Jane Downer
Math 484-01
September 25, 2021

Homework 2

Problem 1 In regression through the origin (RTO) model $Y = \beta_1 X + \epsilon$, it is assumed $\epsilon \sim N(0, \sigma^2)$. If the common variance σ^2 is unknown, a possible estimator for σ^2 is $\hat{\sigma}^2 = SSE/(n-1)$. Prove that σ^2 is an unbiased estimator for σ^2.

i.e. Prove $E(\hat{\sigma}^2) = \sigma^2$

$$E(\hat{\sigma}^2) = E\left(\frac{SSE}{n-1}\right)$$

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$= \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 X_i)^2$$

$$= \sum_{i=1}^{n} Y_i^2 - \sum_{i=1}^{n} (\hat{\beta}_1 X_i)^2$$

$$E(SSE) = E\left(\sum_{i=1}^{n} Y_i^2\right) - E\left(\sum_{i=1}^{n} (\hat{\beta}_1 X_i)^2\right)$$

$$= \sum_{i=1}^{n} E(Y_i^2) - \sum_{i=1}^{n} E(\hat{\beta}_1 X_i)^2$$

$$= \sum_{i=1}^{n} \left(\mathrm{var}(Y_i) + (E(Y_i))^2\right) - \sum_{i=1}^{n} \left(\mathrm{var}(\hat{\beta}_1 X_i) + (E(\hat{\beta}_1 X_i))^2\right)$$

$$= \sum_{i=1}^{n} \left(\sigma^2 + (E(Y_i))^2\right) - \sum_{i=1}^{n} \left(X_i^2 \cdot \frac{\sigma^2}{\sum X_i^2} + (E(\hat{\beta}_1 X_i))^2\right)$$

$$= n\sigma^2 - \sum_{i=1}^{n} \left(X_i^2 \cdot \frac{\sigma^2}{\sum X_i^2}\right)$$

$$= n\sigma^2 - \sigma^2$$

$$= (n-1)\sigma^2$$

$$\longrightarrow E(SSE) = (n-1)\sigma^2$$

$$\therefore E\left(\frac{SSE}{n-1}\right) = \sigma^2$$

✓✓

Problem 2 Consider the *Supervisor Performance Data* in Table 3.3 on page 60 of the TEXT (Table 3.3 attached).

1) Estimate the regression coefficients vector $\hat{\beta}$.
2) Verify that $\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$ for this dataset.
3) Does $\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$ hold true in general for multiple linear regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$? Prove or disprove it.
4) Now consider $p = 2$, and only use X_3 and X_4 two predictors. The model becomes

$$Y = \beta_0 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

Use the 3-step method described on page 63 to obtain the coefficient for X_3, and compare it with the coefficient of X_3 by regressing Y on X_3 and X_4 using the 2-predictor model above. Are they the same? Explain why or why not.

(1) Regression coefficient vector $\hat{\beta} = (X^T X)^{-1} X^T Y$

where $X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & x_{14} & x_{15} & x_{16} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & x_{n4} & x_{n5} & x_{n6} \end{bmatrix}$

$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

n = # observations

Matrix multiplication via Numpy (Python):

$$\boxed{\hat{\beta} = \begin{bmatrix} 10.787 \\ 0.613 \\ -0.073 \\ 0.320 \\ 0.082 \\ 0.038 \\ -0.217 \end{bmatrix}}$$

(2)

$$\hat{Y} = X\hat{\beta}$$

Let: $H = X(X^T X)^{-1} X^T$

Then: $X\hat{\beta} = X(X^T X)^{-1} X^T Y$

So: $\hat{Y} = HY$

Matrix multiplication via Numpy (python):

$$\hat{Y} = \begin{bmatrix} 51.11 \\ 61.353 \\ \vdots \\ 76.878 \end{bmatrix}$$

Already given:

$$Y = \begin{bmatrix} 43 \\ 63 \\ 61 \\ 43 \\ 71 \\ 67 \\ 67 \\ 68 \\ 81 \\ 65 \\ 50 \\ 64 \\ 40 \\ 66 \\ 48 \\ 82 \end{bmatrix}$$

$$\sum_{i=1}^{n} \hat{Y}_i = 1939$$

$$\sum_{i=1}^{n} Y_i = 1939$$

✓✓

(3) Yes. Proof:

$$\hat{Y} = HY$$

$$= \begin{bmatrix} h_{11}y_1 + h_{22}y_2 + \cdots + h_{nn}y_n \\ \vdots \qquad\qquad \vdots \\ h_{n1}y_n + h_{n2}y_2 + \cdots + h_{nn}y_n \end{bmatrix}$$

$$\Sigma \hat{Y}_i = \sum_{i=1}^{n} h_{i1}y_1 + \sum_{i=1}^{n} h_{i2}y_2 + \cdots + \sum_{i=1}^{n} h_{in}y_n$$

We can show that $\sum_j h_{ij} = 1 \quad \forall j$ and $\sum_i h_{ij} = 1 \quad \forall i$:

$$H = X(X^TX)^{-1}X^T$$

$$HX = X\cancel{(X^TX)^{-1}}\cancel{(X^TX)}$$

$$= X$$

$$= \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ \vdots & & & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix}_{n \times (p+1)}$$

call the columns U_i $\Rightarrow$ $HU = U$

$$H * \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$(HU)_i = U_i$$

$\sum_j h_{ij} = 1 \quad \forall j$ — sum of row

$\sum_i h_{ij} = 1 \quad \forall i$ — sum of column

So: $$\Sigma \hat{Y}_i = \sum_{i=1}^{n} h_{i1}y_1 + \sum_{i=1}^{n} h_{i2}y_2 + \cdots + \sum_{i=1}^{n} h_{in}y_n$$

$$= y_1 + y_2 + \cdots + y_n$$

$$= \Sigma Y_i \quad \checkmark\checkmark$$

$\beta_0 = -6.136$

$\beta_3 = 0.234$

$\beta_4 = 0.031$

$p = 2$

$Y = \beta_0 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$

$\hookrightarrow \beta_3 = ?$

3-step method:

i. $\hat{Y} = \hat{\beta}_{01} + \hat{\beta}_{11} X_4$

$$\hat{\beta}_{11} = \frac{\Sigma (X_{4i} - \bar{X}_4)(y_i - \bar{y})}{\Sigma (X_{4i} - \bar{X}_4)^2} = 0.691$$

$$\hat{\beta}_{01} = \bar{Y} - \hat{\beta}_{11} \bar{X}_4 = 19.972$$

$e_{Y \cdot X_4}$: $\begin{bmatrix} -19.123 \\ -0.505 \\ -1.488 \\ 3.714 \\ \vdots \\ -14.780 \\ 2.748 \\ -7.213 \\ 11.821 \\ 17.804 \end{bmatrix}$

$\bar{e}_{Y \cdot X_4} \approx 0$

ii. $\hat{X}_3 = \hat{\beta}_{02} + \hat{\beta}_{12} X_4$

$$\hat{\beta}_{12} = \frac{\Sigma (X_{4i} - \bar{X}_4)(X_{3i} - \bar{X}_3)}{\Sigma (X_{4i} - \bar{X}_4)^2} = 0.723$$

$$\hat{\beta}_{02} = \bar{X}_3 - \hat{\beta}_{12} \bar{X}_4 = 9.637$$

$e_{X3 \cdot X4}$: $\begin{bmatrix} -14.74 \\ -1.186 \\ 4.415 \\ -1.679 \\ \vdots \\ -10.261 \\ 6.523 \\ -1.51 \\ 5.692 \\ 3.091 \end{bmatrix}$

$\bar{e}_{X3 \cdot X4} \approx 0$

iii. $\hat{e}_{Y \cdot X_4} = \hat{\beta}_{0e} + \hat{\beta}_{1e} e_{X3 \cdot X4}$

$$\hat{\beta}_{1e} = \frac{\Sigma (e_{X3 \cdot X4, i} - \bar{e}_{X3 \cdot X4})(e_{Y \cdot X4, i} - \bar{e}_{Y \cdot X4})}{\Sigma (e_{X3 \cdot X4, i} - \bar{e}_{X3 \cdot X4})^2} = 0.432$$

$$\hat{\beta}_{0e} = \bar{e}_{Y \cdot X_4} - \hat{\beta}_{1e} \cdot e_{X3 \cdot X4} = 0.0$$

compare against results from second method:

Regressing Y on X_3 and X_4:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4$$

Regression coefficient vector $\hat{\beta} = (X^T X)^{-1} X^T Y$

where $X = \begin{bmatrix} 1 & x_{13} & x_{14} \\ \vdots & \vdots & \vdots \\ 1 & x_{n3} & x_{n4} \end{bmatrix}$

$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

n = # observations

Both methods yield the same estimate for the coefficient of X_3

Matrix multiplication via Numpy (Python):

$$\hat{\beta} = \begin{bmatrix} 15.809 \\ 0.432 \\ 0.379 \end{bmatrix} \begin{matrix} \leftarrow \hat{\beta}_0 \\ \leftarrow \hat{\beta}_3 \\ \leftarrow \hat{\beta}_4 \end{matrix}$$

Explanation: The residuals from step 1 are the parts of Y not linearly related to X_4.
The residuals from step 2 are the parts of X_3 not linearly related to X_4.
The linear relationship between these two arrays is the effect of X_3 on Y after removing the effects of X_4.

Problem 3

3.3 Table 3.10 shows the scores in the final examination F and the scores in two preliminary examinations P_1 and P_2 for 22 students in a statistics course. The data can be found at the book's Website.

(a) Fit each of the following models to the data:

$$\begin{aligned}
&\text{Model 1:} \quad F = \beta_0 + \beta_1 P_1 \qquad\qquad + \varepsilon \\
&\text{Model 2:} \quad F = \beta_0 \qquad\qquad + \beta_2 P_2 + \varepsilon \\
&\text{Model 3:} \quad F = \beta_0 + \beta_1 P_1 + \beta_2 P_2 + \varepsilon
\end{aligned}$$

(b) Test whether $\beta_0 = 0$ in each of the three models.

(c) Which variable individually, P_1 or P_2, is a better predictor of F?

(d) Which of the three models would you use to predict the final examination scores for a student who scored 78 and 85 on the first and second preliminary examinations, respectively? What is your prediction in this case?

(a) General SLR formulas:

$$\hat{\beta}_1 = \frac{\sum(\text{predictor}_i - \overline{\text{predictor}})(\text{response}_i - \overline{\text{response}})}{\sum(\text{response}_i - \overline{\text{response}})^2}$$

$$\hat{\beta}_0 = \overline{\text{response}} - \hat{\beta}_1 \cdot \text{predictor}$$

Model 1: $\hat{\beta}_0 = -22.382$

$\hat{\beta}_1 = 1.261$

Model 2: $\hat{\beta}_0 = -1.831$

$\hat{\beta}_1 = 1.004$

Model 3: $\hat{B} = \begin{bmatrix} -14.5 \\ 0.488 \\ 0.672 \end{bmatrix}$

(b)

$$\begin{cases} H_0: \beta_0 = 0 \\ H_a: \beta_0 \neq 0 \end{cases} \qquad t = \frac{\hat{\beta}_0 - 0}{s.e.(\hat{\beta}_0)} \qquad t_{0.025,20} \approx t_{.025,19} \approx 2.09$$

Model 1:

$$SSE = \Sigma(F_i - \hat{F}_i)^2 = 516.343$$

$$\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{516.343}{20} = 25.817$$

$$\hat{\sigma} = 5.08$$

$$s.e.(\beta_0) = \hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x_i - \bar{x})^2}\right] = 133.725$$

$$t = \frac{-22.382 - 0}{133.725} \approx \underline{-0.167}$$

Model 2:

$$SSE = \Sigma(F_i - \hat{F}_i)^2 = 365.46$$

$$\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{365.46}{20} = 18.273$$

$$\hat{\sigma} = 4.275$$

$$s.e.(\beta_0) = \hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x_i - \bar{x})^2}\right] = 57.181$$

$$t = \frac{-1.831 - 0}{133.725} \approx \underline{-0.032}$$

None of these fall in rejection. We fail to reject H_0 in all three models.

MLR:

$$P = \begin{bmatrix} 1 & P_{11} & P_{21} \\ \vdots & \vdots & \vdots \\ 1 & P_{1n} & P_{2n} \end{bmatrix}$$

$$C = (P^T P)^{-1} \longrightarrow C_{00} = 22$$

$$SSE = \Sigma(F_i - \hat{F}_i)^2 = 296.83$$

$$\hat{\sigma}^2 = \frac{SSE}{n-p-1} = \frac{296.83}{19} = 15.623$$

$$\hat{\sigma} = 3.953$$

$$s.e.(\hat{\beta}_0) = \hat{\sigma}\sqrt{C_{00}} = 18.541$$

$$t = \frac{-14.5 - 0}{18.541} = \underline{-0.782}$$

(c) SSE was lower when using P2 as a predictor.

P2 is a better predictor than P1.

(d) I would use Model 3.

$$\hat{F} = -14.5 + 78 * 0.488 + 85 * 0.672$$

$$= \boxed{80.7}$$

Problem 4 Verify that for the multiple linear regression model $Y = X\boldsymbol{\beta} + \epsilon$ with one predictor variable, the least square estimate of $\boldsymbol{\beta}$ using the matrix form gives the same result as in SLR. Then use the matrix form to derive the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ and verify the variances $\text{Var}(\hat{\beta}_0)$ and $\text{Var}(\hat{\beta}_1)$ are the same as their counterparts derived in SLR.

$$Q = (Y - XB)^T (Y - XB)$$

$$= Y^T Y + B^T X^T X B - 2B^T X^T Y$$

$$\frac{\partial Q}{\partial \hat{B}} = 2X^T X B - 2X^T Y$$

set to zero $\longrightarrow$ $2X^T X \hat{B} - 2X^T Y = 0$

$$(X^T X)\hat{B} = X^T Y$$

$$\hat{B} = (X^T X)^{-1} X^T Y$$

equivalent

SLR:

$$Y = \beta_0 + \beta_1 X$$

$$\Sigma Y = n \cdot \beta_0 + (\beta_1 X_1 + \dots + \beta_1 X_n)$$

$$= \beta_0 n + \beta_1 \Sigma X_i$$

$$\Sigma X^T Y = \beta_0 \Sigma X_i + \beta_1 \Sigma X_i^2$$

$$\begin{bmatrix} \Sigma \hat{Y} \\ \Sigma X^T Y \end{bmatrix} = \begin{bmatrix} n & \Sigma X_i \\ \Sigma X_i & \Sigma X_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \Longrightarrow X^T Y = (X^T X)\hat{\beta}$$

$$\downarrow$$

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

continued
↓

$$X^T X = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

$$C = (X^T X)^{-1} = \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} & \frac{-\bar{x}}{S_{xx}} \\ \frac{-\bar{x}}{S_{xx}} & \frac{1}{S_{xx}} \end{bmatrix}$$

$$\text{var}(\beta_0) = \sigma^2 \cdot C_{00} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) \leftarrow$$

same as in SLR

$$\text{var}(\beta_1) = \sigma^2 \cdot C_{11} = \sigma^2 \cdot \frac{1}{S_{xx}} \leftarrow$$

Problem 5

Table 3.14 Regression Output When Salary is Related to Four Predictor Variables

ANOVA Table				
Source	Sum of Squares	df	Mean Square	F-Test
Regression	23665352	4	5916338	22.98
Residuals	22657938	88	257477	

Coefficients Table				
Variable	Coefficient	s.e.	t-Test	p-value
Constant	3526.4	327.7	10.76	0.000
Gender	722.5	117.8	6.13	0.000
Education	90.02	24.69	3.65	0.000
Experience	1.2690	0.5877	2.16	0.034
Months	23.406	5.201	4.50	0.000
$n = 93$	$R^2 = 0.515$	$R_a^2 = 0.489$	$\hat{\sigma} = 507.4$	df = 88

Gender	An indicator variable (1 = man and 0 = woman)
Education	Years of schooling at the time of hire
Experience	Number of months of previous work experience
Months	Number of months with the company

Consider the regression model that generated the output in Table 3.14 to be a full model. Now consider the reduced model in which Salary is regressed on only Education. The ANOVA table obtained when fitting this model is shown in Table 3.15. Conduct a single test to compare the full and reduced models. What conclusion can be drawn from the result of the test? (Use $\alpha = 0.05$.)

$$\begin{cases} H_0: \beta_{Gender} = \beta_{experience} = \beta_{months} = 0 \\ H_a: \text{at least one is not zero} \end{cases}$$

$F(p-q, n-p-1) = F(4,88) \approx 2.5$

$$F^* = \frac{\text{sum of squares (regression)} / df_{regression}}{\text{sum of squares (residuals)} / df_{residuals}}$$

$$= \frac{23665352/4}{22657938/88}$$

$$= 22.98 > 2.5$$

we reject H_0

$\Downarrow$

We prefer the full model