24500 HW6

Bin Yu

Feb 20, 2025

Question 1

We have the linear model

$$y_i \sim \mathcal{N}\big(\beta_0 + \beta_1 x_i,\ \sigma^2\big), \quad i = 1, 2, 3, 4,$$

and we wish to choose $x_1, x_2, x_3, x_4 \in [-1, 1]$ *before* observing y_i, so that the MLE (or LSE) of β_1 is as accurate as possible.

The ordinary least squares (OLS) estimator for β_1 is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \text{where } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

For $n = 4$, we have four points (x_i, y_i).

By definition, the mean squared error of $\hat{\beta}_1$ is

$$\text{MSE}(\hat{\beta}_1) \ = \ E\big[(\hat{\beta}_1 - \beta_1)^2\big] \ = \ \text{Var}(\hat{\beta}_1) \ + \ \big[\text{Bias}(\hat{\beta}_1)\big]^2.$$

and we have proved in class that $\hat{\beta}_1$ is unbiased for β_1, i.e.,

$$E[\hat{\beta}_1] \ = \ \beta_1,$$

which implies $\text{Bias}(\hat{\beta}_1) = 0$. Hence,

$$\text{MSE}(\hat{\beta}_1) \ = \ \text{Var}(\hat{\beta}_1).$$

$$\text{Var}(\hat{\beta}_1) \ = \ \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Therefore, to minimize $\text{Var}(\hat{\beta}_1)$, we must maximize the quantity $\sum_{i=1}^{n}(x_i - \bar{x})^2$. For a fixed sample size n, we want to spread out the x_i's as much as possible around their mean $\bar{x}$.

Since each x_i must lie in the interval $[-1, 1]$, we aim to:

- Force $\bar{x} = \frac{1}{n}\sum x_i = 0$ so that the points are centered.
- Place the x_i's at the extreme values to maximize $\sum(x_i - \bar{x})^2$.

For $n = 4$, the optimal choice is to set two of the x_i's to -1 and two to $+1$. For instance:

$$x_1 = -1, \quad x_2 = -1, \quad x_3 = 1, \quad x_4 = 1.$$

Then the sample mean $\bar{x} = 0$, and

$$\sum_{i=1}^{4}(x_i - \bar{x})^2 = (-1-0)^2 + (-1-0)^2 + (1-0)^2 + (1-0)^2 = 4.$$

This is the maximum possible under $|x_i| \leq 1$. Consequently,

$$\mathrm{Var}(\hat{\beta}_1) \;=\; \frac{\sigma^2}{4}$$

is as small as it can be in that domain, and hence $\mathrm{MSE}(\hat{\beta}_1)$ is also minimized.

Therefore, to achieve the smallest MSE (which equals the variance here, since $\hat{\beta}_1$ is unbiased), we must choose the x_i's so as to maximize their spread around zero. When $x_i \in [-1, 1]$ and $n = 4$, the best choice is to place two points at -1 and two points at $+1$. This yields the most precise estimation of β_1.

Question 2

Let

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

where X is the $n \times 2$ design matrix

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},$$

and $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$.

First, note that

$$X^T = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix}.$$

Hence,

$$X^T X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \displaystyle\sum_{i=1}^{n} x_i \\ \displaystyle\sum_{i=1}^{n} x_i & \displaystyle\sum_{i=1}^{n} x_i^2 \end{bmatrix}.$$

For a 2×2 matrix $\begin{bmatrix} a & b \\ b & c \end{bmatrix}$, its inverse is

$$\frac{1}{ac - b^2} \begin{bmatrix} c & -b \\ -b & a \end{bmatrix}.$$

here,

$$a = n, \quad b = \sum_{i=1}^{n} x_i, \quad c = \sum_{i=1}^{n} x_i^2.$$

Therefore,

$$(X^T X)^{-1} = \frac{1}{n \sum_{i=1}^{n} x_i^2 \;-\; \left(\sum_{i=1}^{n} x_i\right)^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}.$$

$$X^T y = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \displaystyle\sum_{i=1}^n y_i \\ \displaystyle\sum_{i=1}^n x_i\, y_i \end{bmatrix}.$$

Thus,

$$\hat{\beta} = (X^T X)^{-1}\, X^T y = \frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i\, y_i \end{bmatrix}.$$

$$\hat{\beta}_0 = \frac{\Bigl(\displaystyle\sum_{i=1}^n x_i^2\Bigr)\Bigl(\displaystyle\sum_{i=1}^n y_i\Bigr) \;-\; \Bigl(\displaystyle\sum_{i=1}^n x_i\Bigr)\Bigl(\displaystyle\sum_{i=1}^n x_i\, y_i\Bigr)}{n\displaystyle\sum_{i=1}^n x_i^2 \;-\; \Bigl(\displaystyle\sum_{i=1}^n x_i\Bigr)^2}$$

$$\hat{\beta}_1 = \frac{-\Bigl(\displaystyle\sum_{i=1}^n x_i\Bigr)\Bigl(\displaystyle\sum_{i=1}^n y_i\Bigr) \;+\; n\displaystyle\sum_{i=1}^n x_i\, y_i}{n\displaystyle\sum_{i=1}^n x_i^2 \;-\; \Bigl(\displaystyle\sum_{i=1}^n x_i\Bigr)^2}.$$

To simplify β_1, β_0:

Let

$$\bar{x} \;=\; \frac{1}{n}\sum_{i=1}^n x_i, \quad \bar{y} \;=\; \frac{1}{n}\sum_{i=1}^n y_i.$$

By definition,

$$\sum_{i=1}^n (x_i - \bar{x})^2 \;=\; \sum_{i=1}^n \bigl[x_i^2 - 2\,x_i\,\bar{x} + \bar{x}^2\bigr] \;=\; \sum_{i=1}^n x_i^2 \;-\; 2\,\bar{x}\sum_{i=1}^n x_i \;+\; n\,\bar{x}^2.$$

Since $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, it follows that $\sum_{i=1}^n x_i = n\,\bar{x}$. Substituting gives:

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 \;-\; 2\,\bar{x}\,(n\,\bar{x}) \;+\; n\,\bar{x}^2 = \sum_{i=1}^n x_i^2 \;-\; n\,\bar{x}^2.$$

We have $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, so $\bar{x}^2 = \frac{(\sum x_i)^2}{n^2}$. Thus

$$n\,\bar{x}^2 = n \cdot \frac{(\sum_{i=1}^n x_i)^2}{n^2} = \frac{(\sum_{i=1}^n x_i)^2}{n}.$$

Hence

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 \;-\; \frac{\bigl(\sum_{i=1}^n x_i\bigr)^2}{n}.$$

$$n\sum_{i=1}^n x_i^2 \;-\; \Bigl(\sum_{i=1}^n x_i\Bigr)^2 = n\sum_{i=1}^n (x_i - \bar{x})^2.$$

For the numerator:

$$(x_i - \bar{x})(y_i - \bar{y}) \;=\; x_i\,y_i \;-\; x_i\,\bar{y} \;-\; \bar{x}\,y_i \;+\; \bar{x}\,\bar{y}.$$

Summing from $i = 1$ to n gives

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i\,y_i \;-\; \bar{y}\sum_{i=1}^n x_i \;-\; \bar{x}\sum_{i=1}^n y_i \;+\; n\,\bar{x}\,\bar{y}.$$

We know $\sum_{i=1}^{n} x_i = n\,\bar{x}$ and $\sum_{i=1}^{n} y_i = n\,\bar{y}$. Substituting,

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n} x_i\,y_i \;-\; \bar{y}\,\big(n\,\bar{x}\big) \;-\; \bar{x}\,\big(n\,\bar{y}\big) \;+\; n\,\bar{x}\,\bar{y}.$$

$$= \sum_{i=1}^{n} x_i\,y_i \;-\; n\,\bar{x}\,\bar{y}.$$

Thus,

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n} x_i\,y_i \;-\; n\,\bar{x}\,\bar{y}.$$

$$= \sum_{i=1}^{n} x_i\,y_i \;-\; \frac{\big(\sum_{i=1}^{n} x_i\big)\big(\sum_{i=1}^{n} y_i\big)}{n}$$

Putting numerator and denominator together, we obtain:

$$\hat{\beta}_1 = \frac{-\Big(\sum_{i=1}^{n} x_i\Big)\Big(\sum_{i=1}^{n} y_i\Big) \;+\; n\sum_{i=1}^{n} x_i\,y_i}{n\sum_{i=1}^{n} x_i^2 \;-\; \Big(\sum_{i=1}^{n} x_i\Big)^2}.$$

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

To Simplify $\hat{\beta}_0$:

We start with

$$\hat{\beta}_0 = \frac{\Big(\sum_{i=1}^{n} x_i^2\Big)\Big(\sum_{i=1}^{n} y_i\Big) \;-\; \Big(\sum_{i=1}^{n} x_i\Big)\Big(\sum_{i=1}^{n} x_i\,y_i\Big)}{n\sum_{i=1}^{n} x_i^2 \;-\; \Big(\sum_{i=1}^{n} x_i\Big)^2}.$$

We already have:

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

Hence,

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

Since $\bar{y} = \frac{1}{n}\sum y_i$

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

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

Since

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n} x_i^2 \;-\; \frac{\big(\sum x_i\big)^2}{n}.$$

Similarly,

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n} x_i\,y_i \;-\; n\,\bar{x}\,\bar{y} = \sum_{i=1}^{n} x_i\,y_i \;-\; \frac{\big(\sum_{i=1}^{n} x_i\big)\big(\sum_{i=1}^{n} y_i\big)}{n}.$$

Hence

$$\bar{y}\sum_{i=1}^{n}(x_i-\bar{x})^2=\Big(\frac{\sum_{i=1}^{n}y_i}{n}\Big)\Big(\sum_{i=1}^{n}x_i^2-\frac{(\sum_{i=1}^{n}x_i)^2}{n}\Big),$$

$$\bar{x}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\Big(\frac{\sum_{i=1}^{n}x_i}{n}\Big)\Big(\sum_{i=1}^{n}x_i\,y_i-\frac{(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)}{n}\Big).$$

$$\hat{\beta}_0=\frac{\Big(\sum_{i=1}^{n}x_i^2\Big)\Big(\sum_{i=1}^{n}y_i\Big)\;-\;\Big(\sum_{i=1}^{n}x_i\Big)\Big(\sum_{i=1}^{n}x_i\,y_i\Big)}{n\sum_{i=1}^{n}x_i^2\;-\;\Big(\sum_{i=1}^{n}x_i\Big)^2}.$$

Substitute the expansions, the numerator of $\bar{y}-\hat{\beta}_1\bar{x}$ is

$$\Big(\frac{\sum_{i=1}^{n}y_i}{n}\Big)\Big(\sum_{i=1}^{n}x_i^2-\frac{(\sum_{i=1}^{n}x_i)^2}{n}\Big)\;-\;\Big(\frac{\sum_{i=1}^{n}x_i}{n}\Big)\Big(\sum_{i=1}^{n}x_i\,y_i-\frac{(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)}{n}\Big).$$

$$=\frac{1}{n}\Big[(\sum_{i=1}^{n}y_i)\Big(\sum_{i=1}^{n}x_i^2-\frac{(\sum_{i=1}^{n}x_i)^2}{n}\Big)\;-\;(\sum_{i=1}^{n}x_i)\Big(\sum_{i=1}^{n}x_i\,y_i-\frac{(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)}{n}\Big)\Big].$$

$$=\frac{1}{n}[(\sum_{i=1}^{n}y_i)\Big(\sum_{i=1}^{n}x_i^2\Big)\;-\;(\sum_{i=1}^{n}y_i)\,\frac{(\sum_{i=1}^{n}x_i)^2}{n}\;-\;(\sum_{i=1}^{n}x_i)\Big(\sum_{i=1}^{n}x_i\,y_i\Big)\;+\;(\sum_{i=1}^{n}x_i)\,\frac{(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)}{n}]$$

Notice the second and last terms both contain $\frac{1}{n}$:

$$-(\sum_{i=1}^{n}y_i)\,\frac{(\sum_{i=1}^{n}x_i)^2}{n}+(\sum_{i=1}^{n}x_i)\,\frac{(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)}{n}=\frac{(\sum_{i=1}^{n}y_i)}{n}\Big[-(\sum_{i=1}^{n}x_i)^2+(\sum_{i=1}^{n}x_i)^2\Big]=0.$$

So those two terms cancel

Hence the numerator is:

$$\frac{1}{n}\Big[(\sum_{i=1}^{n}x_i^2)(\sum_{i=1}^{n}y_i)\;-\;(\sum_{i=1}^{n}x_i)\,\sum_{i=1}^{n}x_i\,y_i\Big].$$

And the denominator is $\sum(x_i-\bar{x})^2$, we have proved that:

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-\frac{(\sum_{i=1}^{n}x_i)^2}{n}\;=\;\frac{1}{n}\Big[n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2\Big].$$

Hence,

$$\bar{y}-\hat{\beta}_1\bar{x}=\frac{\frac{1}{n}\big[(\sum_{i=1}^{n}x_i^2)(\sum_{i=1}^{n}y_i)-(\sum_{i=1}^{n}x_i)\sum_{i=1}^{n}x_i\,y_i\big]}{\frac{1}{n}\big[n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2\big]}=\frac{(\sum_{i=1}^{n}x_i^2)(\sum_{i=1}^{n}y_i)-(\sum_{i=1}^{n}x_i)\sum_{i=1}^{n}x_i\,y_i}{n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2}.$$

Therefore,

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\,\bar{x}=\frac{\sum_{i=1}^{n}x_i^2\,\sum_{i=1}^{n}y_i-(\sum_{i=1}^{n}x_i)\,\sum_{i=1}^{n}(x_i\,y_i)}{n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2}.$$

Hence, from the matrix solution $\hat{\beta}=(X^TX)^{-1}X^Ty$ for the two-parameter case, we obtain the same estimates for simple linear regression estimates that we had derived from class:

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2},\quad \hat{\beta}_0=\bar{y}-\hat{\beta}_1\,\bar{x}.$$

Question 3

A standard result in linear regression states that

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

Hence,

$$E[\hat{\beta}] = \beta$$
$$E[\hat{\beta}_0] = \beta_0, \quad E[\hat{\beta}_1] = \beta_1,$$

The covariance matrix of $\hat{\beta}$ is

$$\mathrm{Var}(\hat{\beta}) = \sigma^2\,(X^TX)^{-1}.$$

In the simple linear regression setting ($p = 2$), note that

$$X^T = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix}.$$

Hence,

$$X^TX = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \displaystyle\sum_{i=1}^n x_i \\ \displaystyle\sum_{i=1}^n x_i & \displaystyle\sum_{i=1}^n x_i^2 \end{bmatrix}.$$

For a 2×2 matrix $\begin{bmatrix} a & b \\ b & c \end{bmatrix}$, its inverse is

$$\frac{1}{ac - b^2}\begin{bmatrix} c & -b \\ -b & a \end{bmatrix}.$$

here,

$$a = n, \quad b = \sum_{i=1}^n x_i, \quad c = \sum_{i=1}^n x_i^2.$$

$$(X^TX)^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2\ -\ \big(\sum_{i=1}^n x_i\big)^2}\begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix}.$$

Therefore, expand $\mathrm{Var}(\hat{\beta}) = \sigma^2\,(X^TX)^{-1}$.:

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2\,n}{n\sum_{i=1}^n x_i^2 - \Big(\sum_{i=1}^n x_i\Big)^2}$$

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2\,\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n x_i^2 - \Big(\sum_{i=1}^n x_i\Big)^2}$$

$$\mathrm{Cov}(\hat{\beta}_0,\ \hat{\beta}_1) = -\frac{\sigma^2\,\sum_{i=1}^n x_i}{n\sum_{i=1}^n x_i^2 - \Big(\sum_{i=1}^n x_i\Big)^2}.$$

Let

$$\bar{x}\ =\ \frac{1}{n}\sum_{i=1}^n x_i \quad \text{and} \quad V\ =\ \frac{1}{n}\sum_{i=1}^n \big(x_i - \bar{x}\big)^2.$$

By definition,

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 \;=\; \sum_{i=1}^{n}\Big(x_i^2 - 2\,x_i\,\bar{x} + \bar{x}^2\Big) \;=\; \sum_{i=1}^{n} x_i^2 \;-\; 2\,\bar{x}\sum_{i=1}^{n} x_i \;+\; n\,\bar{x}^2.$$

Since $\sum_{i=1}^{n} x_i = n\,\bar{x}$, the middle term becomes $-\,2\,\bar{x}\,(n\,\bar{x}) = -\,2n\,\bar{x}^2$. Hence

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n} x_i^2 \;-\; 2n\,\bar{x}^2 \;+\; n\,\bar{x}^2 = \sum_{i=1}^{n} x_i^2 \;-\; n\,\bar{x}^2.$$

We have

$$\bar{x}^2 = \Big(\tfrac{1}{n}\sum_{i=1}^{n} x_i\Big)^2 = \frac{\big(\sum_{i=1}^{n} x_i\big)^2}{n^2}.$$

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n} x_i^2 \;-\; \frac{\big(\sum_{i=1}^{n} x_i\big)^2}{n}.$$

$$n\sum_{i=1}^{n}(x_i-\bar{x})^2 = n\sum_{i=1}^{n} x_i^2 \;-\; \big(\sum_{i=1}^{n} x_i\big)^2.$$

$$n\sum_{i=1}^{n} x_i^2 \;-\; \Big(\sum_{i=1}^{n} x_i\Big)^2 = n\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

Since

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = n\,V,$$

so

$$n\sum_{i=1}^{n}(x_i-\bar{x})^2 = n\,(nV) = n^2\,V.$$

Thus

$$n\sum_{i=1}^{n} x_i^2 \;-\; \Big(\sum_{i=1}^{n} x_i\Big)^2 = n^2\,V.$$

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2\,n}{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{\sigma^2\,n}{n^2\,V} = \frac{\sigma^2}{n\,V}.$$

We have

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2\,\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{\sigma^2\,\sum_{i=1}^{n} x_i^2}{n^2\,V}.$$

Now expand $\sum_{i=1}^{n} x_i^2$ in terms of $\bar{x}$ and $\sum(x_i-\bar{x})^2$:

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n}\big[(x_i-\bar{x})+\bar{x}\big]^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 \;+\; 2\bar{x}\,\sum_{i=1}^{n}(x_i-\bar{x}) \;+\; n\bar{x}^2.$$

Because $\sum_{i=1}^{n}(x_i-\bar{x}) = 0$, this reduces to

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 + n\,\bar{x}^2 = nV + n\bar{x}^2 = n\,\big(V+\bar{x}^2\big).$$

Thus

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2\cdot n(V+\bar{x}^2)}{n^2\,V} = \frac{\sigma^2}{n\,V}\,\big[V+\bar{x}^2\big] = \frac{\sigma^2}{n}\,\Big(\frac{V}{V}+\frac{\bar{x}^2}{V}\Big) = \frac{\sigma^2}{n}\Big(1+\frac{\bar{x}^2}{V}\Big).$$

$$\mathrm{Cov}(\hat{\beta}_0,\;\hat{\beta}_1) = -\,\frac{\sigma^2\,\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = -\,\frac{\sigma^2\,\sum_{i=1}^{n} x_i}{n^2\,V}.$$

Since $\sum_{i=1}^{n} x_i = n\,\bar{x}$

$$\mathrm{Cov}(\hat{\beta}_0,\ \hat{\beta}_1) = -\,\frac{\sigma^2\, n\bar{x}}{n^2\, V} = -\,\frac{\sigma^2\,\bar{x}}{n\,V}.$$

Putting it all together, we arrive at the simpler expressions in terms of $\bar{x}$ and $V = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2$:

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{n\,V}, \quad \mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2}{n}\Big(1 + \frac{\bar{x}^2}{V}\Big), \quad \mathrm{Cov}(\hat{\beta}_0,\ \hat{\beta}_1) = -\,\frac{\sigma^2\,\bar{x}}{n\,V}.$$

This matches the forms we arrived in class.

Question 4

(a)

Define

$$\hat{\sigma}_c^2 \;=\; c\,\|y - X\hat{\beta}\|^2.$$

Since $\|y - X\hat{\beta}\|^2 \sim \sigma^2\,\chi^2_{n-p}$, and $E(\chi^2_{n-p}) = n - p$, we know

$$E\big[\|y - X\hat{\beta}\|^2\big] = \sigma^2\,E[\chi^2_{n-p}] = \sigma^2\,(n-p).$$

Thus,

$$E[\hat{\sigma}_c^2] = c\,E\big[\|y - X\hat{\beta}\|^2\big] = c\,\sigma^2\,(n-p).$$

For $\hat{\sigma}_c^2$ to be *unbiased*, we need $E[\hat{\sigma}_c^2] = \sigma^2$. Hence

$$c\,\sigma^2\,(n-p) = \sigma^2 \quad\Longrightarrow\quad c = \frac{1}{n-p}.$$

Therefore,

$$c = \frac{1}{n-p}$$

(b)

We assume the usual normal linear regression setting:

$$y \;\sim\; \mathcal{N}\big(X\beta,\ \sigma^2 I_n\big),$$

and let $\hat{\beta}$ be the least squares estimator (LSE) of β. We want to find the maximum likelihood estimator (MLE) of σ^2.

In the linear model

$$y \;\sim\; \mathcal{N}(X\beta,\ \sigma^2 I_n),$$

To find the MLEs, we take partial derivatives with respect to β and σ^2 and set them to zero. By differentiating w.r.t. β, the MLE for β coincides with the least squares estimator $\hat{\beta}$

$$\hat{\beta} \;=\; \operatorname*{argmin}_{\beta}\,\|y - X\beta\|^2 \;=\; \big(X^T X\big)^{-1} X^T y.$$

Once $\hat{\beta}$ is obtained, for any fixed σ^2, the likelihood is largest at $\beta = \hat{\beta}$. Hence the maximization problem for σ^2 reduces to a function only of σ^2. We then differentiate this expression w.r.t. σ^2 and solve for the critical point to get the MLE for σ^2.

The likelihood for σ^2, after substituting $\beta = \hat{\beta}$, is given by

$$L(\sigma^2) \;=\; \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\!\Big(-\frac{1}{2\sigma^2}\,\|\,y - X\hat{\beta}\|^2\Big).$$

$$L(\sigma^2) = \big(2\pi\,\sigma^2\big)^{-\frac{n}{2}} \exp\!\Big(-\tfrac{1}{2\sigma^2}\,\|y - X\hat{\beta}\|^2\Big),$$

Taking the natural log of $L(\sigma^2)$

$$\ell(\sigma^2) \;=\; -\frac{n}{2}\ln(2\pi) \;-\; \frac{n}{2}\ln(\sigma^2) \;-\; \frac{1}{2\sigma^2}\,\|y - X\hat{\beta}\|^2.$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\,\frac{n}{2}\,\frac{1}{\sigma^2} \;+\; \frac{1}{2(\sigma^2)^2}\,\|y - X\hat{\beta}\|^2.$$

Set this derivative to zero to find the critical point:

$$-\,\frac{n}{2\,\sigma^2} \;+\; \frac{1}{2(\sigma^2)^2}\,\|y - X\hat{\beta}\|^2 \;=\; 0.$$

$$-\,n\,\sigma^2 \;+\; \|\,y - X\hat{\beta}\|^2 \;=\; 0.$$

Hence, the maximum likelihood solution is

$$\hat{\sigma}^2_{\text{MLE}} \;=\; \frac{1}{n}\,\|y - X\hat{\beta}\|^2.$$

$$c = \frac{1}{n}$$

(c)

define $\hat{\sigma}_c^{\,2} = c\,\|y - X\hat{\beta}\|^2$. The mean squared error (MSE) is

$$\text{MSE}\big(\hat{\sigma}_c^{\,2}\big) = E\Big[\big(\hat{\sigma}_c^{\,2} - \sigma^2\big)^2\Big].$$

let $\|y - X\hat{\beta}\|^2 = \sigma^2\,T$ with $T \sim \chi^2_{n-p}$. Then

$$\hat{\sigma}_c^{\,2} = c\,\sigma^2\,T, \quad \hat{\sigma}_c^{\,2} - \sigma^2 = \sigma^2\big(c\,T - 1\big).$$

Hence

$$\text{MSE}\big(\hat{\sigma}_c^{\,2}\big) = E\Big[\big(\hat{\sigma}_c^{\,2} - \sigma^2\big)^2\Big] = \sigma^4\,E\big[(c\,T - 1)^2\big].$$

$$E\big[(c\,T - 1)^2\big] = E\big[c^2\,T^2 - 2c\,T + 1\big] = c^2\,E[T^2] - 2c\,E[T] + 1.$$

For a χ^2_k random variable T, $E[T] = k$ and $\text{Var}(T) = 2k$. Hence $E[T^2] = \text{Var}(T) + \big(E[T]\big)^2 = 2k + k^2$. here $k = n - p$. Thus

$$E[T^2] = 2(n-p) + (n-p)^2.$$

$$E\big[(c\,T - 1)^2\big] = c^2\big[2(n-p) + (n-p)^2\big] \;-\; 2c\,(n-p) \;+\; 1.$$

Let

$$f(c) = c^2\big[2(n-p) + (n-p)^2\big] \;-\; 2c\,(n-p) \;+\; 1.$$

Taking derivative and setting it to zero:

$$\frac{d}{dc}f(c) = 2c\big[2(n-p) + (n-p)^2\big] - 2(n-p) = 0.$$

$$2c\big[2(n-p) + (n-p)^2\big] = 2(n-p)$$

$$c = \frac{n-p}{2(n-p)+(n-p)^2} = \frac{n-p}{(n-p)\big((n-p)+2\big)} = \frac{1}{n-p+2}.$$

Therefore, the value of c that minimizes the MSE is

$$c = \frac{1}{n-p+2}$$

(d)

we have the statistic

$$\| y - X\hat{\beta}\|^2 \;\sim\; \sigma^2\,\chi^2_{n-p},$$

and we define

$$\hat{\sigma}_c^{\,2} \;=\; c\,\|y - X\hat{\beta}\|^2.$$

Then

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

$$\frac{\frac{1}{c}\,\hat{\sigma}_c^{\,2}}{\sigma^2} \;=\; \frac{\hat{\sigma}_c^{\,2}}{c\,\sigma^2} \;\sim\; \chi^2_{n-p}.$$

Hence, a $(1-\alpha)$-level confidence interval for σ^2:

$$P\Big(\chi^2_{n-p,\alpha/2} \;\le\; \frac{\hat{\sigma}_c^{\,2}}{c\,\sigma^2} \;\le\; \chi^2_{n-p,1-\alpha/2}\Big) = 1-\alpha.$$

Solving for σ^2 gives

$$\frac{\hat{\sigma}_c^{\,2}}{c\,\chi^2_{n-p,1-\alpha/2}} \;\le\; \sigma^2 \;\le\; \frac{\hat{\sigma}_c^{\,2}}{c\,\chi^2_{n-p,\alpha/2}}.$$

Therefore, the $(1-\alpha)$-level confidence interval is

$$\left[\frac{\hat{\sigma}_c^{\,2}}{c\,\chi^2_{n-p,1-\alpha/2}},\; \frac{\hat{\sigma}_c^{\,2}}{c\,\chi^2_{n-p,\alpha/2}}\right].$$

it can also be written as:

$$\left[\frac{\|y - X\hat{\beta}\|^2}{\chi^2_{n-p,1-\alpha/2}},\; \frac{\|y - X\hat{\beta}\|^2}{\chi^2_{n-p,\alpha/2}}\right].$$

Question 5

(a)

From Question 3 we have known:

$$\hat{\beta} \;\sim\; \mathcal{N}\Big(\beta,\; \sigma^2\,\big(X^TX\big)^{-1}\Big).$$

Now consider the linear combination $(x^*)^T\hat{\beta}$. Since $\hat{\beta}$ is multivariate normal, any linear combination of its components is also normally distributed

$$E\big[(x^*)^T\hat{\beta}\big] \;=\; (x^*)^T\, E[\hat{\beta}] \;=\; (x^*)^T\,\beta,$$

because $E[\hat{\beta}] = \beta$.

$$\mathrm{Var}\big((x^*)^T\hat{\beta}\big) \;=\; (x^*)^T\,\mathrm{Var}(\hat{\beta})\,(x^*) \;=\; (x^*)^T\big[\sigma^2(X^TX)^{-1}\big](x^*) \;=\; \sigma^2\,(x^*)^T(X^TX)^{-1}x^*.$$

Hence,

$$(x^*)^T\hat{\beta} \;\sim\; \mathcal{N}\Big((x^*)^T\beta,\; \sigma^2\,(x^*)^T(X^TX)^{-1}x^*\Big).$$

(b)

We want a $(1-\alpha)$-level confidence interval for $(x^*)^T\beta$, using the fact that $\hat{\beta}$ is the MLE (or LSE) under the model

$$y \;\sim\; \mathcal{N}\big(X\beta,\; \sigma^2 I_n\big).$$

Let

$$x^* \;=\; \big(1,\; x_1^*,\; x_2^*,\; \ldots,\; x_{p-1}^*\big)^T,$$

and consider the prediction $(x^*)^T\hat{\beta}$.

Since $\hat{\beta} \sim \mathcal{N}(\beta,\; \sigma^2(X^TX)^{-1})$, the linear combination

$$(x^*)^T\hat{\beta} - (x^*)^T\beta$$

is normally distributed with mean 0 and variance $\sigma^2\,(x^*)^T(X^TX)^{-1}x^*$. That is,

$$(x^*)^T\hat{\beta} - (x^*)^T\beta \;\sim\; \mathcal{N}\Big(0,\; \sigma^2\,\big[(x^*)^T(X^TX)^{-1}x^*\big]\Big).$$

$$\frac{(x^*)^T\hat{\beta} - (x^*)^T\beta}{\sqrt{\sigma^2\big[(x^*)^T(X^TX)^{-1}x^*\big]}} \;\sim\; \mathcal{N}(0,1)$$

since σ is unknown, need to use:

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

and this quantity is independent of $\hat{\beta}$.

Hence, linear combination of $\hat{\beta}$, which is $\frac{(x^*)^T\hat{\beta}-(x^*)^T\beta}{\sqrt{\sigma^2[(x^*)^T(X^TX)^{-1}x^*]}}$ is independent of $\frac{\|\,y-X\hat{\beta}\|^2}{\sigma^2}$. Therefore,

$$T = \frac{\frac{(x^*)^T\hat{\beta}-(x^*)^T\beta}{\sqrt{\sigma^2[(x^*)^T(X^TX)^{-1}x^*]}}}{\sqrt{\frac{\|\,y-X\hat{\beta}\|^2}{\sigma^2}}\sqrt{\frac{1}{n-p}}} \sim t_{n-p}$$

$$T \;=\; \frac{(x^*)^T\hat{\beta} - (x^*)^T\beta}{\sqrt{(x^*)^T(X^TX)^{-1}x^*}\;\;\big(\|y - X\hat{\beta}\|/\sqrt{n-p}\big)} \;\sim\; t_{\,n-p}.$$

For a $(1-\alpha)$ level confidence interval, note that

$$P\Big(-t_{n-p,\,1-\frac{\alpha}{2}} \;\le\; T \;\le\; t_{n-p,\,1-\frac{\alpha}{2}}\Big) = 1-\alpha.$$

Rewriting in terms of $(x^*)^T\beta$, we obtain

$$P\Big((x^*)^T\hat{\beta} - t_{n-p,\,1-\frac{\alpha}{2}}\,\frac{\|y - X\hat{\beta}\|}{\sqrt{n-p}}\,\sqrt{(x^*)^T(X^TX)^{-1}x^*} \;\;\le\; (x^*)^T\beta \;\le\; (x^*)^T\hat{\beta} + t_{n-p,\,1-\frac{\alpha}{2}}\,\frac{\|y - X\hat{\beta}\|}{\sqrt{n-p}}\,\sqrt{(x^*)^T(X^TX)^{-1}x^*}\Big)$$

$$= 1 - \alpha.$$

Thus, the $(1-\alpha)$-level confidence interval for $(x^*)^T\beta$ is

$$\Big[(x^*)^T\hat{\beta} \;-\; t_{n-p,\,1-\frac{\alpha}{2}}\,\frac{\|y - X\hat{\beta}\|}{\sqrt{n-p}}\,\sqrt{(x^*)^T(X^TX)^{-1}x^*},\;\; (x^*)^T\hat{\beta} \;+\; t_{n-p,\,1-\frac{\alpha}{2}}\,\frac{\|y - X\hat{\beta}\|}{\sqrt{n-p}}\,\sqrt{(x^*)^T(X^TX)^{-1}x^*}\Big].$$

Question 6

(a)

Starting with

$$E\big[(\widehat{\theta} - \theta)^2\big],$$

we can add and subtract $E[\widehat{\theta}]$ inside the parentheses, then expand:

$$(\widehat{\theta} - \theta)^2 \;=\; \big(\widehat{\theta} - E[\widehat{\theta}] + E[\widehat{\theta}] - \theta\big)^2.$$

Taking expectation on both sides and expanding the square:

$$E\big[(\widehat{\theta} - \theta)^2\big] \;=\; E\Big[\big(\widehat{\theta} - E[\widehat{\theta}]\big)^2\Big] \;+\; 2\,E\Big[\big(\widehat{\theta} - E[\widehat{\theta}]\big)\,\big(E[\widehat{\theta}] - \theta\big)\Big] \;+\; E\Big[\big(E[\widehat{\theta}] - \theta\big)^2\Big].$$

Here for each term:

$\big(E[\widehat{\theta}] - \theta\big)$ is a constant (it does not depend on the random variable $\widehat{\theta}$).

So $E\Big[\big(E[\widehat{\theta}] - \theta\big)^2\Big] = \big(E[\widehat{\theta}] - \theta\big)^2$, since $\big(E[\widehat{\theta}] - \theta\big)$ is a constant.

For the term $2E\Big[\big(\widehat{\theta} - E[\widehat{\theta}]\big)\,\big(E[\widehat{\theta}] - \theta\big)\Big]$

$\big(E[\widehat{\theta}] - \theta\big)$ is a constant (not random), so it can be factored out the expectation:

$$2E\Big[\big(\widehat{\theta} - E[\widehat{\theta}]\big)\,\big(E[\widehat{\theta}] - \theta\big)\Big] = 2\big(E[\widehat{\theta}] - \theta\big)\,E\big[\widehat{\theta} - E[\widehat{\theta}]\big].$$

By definition,

$$E\big[\widehat{\theta} - E[\widehat{\theta}]\big] = E[\widehat{\theta}] - E[\widehat{\theta}] = 0.$$

Hence the entire product is 0:

$$2\big(E[\widehat{\theta}] - \theta\big)\,E\big[\widehat{\theta} - E[\widehat{\theta}]\big] = 2\big(E[\widehat{\theta}] - \theta\big) \times 0 = 0.$$

$E\big[\big(\widehat{\theta} - E[\widehat{\theta}]\big)^2\big]$ is $\mathrm{Var}(\widehat{\theta})$ by definition.

Therefore:

$$E\big[(\widehat{\theta} - \theta)^2\big] \;=\; E\Big[\big(\widehat{\theta} - E[\widehat{\theta}]\big)^2\Big] \;+\; 2\,E\Big[\big(\widehat{\theta} - E[\widehat{\theta}]\big)\,\big(E[\widehat{\theta}] - \theta\big)\Big] \;+\; E\Big[\big(E[\widehat{\theta}] - \theta\big)^2\Big].$$

$$E\big[(\widehat{\theta} - \theta)^2\big] \;=\; \mathrm{Var}(\widehat{\theta}) \;+\; 0 + \big(E[\widehat{\theta}] - \theta\big)^2.$$

$$E\big[(\widehat{\theta} - \theta)^2\big] \;=\; \mathrm{Var}(\widehat{\theta}) \;+\; \big(E[\widehat{\theta}] - \theta\big)^2.$$

(b)

Write each $x_i - \theta$ as $(x_i - \bar{x}) + (\bar{x} - \theta)$. Then

$$(x_i - \theta)^2 = (x_i - \bar{x})^2 \; + \; 2(x_i - \bar{x})(\bar{x} - \theta) \; + \; (\bar{x} - \theta)^2.$$

Summing over i:

$$\sum_{i=1}^{n}(x_i - \theta)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n} 2(\bar{x} - \theta)(x_i - \bar{x}) + \sum_{i=1}^{n}(\bar{x} - \theta)^2.$$

$$\sum_{i=1}^{n}(x_i - \theta)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2(\bar{x} - \theta) \sum_{i=1}^{n}(x_i - \bar{x}) + n\,(\bar{x} - \theta)^2.$$

Since $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, Thus,

$$\sum_{i=1}^{n}(x_i - \theta)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n\,(\bar{x} - \theta)^2.$$

(c)

For each i, write

$$X_i - \theta \; = \; \big(X_i - \overline{X}\big) \; + \; \big(\overline{X} - \theta\big).$$

Then expand the squared norm:

$$\|\, X_i - \theta \,\|^2 \; = \; \|\, X_i - \overline{X} \,\|^2 \; + \; 2\,\big(X_i - \overline{X}\big)^T\big(\overline{X} - \theta\big) \; + \; \|\overline{X} - \theta\|^2.$$

Summing the above expression from $i = 1$ to n yields:

$$\sum_{i=1}^{n} \|\, X_i - \theta \,\|^2 \; = \; \sum_{i=1}^{n} \|\, X_i - \overline{X} \,\|^2 \; + \; 2\sum_{i=1}^{n}\big(X_i - \overline{X}\big)^T\big(\overline{X} - \theta\big) \; + \; \sum_{i=1}^{n} \|\overline{X} - \theta\|^2.$$

We observe:

$$\sum_{i=1}^{n}\big(X_i - \overline{X}\big) \; = \; \sum_{i=1}^{n} X_i - n\,\overline{X} \; = \; n\,\overline{X} - n\,\overline{X} \; = \; 0.$$

Hence

$$\sum_{i=1}^{n}\big(X_i - \overline{X}\big)^T\big(\overline{X} - \theta\big) \; = \; (\overline{X} - \theta)^T \sum_{i=1}^{n}\big(X_i - \overline{X}\big) \; = \; (\overline{X} - \theta)^T \cdot 0 \; = \; 0.$$

Since the cross term is zero, the remaining sum is

$$\sum_{i=1}^{n} \|\, X_i - \theta \,\|^2 \; = \; \sum_{i=1}^{n} \|\, X_i - \overline{X} \,\|^2 \; + \; \sum_{i=1}^{n} \|\overline{X} - \theta\|^2.$$

And $\|\overline{X} - \theta\|^2$ does not depend on i, so this last sum equals $n\,\|\overline{X} - \theta\|^2$. Therefore,

$$\sum_{i=1}^{n} \|\, X_i - \theta \,\|^2 \; = \; \sum_{i=1}^{n} \|\, X_i - \overline{X} \,\|^2 \; + \; n\,\|\overline{X} - \theta\|^2,$$

(d)

We write

$$y - X\beta \; = \; \big(y - X\hat{\beta}\big) \; + \; \big(X\hat{\beta} - X\beta\big).$$

Hence, by expanding the squared norm,

$$\| y - X\beta \|^2 \;=\; \left\| (y - X\hat{\beta}) + (X\hat{\beta} - X\beta) \right\|^2 \;=\; \| y - X\hat{\beta}\|^2 \;+\; 2\,(y - X\hat{\beta})^T\,(X\hat{\beta} - X\beta) \;+\; \| X\hat{\beta} - X\beta \|^2.$$

Consider the term

$$(y - X\hat{\beta})^T\,(X\hat{\beta} - X\beta).$$

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

$$y - X\hat{\beta} \;=\; y - X\,(X^TX)^{-1}X^T\,y \;=\; \big(I_n - X(X^TX)^{-1}X^T\big)\,y.$$

$$X\hat{\beta} - X\beta = X(\hat{\beta} - \beta).$$

Hence,

$$(y - X\hat{\beta})^T\,(X\hat{\beta} - X\beta) \;=\; y^T\big(I_n - X(X^TX)^{-1}X^T\big)^T\,X(\hat{\beta} - \beta).$$

And $\big(I_n - X(X^TX)^{-1}X^T\big)$ is symmetric:

$$y^T\big(I_n - X(X^TX)^{-1}X^T\big)X(\hat{\beta} - \beta) = y^T\big(X - X(X^TX)^{-1}X^TX\big)(\hat{\beta} - \beta) = y^T(X - X)(\hat{\beta} - \beta) = 0$$

$$(y - X\hat{\beta})^T\,(X\hat{\beta} - X\beta) = 0.$$

Since the cross term vanishes, we are left with

$$\| y - X\beta \|^2 \;=\; \| y - X\hat{\beta}\|^2 \;+\; \| X\hat{\beta} - X\beta \|^2,$$

(e)

All the identities in parts (a)–(d) can be viewed as special cases of a single projection principle. The essential idea is whenever we decompose a vector (or random variable) into two orthogonal or uncorrelated parts, the squared norm (or variance) of the original object equals the sum of the squared norms (variances) of the two parts.

Parts (a)–(c): Orthogonal Decomposition onto a One-Dimensional Subspace.

- **(a) Random variable version.** We can think of θ (a random variable) as a point in an infinite-dimensional function space. The decomposition

  $$E\big[(\theta - \theta_0)^2\big] \;=\; \mathrm{var}(\theta) \;+\; \big(E[\theta] - \theta_0\big)^2$$

  arises from "projecting" θ onto the constant random variable $E[\theta]$. The difference $\theta - E[\theta]$ is then uncorrelated (orthogonal in an L^2 sense) to that constant part. Hence we get a sum of squares decomposition.

- **(b) Real numbers $x_1, \ldots, x_n$.** Here, we view $\big(x_1,\, x_2, \ldots, x_n\big)$ as a vector in R^n. The "projection" is onto the subspace spanned by the vector of all ones. Minimizing $\sum_i (x_i - \theta)^2$ yields $\theta = \bar{x}$. The difference $(x_i - \bar{x})$ is orthogonal to that constant vector. Thus,

  $$\sum_{i=1}^{n}(x_i - \theta)^2 \;=\; \sum_{i=1}^{n}(x_i - \bar{x})^2 \;+\; n\,(\bar{x} - \theta)^2.$$

- **(c) Vectors $X_1, \ldots, X_n \in R^d$.** This is the same idea as (b) but in higher-dimensional space. Now each X_i is itself a vector, and $\bar{X}$ is the centroid (mean). We again split

  $$X_i - \theta = (X_i - \bar{X}) + (\bar{X} - \theta),$$

  and rely on the orthogonality of the "centered vectors" to the constant shift $\bar{X} - \theta$. The net result is the same Pythagorean sum of squared norms.

Part (d): Orthogonal Projection in Regression. In linear regression,

$$y \;\in\; R^n, \quad X \;\in\; R^{n\times p}, \quad \hat{\beta} = (X^TX)^{-1}X^Ty,$$

we interpret $X\hat{\beta}$ as the orthogonal *projection* of y onto the column space of X. The residual $y - X\hat{\beta}$ is orthogonal to that column space. Geometrically, one can visualize y in R^n, the subspace spanned by the columns of X, and the perpendicular drop from y to that subspace is $y - X\hat{\beta}$. Thus:

$$\|\, y - X\beta \,\|^2 \;=\; \|\, y - X\hat{\beta}\|^2 \;+\; \|\, X\hat{\beta} - X\beta \,\|^2,$$

mirroring the same Pythagorean relationship.

In each scenario, we are splitting a "distance" (or squared difference) into: The distance between the data and the prediction (or estimated value), and the distance between the prediction using the real parameter and the prediction using the estimated parameter.

Because of the orthogonality (or uncorrelatedness) between these two components, the total distance is the sum of their squares.