Homework 2

Brian Cervantes Alvarez
January 26, 2024
ST552: Statistical Methods

Question 1

Part A

Here is the simple linear model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

The linear regression model in matrix form represented as:

$$Y = X\beta + \epsilon$$

where:

- Y is the vector of responses y_i for $i = 1$ to n.
- X is the matrix of predictors with each row corresponding to an observation and each column to a predictor.
- β is the vector of coefficients $\beta_0, \beta_1, \ldots, \beta_p$ where p is the number of predictors; in our case, $p = 1, \{\beta_0, \beta_1\}$.
- ϵ is the vector of random errors.

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Therefore, the matrices and vectors become:

$$Y = X\beta + \epsilon, \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Part B

Solution for X^TX

We know X, and X^T can be easily transformed as the following:

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

Now, we multiply X^T by X. Then, by using matrix multiplication, it becomes a $2x2$ square matrix, we get our solution:

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

Solution for X^TY

Given X and Y, we can solve for X^TY as follows::

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \Rightarrow X^TY = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_iy_i \end{bmatrix} = \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^n x_iy_i \end{bmatrix}$$

Solution for $(X^TX)^{-1}$

Recall that the inverse of a 2x2 matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is given by:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Given we know X^TX, we can use that to solve for $(X^TX)^{-1}$:

$$(X^TX)^{-1} = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{bmatrix}^{-1} = \frac{1}{nS_{xx}} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} = \frac{1}{S_{xx}} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

Part C

Least squares estimates $\hat{\beta}$ would be computed like this:

$$\begin{aligned}
\hat{\beta} &= (X^T X)^{-1} X^T Y \\
&= \frac{1}{S_{xx}} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^n x_i y_i \end{bmatrix} \\
&= \frac{1}{S_{xx}} \begin{bmatrix} \bar{y}\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} \end{bmatrix}
\end{aligned}$$

Part D

Show that the least squares estimates in are equivalent to the usual form for the estimates in simple linear regression. We can show it by simplifying from part c

$$
\begin{aligned}
&= \frac{1}{S_{xx}} \begin{bmatrix} \bar{y}\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\bar{y} + n\bar{x}^2\bar{y} - \bar{x}\sum_{i=1}^{n} x_i y_i \\ S_{xy} \end{bmatrix} \\
&= \frac{1}{S_{xx}} \begin{bmatrix} \bar{y}(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) - \bar{x}(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}) \\ S_{xy} \end{bmatrix} \\
&= \frac{1}{S_{xx}} \begin{bmatrix} \bar{y}S_{xx} - \bar{x}S_{xy} \\ S_{xy} \end{bmatrix} \\
&= \begin{bmatrix} \bar{y} - \hat{\beta}_1\bar{x} \\ \hat{\beta}_1 \end{bmatrix}
\end{aligned}
$$

Problem 2

Part A

```r
library(faraway)
library(ggplot2)
data(teengamb)
ds <- teengamb

# Construct matrix X and response vector Y
X <- cbind(1, ds$sex, ds$status, ds$income, ds$verbal)
Y <- teengamb$gamble

head(X,5)
```

```
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    1   51  2.0    8
[2,]    1    1   28  2.5    8
[3,]    1    1   37  2.0    6
[4,]    1    1   28  7.0    4
[5,]    1    1   65  2.0    8
```

```r
head(Y, 5)
```

```
[1]  0.0  0.0  0.0  7.3 19.6
```

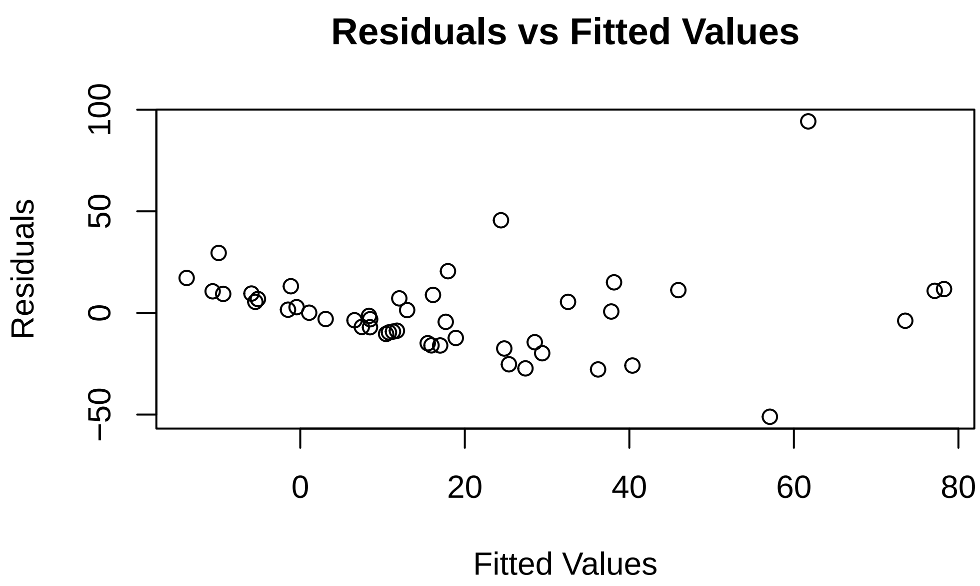
Part B

```r
# Find the least squares estimates of the regression coefficients
betaHat <- solve(t(X) %*% X) %*% t(X) %*% Y
betaHat
```

```
             [,1]
[1,]  22.55565063
[2,] -22.11833009
[3,]   0.05223384
[4,]   4.96197922
[5,]  -2.95949350
```

Part C

```
# Find the fitted values and residuals
fittedVals <- X %*% betaHat
residuals <- Y - fittedVals

# Plot residuals against fitted values
plot(fittedVals, residuals, main = "Residuals vs Fitted Values",
     xlab = "Fitted Values", ylab = "Residuals")
```



Part D

In the model, the coefficient for the "sex" variable (male or female) is -22.11833, indicating that, with all other predictors held constant, the predicted expenditure on gambling for males is expected to be approximately $22.12 less than for females. The negative sign implies a decrease in predicted expenditure for males. The associated p-value of 0.0101 suggests that this gender difference is statistically significant.

```r
# Fit the model
model <- lm(data = ds, gamble ~ sex + status + income + verbal)
summary(model)
```

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

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
```

Part E

The coefficient for the “income” is 4.96198, which means that a single unit (or an increase of 1 unit) in income yields a $4.96 increase in gambling expenditure. This positive relationship implies that higher income individuals are expected to spend more on gambling (I mean, if I was strapped with cash and was like 65 years old, I would do the same). The p-value less than 0.001, which is statistically significant and reinforces the idea of this association.

Problem 3

Part A

Here is the model that we are going to use:

$$\text{Weekly Wages} = \beta_0 + \beta_1 \times \text{Education} + \beta_2 \times \text{Experience}$$

Part B

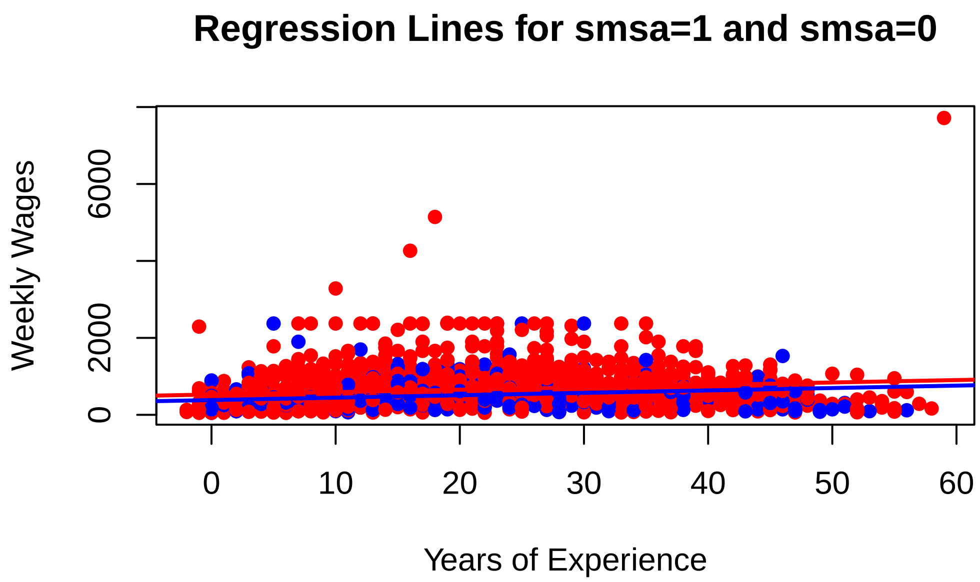
In basic terms, the model forecasts salaries by considering education and years of experience. The intercept, set at -242.7994, represents the projected initial wage (not realistic of course, but that's the model's y-int). Meanwhile, the coefficients for education (51.1753) and experience (9.7748) indicate the anticipated wage adjustment for each additional year of education and experience. Given that 0.1351 or 13.51% variability being explained by the Multiple R-squared, it is evident that these two variables alone are insufficient for accurately modeling wages.

```r
data(uswages)
ds <- uswages

# Fit the model
expEducationModel <- lm(data = ds, wage ~ educ + exper)
summary(expEducationModel)
```

```
Call:
lm(formula = wage ~ educ + exper, data = ds)

Residuals:
    Min      1Q  Median      3Q     Max
-1018.2  -237.9   -50.9   149.9  7228.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -242.7994    50.6816  -4.791 1.78e-06 ***
educ          51.1753     3.3419  15.313  < 2e-16 ***
exper          9.7748     0.7506  13.023  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 427.9 on 1997 degrees of freedom
Multiple R-squared:  0.1351,    Adjusted R-squared:  0.1343
F-statistic:   156 on 2 and 1997 DF,  p-value: < 2.2e-16
```

Part C

```r
# Fit the model with years of experience and smsa as explanatory variables
expExperienceModel <- lm(data = ds, wage ~ exper + smsa)

# Create a plot
plot(ds$exper, ds$wage,
     col = ifelse(ds$smsa == 1, "red", "blue"),
     pch = 16,
     main = "Regression Lines for smsa=1 and smsa=0",
     xlab = "Years of Experience",
     ylab = "Weekly Wages")

# Add regression lines
abline(coef(expExperienceModel)[1],
       coef(expExperienceModel)[2],
       col = "blue", lwd = 2)
abline(coef(expExperienceModel)[1] + coef(expExperienceModel)[3],
       coef(expExperienceModel)[2],
       col = "red", lwd = 2)
```




```r
# Calculate the vertical distance between the lines
vertical_distance <- coef(expExperienceModel)[3]
cat("Vertical distance between the lines:", vertical_distance, "\n")
```

```
Vertical distance between the lines: 144.2175
```

```r
# Display the model summary
summary(expExperienceModel)
```

```

Call:
lm(formula = wage ~ exper + smsa, data = ds)

Residuals:
   Min     1Q Median     3Q    Max 
-789.2 -274.3  -73.3  156.3 6818.7 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 383.8832    24.4398  15.707  < 2e-16 ***
exper         6.2576     0.7492   8.353  < 2e-16 ***
smsa        144.2175    23.3256   6.183 7.61e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 448 on 1997 degrees of freedom
Multiple R-squared:  0.05171,	Adjusted R-squared:  0.05077 
F-statistic: 54.45 on 2 and 1997 DF,  p-value: < 2.2e-16
```