Tyler Singleton
May 28, 2025

Problem 1:

Part A

To show that ordinary least–squares estimation in model M1 behaves in the usual way, introduce the centred regressor $z_i = ax_i + b$. Since $a > 0$ and b are known constants, fitting

$$y_i = \beta_0 + \beta_1 z_i + \varepsilon_i, \qquad E(\varepsilon_i \mid \mathbf{x}) = 0,\ \mathrm{Var}(\varepsilon_i \mid \mathbf{x}) = \sigma^2$$

This is mathematically equivalent to an ordinary simple-regression problem with covariates z. The least–squares slope is:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})^2}$$

Because $z_i - \bar{z} = a(x_i - \bar{x})$ and:

$$\sum (z_i - \bar{z})(y_i - \bar{y}) = a \sum (x_i - \bar{x})(y_i - \bar{y})$$

While:

$$\sum (z_i - \bar{z})^2 = a^2 \sum (x_i - \bar{x})^2$$

It follows that:

$$\widehat{\beta}_1 = \frac{1}{a} \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (wx_i - \bar{x})^2}$$

Which is a form that contains the scale factor $1/a$ but no occurrence of b. So, changing b leaves $\widehat{\beta}_1$ unchanged.

Part B

Conditional on the observed x-values, write $y_i = \beta_0 + \beta_1 z_i + \varepsilon_i$. Re-expressing the numerator of $\widehat{\beta}_1$ gives

$$\sum (z_i - \bar{z})\varepsilon_i$$

Where the conditional expectation is zero because the errors have mean zero and are independent of x. So, $E(\widehat{\beta}_1 \mid \mathbf{x}) = \beta_1$, and the estimator is unbiased.

Part C The fitted intercept is:

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{z} = \bar{y} - \widehat{\beta}_1 (a\bar{x} + b)$$

Clearly, this depends explicitly on b. Altering b shifts $\bar{z}$ and also shifts $\widehat{\beta}_0$ by $-\widehat{\beta}_1 \, \Delta b$.

Part D

Because $\widehat{\beta}_1$ is linear in errors, its conditional variance is:

$$\mathrm{Var}(\widehat{\beta}_1 \mid \mathbf{x}) = \frac{\sigma^2}{\sum_i (z_i - \bar{z})^2} = \frac{\sigma^2}{a^2 \sum_i (x_i - \bar{x})^2}$$

We can replace the unknown σ^2 by the usual unbiased residual estimator:

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 z_i)^2$$

Which results in the the unbiased variance estimator:

$$\widehat{\mathrm{Var}}(\widehat{\beta}_1 \mid \mathbf{x}) = \frac{\widehat{\sigma}^2}{\sum_i (z_i - \bar{z})^2}$$

Unbiasedness follows from $E(\widehat{\sigma}^2 \mid \mathbf{x}) = \sigma^2$.

Part E

Because $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{z}$, linearity of covariance yields:

$$\mathrm{Cov}(\widehat{\beta}_1, \widehat{\beta}_0 \mid \mathbf{x}) = -\bar{z}\ \mathrm{Var}(\widehat{\beta}_1 \mid \mathbf{x}) = -\frac{(a\bar{x} + b)\,\sigma^2}{a^2 \sum_i (x_i - \bar{x})^2}$$

Part F

The conditional correlation vanishes precisely when the covariance vanishes; this occurs when $a\bar{x} + b = 0$. So, if we have:

$$b = -a\,\bar{x}$$

Centering the transformed regressor at zero renders $\widehat{\beta}_0$ and $\widehat{\beta}_1$ is uncorrelated given the design.

Part G

The residual mean-square just defined:

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 z_i)^2$$

Which is unbiased for σ^2 because, under the model assumptions, the numerator is a sum of $n-2$ independent squared centered errors with expectation $(n-2)\sigma^2$. Soo, $E(\widehat{\sigma}^2 \mid \mathbf{x}) = \sigma^2$.

Problem 2:

Let the design points $x_1, \ldots, x_n$ be fixed constants and impose the usual homoskedastic linear-model assumptions: the response is generated by $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with independent errors that satisfy $E(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2 < \infty$. No distributional form for the errors is stipulated unless stated explicitly. Because the x_i are non-random, ordinary least squares treats the vector x as conditioning information. The slope estimator can be written:

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \beta_1 + \frac{\sum_i (x_i - \bar{x})\varepsilon_i}{\sum_i (x_i - \bar{x})^2}$$

Taking expectation over the error distribution while holding the design fixed yields $E(\hat{\beta}_1) = \beta_1$; so $\hat{\beta}_1$ is always unbiased, regardless of the value of β_0. The Gauss–Markov theorem asserts that, under the stated assumptions, $\hat{\beta}_1$ has the smallest conditional variance among all linear unbiased estimators of β_1. If we restrict attention to this linear class of models, no other unbiased rule (regardless of how it might involve β_0) can outperform

OLS.

Problem 3:

Using a random number generator to create a vector of 20 indices between 1 and 32 was generated:

Index	**Car**	y **(MPG)**	x_1 **(cubic in.)**
1	Apollo	18.90	350.0
2	Omega	17.00	350.0
4	Monarch	18.25	351.0
5	Duster	20.07	225.0
6	Jenson Conv.	11.20	440.0
8	Monza	21.47	262.0
10	Corolla SR-5	30.40	96.9
11	Camaro	16.50	350.0
12	Datson B210	36.50	85.3
14	Pacer	19.70	258.0
15	Babcat	20.30	140.0
18	Imperial	14.89	440.0
21	Starfire	23.54	231.0
23	Trans AM	16.59	400.0
24	Corolla E-5	31.90	96.9
25	Astre	29.40	140.0
26	Mark IV	13.27	460.0
28	Charger SE	19.73	318.0
29	Cougar	13.90	351.0
30	Elite	13.27	351.0

Table 1: $(n = 20)$ random observations.

Part A

Sample statistics:

$$\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{1i} = 284.8050, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 20.3390$$

$$S_{xx} = \sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2 = 279\,198.5495, \quad S_{xy} = \sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(y_i - \bar{y}) = -14\,307.0679$$

So:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = -0.05124, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1 = 34.93336$$

Which results in the fitted model:

$$\hat{y} = 34.9334 - 0.05124\,x_1$$

Part B

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 158.3883, \qquad \text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 733.1420$$

$$\text{SST} = \text{SSR} + \text{SSE} = 891.5304, \qquad \text{MSE} = \frac{\text{SSE}}{n-2} = 8.7994$$

$$\text{MSR} = \frac{\text{SSR}}{1} = 733.1420, \qquad F = \frac{\text{MSR}}{\text{MSE}} = 83.32$$

Source	**df**	**SS**	**MS**	F
Regression	1	733.1420	733.1420	83.32
Residual	18	158.3883	8.7994	
Total	19	891.5304		

The critical value $F_{1,18;\,0.05} = 4.41$. Because $F = 83.32 > 4.41$, engine displacement explains a significant portion of the variability in gas mileage at the 5% level.

Part C

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{733.1420}{891.5304} = 0.8223$$

So, approximately 82.2% of the variation in y is accounted for by x_1 in this subsample.

Part D

$$\hat{y}(275) = 34.9334 - 0.05124(275) = 20.8414, \quad s = \sqrt{\text{MSE}} = 2.9664$$

$$\text{SE}_{\text{mean}} = s\sqrt{\frac{1}{n} + \frac{(275 - \bar{x}_1)^2}{S_{xx}}} = 0.6656, \quad t_{18;\,0.025} = 2.101$$

Then:

$$19.44 \le E[Y \mid x_1 = 275] \le 22.24$$

Part E

$$\text{SE}_{\text{pred}} = s\sqrt{1 + \frac{1}{n} + \frac{(275 - \bar{x}_1)^2}{S_{xx}}} = 3.0401$$

$$14.45 \le Y_{\text{new}} \le 27.23$$

Part F

The confidence interval width is $22.24 - 19.44 = 2.80$ MPG, whereas the prediction interval width is $27.23 - 14.45 = 12.77$ MPG. The prediction interval is wider because it incorporates the random variability of an individual future observation in addition to uncertainty in estimating the conditional mean response.