Chapter 13 Solutions

Andrew Wu
Wasserman: All of Statistics

July 29, 2025

Problem 13.1. Prove that the least squares estimates are given by

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^n (X_i - \overline{X}_n)^2}$$
$$\widehat{\beta}_0 = \overline{Y}_n - \widehat{\beta}_1 \overline{X}_n,$$

and that an unbiased estimate of σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \widehat{\epsilon}_i^2.$$

Solution. We need to find $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that minimize

$$\begin{aligned}
\sum_{i=1}^n \widehat{\epsilon}_i^2 &= \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2 \\
&= \sum_{i=1}^n (Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i))^2 \\
&= \sum_{i=1}^n \left[Y_i^2 - 2Y_i(\widehat{\beta}_0 + \widehat{\beta}_1 X_i) + (\widehat{\beta}_0 + \widehat{\beta}_1 X_i)^2 \right] \\
&= \sum_{i=1}^n Y_i^2 - 2\sum_{i=1}^n Y_i \widehat{\beta}_0 - 2\sum_{i=1}^n X_i Y_i \widehat{\beta}_1 + \sum_{i=1}^n (\widehat{\beta}_0^2 + 2\widehat{\beta}_0\widehat{\beta}_1 X_i + \widehat{\beta}_1^2 X_i^2).
\end{aligned}$$

Note that $\sum_{i=1}^n X_i = n\overline{X}$ and $\sum_{i=1}^n Y_i = n\overline{Y}$. Then the right-hand side simplifies as follows:

$$\sum_{i=1}^n Y_i^2 - 2n\widehat{\beta}_0 \overline{Y}_n - 2\widehat{\beta}_1 \sum_{i=1}^n X_i Y_i + n\widehat{\beta}_0^2 + 2n\widehat{\beta}_0\widehat{\beta}_1\overline{X}_n + \widehat{\beta}_1^2 \sum_{i=1}^n X_i^2.$$

Let this expression be S. Then

$$\frac{\partial S}{\partial \widehat{\beta}_1} = -2\sum_{i=1}^n X_i Y_i + 2n\widehat{\beta}_0 \overline{X}_n + 2\widehat{\beta}_1 \sum_{i=1}^n X_i^2$$

and

$$\frac{\partial S}{\partial \widehat{\beta}_0} = -2n\overline{Y}_n + 2n\widehat{\beta}_0 + 2n\widehat{\beta}_1 \overline{X}_n.$$

Setting $\frac{\partial S}{\partial \widehat{\beta}_0} = 0$ yields

$$\widehat{\beta}_0 = \overline{Y}_n - \widehat{\beta}_1 \overline{X}_n$$

and setting $\frac{\partial S}{\partial \widehat{\beta}_1} = 0$ and doing some initial manipulations yields

$$\sum_{i=1}^{n} X_i Y_i = n(\overline{Y}_n - \widehat{\beta}_1 \overline{X}_n)\overline{X}_n + \widehat{\beta}_1 \sum_{i=1}^{n} X_i^2.$$

So

$$\begin{aligned}\sum_{i=1}^{n} X_i Y_i - n\overline{X}_n \overline{Y}_n &= -n\overline{X}_n^2 \widehat{\beta}_1 + \widehat{\beta}_1 \sum_{i=1}^{n} X_i^2 \\ &= \widehat{\beta}_1 \left(-n\overline{X}_n^2 + \sum_{i=1}^{n} X_i^2 \right)\end{aligned}$$

so

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}_n \overline{Y}_n}{\sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2}.$$

But we know that

$$\begin{aligned}\sum_{i=1}^{n} (X_i - \overline{X}_n)^2 &= \sum_{i=1}^{n} X_i^2 - 2n\overline{X}_n^2 + n\overline{X}_n^2 \\ &= \sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2\end{aligned}$$

and

$$\begin{aligned}\sum_{i=1}^{n} (X_i - \overline{X}_n)(Y_i - \overline{Y}_n) &= \sum_{i=1}^{n} X_i Y_i - n\overline{X}_n \overline{Y}_n - n\overline{X}_n \overline{Y}_n + n\overline{X}_n \overline{Y}_n \\ &= \sum_{i=1}^{n} X_i Y_i - n\overline{X}_n \overline{Y}_n\end{aligned}$$

where we used that $n\overline{X}_n = \sum_{i=1}^{n} X_i$ and $n\overline{Y}_n = \sum_{i=1}^{n} Y_i$.

Now, to prove $\widehat{\sigma}^2$ is an unbiased estimator, we want to show that $\mathbb{E}(\widehat{\sigma}^2) = \sigma^2$.

We can write $Y_i = \begin{bmatrix} 1 & X_i \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \epsilon_i$. Therefore we can write

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

where

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Let $H = X(X^T X)^{-1} X^T$ be the hat matrix, the matrix which maps $\mathbf{Y}$ to $\mathbf{X}\widehat{\beta}$. Then the vector of residuals is

$$\widehat{\epsilon} = \mathbf{Y} - H\mathbf{Y} = (I - H)\mathbf{Y}.$$

Note that the residual sum of squares $\sum_{i=1}^{n} \widehat{\epsilon}_i^2$ can also be computed as

$$\begin{bmatrix} \widehat{\epsilon}_1 & \ldots \widehat{\epsilon}_n \end{bmatrix} \begin{bmatrix} \widehat{\epsilon}_1 \\ \vdots \\ \widehat{\epsilon}_n \end{bmatrix} = \widehat{\epsilon}^T \widehat{\epsilon}$$

so it follows that

$$\begin{aligned}
\sum_{i=1}^n \widehat{\epsilon}_i^2 &= \widehat{\epsilon}^T \widehat{\epsilon} \\
&= ((I-H)\mathbf{Y})^T((I-H)\mathbf{Y}) \\
&= \mathbf{Y}^T(I-H)^T(I-H)\mathbf{Y}.
\end{aligned}$$

Now we note two facts: firstly, $I-H$ is symmetric and idempotent. That $I-H$ is symmetric follows from the fact that H, being an orthogonal projection matrix, must be symmetric; also, $(I-H)^2 = I^2 - HI - IH + H^2 = I - 2H + H = I - H$, as orthogonal projection matrices are idempotent.

Next, we note that $(I-H)X = 0$. This follows from the fact that $HX = X$, as H is a matrix that projects onto the column space of X.

That means that

$$\begin{aligned}
\sum_{i=1}^n \widehat{\epsilon}_i^2 &= \mathbf{Y}^T(I-H)\mathbf{Y} \\
&= (\mathbf{X}\beta+\epsilon)^T(I-H)(\mathbf{X}\beta+\epsilon) \\
&= ((I-H)(\mathbf{X}\beta+\epsilon))^T(\mathbf{X}\beta+\epsilon) \\
&= ((I-H)\epsilon)^T(\mathbf{X}\beta+\epsilon) \\
&= \epsilon^T(I-H)(\mathbf{X}\beta+\epsilon) \\
&= \epsilon^T(I-H)\epsilon.
\end{aligned}$$

Thus $\mathbb{E}\left(\sum_{i=1}^n \widehat{\epsilon}_i^2\right) = (n-2)\sigma^2$, as $\operatorname{tr}(I-H) = n-2$, and the result follows. □

Problem 13.2. Let $\widehat{\beta} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix}$ denote the least squares estimators.

Prove that

$$\begin{aligned}
\mathbb{E}(\widehat{\beta}|X^n) &= \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \\
\mathbb{V}(\widehat{\beta}|X^n) &= \frac{\sigma^2}{n s_X^2} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n X_i^2 & -\overline{X}_n \\ -\overline{X}_n & 1 \end{bmatrix}
\end{aligned}$$

with $s_X^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$.

Moreover, show that

$$\begin{aligned}
\widehat{\operatorname{se}}(\widehat{\beta}_0) &= \frac{\widehat{\sigma}}{s_X\sqrt{n}}\sqrt{\frac{\sum_{i=1}^n X_i^2}{n}} \\
\widehat{\operatorname{se}}(\widehat{\beta}_1) &= \frac{\widehat{\sigma}}{s_X\sqrt{n}}.
\end{aligned}$$

Solution. In general throughout this solution, assume expectations and variances condition on X^n.

We know that

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^n (X_i - \overline{X}_n)^2}.$$

Observe that the numerator comes out as

$$\begin{aligned}
\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n) &= \sum_{i=1}^n (X_i - \overline{X}_n)Y_i - \overline{Y}_n \sum_{i=1}^n (X_i - \overline{X}_n) \\
&= \sum_{i=1}^n (X_i - \overline{X}_n)Y_i - \overline{Y}_n(-n\overline{X}_n + \sum_{i=1}^n X_i) \\
&= \sum_{i=1}^n (X_i - \overline{X}_n)Y_i.
\end{aligned}$$

With this simplification and $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, we obtain

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n) &= \sum_{i=1}^{n}(X_i - \overline{X}_n)(\beta_0 + \beta_1 X_i + \epsilon_i) \\
&= \sum_{i=1}^{n}(X_i - \overline{X}_n)\beta_0 + \sum_{i=1}^{n}(X_i - \overline{X}_n)(\beta_1 X_i) + \sum_{i=1}^{n}(X_i - \overline{X}_n)\epsilon_i \\
&= \beta_1 \sum_{i=1}^{n}(X_i - \overline{X}_n)X_i + \sum_{i=1}^{n}(X_i - \overline{X}_n)\epsilon_i
\end{aligned}$$

Thus

$$\begin{aligned}
\mathbb{E}(\widehat{\beta}_1) &= \mathbb{E}\left(\frac{\beta_1 \sum_{i=1}^{n}(X_i - \overline{X}_n)X_i + \sum_{i=1}^{n}(X_i - \overline{X}_n)\epsilon_i}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}\right) \\
&= \mathbb{E}\left(\frac{\beta_1 \sum_{i=1}^{n}(X_i - \overline{X}_n)X_i}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}\right) + \mathbb{E}\left(\frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)\epsilon_i}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}\right) \\
&= \mathbb{E}\left(\frac{\beta_1 \sum_{i=1}^{n}(X_i - \overline{X}_n)X_i}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}\right)
\end{aligned}$$

because we are to treat the X_is as constants (we're conditioning on $X^n = (X_1, \ldots, X_n)$) and because $\mathbb{E}(\epsilon_i) = 0$.

Next, we have

$$\begin{aligned}
\mathbb{E}\left(\frac{\beta_1 \sum_{i=1}^{n}(X_i - \overline{X}_n)X_i}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}\right) &= \beta_1 \mathbb{E}\left(\frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)X_i}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}\right) \\
&= \beta_1 \mathbb{E}\left(\frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}\right) \\
&= \beta_1
\end{aligned}$$

because we know that $\sum_{i=1}^{n}(X_i - \overline{X}_n)\overline{X}_n = 0$.

Then,

$$\begin{aligned}
\mathbb{E}(\widehat{\beta}_0) &= \mathbb{E}(\overline{Y}_n - \widehat{\beta}_1 \overline{X}_n) \\
&= \mathbb{E}(\overline{Y}_n) - \beta_1 \overline{X}_n \\
&= \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} Y_i\right) - \beta_1 \overline{X}_n \\
&= \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n}(\beta_0 + \beta_1 X_i + \epsilon_i)\right) - \beta_1 \overline{X}_n.
\end{aligned}$$

But

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{n}(\beta_0 + \beta_1 X_i + \epsilon_i)\right) &= \mathbb{E}\left(n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} \epsilon_i\right) \\
&= n\beta_0 + \beta_1 n \overline{X}_n
\end{aligned}$$

where we again use that $\mathbb{E}(\epsilon_i) = 0$.

It follows that

$$\begin{aligned}
\mathbb{E}(\widehat{\beta}_0) &= \frac{1}{n}(n\beta_0 + \beta_1 n \overline{X}_n) - \beta_1 \overline{X}_n \\
&= \beta_0
\end{aligned}$$

as desired.

Now, to compute $\mathbb{V}(\widehat{\beta})$, we'll compute $\mathbb{V}(\widehat{\beta}_1), \mathbb{V}(\widehat{\beta}_0)$, and $\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1)$.

We have

$$\begin{aligned}
\mathbb{V}(\widehat{\beta}_1) &= \mathbb{V}\left(\frac{\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^n (X_i - \overline{X}_n)^2}\right) \\
&= \frac{1}{\left(\sum_{i=1}^n (X_i - \overline{X}_n)^2\right)^2} \mathbb{V}\left(\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)\right) \\
&= \frac{1}{(ns_X^2)^2} \mathbb{V}\left(\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)\right).
\end{aligned}$$

Now we have

$$\begin{aligned}
\mathbb{V}\left(\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)\right) &= \sum_{i=1}^n \mathbb{V}\left[(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)\right] \\
&= \sum_{i=1}^n \left((X_i - \overline{X}_n)^2 \mathbb{V}(Y_i - \overline{Y}_n)\right).
\end{aligned}$$

Let's deal with the $\mathbb{V}(Y_i - \overline{Y}_n)$ term. We will compute $\mathbb{V}(Y_1 - \overline{Y}_n)$, which we'll extend to the other cases, without loss of generality. Note also that the variances $\mathbb{V}(Y_i)$ are equal for all i.

$$\begin{aligned}
\mathbb{V}(Y_1 - \overline{Y}_n) &= \mathbb{V}\left(Y_1 - \frac{1}{n}\sum_{i=1}^n Y_i\right) \\
&= \mathbb{V}\left(\frac{n-1}{n} Y_1 - \frac{1}{n}\sum_{i=2}^n Y_i\right) \\
&= \left(\frac{n-1}{n}\right)^2 \mathbb{V}(Y_1) + \frac{1}{n^2}\mathbb{V}\left(\sum_{i=2}^n Y_i\right) \\
&= \left(\frac{n-1}{n}\right)^2 \mathbb{V}(Y_1) + \frac{1}{n^2}\sum_{i=2}^n \mathbb{V}(Y_i) \\
&= \left(\left(\frac{n-1}{n}\right)^2 + \frac{n-1}{n^2}\right) \mathbb{V}(Y_1) \\
&= \frac{n-1}{n}\mathbb{V}(Y_1).
\end{aligned}$$

So $\mathbb{V}(Y_i - \overline{Y}_n) = \frac{n-1}{n}\mathbb{V}(Y_i)$. But we also have

$$\begin{aligned}
\mathbb{V}(Y_i) &= \mathbb{V}(\beta_0 + \beta_1 X_i + \epsilon_i) \\
&= \mathbb{V}(\epsilon_i) = \sigma^2
\end{aligned}$$

because we are treating X_is as constants. It follows that

$$\begin{aligned}
\sum_{i=1}^n \left((X_i - \overline{X}_n)^2 \mathbb{V}(Y_i - \overline{Y}_n)\right) &= \sum_{i=1}^n (X_i - \overline{X}_n)^2 \sigma^2 \\
&= \sigma^2 n s_X^2
\end{aligned}$$

so $\mathbb{V}(\widehat{\beta}_1) = \frac{\sigma^2}{ns_X^2}$.

Next, we have

$$\begin{aligned}
\mathbb{V}(\widehat{\beta}_0) &= \mathbb{V}(\overline{Y}_n - \widehat{\beta}_1 \overline{X}_n) \\
&= \mathbb{V}(\overline{Y}_n) + (\overline{X}_n)^2 \mathbb{V}(\widehat{\beta}_1) - 2\overline{X}_n \mathrm{Cov}(\overline{Y}_n, \widehat{\beta}_1).
\end{aligned}$$

But

$$
\begin{aligned}
\mathbb{V}(\overline{Y}_n) &= \frac{1}{n^2}\mathbb{V}\left(\sum_{i=1}^n Y_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}\left(\beta_0+\beta_1 X_i+\epsilon_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}(\epsilon_i) \\
&= \frac{\sigma^2}{n}.
\end{aligned}
$$

Moreover, we previously computed $\mathbb{V}(\widehat{\beta}_1) = \frac{\sigma^2}{ns_X^2}$. Thus all that remains is to compute $\mathrm{Cov}(\overline{Y}_n, \widehat{\beta}_1)$.

We have

$$
\begin{aligned}
\mathrm{Cov}(\overline{Y}_n, \widehat{\beta}_1) &= \mathbb{E}\left((\overline{Y}_n - \mathbb{E}(\overline{Y}_n))(\widehat{\beta}_1 - \mathbb{E}(\widehat{\beta}_1))\right) \\
&= \mathbb{E}\left(\left(\overline{Y}_n - \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n(\beta_0+\beta_1X_i+\epsilon_i)\right)\right)(\widehat{\beta}_1-\beta_1)\right) \\
&= \mathbb{E}\left(\left(\overline{Y}_n - \frac{1}{n}\sum_{i=1}^n(\beta_0+\beta_1X_i)\right)\left(\widehat{\beta}_1-\beta_1\right)\right).
\end{aligned}
$$

But we have

$$
\begin{aligned}
\overline{Y}_n - \frac{1}{n}\sum_{i=1}^n(\beta_0+\beta_1X_i) &= \frac{1}{n}\sum_{i=1}^n(\beta_0+\beta_1X_i+\epsilon_i) - \frac{1}{n}\sum_{i=1}^n(\beta_0+\beta_1X_i) \\
&= \frac{1}{n}\sum_{i=1}^n \epsilon_i.
\end{aligned}
$$

Thus it follows that

$$
\begin{aligned}
\mathbb{E}\left(\left(\overline{Y}_n - \frac{1}{n}\sum_{i=1}^n(\beta_0+\beta_1X_i)\right)\left(\widehat{\beta}_1-\beta_1\right)\right) &= \frac{1}{n}\mathbb{E}\left(\left(\sum_{i=1}^n \epsilon_i\right)\left(\widehat{\beta}_1-\beta_1\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\epsilon_i\cdot(\widehat{\beta}_1-\beta_1)\right).
\end{aligned}
$$

Using our previous work dealing with the numerator of the expression for $\widehat{\beta}_1$, we obtain

$$
\begin{aligned}
\mathbb{E}(\epsilon_1(\widehat{\beta}_1-\beta_1)) &= \mathbb{E}\left(\epsilon_1\left(\frac{\beta_1\sum_{i=1}^n(X_i-\overline{X}_n)X_i + \sum_{i=1}^n(X_i-\overline{X}_n)\epsilon_i}{\sum_{i=1}^n(X_i-\overline{X}_n)^2} - \beta_1\right)\right) \\
&= \mathbb{E}\left(\epsilon_1\left(\frac{\beta_1\sum_{i=1}^n(X_i-\overline{X}_n)X_i + \sum_{i=1}^n(X_i-\overline{X}_n)\epsilon_i}{\sum_{i=1}^n(X_i-\overline{X}_n)X_i} - \beta_1\right)\right) \\
&= \mathbb{E}\left(\epsilon_1\left(\frac{\sum_{i=1}^n(X_i-\overline{X}_n)\epsilon_i}{\sum_{i=1}^n(X_i-\overline{X}_n)X_i}\right)\right) \\
&= \frac{1}{ns_X^2}\mathbb{E}\left(\epsilon_1\left(\sum_{i=1}^n(X_i-\overline{X}_n)\epsilon_i\right)\right)
\end{aligned}
$$

Note that $\mathbb{E}(\epsilon_1^2) = \mathbb{V}(\epsilon_1)+\mathbb{E}(\epsilon_1)^2$, so $\mathbb{E}(\epsilon_1^2)=\sigma^2$. Moreover, when $i \neq j$, we have $\mathbb{E}(\epsilon_i\epsilon_j) = \mathbb{E}(\epsilon_i)\mathbb{E}(\epsilon_j) = 0$, as ϵ_i and ϵ_j are independent. It follows that

$$\begin{aligned}
\frac{1}{ns_X^2}\mathbb{E}\left(\epsilon_1\left(\sum_{i=1}^{n}(X_i-\overline{X}_n)\epsilon_i\right)\right) &= \frac{1}{ns_X^2}\mathbb{E}\left(\sum_{i=1}^{n}\epsilon_1\epsilon_i(X_i-\overline{X}_n)\right)\\
&= \frac{1}{ns_X^2}\sum_{i=1}^{n}\mathbb{E}(\epsilon_1\epsilon_i(X_i-\overline{X}_n))\\
&= \frac{1}{ns_X^2}\sum_{i=1}^{n}\left(\mathbb{E}(\epsilon_1\epsilon_i X_i)-\mathbb{E}(\epsilon_1\epsilon_i\overline{X}_n)\right)\\
&= \frac{1}{ns_X^2}\left(\mathbb{E}(\epsilon_1^2 X_i)-\mathbb{E}(\epsilon_1^2\overline{X}_n)\right)\\
&= \frac{1}{ns_X^2}\left(\sigma^2(X_i-\overline{X}_n)\right).
\end{aligned}$$

Therefore

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\epsilon_i\cdot(\widehat{\beta}_1-\beta_1)\right) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{ns_X^2}\left(\sigma^2(X_i-\overline{X}_n)\right)\\
&= \frac{\sigma^2}{n^2 s_X^2}\sum_{i=1}^{n}(X_i-\overline{X}_n)=0.
\end{aligned}$$

So we obtain, in the end,

$$\begin{aligned}
\mathbb{V}(\widehat{\beta}_0) &= \mathbb{V}(\overline{Y}_n)+(\overline{X}_n)^2\mathbb{V}(\widehat{\beta}_1)\\
&= \frac{\sigma^2}{n}+(\overline{X}_n)^2\frac{\sigma^2}{ns_X^2}\\
&= \frac{\sigma^2}{ns_X^2}\left(s_X^2+(\overline{X}_n)^2\right)\\
&= \frac{\sigma^2}{ns_X^2}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2+(\overline{X}_n)^2\right)\\
&= \frac{\sigma^2}{ns_X^2}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2-\frac{1}{n}\sum_{i=1}^{n}2X_i\overline{X}_n+\frac{1}{n}\sum_{i=1}^{n}\overline{X}_n^2+\overline{X}_n^2\right)\\
&= \frac{\sigma^2}{ns_X^2}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2-\frac{1}{n}\overline{X}_n\cdot 2n\overline{X}_n+2\overline{X}_n^2\right)\\
&= \frac{\sigma^2}{ns_X^2}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right).
\end{aligned}$$

Now we'll compute $\mathrm{Cov}(\widehat{\beta}_0,\widehat{\beta}_1)$. We have

$$\begin{aligned}
\mathrm{Cov}(\widehat{\beta}_0,\widehat{\beta}_1) &= \mathbb{E}\left((\widehat{\beta}_0-\mathbb{E}(\widehat{\beta}_0))(\widehat{\beta}_1-\mathbb{E}(\widehat{\beta}_1))\right)\\
&= \mathbb{E}\left((\widehat{\beta}_0-\beta_0)(\widehat{\beta}_1-\beta_1)\right)\\
&= \mathbb{E}(\widehat{\beta}_0\widehat{\beta}_1)-\beta_0\beta_1.
\end{aligned}$$

We have

$$\begin{aligned}
\widehat{\beta}_0\widehat{\beta}_1 &= \left(\overline{Y}_n-\widehat{\beta}_1\overline{X}_n\right)\left(\widehat{\beta}_1\right)\\
&= \overline{Y}_n\widehat{\beta}_1-\overline{X}_n\widehat{\beta}_1^2
\end{aligned}$$

so $\mathbb{E}(\widehat{\beta}_0\widehat{\beta}_1) = \mathbb{E}(\overline{Y}_n\widehat{\beta}_1) - \mathbb{E}(\overline{X}_n\widehat{\beta}_1^2)$. But

$$\begin{aligned}
\mathbb{E}(\overline{X}_n\widehat{\beta}_1^2) &= \overline{X}_n\mathbb{E}(\widehat{\beta}_1^2) \\
&= \overline{X}_n\left(\mathbb{V}(\widehat{\beta}_1) + \mathbb{E}(\widehat{\beta}_1)^2\right) \\
&= \overline{X}_n\left(\frac{\sigma^2}{ns_X^2} + \beta_1^2\right).
\end{aligned}$$

Next,

$$\begin{aligned}
\overline{Y}_n\widehat{\beta}_1 &= \frac{1}{n}\sum_{i=1}^n Y_i \cdot \widehat{\beta}_1 \\
&= \frac{1}{n}\sum_{i=1}^n (\beta_0 + \beta_1 X_i + \epsilon_i) \cdot \widehat{\beta}_1 \\
&= \widehat{\beta}_1\left(\beta_0 + \beta_1\overline{X}_n + \frac{1}{n}\sum_{i=1}^n \epsilon_i\right)
\end{aligned}$$

so we have

$$\begin{aligned}
\mathbb{E}(\overline{Y}_n\widehat{\beta}_1) &= \mathbb{E}\left(\widehat{\beta}_1\left(\beta_0 + \beta_1\overline{X}_n + \frac{1}{n}\sum_{i=1}^n \epsilon_i\right)\right) \\
&= \mathbb{E}(\widehat{\beta}_1\beta_0) + \mathbb{E}(\widehat{\beta}_1\beta_1\overline{X}_n) + \mathbb{E}\left(\widehat{\beta}_1 \cdot \frac{1}{n}\sum_{i=1}^n \epsilon_i\right) \\
&= \beta_0\beta_1 + \beta_1^2\overline{X}_n + \frac{1}{n}\sum_{i=1}^n \mathbb{E}(\widehat{\beta}_1\epsilon_i).
\end{aligned}$$

But we computed previously that $\frac{1}{n}\sum_{i=1}^n \mathbb{E}(\widehat{\beta}_1\epsilon_i) = 0$ in our calculation of $\mathrm{Cov}(\overline{Y}_n, \widehat{\beta}_1)$. Thus it follows that

$$\begin{aligned}
\mathbb{E}(\widehat{\beta}_0\widehat{\beta}_1) &= \beta_0\beta_1 + \beta_1^2\overline{X}_n - \overline{X}_n\left(\frac{\sigma^2}{ns_X^2} + \beta_1^2\right) \\
&= \beta_0\beta_1 - \overline{X}_n\frac{\sigma^2}{ns_X^2}
\end{aligned}$$

and thus $\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = -\overline{X}_n \cdot \frac{\sigma^2}{ns_X^2}$. The variance-covariance matrix thus is exactly as stated in the problem.

Then, the estimated standard errors $\widehat{\mathrm{se}}(\widehat{\beta}_0)$ and $\widehat{\mathrm{se}}(\widehat{\beta}_1)$ are the square roots of the diagonal terms in $\mathbb{V}(\widehat{\beta}|X^n)$, with σ replaced by $\widehat{\sigma}$. So

$$\begin{aligned}
\widehat{\mathrm{se}}(\widehat{\beta}_0) &= \frac{\widehat{\sigma}}{s_X\sqrt{n}}\sqrt{\frac{\sum_{i=1}^n X_i^2}{n}} \\
\widehat{\mathrm{se}}(\widehat{\beta}_1) &= \frac{\widehat{\sigma}}{s_X\sqrt{n}}.
\end{aligned}$$

□

Problem 11.3. Consider the regression through the origin model: $Y_i = \beta X_i + \epsilon$. Find the least squares estimate for β. Find the standard error of the estimate. Find conditions that guarantee that the estimate is consistent.

Solution. The least squares estimate must minimize the RSS, $\sum_{i=1}^n \widehat{\epsilon}_i^2$. We'll expand this expression in terms of an estimator $\widehat{\beta}$ and then take the derivative with respect to $\widehat{\beta}$.

We have

$$\begin{aligned}
\sum_{i=1}^{n} \widehat{\epsilon}_i^2 &= \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 \\
&= \sum_{i=1}^{n} (Y_i - (\widehat{\beta} X_i))^2 \\
&= \sum_{i=1}^{n} \left[Y_i^2 - 2\widehat{\beta} X_i Y_i + (\widehat{\beta} X_i)^2 \right] \\
&= \sum_{i=1}^{n} Y_i^2 - \widehat{\beta} \sum_{i=1}^{n} 2 X_i Y_i + \widehat{\beta}^2 \sum_{i=1}^{n} X_i^2
\end{aligned}$$

and now we take the derivative with respect to $\widehat{\beta}$. So it follows that

$$\frac{d}{d\widehat{\beta}} \sum_{i=1}^{n} \widehat{\epsilon}_i^2 = -\sum_{i=1}^{n} 2 X_i Y_i + 2\widehat{\beta} \sum_{i=1}^{n} X_i^2.$$

Setting this equal to 0, we obtain

$$\sum_{i=1}^{n} X_i Y_i = \widehat{\beta} \sum_{i=1}^{n} X_i^2$$

and thus

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}.$$

The standard error of $\widehat{\beta}$ is $\sqrt{\mathbb{V}(\widehat{\beta})}$. Here, as usual, we condition on a fixed set of X_is, and we let $\mathbb{V}(\epsilon_i) = \sigma^2$. We have

$$\begin{aligned}
\widehat{\beta} &= \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \\
&= \frac{\sum_{i=1}^{n} X_i (\beta X_i + \epsilon_i)}{\sum_{i=1}^{n} X_i^2}
\end{aligned}$$

so

$$\begin{aligned}
\mathbb{V}(\widehat{\beta}) &= \mathbb{V}\left(\frac{\sum_{i=1}^{n} X_i (\beta X_i + \epsilon_i)}{\sum_{i=1}^{n} X_i^2} \right) \\
&= \frac{1}{\left(\sum_{i=1}^{n} X_i^2\right)^2} \left(\sum_{i=1}^{n} \mathbb{V}(X_i(\beta X_i + \epsilon_i)) \right) \\
&= \frac{1}{\left(\sum_{i=1}^{n} X_i^2\right)^2} \left(\sum_{i=1}^{n} X_i^2 \mathbb{V}(\epsilon_i) \right) \\
&= \frac{\sigma^2}{\sum_{i=1}^{n} X_i^2}.
\end{aligned}$$

Thus the standard error is

$$\frac{\sigma}{\sqrt{\sum_{i=1}^{n} X_i^2}}.$$

Now we need to find the conditions under which $\widehat{\beta}$ is consistent; that is, the conditions under which

$\widehat{\beta} \xrightarrow{P} \beta$ holds. Note that

$$\begin{aligned}
\widehat{\beta} &= \frac{\sum_{i=1}^n X_i(\beta X_i + \epsilon_i)}{\sum_{i=1}^n X_i^2} \\
&= \frac{\beta \sum_{i=1}^n X_i^2 + \sum_{i=1}^n X_i \epsilon_i}{\sum_{i=1}^n X_i^2} \\
&= \beta + \frac{\sum_{i=1}^n X_i \epsilon_i}{\sum_{i=1}^n X_i^2}
\end{aligned}$$

so we need conditions under which

$$\mathbb{P}\left(\left| \frac{\sum_{i=1}^n X_i \epsilon_i}{\sum_{i=1}^n X_i^2} \right| > \epsilon \right) \to 0$$

as $n \to \infty$.

Assume the conditions necessary to use the Law of Large Numbers. Note that

$$\left| \frac{\sum_{i=1}^n X_i \epsilon_i}{\sum_{i=1}^n X_i^2} \right| = \left| \frac{\frac{1}{n} \sum_{i=1}^n X_i \epsilon_i}{\frac{1}{n} \sum_{i=1}^n X_i^2} \right|$$

so we want the following conditions: $\mathbb{E}(\epsilon_i) = 0$ forces the numerator to converge to 0, and $\mathbb{E}(X_i^2) = c > 0$ forces the denominator to converge to something nonzero. Note that $\mathbb{E}(X_i^2) > 0$ follows from $\mathbb{V}(X_i) > 0$, as $\mathbb{V}(X_i) = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2$, so we want $\mathbb{E}(\epsilon_i) = 0$ and $\mathbb{V}(X_i) > 0$. □

Problem 11.4. Prove that

$$\text{bias}(\widehat{R}_{tr}(S)) = \mathbb{E}(\widehat{R}_{tr}(S)) - R(S) = -2 \sum_{i=1}^n \text{Cov}(\widehat{Y}_i, Y_i)$$

where $\widehat{R}_{tr}(S) = \sum_{i=1}^n (\widehat{Y}_i(S) - Y_i)^2$ is the training error and $R(S) = \sum_{i=1}^n \mathbb{E}((\widehat{Y}_i(S) - Y_i^*)^2)$ is the prediction risk.

Solution. We have

$$\begin{aligned}
\mathbb{E}(\widehat{R}_{tr}(S)) &= \mathbb{E}\left(\sum_{i=1}^n (\widehat{Y}_i(S) - Y_i)^2 \right) \\
&= \sum_{i=1}^n \mathbb{E}((\widehat{Y}_i(S) - Y_i)^2)
\end{aligned}$$

so

$$\begin{aligned}
\mathbb{E}(\widehat{R}_{tr}(S)) - R(S) &= \sum_{i=1}^n \mathbb{E}((\widehat{Y}_i(S) - Y_i)^2) - \sum_{i=1}^n \mathbb{E}((\widehat{Y}_i(S) - Y_i^*)^2) \\
&= \sum_{i=1}^n \left[\mathbb{E}((\widehat{Y}_i(S) - Y_i)^2) - \mathbb{E}((\widehat{Y}_i(S) - Y_i^*)^2) \right] \\
&= \sum_{i=1}^n \left[\mathbb{E}(\widehat{Y}_i(S)^2 - 2Y_i\widehat{Y}_i(S) + Y_i^2) - \mathbb{E}(\widehat{Y}_i(S)^2 - 2Y_i^*\widehat{Y}_i(S) + (Y_i^*)^2) \right] \\
&= \sum_{i=1}^n \left[\mathbb{E}(Y_i^2) + \mathbb{E}(2Y_i^*\widehat{Y}_i(S)) - \mathbb{E}(2Y_i\widehat{Y}_i(S)) - \mathbb{E}((Y_i^*)^2) \right].
\end{aligned}$$

Here we note that $Y_i = X_i^T\beta + \epsilon_i$ and $Y_i^* = X_i^T\beta + \epsilon_i^*$, where ϵ_i and ϵ_i^* are both draws from $N(0, \sigma^2)$. Thus Y_i and Y_i^* are draws from the same distribution, and $\mathbb{E}(Y_i^2) = \mathbb{E}((Y_i^*)^2)$.

Thus

$$\begin{aligned}\mathbb{E}(\widehat{R}_{tr}(S)) - R(S) &= \sum_{i=1}^{n}\left[\mathbb{E}(2Y_i^*\widehat{Y}_i(S)) - \mathbb{E}(2Y_i\widehat{Y}_i(S))\right]\\ &= -2\sum_{i=1}^{n}\left[\mathbb{E}(Y_i\widehat{Y}_i(S)) - \mathbb{E}(Y_i^*\widehat{Y}_i(S))\right]\\ &= -2\sum_{i=1}^{n}\left[\mathbb{E}(Y_i\widehat{Y}_i(S)) - \mathbb{E}(Y_i^*)\mathbb{E}(\widehat{Y}_i(S))\right]\end{aligned}$$

where the last step follows from the fact that Y_i^* and $\widehat{Y}_i(S)$ are independent, as Y_i^* is some future draw and $\widehat{Y}_i(S)$ is built from the observed values.

But we have

$$\mathrm{Cov}(\widehat{Y}_i, Y_i) = \mathbb{E}(\widehat{Y}_iY_i) - \mathbb{E}(\widehat{Y}_i)\mathbb{E}(Y_i)$$

so

$$-2\sum_{i=1}^{k}\mathrm{Cov}(\widehat{Y}_i, Y_i) = -2\sum_{i=1}^{k}\left[\mathbb{E}(\widehat{Y}_iY_i) - \mathbb{E}(\widehat{Y}_i)\mathbb{E}(Y_i)\right]$$

so using $\mathbb{E}(Y_i) = \mathbb{E}(Y_i^*)$, we may conclude. □

Problem 13.5. In the simple linear regression model, construct a Wald test for $H_0 : \beta_1 = 17\beta_0$ versus $H_1 : \beta_1 \neq 17\beta_0$.

Solution. Let $\gamma = \beta_1 - 17\beta_0$. We want to test $H_0 : \gamma = 0$ versus $H_1 : \gamma \neq 0$.

We have $\widehat{\gamma} = \widehat{\beta}_1 - 17\widehat{\beta}_0$ so

$$\begin{aligned}\mathbb{V}(\widehat{\gamma}) &= \mathbb{V}(\widehat{\beta}_1 - 17\widehat{\beta}_0)\\ &= \mathbb{V}(\widehat{\beta}_1 - 17(\overline{Y}_n - \widehat{\beta}_1\overline{X}_n))\\ &= \mathbb{V}(\widehat{\beta}_1 + 17\widehat{\beta}_1\overline{X}_n - 17\overline{Y}_n).\end{aligned}$$

But

$$\begin{aligned}\overline{Y}_n &= \frac{1}{n}\sum_{i=1}^{n}Y_i\\ &= \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1X_i + \epsilon_i)\\ &= \beta_0 + \beta_1\overline{X}_n + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i.\end{aligned}$$

So

$$\begin{aligned}\mathbb{V}(\widehat{\beta}_1 + 17\widehat{\beta}_1\overline{X}_n - 17\overline{Y}_n) &= \mathbb{V}\left(\widehat{\beta}_1(1 + 17\overline{X}_n) - 17\left(\beta_0 + \beta_1\overline{X}_n + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right)\right)\\ &= \mathbb{V}\left(\widehat{\beta}_1(1 + 17\overline{X}_n) - \frac{17}{n}\sum_{i=1}^{n}\epsilon_i\right)\\ &= \mathbb{V}(\widehat{\beta}_1(1 + 17\overline{X}_n)) + \mathbb{V}\left(\frac{17}{n}\sum_{i=1}^{n}\epsilon_i\right) - 2\mathrm{Cov}\left(\widehat{\beta}_1(1 + 17\overline{X}_n), \frac{17}{n}\sum_{i=1}^{n}\epsilon_i\right).\end{aligned}$$

The first term we can compute as follows:

$$\mathbb{V}(\widehat{\beta}_1(1+17\overline{X}_n)) = (1+17\overline{X}_n)^2\frac{\sigma^2}{ns_X^2}.$$

The second term:

$$\begin{aligned}
\mathbb{V}\left(\frac{17}{n}\sum_{i=1}^n \epsilon_i\right) &= \left(\frac{17}{n}\right)^2 \mathbb{V}\left(\sum_{i=1}^n \epsilon_i\right) \\
&= \left(\frac{17}{n}\right)^2 \sum_{i=1}^n \mathbb{V}(\epsilon_i) \\
&= \left(\frac{17}{n}\right)^2 \cdot n\sigma^2 \\
&= \frac{289\sigma^2}{n}.
\end{aligned}$$

Finally, we'll compute $\mathrm{Cov}\left(\widehat{\beta}_1(1+17\overline{X}_n), \frac{17}{n}\sum_{i=1}^n \epsilon_i\right)$. Note that $\mathrm{Cov}\left(\widehat{\beta}_1(1+17\overline{X}_n), \frac{17}{n}\sum_{i=1}^n \epsilon_i\right) = \frac{17}{n}(1+17\overline{X}_n)\mathrm{Cov}(\widehat{\beta}_1, \sum_{i=1}^n \epsilon_i)$. We have

$$\begin{aligned}
\mathrm{Cov}\left(\widehat{\beta}_1, \sum_{i=1}^n \epsilon_i\right) &= \sum_{i=1}^n \mathrm{Cov}(\widehat{\beta}_1, \epsilon_i) \\
&= \sum_{i=1}^n \left[\mathbb{E}(\widehat{\beta}_1\epsilon_i) - \mathbb{E}(\widehat{\beta}_1)\mathbb{E}(\epsilon_i)\right] \\
&= \sum_{i=1}^n \mathbb{E}(\widehat{\beta}_1\epsilon_i).
\end{aligned}$$

But this comes out to 0; see our intermediate steps from problem 13.2. Thus it follows that

$$\widehat{\mathrm{se}}(\widehat{\gamma}) = \sqrt{(1+17\overline{X}_n)^2\frac{\sigma^2}{ns_X^2} + \frac{289\sigma^2}{n}}$$

and the Wald test statistic is

$$W = \frac{\widehat{\gamma}}{\widehat{\mathrm{se}}(\widehat{\gamma})}.$$

We reject H_0 when $|W| > z_{\alpha/2}$. □

Problem 13.8. Assume a linear regression model with Normal errors. Take σ known. Show that the model with the highest AIC is the model with the lowest Mallows C_p statistic.

Solution. Suppose there are k possible covariates and n observations. The AIC is $\ell_S - |S|$, and we have

$$\begin{aligned}
\ell_S &= \log\left(\prod_{i=1}^n f_{Y|X}(Y_i|X_i)\right) \\
&= \sum_{i=1}^n \log f_{Y|X}(Y_i|X_i).
\end{aligned}$$

But, letting f be the distribution function for $N(0, \sigma^2)$, we have

$$\begin{aligned}
\sum_{i=1}^{n} \log f_{Y|X}(Y_i|X_i) &= \sum_{i=1}^{n} \log f(\widehat{\epsilon}_i) \\
&= \sum_{i=1}^{n} \log\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\widehat{\epsilon}_i^2}{2\sigma^2}\right)\right) \\
&= n \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \sum_{i=1}^{n}\left(\frac{\widehat{\epsilon}_i^2}{2\sigma^2}\right) \\
&= -n\log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\widehat{\epsilon}_i^2.
\end{aligned}$$

Thus $\ell_S - |S| = -n\log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\widehat{\epsilon}_i^2 - |S|$. It follows that maximizing the AIC is equivalent to minimizing $\frac{1}{2\sigma^2}\sum_{i=1}^{n}\widehat{\epsilon}_i^2 + |S|$ across possible choices of the covariates.

Mallow's C_p statistic is

$$\begin{aligned}
\widehat{R}(S) &= \widehat{R}_{tr}(S) + 2|S|\widehat{\sigma}^2 \\
&= \sum_{i=1}^{n}(\widehat{Y}_i(S) - Y_i)^2 + 2|S|\widehat{\sigma}^2 \\
&= \sum_{i=1}^{n}\widehat{\epsilon}_i^2 + 2|S|\widehat{\sigma}^2.
\end{aligned}$$

But with σ known, this is just $2\sigma^2$ multiplied by the expression we wanted to minimize regarding the AIC, and we are done. $\square$

Problem 13.9. Let $X_1, \ldots, X_n$ be IID observations. Consider two models $\mathcal{M}_0, \mathcal{M}_1$, where under $\mathcal{M}_0$ the data are assumed to be $N(0,1)$ and under $\mathcal{M}_1$ the data are assumed to be $N(\theta, 1)$ for some unknown $\theta \in \mathbb{R}$.

This is another way to view the hypothesis testing problem: $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. Let $\ell_n(\theta)$ be the log-likelihood function. The AIC score for a model is the log-likelihood at the MLE minus the number of parameters, so the AIC score for $\mathcal{M}_0$ is $AIC_0 = \ell_n(0)$ and the AIC score for $\mathcal{M}_1$ is $AIC_1 = \ell_n(\widehat{\theta}) - 1$. Suppose we choose the model with the highest AIC score, and let J_n denote the selected model. So $J_n = 0$ if $AIC_0 > AIC_1$, and $J_n = 1$ otherwise.

a) Suppose that $\mathcal{M}_0$ is the true model. Find $\lim_{n\to\infty} \mathbb{P}(J_n = 0)$. Then compute $\lim_{n\to\infty} \mathbb{P}(J_n = 0)$ when $\theta \neq 0$.

b) Let $\phi_\theta(x)$ denote a Normal density function with mean θ and variance 1. Define

$$\widehat{f}_n(x) = \begin{cases} \phi_0(x) & \text{if } J_n = 0 \\ \phi_{\widehat{\theta}}(x) & \text{if } J_n = 1. \end{cases}$$

Show that $D(\phi_\theta, \widehat{f}_n) \xrightarrow{P} 0$ whether $\theta = 0$ or $\theta \neq 0$, where D is the Kullback-Leibler distance.

c) Repeat this analysis for the BIC.

Solution. We have

$$\begin{aligned}
\mathbb{P}(J_n = 0) &= \mathbb{P}(AIC_0 > AIC_1) \\
&= \mathbb{P}(\ell_n(0) > \ell_n(\widehat{\theta}) - 1) \\
&= \mathbb{P}(\ell_n(0) > \ell_n(\overline{X}_n) - 1)
\end{aligned}$$

as the MLE is $\overline{X}_n$.

We have

$$\begin{aligned}
\ell_n(\theta) &= \log\left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i-\theta)^2\right)\right) \\
&= \sum_{i=1}^n \log\left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i-\theta)^2\right)\right) \\
&= n\log\frac{1}{\sqrt{2\pi}} - \frac{1}{2}\sum_{i=1}^n (x_i-\theta)^2.
\end{aligned}$$

So $AIC_0 = \ell_n(0) = n\log\frac{1}{\sqrt{2\pi}} - \frac{1}{2}\sum_{i=1}^n X_i^2$ and $AIC_1 = \ell_n(\overline{X}_n) - 1 = n\log\frac{1}{\sqrt{2\pi}} - \frac{1}{2}\sum_{i=1}^n (X_i - \overline{X}_n)^2 - 1$. So

$$\begin{aligned}
\mathbb{P}(\ell_n(0) > \ell_n(\overline{X}_n) - 1) &= \mathbb{P}\left(n\log\frac{1}{\sqrt{2\pi}} - \frac{1}{2}\sum_{i=1}^n X_i^2 > n\log\frac{1}{\sqrt{2\pi}} - \frac{1}{2}\sum_{i=1}^n (X_i-\overline{X}_n)^2 - 1\right) \\
&= \mathbb{P}\left(\sum_{i=1}^n (X_i - \overline{X}_n)^2 > \sum_{i=1}^n X_i^2 - 2\right) \\
&= \mathbb{P}\left(-2\sum_{i=1}^n X_i\overline{X}_n + \sum_{i=1}^n \overline{X}_n^2 > -2\right) \\
&= \mathbb{P}\left(-2n\overline{X}_n^2 + n\overline{X}_n^2 > -2\right) \\
&= \mathbb{P}\left(n\overline{X}_n^2 < 2\right) = \mathbb{P}\left(\left|\overline{X}_n\right| < \sqrt{\frac{2}{n}}\right).
\end{aligned}$$

Assume $\mathcal{M}_0$ is the true model. By the Central Limit Theorem, we have $\sqrt{n}(\overline{X}_n) \xrightarrow{\text{d}} Z$. So

$$\mathbb{P}\left(\left|\overline{X}_n\right| < \sqrt{\frac{2}{n}}\right) = \mathbb{P}\left(\sqrt{n}|\overline{X}_n| < \sqrt{2}\right)$$

so given the distribution converge highlighted above, $\mathbb{P}(J_n = 0) \approx 0.84$.

Now assume $\mathcal{M}_0$ is not the true model. Then $\sqrt{n}(\overline{X}_n - \theta) \xrightarrow{\text{d}} Z$, or $\overline{X}_n \xrightarrow{\text{d}} N(\theta, \frac{1}{n})$, again by the Central Limit Theorem. But that means that $\mathbb{P}\left(|\overline{X}_n| < \sqrt{\frac{2}{n}}\right)$ vanishes as $n \to \infty$, because $\overline{X}_n$ converges to $\theta \neq 0$ and $\sqrt{\frac{2}{n}} \to 0$.

For the next part of the problem, the Kullback-Leibler distance is given by

$$D(f,g) = \int f(x)\log\left(\frac{f(x)}{g(x)}\right)dx.$$

We have

$$\mathbb{P}(|D(\phi_\theta, \widehat{f}_n)| > \epsilon) = \mathbb{P}\left(\left|\int \phi_\theta(x)\log\left(\frac{\phi_\theta(x)}{\widehat{f}_n(x)}\right)dx\right| > \epsilon\right).$$

So, with $\theta = 0$, in the case that $\widehat{f}_n(x) = \phi_0(x)$, we have

$$\begin{aligned}
\mathbb{P}(|D(\phi_0, \widehat{f}_n)| > \epsilon) &= \mathbb{P}\left(\left|\int \phi_0(x)\log\left(\frac{\phi_0(x)}{\phi_0(x)}\right)dx\right| > \epsilon\right) \\
&= \mathbb{P}(0 > \epsilon) = 0.
\end{aligned}$$

Therefore we just need to deal with the case that $\widehat{f}_n(x) = \phi_{\widehat{\theta}}(x)$. Here we have

$$\begin{aligned}
D(\phi_0, \widehat{f}_n) &= \int \phi_0(x) \log\left(\frac{\phi_0(x)}{\phi_{\widehat{\theta}}(x)}\right) dx \\
&= \int \phi_0(x) \log\left(\frac{\exp\left(-\frac{1}{2}x^2\right)}{\exp\left(-\frac{1}{2}(x-\widehat{\theta})^2\right)}\right) dx \\
&= \int \phi_0(x) \left(-\frac{1}{2}x^2 - \left(-\frac{1}{2}(x-\widehat{\theta})^2\right)\right) dx \\
&= \frac{1}{2}\int \phi_0(x) \left((x-\widehat{\theta})^2 - x^2\right) dx \\
&= \frac{1}{2}\left[\int \phi_0(x)(-2x\widehat{\theta})dx + \int \phi_0(x)\widehat{\theta}^2 dx\right] \\
&= \frac{1}{2}\left[\mathbb{E}(X)(-2\widehat{\theta}) + \widehat{\theta}^2\right]
\end{aligned}$$

where $X \sim N(0,1)$. But then $\mathbb{E}(X) = 0$, so $D(\phi_0, \widehat{f}_n) = \frac{1}{2}\widehat{\theta}^2$. But then as $\widehat{\theta} = \overline{X}_n$, we observe by the Law of Large Numbers that $D(\phi_0, \widehat{f}_n) \xrightarrow{P} 0$, as desired.

Now take $\theta \neq 0$. We analyze first the case where $J_n = 1$.

$$\begin{aligned}
D(\phi_\theta, \widehat{f}_n) &= D(\phi_\theta, \phi_{\widehat{\theta}}) \\
&= \int \phi_\theta(x) \log\left(\frac{\phi_\theta(x)}{\phi_{\widehat{\theta}}(x)}\right) dx \\
&= \int \phi_\theta(x) \log\left(\frac{\exp\left(-\frac{1}{2}(x-\theta)^2\right)}{\exp\left(-\frac{1}{2}(x-\widehat{\theta})^2\right)}\right) dx \\
&= \int \phi_\theta(x) \left(-\frac{1}{2}(x-\theta)^2 + \frac{1}{2}(x-\widehat{\theta})^2\right) dx \\
&= \frac{1}{2}\int \phi_\theta(x) \left((x-\widehat{\theta})^2 - (x-\theta)^2\right) dx \\
&= \frac{1}{2}\int \phi_\theta(x) \left(\widehat{\theta}^2 - 2x\widehat{\theta} - \theta^2 + 2x\theta\right) dx \\
&= \frac{1}{2}\left[\mathbb{E}(X)(-2\widehat{\theta} + 2\theta) + \widehat{\theta}^2 - \theta^2\right]
\end{aligned}$$

where $X \sim N(\theta, 1)$. But this comes out to $\frac{1}{2}\left[\theta^2 - 2\theta\widehat{\theta} + \widehat{\theta}^2\right] = \frac{1}{2}(\theta - \widehat{\theta})^2$. Moreover, we know already that $\widehat{\theta} = \overline{X} \to \theta$ if $\theta \neq 0$, so $D(\phi_\theta, \widehat{f}_n) \xrightarrow{P} 0$ in this case.

Finally, if $J_n = 0$, then

$$\begin{aligned}
D(\phi_\theta, \widehat{f}_n) &= D(\phi_\theta, \phi_0) \\
&= \int \phi_\theta(x) \log\left(\frac{\phi_\theta(x)}{\phi_0(x)}\right) dx \\
&= \int \phi_\theta(x) \log\left(\frac{\exp\left(-\frac{1}{2}(x-\theta)^2\right)}{\exp\left(-\frac{1}{2}x^2\right)}\right) dx \\
&= \int \phi_\theta(x) \left(\frac{1}{2}x^2 - \frac{1}{2}(x-\theta)^2\right) dx \\
&= \frac{1}{2}\int \phi_\theta(x) \left(2x\theta - \theta^2\right) dx \\
&= \frac{1}{2}\left[2\mathbb{E}(X)\theta - \theta^2\right]
\end{aligned}$$

with $X \sim N(\theta, 1)$. But then $\mathbb{E}(X) = \theta$, so $D(\phi_\theta, \widehat{f}_n) = \frac{\theta^2}{2}$. However, this term vanishes in probability, so $D(\phi_\theta, \widehat{f}_n) \xrightarrow{P} 0$, as desired.

Let us now analyze the BIC. The BIC score for $\mathcal{M}_0$ is $\ell_n(0)$ and the BIC score for $\mathcal{M}_1$ is $\ell_n(\widehat{\theta}) - \frac{1}{2}\log n$. We have

$$\begin{aligned}\mathbb{P}(J_n = 0) &= \mathbb{P}\left(\ell_n(0) > \ell_n(\widehat{\theta}) - \frac{1}{2}\log n\right)\\ &= \mathbb{P}\left(\ell_n(0) > \ell_n(\overline{X}_n) - \frac{1}{2}\log n\right).\end{aligned}$$

Moreover, we computed previously that

$$\ell_n(\theta) = n \log \frac{1}{\sqrt{2\pi}} - \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2.$$

Therefore

$$\begin{aligned}\mathbb{P}\left(\ell_n(0) > \ell_n(\overline{X}_n) - \frac{1}{2}\log n\right) &= \mathbb{P}\left(-\frac{1}{2}\sum_{i=1}^{n} X_i^2 > -\frac{1}{2}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 - \frac{1}{2}\log n\right)\\ &= \mathbb{P}\left(\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 > \sum_{i=1}^{n} X_i^2 - \log n\right)\\ &= \mathbb{P}(-n\overline{X}_n^2 > -\log n)\\ &= \mathbb{P}\left(\log n > n\overline{X}_n^2\right) = \mathbb{P}\left(\sqrt{\frac{\log n}{n}} > |\overline{X}_n|\right).\end{aligned}$$

Assume $\mathcal{M}_0$ is the true model. Again, by the Central Limit Theorem, we have $\sqrt{n}(\overline{X}_n) \xrightarrow{\text{d}} Z$. So

$$\mathbb{P}\left(\sqrt{\frac{\log n}{n}} > |\overline{X}_n|\right) = \mathbb{P}\left(\sqrt{\log n} > |\overline{X}_n|\sqrt{n}\right)$$

so by the distribution convergence highlighted above, we conclude that as $n \to \infty$, $\mathbb{P}\left(\sqrt{\frac{\log n}{n}} > |\overline{X}_n|\right) \to 1$.

Now assume $\mathcal{M}_0$ is not the true model. Then $\overline{X}_n \xrightarrow{\text{d}} N(\theta, \frac{1}{n})$. But $\sqrt{\frac{\log n}{n}} \to 0$ as $n \to \infty$, so this probability also vanishes as $n \to \infty$. $\square$

Problem 13.10. Let $\theta = \beta_0 + \beta_1 X_*$ and let $\widehat{\theta} = \widehat{\beta}_0 + \widehat{\beta}_1 X_*$, so that $Y_* = \theta + \epsilon$ and $\widehat{Y}_* = \widehat{\theta}$.

a) Let $s = \sqrt{\mathbb{V}(\widehat{Y}_*)}$. Show that

$$\begin{aligned}\mathbb{P}(\widehat{Y}_* - 2s < Y_* < \widehat{Y}_* + 2s) &\approx \mathbb{P}(-2 < N(0, 1 + \sigma^2/s^2) < 2)\\ &\neq 0.95.\end{aligned}$$

b) Define

$$\xi_n^2 = \mathbb{V}(\widehat{Y}_*) + \sigma^2 = \left[\frac{\sum_i (x_i - x_*)^2}{n\sum_i (x_i - \overline{x})^2} + 1\right]\sigma^2,$$

where in practice, we substitute $\widehat{\sigma}$ for σ, and denote the resulting quantity by $\widehat{\xi}_n$. Now show that $\widehat{Y}_n \pm 2\widehat{\xi}_n$ is a 95% confidence interval.

Solution. We would like to show that $\frac{Y_* - \widehat{Y}_*}{s} \approx N(0, 1 + \sigma^2/s^2)$, or equivalently that $Y_* - \widehat{Y}_* \approx N(0, s^2 + \sigma^2)$.

Note that $\mathbb{E}(Y_*) = \mathbb{E}(\theta + \epsilon) = \theta$, as $\mathbb{E}(\epsilon) = 0$; moreover, $\mathbb{E}(\widehat{Y}_*) = \mathbb{E}(\widehat{\theta}) = \beta_0 + \beta_1 X_* = \theta$. Thus $\mathbb{E}(Y_* - \widehat{Y}_*) = 0$.

Moreover, we were given $\mathbb{V}(\widehat{Y}_*) = s^2$, and we know that $\mathbb{V}(Y_*) = \mathbb{V}(\theta + \epsilon) = \mathbb{V}(\epsilon) = \sigma^2$. Finally, conditioned on the training data, $\widehat{Y}_*$ and Y_* are independent, so $Y_* - \widehat{Y}_* \approx N(0, s^2 + \sigma^2)$ and we are done.

Next, note that $Y_* - \widehat{Y}_* \approx N(0, s^2 + \sigma^2) = N(0, \xi_n^2)$. So it suffices to show that $\mathbb{V}(\widehat{Y}_*) = \left[\frac{\sum_i (x_i - x_*)^2}{n \sum_i (x_i - \overline{x})^2}\right] \sigma^2$.

Using problem 13.2, we have

$$\begin{aligned}
\mathbb{V}(\widehat{Y}_*) &= \mathbb{V}\left(\widehat{\beta}_0 + \widehat{\beta}_1 X_*\right) \\
&= \mathbb{V}(\widehat{\beta}_0) + \mathbb{V}(\widehat{\beta}_1 X_*) + 2\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1 X_*) \\
&= \frac{\sigma^2}{n s_X^2} \cdot \frac{1}{n} \sum_{i=1}^n X_i^2 + \frac{\sigma^2}{n s_X^2} \cdot X_*^2 - \frac{\sigma^2}{n s_X^2} \cdot 2 X_* \overline{X}_n \\
&= \frac{\sigma^2}{n s_X^2} \left[\frac{1}{n} \sum_{i=1}^n X_i^2 + X_*^2 - 2 X_* \overline{X}_n\right].
\end{aligned}$$

Thus, cancelling some terms, we'll want to show that

$$\frac{1}{n} \sum_i (x_i - x_*)^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 + X_*^2 - 2 X_* \overline{X}_n.$$

But

$$\begin{aligned}
\frac{1}{n} \sum_i (x_i - x_*)^2 &= \frac{1}{n} \sum_{i=1}^n (X_i^2 - 2 X_* X_i + X_*^2) \\
&= \frac{1}{n} \sum_{i=1}^n X_i^2 - 2 X_* \overline{X}_n + X_*^2
\end{aligned}$$

and we are done. The analysis for KL-distance is the same as in part (b). □