Homework 3

Brian Cervantes Alvarez
February 3, 2024
ST552 Statistical Methods

Problem 1

Part A

The design matrix X for $J = 3$ and $K = 3$ is as follows:

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

Part B

```r
y <- rnorm(9)
beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y
beta_hat
```

```
           [,1]
[1,] -0.3238472
[2,] -0.6446748
[3,] -0.7540890
```

Part C

```r
sigma2 <- var(y)
var_beta_hat <- sigma2 * solve(t(X) %*% X)
var_beta_hat
```

```
          [,1]      [,2]      [,3]
[1,] 0.1074796 0.0000000 0.0000000
[2,] 0.0000000 0.1074796 0.0000000
[3,] 0.0000000 0.0000000 0.1074796
```

Problem 2

Part A

Given X is all 1's, the ordinary least squares estimations of β_0 and β_1 are obtained by minimizing the sum of squared residuals:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

The condition for the minimum requires that the partial derivatives of this sum with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ are zero:

$$\frac{\partial}{\partial \hat{\beta}_0} \sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 = -2 \sum (y_i - \hat{y}_i) = -2 \sum e_i = 0$$

Hence, $\sum_{i=1}^{n} e_i = 0$.

Part B

The fitted values are given by $\hat{y} = X\hat{\beta}$, and the residuals are $e = y - \hat{y}$.

The OLS estimation aims to minimize the sum of squared residuals, $e^T e$. The condition for the minimum is obtained by setting the derivative of $e^T e$ with respect to $\hat{\beta}$ to zero:

$$\frac{\partial}{\partial \hat{\beta}} e^T e = \frac{\partial}{\partial \hat{\beta}} (y - X\hat{\beta})^T (y - X\hat{\beta}) = X^T (y - X\hat{\beta}) = 0$$

Since the first column of X is all 1's, this implies the first row of X^T will be multiplied by all residuals, summing them:

$$\sum_{i=1}^{n} e_i = 0$$

Part C

The condition that the total true error term ϵ_i sums to zero is not guaranteed in population regression. This is because ϵ_i represents the random and unobserved variations in the data, unlike the residuals e_i in OLS regression, which are forced to sum to zero to minimize error. The true errors' sum not equaling zero reflects the inherent randomness in the data, rather than a model's accuracy.

Problem 3

Part A

To find the $\hat{\sigma}$ we can use this formula where n is the number of observations, k is the number of predictors including the intercept, and e_i are the residuals:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-k}}$$

```
library(IRdisplay)
library(faraway)
data(teengamb)

xMatrix <- as.matrix(cbind(1, teengamb$sex, teengamb$status, teengamb$income, teengamb$verba

yVector <- teengamb$gamble

betaHat <- solve(t(xMatrix) %*% xMatrix) %*% t(xMatrix) %*% yVector

yPredicted <- xMatrix %*% betaHat

residualsVector <- yVector - yPredicted

# Calculate the estimator
nObservations <- length(yVector)
kPredictors <- 5
sigmaHat <- sqrt(sum(residualsVector^2) / (nObservations - kPredictors))
print(sigmaHat)
```

```
[1] 22.69034
```

$$\hat{\sigma} = 22.69$$

In the context of the data, it provides a measure of the typical amount by which the actual gambling expenditures (gamble) deviate from the values predicted by the model based on sex, status, income, and verbal score.

Part B

```r
library(faraway)
data(teengamb)

model <- lm(gamble ~ sex + status + income + verbal, data = teengamb)
summary(model)
```

```
Call:
lm(formula = gamble ~ sex + status + income + verbal, data = teengamb)

Residuals:
    Min      1Q  Median      3Q     Max 
-51.082 -11.320  -1.451   9.452  94.252 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  22.55565   17.19680   1.312   0.1968    
sex         -22.11833    8.21111  -2.694   0.0101 *  
status        0.05223    0.28111   0.186   0.8535    
income        4.96198    1.02539   4.839 1.79e-05 ***
verbal       -2.95949    2.17215  -1.362   0.1803    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.69 on 42 degrees of freedom
Multiple R-squared:  0.5267,	Adjusted R-squared:  0.4816 
F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

The Residual standard Error = 22.69 which matches what we got from part a.

Part C

$$\mathsf{Var}(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1}$$

```
vcov(model)
```

```
            (Intercept)        sex      status      income      verbal
(Intercept)  295.730049 -72.731725 -2.39537944 -9.88839252 -15.1841234
sex          -72.731725  67.422402  1.27368110  2.46514518  -3.5408987
status        -2.395379   1.273681  0.07902370  0.09665741  -0.3217548
income        -9.888393   2.465145  0.09665741  1.05142939  -0.0542087
verbal       -15.184123  -3.540899 -0.32175476 -0.05420870   4.7182371
```

This function returns the covariance matrix of the model's coefficient estimates.

Problem 4

Part A

$$\text{Energy} = \beta_0 + \beta_1 \text{Mass} + \beta_2 I_{\text{noEchoBat}} + \beta_3 I_{\text{noEchoBird}} + \beta_4 (\text{Mass} \times I_{\text{noEchoBat}})$$
$$+ \beta_5 (\text{Mass} \times I_{\text{noEchoBird}}) + \beta_6 I_{\text{echoBat}} + \beta_7 (\text{Mass} \times I_{\text{echoBat}}) + \epsilon$$

Part B

```r
# Load necessary packages
library(Sleuth3)
library(dplyr)
library(ggplot2)

data(case1002, package = "Sleuth3")

case1002$Type <- as.factor(case1002$Type)
unique(case1002$Type)
```

```
[1] non-echolocating bats  non-echolocating birds echolocating bats
Levels: echolocating bats non-echolocating bats non-echolocating birds
```

```r
model <- lm(Energy ~ Mass * Type, data = case1002)
summary(model)
```

```

Call:
lm(formula = Energy ~ Mass * Type, data = case1002)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.0486 -2.2709 -0.0822  0.9937 12.4601 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      0.49398    3.19470   0.155    0.879
Mass                             0.08964    0.06804   1.317    0.209
Typenon-echolocating bats       10.73340    7.06170   1.520    0.151
Typenon-echolocating birds       2.82276    4.26463   0.662    0.519
Mass:Typenon-echolocating bats  -0.04959    0.06904  -0.718    0.484
Mass:Typenon-echolocating birds -0.02186    0.06866  -0.318    0.755

Residual standard error: 5.041 on 14 degrees of freedom
Multiple R-squared:  0.9045,	Adjusted R-squared:  0.8703 
F-statistic:  26.5 on 5 and 14 DF,  p-value: 1.136e-06
```

To calculate the mean Energy expenditure for each type when Mass is held at 0: For non-echolocating bats, it is $\beta_0 + 10.73340 = 0.49398 + 10.73340 = 11.22738$. For non-echolocating birds, it is $\beta_0 + 2.82276 = 0.49398 + 2.82276 = 3.31674$. For echolocating bats (assuming they are the reference category), it is simply $\beta_0 = 0.49398$.

Part C

The interaction term between Mass and non-echolocating bats, with a coefficient of $\hat{\beta}_j = -0.04959$, indicates that the relationship between Mass and Energy expenditure for non-echolocating bats decreases by 0.04959 units for each unit increase in Mass, compared to the reference category. This suggests that for non-echolocating bats, as their mass increases, the expected increase in energy expenditure is slightly less than what is observed for the baseline category (presumed to be echolocating bats) by this amount. Essentially, this term quantifies the unique influence of Mass on Energy expenditure among non-echolocating bats, differentiating it from the pattern seen in the reference group.