\sigma^2}||\boldsymbol{E}||^2$.

Starting with the left-hand side:

$$\frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top \mathbf{B}_2\boldsymbol{\varepsilon} = \frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top(\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}$$

Recall that $\mathbf{E} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y} = (\mathbf{I}_n - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})$

Since $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \mathbf{0}$, we have:

$$\mathbf{E} = (\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}$$

Therefore:

$$\frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top(\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon} = \frac{1}{\sigma^2}[(\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}]^\top[\frac{\boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}}\cdot(\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}]$$

Since $\mathbf{I}_n - \mathbf{H}$ is symmetric and idempotent:

$$\frac{1}{\sigma^2}[(\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}]^\top[(\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}] = \frac{1}{\sigma^2}\|(\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}\|^2 = \frac{1}{\sigma^2}\|\mathbf{E}\|^2$$

Therefore:

$$\frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top \mathbf{B}_2\boldsymbol{\varepsilon} = \frac{1}{\sigma^2}\|\mathbf{E}\|^2$$

(s) [harder] In what scenarios is $\boldsymbol{\mathcal{E}}^\top \boldsymbol{B}_1 \boldsymbol{\mathcal{E}} > \boldsymbol{\mathcal{E}}^\top \boldsymbol{B}_2 \boldsymbol{\mathcal{E}}$?

(t) [harder] Draw an illustration of $\boldsymbol{\mathcal{E}}$ being orthogonally projected onto colsp $[\boldsymbol{X}]$ via projection matrix $\boldsymbol{H}$. Use the previous answers to denote the quantities of the projection and the error of the projection.

(u) [difficult] A good linear model has a large or a small projection of the error? Discuss.

(v) [easy] Find $\mathbb{E}\left[\frac{1}{\sigma^2}||\boldsymbol{E}||^2\right]$.

From part (o), we know that:

$$\frac{1}{\sigma^2}\|\mathbf{E}\|^2 = \frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top \mathbf{B}_2 \boldsymbol{\varepsilon} \sim \chi^2_{n-(p+1)}$$

For a chi-square random variable with k degrees of freedom, its expected value is k.

Therefore:

$$E[\frac{1}{\sigma^2}\|\mathbf{E}\|^2] = n - (p+1)$$

(w) [easy] Show that $\frac{||\boldsymbol{E}||^2}{n-(p+1)}$ is an unbiased estimate of σ^2.

From part (u), we have:

$$E[\frac{1}{\sigma^2}\|\mathbf{E}\|^2] = n - (p+1)$$

Multiplying both sides by σ^2:

$$E[\|\mathbf{E}\|^2] = \sigma^2 \cdot (n - (p+1))$$

Dividing both sides by $n - (p+1)$:

$$E[\frac{\|\mathbf{E}\|^2}{n - (p+1)}] = \sigma^2$$

Therefore, $\frac{\|\mathbf{E}\|^2}{n-(p+1)}$ is an unbiased estimator of σ^2.

(x) [easy] Prove that $\dfrac{\sqrt{n-(p+1)}(B_j - \beta_j)}{||\boldsymbol{E}||\sqrt{\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}_{j,j}}} \sim T_{n-(p+1)}$.

From earlier results, we know: 1. $B_j \sim \mathcal{N}(\beta_j, \sigma^2[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj})$ 2. $\frac{(n-(p+1))s^2}{\sigma^2} = \frac{\|\mathbf{E}\|^2}{\sigma^2} \sim \chi^2_{n-(p+1)}$ 3. B_j and $\mathbf{E}$ are independent

The t-distribution with ν degrees of freedom can be defined as:

$$T_\nu = \frac{Z}{\sqrt{U/\nu}}$$

where $Z \sim \mathcal{N}(0,1)$ and $U \sim \chi^2_\nu$ are independent.

In our case: - $Z = \frac{B_j - \beta_j}{\sigma\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}}} \sim \mathcal{N}(0,1)$ - $U = \frac{\|\mathbf{E}\|^2}{\sigma^2} \sim \chi^2_{n-(p+1)}$ - $\nu = n-(p+1)$

Therefore:

$$\frac{B_j - \beta_j}{\sigma\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}}} \cdot \sqrt{\frac{n-(p+1)}{\|\mathbf{E}\|^2/\sigma^2}} = \frac{\sqrt{n-(p+1)}(B_j - \beta_j)}{\|\mathbf{E}\|\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}}} \sim T_{n-(p+1)}$$

(y) [easy] Let $H_0 : \beta_j = 0$. Find the test statistic using the fact from the previous question. Let s_e denote $RMSE := \sqrt{MSE} := \sqrt{SSE/(n-(p+1)} = \sqrt{||e||^2/(n-(p+1)}$.

From the previous question, we have:

$$\frac{\sqrt{n-(p+1)}(B_j - \beta_j)}{\|\mathbf{E}\|\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}}} \sim T_{n-(p+1)}$$

Under the null hypothesis $H_0 : \beta_j = 0$, this becomes:

$$\frac{\sqrt{n-(p+1)}B_j}{\|\mathbf{E}\|\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}}} \sim T_{n-(p+1)}$$

Substituting $s_e = \sqrt{\|\mathbf{E}\|^2/(n-(p+1))}$, we get:

$$\frac{B_j}{s_e\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}}} \sim T_{n-(p+1)}$$

This is the t-statistic for testing $H_0 : \beta_j = 0$.

(z) [easy] Consider a new parameter of interest $\mu_\star := \mathbb{E}[Y_\star] = \boldsymbol{x}_\star\boldsymbol{\beta}$, this is the expected response for a unit with measurements given in row vector $\boldsymbol{x}_\star$ whose first entry is 1. Prove that $\dfrac{\hat{Y}_\star - \mu_\star}{\sigma\sqrt{\boldsymbol{x}_\star\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_\star^\top}} \sim \mathcal{N}(0,1)$.

For a new point $\mathbf{x}_\star$, the predicted value is:

$$\hat{Y}_\star = \mathbf{x}_\star\mathbf{B} = \mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$$

The true expected value is:

$$\mu_\star = \mathbf{x}_\star\boldsymbol{\beta}$$

The difference is:

$$\hat{Y}_\star - \mu_\star = \mathbf{x}_\star(\mathbf{B} - \boldsymbol{\beta})$$

We know $\mathbf{B} - \boldsymbol{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}$, so:

$$\hat{Y}_\star - \mu_\star = \mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}$$

Since this is a linear combination of normally distributed errors, $\hat{Y}_\star - \mu_\star$ follows a normal distribution.

The mean is:

$$E[\hat{Y}_\star - \mu_\star] = \mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top E[\boldsymbol{\varepsilon}] = \mathbf{0}$$

The variance is:

$$\begin{aligned} Var[\hat{Y}_\star - \mu_\star] &= \mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Var[\boldsymbol{\varepsilon}]\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top \\ &= \mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\sigma^2\mathbf{I}_n)\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top \\ &= \sigma^2\mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top \\ &= \sigma^2\mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top \end{aligned}$$

Therefore:

$$\hat{Y}_\star - \mu_\star \sim \mathcal{N}(0, \sigma^2\mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top)$$

Standardizing:

$$\frac{\hat{Y}_\star - \mu_\star}{\sigma\sqrt{\mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top}} \sim \mathcal{N}(0,1)$$

(aa) [easy] Prove that $\dfrac{\sqrt{n-(p+1)}(\hat{Y}_\star - \mu_\star)}{||\boldsymbol{E}||\sqrt{\boldsymbol{x}_\star\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_\star^\top}} \sim T_{n-(p+1)}$.

From part (y), we have:

$$\frac{\hat{Y}_\star - \mu_\star}{\sigma\sqrt{\mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top}} \sim \mathcal{N}(0,1)$$

We also know from earlier results:

$$\frac{\|\mathbf{E}\|^2}{\sigma^2} \sim \chi^2_{n-(p+1)}$$

And importantly, $\hat{Y}_\star - \mu_\star$ and $\mathbf{E}$ are independent.

Using the same approach as in part (w), we can construct a t-distribution:

$$\frac{\hat{Y}_\star - \mu_\star}{\sigma\sqrt{\mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top}} \cdot \sqrt{\frac{n-(p+1)}{\|\mathbf{E}\|^2/\sigma^2}} = \frac{\sqrt{n-(p+1)}(\hat{Y}_\star - \mu_\star)}{\|\mathbf{E}\|\sqrt{\mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top}} \sim T_{n-(p+1)}$$

Therefore:

$$\frac{\sqrt{n-(p+1)}(\hat{Y}_\star - \mu_\star)}{\|\mathbf{E}\|\sqrt{\mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top}} \sim T_{n-(p+1)}$$

(bb) [easy] Let $H_0 : \mu_\star = 17$. Find the test statistic using the fact from the previous question. Let s_e denote the $RMSE$.

From problem aa, we have:

$$\frac{\sqrt{n-(p+1)}(\hat{Y}_\star - \mu_\star)}{\|\mathbf{E}\|\sqrt{\mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top}} \sim T_{n-(p+1)}$$

Under the null hypothesis $H_0 : \mu_\star = 17$, this becomes:

$$\frac{\sqrt{n-(p+1)}(\hat{Y}_\star - 17)}{\|\mathbf{E}\|\sqrt{\mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top}} \sim T_{n-(p+1)}$$

Since $s_e = \frac{\|\mathbf{E}\|}{\sqrt{n-(p+1)}}$, we can rewrite this as:

$$\frac{\hat{Y}_\star - 17}{s_e\sqrt{\mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top}} \sim T_{n-(p+1)}$$

This is the t-statistic for testing $H_0 : \mu_\star = 17$.

(cc) [easy] Consider a new parameter of interest $y_\star = \boldsymbol{x}_\star\boldsymbol{\beta} + \epsilon_\star$, this is the response for a unit with measurements given in row vector $\boldsymbol{x}_\star$ whose first entry is 1. Prove that $\frac{\hat{Y}_\star - y_\star}{\sigma\sqrt{1+\boldsymbol{x}_\star\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_\star^\top}} \sim \mathcal{N}(0,1)$.

For a new observation, we have:

- - The true value: $y_\star = \mathbf{x}_\star\boldsymbol{\beta} + \epsilon_\star$
- - The predicted value: $\hat{Y}_\star = \mathbf{x}_\star\mathbf{B}$

The prediction error is:

$$\hat{Y}_\star - y_\star = \mathbf{x}_\star\mathbf{B} - (\mathbf{x}_\star\boldsymbol{\beta} + \epsilon_\star) = \mathbf{x}_\star(\mathbf{B} - \boldsymbol{\beta}) - \epsilon_\star$$

Since $\mathbf{B} - \boldsymbol{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}$, we have:

$$\hat{Y}_\star - y_\star = \mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon} - \epsilon_\star$$

This is a linear combination of normal random variables ($\boldsymbol{\varepsilon}$ and $\epsilon_\star$), so $\hat{Y}_\star - y_\star$ follows a normal distribution.

Assuming $\epsilon_\star \sim \mathcal{N}(0, \sigma^2)$ and is independent of $\boldsymbol{\varepsilon}$, the mean is:

$$E[\hat{Y}_\star - y_\star] = \mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top E[\boldsymbol{\varepsilon}] - E[\epsilon_\star] = 0 - 0 = 0$$

The variance is:

$$Var[\hat{Y}_\star - y_\star] = Var[\mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}] + Var[\epsilon_\star]$$

We know from problem 1y that:

$$Var[\mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}] = \sigma^2\mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top$$

And $Var[\epsilon_\star] = \sigma^2$, so:

$$Var[\hat{Y}_\star - y_\star] = \sigma^2\mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top + \sigma^2 = \sigma^2(1 + \mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top)$$

Therefore:

$$\hat{Y}_\star - y_\star \sim \mathcal{N}(0, \sigma^2(1 + \mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top))$$

Standardizing:

$$\frac{\hat{Y}_\star - y_\star}{\sigma\sqrt{1 + \mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top}} \sim \mathcal{N}(0,1)$$

(dd) [easy] Prove that $\dfrac{\sqrt{n-(p+1)}(\hat{Y}_\star - y_\star)}{||\boldsymbol{E}||\sqrt{1 + \boldsymbol{x}_\star\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_\star^\top}} \sim T_{n-(p+1)}$.

From problem cc, we have:

$$\frac{\hat{Y}_\star - y_\star}{\sigma\sqrt{1 + \mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top}} \sim \mathcal{N}(0,1)$$

We also know:

$$\frac{\|\mathbf{E}\|^2}{\sigma^2} \sim \chi^2_{n-(p+1)}$$

Since $\hat{Y}_\star - y_\star$ and $\mathbf{E}$ are independent (because $\epsilon_\star$ is independent of $\boldsymbol{\varepsilon}$), we can construct a t-distribution:

$$\frac{\hat{Y}_\star - y_\star}{\sigma\sqrt{1 + \mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top}} \cdot \sqrt{\frac{n-(p+1)}{\|\mathbf{E}\|^2/\sigma^2}} = \frac{\sqrt{n-(p+1)}(\hat{Y}_\star - y_\star)}{\|\mathbf{E}\|\sqrt{1 + \mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top}} \sim T_{n-(p+1)}$$

Therefore:

$$\frac{\sqrt{n-(p+1)}(\hat{Y}_\star - y_\star)}{\|\mathbf{E}\|\sqrt{1 + \mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top}} \sim T_{n-(p+1)}$$

(ee) [easy] Let $H_0 : y_\star = 37$. Find the test statistic using the fact from the previous question. Let s_e denote the $RMSE$.

From problem dd, we have:

$$\frac{\sqrt{n-(p+1)}(\hat{Y}_\star - y_\star)}{\|\mathbf{E}\|\sqrt{1 + \mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top}} \sim T_{n-(p+1)}$$

Under the null hypothesis $H_0 : y_\star = 37$, this becomes:

$$\frac{\sqrt{n-(p+1)}(\hat{Y}_\star - 37)}{\|\mathbf{E}\|\sqrt{1+\mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top}} \sim T_{n-(p+1)}$$

Since $s_e = \frac{\|\mathbf{E}\|}{\sqrt{n-(p+1)}}$, we can rewrite this as:

$$\frac{\hat{Y}_\star - 37}{s_e\sqrt{1+\mathbf{x}_\star(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_\star^\top}} \sim T_{n-(p+1)}$$

This is the t-statistic for testing $H_0 : y_\star = 37$.

(ff) [difficult] Let $S \subseteq \{1, 2, \ldots, p\}$, let $k := |S|$ and let $A = \{0\} \cup S^C$, its complement with zero for the index of the intercept. For convenience, assume you rearrange the columns of the design matrix so that $\boldsymbol{X} = [\boldsymbol{X}_A \mid \boldsymbol{X}_S]$ and the first column is $\mathbf{1}_n$. Let $\boldsymbol{H}_A := \boldsymbol{X}_A(\boldsymbol{X}_A^\top\boldsymbol{X}_A)^{-1}\boldsymbol{X}_A^\top$. It is obvious that $\boldsymbol{H} - \boldsymbol{H}_A$ is symmetric as both $\boldsymbol{H}$ and $\boldsymbol{H}_A$ are symmetric. To prove that $\boldsymbol{H} - \boldsymbol{H}_A$ is an orthogonal projection matrix, prove that it is idempotent. Hint: use the Gram-Schmidt decomposition for both matrices and use block matrix format for $\boldsymbol{H}$.

(gg) [easy] Let $\hat{\boldsymbol{Y}}_A := \boldsymbol{H}_A\boldsymbol{Y}$, the orthogonal projection onto colsp $[\boldsymbol{X}_A]$. Prove that $\frac{(n-(p+1))\left|\left|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_A\right|\right|^2}{k\,||\boldsymbol{E}||^2} \sim F_{k,n-(p+1)}$.

We know: - $\hat{\mathbf{Y}} = \mathbf{HY}$ is the projection onto col$(\mathbf{X})$ - $\hat{\mathbf{Y}}_A = \mathbf{H}_A\mathbf{Y}$ is the projection onto col$(\mathbf{X}_A)$ - $\mathbf{E} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$ is the residual vector

First, let's note that $\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_A = (\mathbf{H} - \mathbf{H}_A)\mathbf{Y}$.

From the hint in problem 1gg, $\mathbf{H} - \mathbf{H}_A$ is an orthogonal projection matrix (symmetric and idempotent). The rank of $\mathbf{H} - \mathbf{H}_A$ equals the difference in ranks between $\mathbf{H}$ and $\mathbf{H}_A$, which is k.

Now:

$$\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_A\|^2 = \mathbf{Y}^\top(\mathbf{H} - \mathbf{H}_A)^\top(\mathbf{H} - \mathbf{H}_A)\mathbf{Y} = \mathbf{Y}^\top(\mathbf{H} - \mathbf{H}_A)\mathbf{Y}$$

Given that $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, we have:

$$\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_A\|^2 = (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})^\top(\mathbf{H} - \mathbf{H}_A)(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})$$

This expands to:

$$\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_A\|^2 = \boldsymbol{\beta}^\top\mathbf{X}^\top(\mathbf{H} - \mathbf{H}_A)\mathbf{X}\boldsymbol{\beta} + 2\boldsymbol{\beta}^\top\mathbf{X}^\top(\mathbf{H} - \mathbf{H}_A)\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^\top(\mathbf{H} - \mathbf{H}_A)\boldsymbol{\varepsilon}$$

Since $\mathbf{HX} = \mathbf{X}$ and assuming $\mathbf{H}_A\mathbf{X}_A = \mathbf{X}_A$, the first term can be simplified but is not zero. However, the key insight is that under the null hypothesis $H_0 : \boldsymbol{\beta}_S = \mathbf{0}_k$, the first two terms become zero, and we have:

$$\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_A\|^2 = \boldsymbol{\varepsilon}^\top(\mathbf{H} - \mathbf{H}_A)\boldsymbol{\varepsilon}$$

Similarly:

$$\|\mathbf{E}\|^2 = \mathbf{Y}^\top(\mathbf{I}_n - \mathbf{H})\mathbf{Y} = \boldsymbol{\varepsilon}^\top(\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}$$

From Cochran's theorem, we know: - $\frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top(\mathbf{H}-\mathbf{H}_A)\boldsymbol{\varepsilon} \sim \chi^2_k$ - $\frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top(\mathbf{I}_n-\mathbf{H})\boldsymbol{\varepsilon} \sim \chi^2_{n-(p+1)}$ - These two chi-square distributions are independent

The F-distribution is defined as the ratio of two independent chi-square distributions, each divided by their degrees of freedom:

$$F_{k,n-(p+1)} = \frac{\chi^2_k/k}{\chi^2_{n-(p+1)}/(n-(p+1))}$$

Therefore:

$$\frac{\boldsymbol{\varepsilon}^\top(\mathbf{H}-\mathbf{H}_A)\boldsymbol{\varepsilon}/k}{\boldsymbol{\varepsilon}^\top(\mathbf{I}_n-\mathbf{H})\boldsymbol{\varepsilon}/(n-(p+1))} = \frac{(n-(p+1))\|\hat{\mathbf{Y}}-\hat{\mathbf{Y}}_A\|^2}{k\|\mathbf{E}\|^2} \sim F_{k,n-(p+1)}$$

(hh) [difficult] Let $\hat{\boldsymbol{E}}_A := (\boldsymbol{I}_n - \boldsymbol{H}_A)\boldsymbol{Y}$, the orthogonal projection onto the colsp $[\boldsymbol{X}_{A_\perp}]$. Prove that $\left|\left|\hat{\boldsymbol{E}}_A\right|\right|^2 - \left|\left|\hat{\boldsymbol{E}}\right|\right|^2 = \left|\left|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_A\right|\right|^2$.

(ii) [easy] Combining the two previous problems, write the test statistic for $H_0 : \boldsymbol{\beta}_S = \mathbf{0}_k$ where β_S denotes the subvector of $\boldsymbol{\beta}$ with indices S. Use the notation $\Delta SSE := SSE_A - SSE$ and MSE.

(jj) [difficult] Prove that the square root of the test statistic in (ii) is the same as t-test statistic from (y) when $k = 1$.

(kk) [harder] The point of this exericse is to demonstrate that the estimator used for the omnibus / global / overall F-test is nothing but a special case of the main result from (gg). Let $S = \{1, 2, \ldots, p\}$ and thus $k = p$ and $A = \{0\}$. Using the result from (gg), show that $\dfrac{(n-(p+1))\left|\left|\hat{\boldsymbol{Y}} - \bar{y}\mathbf{1}_n\right|\right|^2}{p\,||\boldsymbol{E}||^2} \sim F_{p,n-(p+1)}$.

In this case, $S = \{1, 2, \ldots, p\}$ means we're considering all predictors except the intercept. So $A = \{0\}$ refers to only the intercept.

$\mathbf{X}_A$ is just the first column of $\mathbf{X}$, which is $\mathbf{1}_n$.

$\mathbf{H}_A = \mathbf{1}_n(\mathbf{1}_n^\top\mathbf{1}_n)^{-1}\mathbf{1}_n^\top = \mathbf{1}_n(n)^{-1}\mathbf{1}_n^\top = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$

$\mathbf{H}_A\mathbf{Y} = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\mathbf{Y} = \frac{1}{n}\mathbf{1}_n\sum_{i=1}^n Y_i = \bar{y}\mathbf{1}_n$

So $\hat{\mathbf{Y}}_A = \bar{y}\mathbf{1}_n$, which is a vector with all elements equal to the mean of $\mathbf{Y}$.

From problem 1kk, we know:

$$\frac{(n-(p+1))\|\hat{\mathbf{Y}}-\hat{\mathbf{Y}}_A\|^2}{k\|\mathbf{E}\|^2} \sim F_{k,n-(p+1)}$$

Substituting $\hat{\mathbf{Y}}_A = \bar{y}\mathbf{1}_n$ and $k = p$:

$$\frac{(n-(p+1))\|\hat{\mathbf{Y}} - \bar{y}\mathbf{1}_n\|^2}{p\|\mathbf{E}\|^2} \sim F_{p,n-(p+1)}$$

(ll) [easy] Prove that the omnibus / global / overall F-test statistic is $\hat{\hat{F}} = MSR/MSE$ by using the result from (kk).

From problem kk, we have:

$$\frac{(n-(p+1))\|\hat{\mathbf{Y}} - \bar{y}\mathbf{1}_n\|^2}{p\|\mathbf{E}\|^2} \sim F_{p,n-(p+1)}$$

Let's recall the definitions: - $MSR = \frac{\|\hat{\mathbf{Y}}-\bar{y}\mathbf{1}_n\|^2}{p}$ (Mean Square Regression) - $MSE = \frac{\|\mathbf{E}\|^2}{n-(p+1)}$ (Mean Square Error)
Substituting these into the equation:

$$\frac{(n-(p+1))\|\hat{\mathbf{Y}} - \bar{y}\mathbf{1}_n\|^2}{p\|\mathbf{E}\|^2} = \frac{(n-(p+1))}{p} \cdot \frac{\|\hat{\mathbf{Y}} - \bar{y}\mathbf{1}_n\|^2}{\|\mathbf{E}\|^2} = \frac{MSR \cdot (n-(p+1))}{MSE \cdot p} = \frac{MSR}{MSE} \cdot \frac{n-(p+1)}{p}$$

But this doesn't quite match $\hat{\hat{F}} = MSR/MSE$. Let's reconsider:
The F-statistic is actually:

$$F = \frac{\|\hat{\mathbf{Y}} - \bar{y}\mathbf{1}_n\|^2/p}{\|\mathbf{E}\|^2/(n-(p+1))} = \frac{MSR}{MSE}$$

This matches the definition of the omnibus F-test statistic.
Therefore, the omnibus F-test statistic is $\hat{\hat{F}} = MSR/MSE$.

(mm) [difficult] [MA] Prove that the distribution that realizes the R^2 metric (the proportion of response variance explained by the model) is distributed as Beta $\left(\frac{p}{2}, \frac{n-(p+1)}{2}\right)$. This amounts to proving a fact found at the bottom of the F distribution's Wikipedia page .

(nn) [easy] Prove that the maximum likelihood estimate for $\boldsymbol{\beta}$ is $\boldsymbol{b}$, the OLS estimator.

Under the core assumption, $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$, the likelihood function for $\mathbf{Y}$ is:

$$L(\boldsymbol{\beta}, \sigma^2; \mathbf{Y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(\right) \left(-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right)$$

Taking the natural logarithm:

$$\ell(\boldsymbol{\beta}, \sigma^2; \mathbf{Y}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

To find the maximum likelihood estimate for $\boldsymbol{\beta}$, we take the derivative with respect to $\boldsymbol{\beta}$ and set it to zero:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$$

Solving for $\boldsymbol{\beta}$:

$$\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{Y}$$
$$\boldsymbol{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} = \mathbf{B}$$

Therefore, the maximum likelihood estimate for $\boldsymbol{\beta}$ is $\mathbf{B}$, which is the OLS estimator. When realized, this becomes $\mathbf{b}$.

(oo) [harder] Prove that the maximum likelihood estimate for σ^2 is SSE/n.

Continuing from the log-likelihood function:

$$\ell(\boldsymbol{\beta}, \sigma^2; \mathbf{Y}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

To find the maximum likelihood estimate for σ^2, we take the derivative with respect to σ^2 and set it to zero:

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = 0$$

Solving for σ^2:

$$\frac{n}{\sigma^2} = \frac{1}{(\sigma^2)^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$
$$n\sigma^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$
$$\sigma^2 = \frac{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})}{n}$$

Substituting $\boldsymbol{\beta} = \mathbf{B}$, which we proved is the MLE in problem 1oo:

$$\sigma^2 = \frac{(\mathbf{Y} - \mathbf{X}\mathbf{B})^\top(\mathbf{Y} - \mathbf{X}\mathbf{B})}{n} = \frac{\|\mathbf{E}\|^2}{n} = \frac{SSE}{n}$$

Therefore, the maximum likelihood estimate for σ^2 is SSE/n.

(pp) [harder] Find the bias of the maximum likelihood estimator for σ^2 using your answers from (w) and (oo).

Problem 2

This problem is about two types of Bayesian estimation of the slope parameters in linear regression which lead to the ridge and lasso estimates.

(a) [easy] Write the prior assumption about $\boldsymbol{\beta}$ which yields the ridge estimates.

The prior assumption for ridge regression is:

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma_\beta^2 \mathbf{I}_{p+1})$$

This is a multivariate normal prior with mean vector $\mathbf{0}$ and variance-covariance matrix $\sigma_\beta^2 \mathbf{I}_{p+1}$. This prior assumes that all regression coefficients are independent and have the same variance σ_β^2.

(b) [easy] Using the prior and core assumption (which implies a likelihood function for $\boldsymbol{B}$), derive the ridge estimates.

Given:

- Prior: $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma_\beta^2 \mathbf{I}_{p+1})$
- Likelihood (from core assumption): $\mathbf{Y}|\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$

Using Bayes' theorem, the posterior distribution of $\boldsymbol{\beta}$ given $\mathbf{Y}$ is:

$$p(\boldsymbol{\beta}|\mathbf{Y}) \propto p(\mathbf{Y}|\boldsymbol{\beta}) \cdot p(\boldsymbol{\beta})$$

The log-posterior is:

$$\ln p(\boldsymbol{\beta}|\mathbf{Y}) = \ln p(\mathbf{Y}|\boldsymbol{\beta}) + \ln p(\boldsymbol{\beta}) + \text{constant}$$

For the likelihood:

$$\ln p(\mathbf{Y}|\boldsymbol{\beta}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

For the prior:

$$\ln p(\boldsymbol{\beta}) = -\frac{p+1}{2}\ln(2\pi) - \frac{p+1}{2}\ln(\sigma_\beta^2) - \frac{1}{2\sigma_\beta^2}\boldsymbol{\beta}^\top\boldsymbol{\beta}$$

Combining and dropping constants:

$$\ln p(\boldsymbol{\beta}|\mathbf{Y}) \propto -\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) - \frac{1}{2\sigma_\beta^2}\boldsymbol{\beta}^\top\boldsymbol{\beta}$$

To find the maximum a posteriori (MAP) estimate, we maximize this expression, which is equivalent to minimizing:

$$\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})+\frac{1}{2\sigma_\beta^2}\boldsymbol{\beta}^\top\boldsymbol{\beta}$$

Multiplying through by $2\sigma^2$:

$$(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})+\frac{\sigma^2}{\sigma_\beta^2}\boldsymbol{\beta}^\top\boldsymbol{\beta}$$

Let $\lambda=\frac{\sigma^2}{\sigma_\beta^2}$, then we're minimizing:

$$(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})+\lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}$$

This is equivalent to:

$$\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda\|\boldsymbol{\beta}\|^2$$

Taking the derivative with respect to $\boldsymbol{\beta}$ and setting it to zero:

$$-2\mathbf{X}^\top(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})+2\lambda\boldsymbol{\beta}=\mathbf{0}$$

$$\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}+\lambda\boldsymbol{\beta}=\mathbf{X}^\top\mathbf{Y}$$

$$(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I}_{p+1})\boldsymbol{\beta}=\mathbf{X}^\top\mathbf{Y}$$

Solving for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}_{ridge}=(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I}_{p+1})^{-1}\mathbf{X}^\top\mathbf{Y}$$

This is the ridge estimator with regularization parameter $\lambda>0$.

(c) [easy] Write the prior assumption about $\boldsymbol{\beta}$ which yields the lasso estimates.

The prior assumption for lasso regression is a Laplace (double exponential) distribution:

$$p(\boldsymbol{\beta})\propto\exp\left(\,\right)\left(-\frac{\lambda}{\sigma^2}\|\boldsymbol{\beta}\|_1\right)$$

where $\|\boldsymbol{\beta}\|_1=\sum_{j=0}^{p}|\beta_j|$ is the L1-norm of $\boldsymbol{\beta}$.

(d) [easy] Using the prior and core assumption (which implies a likelihood function for $\boldsymbol{B}$), derive the lasso estimates to the point where you need to use a computer to run the optimization.

Given:

- Prior: $p(\boldsymbol{\beta})\propto\exp\left(\,\right)\left(-\frac{\lambda}{\sigma^2}\|\boldsymbol{\beta}\|_1\right)$
- Likelihood (from core assumption): $\mathbf{Y}|\boldsymbol{\beta}\sim\mathcal{N}(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}_n)$

Using Bayes' theorem, the posterior distribution of $\boldsymbol{\beta}$ given $\mathbf{Y}$ is:

$$p(\boldsymbol{\beta}|\mathbf{Y}) \propto p(\mathbf{Y}|\boldsymbol{\beta}) \cdot p(\boldsymbol{\beta})$$

The log-posterior is:

$$\ln p(\boldsymbol{\beta}|\mathbf{Y}) = \ln p(\mathbf{Y}|\boldsymbol{\beta}) + \ln p(\boldsymbol{\beta}) + \text{constant}$$

For the likelihood:

$$\ln p(\mathbf{Y}|\boldsymbol{\beta}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

For the prior:

$$\ln p(\boldsymbol{\beta}) = -\frac{\lambda}{\sigma^2}\|\boldsymbol{\beta}\|_1 + \text{constant}$$

Combining and dropping constants:

$$\ln p(\boldsymbol{\beta}|\mathbf{Y}) \propto -\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) - \frac{\lambda}{\sigma^2}\|\boldsymbol{\beta}\|_1$$

To find the maximum a posteriori (MAP) estimate, we maximize this expression, which is equivalent to minimizing:

$$\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \frac{\lambda}{\sigma^2}\|\boldsymbol{\beta}\|_1$$

Multiplying through by $2\sigma^2$:

$$(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + 2\lambda\|\boldsymbol{\beta}\|_1$$

Let's redefine $\lambda' = 2\lambda$ for simplicity, then we're minimizing:

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda'\|\boldsymbol{\beta}\|_1$$

This is the lasso objective function. Unlike ridge regression, there is no closed-form solution due to the non-differentiability of the L1-norm at zero. The optimization requires numerical methods such as coordinate descent, LARS, or proximal gradient methods.

The solution can be written as:

$$\boldsymbol{\beta}_{lasso} = \arg\min\{\}_{\boldsymbol{\beta}} \left\{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda'\|\boldsymbol{\beta}\|_1\right\}$$

This is the point where computational methods are needed to solve the optimization problem.

(e) [easy] Both ridge and lasso shrink the estimate of $\boldsymbol{\beta}$ towards what vector value?

Both ridge and lasso shrink the estimate of $\boldsymbol{\beta}$ towards the zero vector $\mathbf{0}$.

In ridge regression, the L2 penalty pulls all coefficients toward zero proportionally to their magnitude, but typically doesn't set any exactly to zero.

In lasso regression, the L1 penalty not only pulls coefficients toward zero but can actually set some coefficients exactly to zero when the penalty is sufficiently large.

(f) [easy] Describe what the prestep called "variable selection" is within the modeling enterprise.

Variable selection is a preprocessing step in the modeling process where we decide which predictor variables (features) to include in our final model. The goal is to identify a subset of the available predictors that:

1. Captures the important relationships between predictors and the response 2. Avoids including irrelevant or redundant variables 3. Creates a more interpretable and potentially more stable model 4. Reduces computational complexity and cost 5. Mitigates overfitting by reducing model complexity

Traditional variable selection methods include: - Forward selection: Starting with no variables and sequentially adding the most significant ones - Backward elimination: Starting with all variables and sequentially removing the least significant ones - Stepwise selection: A combination of forward and backward approaches - Best subset selection: Evaluating all possible combinations of variables

The effectiveness of variable selection is typically assessed using criteria such as AIC, BIC, adjusted R^2, cross-validation error, or the significance of individual predictors.

(g) [easy] Describe why Lasso estimation has the added bonus of being able to perform variable selection and ridge does not.

Lasso estimation has the ability to perform variable selection while ridge regression does not, due to fundamental differences in their penalty terms:

(a) **Mathematical Difference**:
- Lasso uses an L1 penalty: $\lambda \sum_{j=0}^{p} |\beta_j|$
- Ridge uses an L2 penalty: $\lambda \sum_{j=0}^{p} \beta_j^2$

(b) **Geometry of the Solutions**:
- The L1 penalty creates corners and edges in the constraint region where solutions tend to occur.
- These corners correspond to points where some coefficients are exactly zero.
- The L2 penalty creates a circular (or spherical in higher dimensions) constraint region with no corners, so solutions rarely have exact zeros.

(c) **Analytical Explanation**:

- For ridge regression, the solution continuously shrinks coefficients toward zero as λ increases, but they typically remain non-zero.
- For lasso regression, when $|\beta_j|$ is small, the derivative of the penalty with respect to β_j is constant (unlike ridge where it approaches zero), creating a threshold effect that pushes weak coefficients exactly to zero.

(d) **Sparsity**:

- Lasso can produce sparse models (models with many coefficients equal to zero).
- Ridge always includes all predictors, just with smaller coefficients.

(e) **Practical Implication**:

- With lasso, unimportant predictors are automatically excluded from the model.
- With ridge, all predictors remain in the model, requiring a separate variable selection step if a simpler model is desired.

This property makes lasso particularly valuable in high-dimensional settings where automatic variable selection is important for model interpretation and reducing overfitting.

Problem 3

This problem is about the specific robust regression methods we studied.

(a) [easy] If we only know that the errors $\mathcal{E}_1, \ldots, \mathcal{E}_n$ are independent, what tried and true method can we employ to get asymptotically valid inference for $\boldsymbol{\beta}$?

When we only know that the errors are independent, the bootstrap method provides asymptotically valid inference for $\boldsymbol{\beta}$.

We resample the data with replacement, the compute the OLS estimate for each bootstrap sample and then Using the empirical distribution of these bootstrap estimates to construct confidence intervals and perform hypothesis tests.

(b) [easy] If we know that the errors $\mathcal{E}_1, \ldots, \mathcal{E}_n$ are iid with expectation zero and variance σ^2 for all values of $\boldsymbol{x}$ (i.e. the errors are "homoskedastic") but the errors are not necessarily normal, what is the asymptotic distribution of $\boldsymbol{B}$?

Under the conditions of iid errors with zero mean and constant variance σ^2, but without assuming normality, we can use the Central Limit Theorem to establish the asymptotic distribution of $\mathbf{B}$.

By the Lindeberg-Levy Central Limit Theorem:

$$\mathbf{B} \overset{a}{\sim} \mathcal{N}_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1})$$

This means that as the sample size n increases, the distribution of $\mathbf{B}$ approaches a multivariate normal distribution with mean $\boldsymbol{\beta}$ and covariance matrix $\sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$, which is the same as in the classical normal error case.

(c) [easy] If we know that the errors $\mathcal{E}_1, \ldots, \mathcal{E}_n \overset{ind}{\sim} \mathcal{N}(0, \sigma_i^2)$ which means the errors are "heteroskedastic", what is the asymptotic distribution of $\boldsymbol{B}$ using the Huber-White estimator?

When the errors are normally distributed but heteroskedastic (i.e., with different variances σ_i^2), the OLS estimator $\mathbf{B}$ is still unbiased but its variance is no longer given by the usual formula.

The Huber-White estimator (also known as the sandwich estimator or robust standard errors) provides a consistent estimate of the covariance matrix of $\mathbf{B}$ under heteroskedasticity.

With heteroskedastic normal errors, the asymptotic distribution of $\mathbf{B}$ is:

$$\mathbf{B} \overset{a}{\sim} \mathcal{N}_{p+1}(\boldsymbol{\beta}, (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Sigma} \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1})$$

Where $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$ is the diagonal matrix of error variances.

The Huber-White estimator of this covariance matrix is:

$$\widehat{\text{Var}}(\mathbf{B}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \text{diag}(e_1^2, e_2^2, \ldots, e_n^2) \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}$$

Where e_i are the OLS residuals.

(d) [easy] If we know that the errors $\mathcal{E}_1, \ldots, \mathcal{E}_n$ are independent with expectation zero and variance σ_i^2 which means the errors are "heteroskedastic", what is the asymptotic distribution of $\boldsymbol{B}$ using the Huber-White estimator?

When the errors are independent (not necessarily normal) with heteroskedastic variances, the asymptotic distribution of $\mathbf{B}$ is still:

$$\mathbf{B} \overset{a}{\sim} \mathcal{N}_{p+1}(\boldsymbol{\beta}, (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Sigma} \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1})$$

Where $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$.

The Huber-White estimator remains the same as in part (c):

$$\widehat{\text{Var}}(\mathbf{B}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \text{diag}(e_1^2, e_2^2, \ldots, e_n^2) \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}$$

The key point is that the Huber-White estimator provides consistent standard errors under heteroskedasticity regardless of whether the errors are normal or not. The asymptotic normality of $\mathbf{B}$ follows from the Central Limit Theorem, and the Huber-White adjustment correctly captures the heteroskedastic structure of the errors.

(e) [easy] Is the F-tests we derived under the core assumption valid in any of the four above scenarios? Yes/no

NO

Problem 4

This problem is about inference for the generalized linear model (glm).

(a) [harder] Let $Y_i \overset{ind}{\sim} \text{Bernoulli}(\theta_i)$ for $i = 1, \ldots, n$ where $\theta_i = \phi(\boldsymbol{x}_i\boldsymbol{\beta})$ and $\boldsymbol{x}_i \in \mathbb{R}^{p+1}$ whose first entry is always 1. For the link function, use the complementary log-log (i.e. the standard Gumbel CDF). Write out the full likelihood below. No need to simplify or take the log.

The Bernoulli probability mass function is:

$$f(y_i|\theta_i) = \theta_i^{y_i}(1-\theta_i)^{1-y_i}$$

The complementary log-log link function is:

$$\theta_i = \phi(\mathbf{x}_i\boldsymbol{\beta}) = 1 - \exp\left(() - \exp\left(()\,\mathbf{x}_i\boldsymbol{\beta}\right)\right)$$

Therefore, the likelihood function is:

$$L(\boldsymbol{\beta}; \mathbf{y}) = \prod_{i=1}^{n} f(y_i|\theta_i) \tag{41}$$

$$= \prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{1-y_i} \tag{42}$$

$$\tag{43}$$

(b) [harder] Given the assumptions in (a), write the likelihood ratio estimate for the omnibus test of $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$.

(c) [harder] Let $Y_i \overset{ind}{\sim} \text{Poisson}(\theta_i)$ for $i = 1, \ldots, n$ where $\theta_i = e^{\boldsymbol{x}_i\boldsymbol{\beta}}$ and $\boldsymbol{x}_i \in \mathbb{R}^{p+1}$ whose first entry is always 1. Write out the likelihood ratio when testing $H_0 : \beta_2 = \beta_3 = 0$.

(d) [harder] Let $Y_i \overset{ind}{\sim} \text{Weibull}(k, \theta_i)$ for $i = 1, \ldots, n$ where $\theta_i = e^{\boldsymbol{x}_i\boldsymbol{\beta}}$ and $\boldsymbol{x}_i \in \mathbb{R}^{p+1}$ whose first entry is always 1. This uses the alternate parameterization so that $\mathbb{E}[Y_i] = \theta_i\Gamma(1 + 1/k)$. There is a censoring vector $\boldsymbol{c}$ which is 1 when censored on the right (meaning the real y_i is $\geq$ to the observed y_i) and 0 when not censored. Write out the likelihood ratio when testing $H_0 : \beta_2 = \beta_3 = 0$.

(e) [difficult] [MA] Let $Y_i \overset{ind}{\sim} \mathcal{N}(\theta_i, \sigma^2)$ for $i = 1, \ldots, n$ where $\theta_i = \boldsymbol{x}_i\boldsymbol{\beta}$ and $\boldsymbol{x}_i \in \mathbb{R}^{p+1}$ whose first entry is always 1. So far, this is the vanilla linear model. However, consider now a wrinkle: there is a censoring vector $\boldsymbol{c}$ which is 1 when censored on the right (meaning the real y_i is $\geq$ to the observed y_i) and 0 when not censored. This is called the Tobit model. Write the likelihood ratio estimate for the omnibus test of $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$.