Statisticial Methods HW 2

2023-01-22

Problem 1)

a)

We know that the n observations and the model for linear regression can be written as:

$$Y_1 = \beta_0 + \beta_1 X_1 + \varepsilon_1$$

$$\vdots = \vdots$$

$$Y_n = \beta_0 + \beta_n X_n + \varepsilon_n$$

Which can be written in matrix form:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Which can be simplified further:

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times 2} \cdot \beta_{2\times 1} + \varepsilon_{n\times 1}$$
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$$

b)

We have already determined: $\boldsymbol{X} = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}$. The transpose would be: $\boldsymbol{X}^T = \begin{pmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{pmatrix}$. So,

$$X^T X = \begin{pmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{pmatrix} \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} = \begin{pmatrix} n & X_1 + \cdots + X_n \\ X_1 + \cdots + X_n & X_1^2 + \cdots + X_n^2 \end{pmatrix}$$

$$X^T X = \begin{pmatrix} n & \sum_{i=0}^n X_i \\ \sum_{i=0}^n X_i & \sum_{i=0}^n X_i^2 \end{pmatrix}$$

Now for $X^T Y$:

$$X^T Y = \begin{pmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} Y_1 + \cdots + Y_n \\ X_1 Y_1 + \cdots X_n Y_n \end{pmatrix}$$

$$X^TY = \begin{pmatrix} \sum_{i=0}^n Y_i \\ \sum_{i=0}^n X_iY_i \end{pmatrix}$$

Finally, $(X^TX)^{-1}$ is the inverse of X^TX, which we found earlier. We can use that when $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ then the inverse is $A^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$. So if $A = X^TX$ then $A^{-1} = (X^TX)^{-1}$. Finally,

$$(X^TX)^{-1} = \frac{1}{n(\sum_{i=0}^n X_i^2) - (\sum_{i=0}^n X_i)(\sum_{i=0}^n X_i)} \begin{pmatrix} \sum_{i=0}^n X_i^2 & -\sum_{i=0}^n X_i \\ -\sum_{i=0}^n X_i & n \end{pmatrix}$$

$$(X^TX)^{-1} = \frac{1}{n(\sum_{i=0}^n X_i^2) - n^2\bar{X}^2} \begin{pmatrix} \sum_{i=0}^n X_i^2 & -\sum_{i=0}^n X_i \\ -\sum_{i=0}^n X_i & n \end{pmatrix}$$

c)

$$\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^TY = \frac{1}{n(\sum_{i=0}^n X_i^2) - n^2\bar{X}^2} \begin{pmatrix} \sum_{i=0}^n X_i^2 & -\sum_{i=0}^n X_i \\ -\sum_{i=0}^n X_i & n \end{pmatrix} \begin{pmatrix} \sum_{i=0}^n Y_i \\ \sum_{i=0}^n X_iY_i \end{pmatrix}$$

MUltiplyign the matrices, and some simplification we get:

$$= \frac{1}{n(\sum_{i=0}^n X_i^2) - n^2\bar{X}^2} \begin{pmatrix} n\bar{Y}\sum_{i=0}^n X_i^2 - n\bar{X}\sum_{i=0}^n X_iY_i \\ n\sum_{i=0}^n X_iY_i - n^2\bar{X}\bar{Y} \end{pmatrix}$$

$$= \frac{1}{n\sum_{i=0}^n (X_i^2 - \bar{X}^2)} \begin{pmatrix} n\bar{Y}\sum_{i=0}^n X_i^2 - n\bar{X}\sum_{i=0}^n X_iY_i \\ n\sum_{i=0}^n X_iY_i - n^2\bar{X}\bar{Y} \end{pmatrix}$$

After a ton of simplification:

$$\hat{\boldsymbol{\beta}} = \frac{1}{\sum_{i=0}^n (X_i - \bar{X})^2} \begin{pmatrix} \bar{Y}\sum_{i=0}^n (X_i - \bar{X})^2 - \bar{X}\sum_{i=0}^n (X_iY_i - \bar{X}\bar{Y}) \\ \sum_{i=0}^n (X_i - \bar{X})(Y_i - \bar{Y}) \end{pmatrix} = \begin{pmatrix} \bar{Y} - \frac{s_{xy}}{s_{xx}}\bar{X} \\ \frac{\sum_{i=0}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=0}^n (X_i - \bar{X})^2} \end{pmatrix}$$

From here we can see some things cancel out, and we finally get:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \bar{Y} - \hat{\beta}_1\bar{X} \\ \hat{\beta}_1 \end{pmatrix}$$

Where $\hat{\beta}_1 = \frac{\sum_{i=0}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=0}^n (X_i - \bar{X})^2}$

d)

Therefore, we have shown what is asked in part (d). Also, I noticed I started the summation at i = 0 when it should have been i = 1 but ran out of time.

Problem 2)

a)

```
library("faraway")
library("matlib")
X <- cbind(rep(1,47), teengamb$sex, teengamb$status, teengamb$income, teengamb$verbal)
Y <- cbind(teengamb$gamble)
```

b)

We can use the formula from the last problem so $\hat{\beta} = (X^T X)^{-1} X^T Y$. And we have everything we need to calculate this.

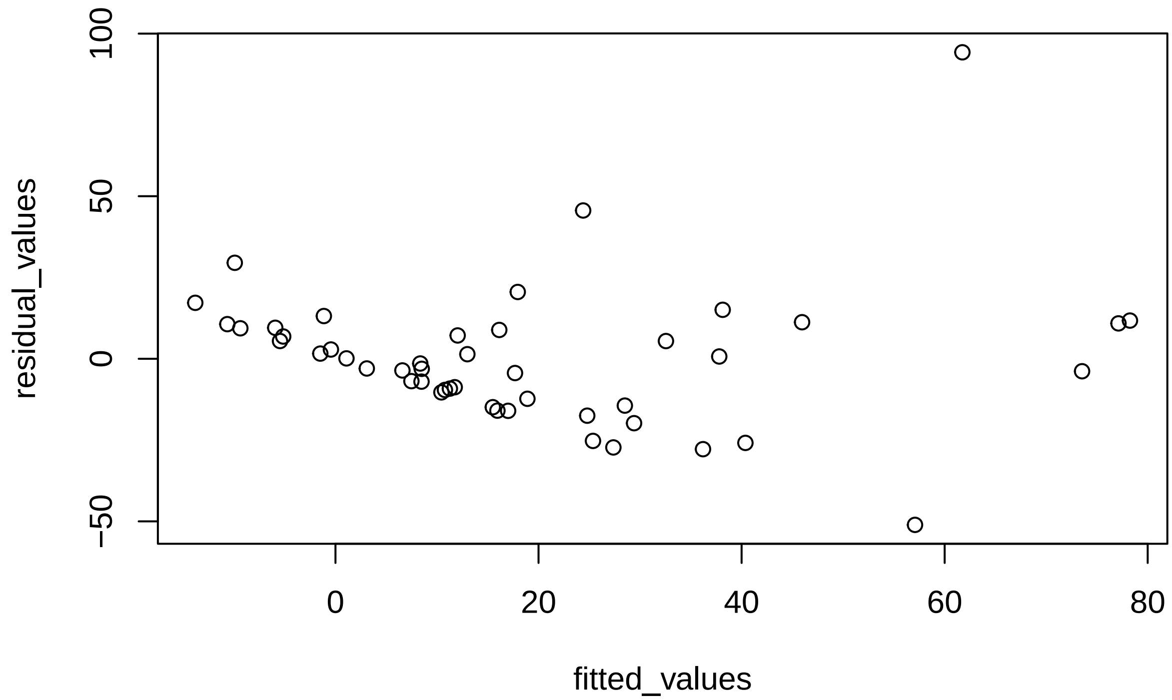
```
Beta_hat <- solve(t(X)%*%X)%*%t(X)%*%Y
Beta_hat
```

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

```
#
```

c)

```
#hat matrix
H <- X%*%solve(t(X)%*%X)%*%t(X)
fitted_values <- H%*%Y #product of H and Y for fitted values
residual_values <- Y - fitted_values #subtracting actual values from predicted values
plot(fitted_values,residual_values)
```



d)

```r
mod <- lm(teengamb$gamble ~ teengamb$sex +teengamb$status+teengamb$income+teengamb$verbal)
mod2 <- lm(Y~X)
#same thing
summary(mod)
```

```
##
## Call:
## lm(formula = teengamb$gamble ~ teengamb$sex + teengamb$status +
##     teengamb$income + teengamb$verbal)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      22.55565   17.19680   1.312   0.1968
## teengamb$sex    -22.11833    8.21111  -2.694   0.0101 *
## teengamb$status   0.05223    0.28111   0.186   0.8535
## teengamb$income   4.96198    1.02539   4.839 1.79e-05 ***
## teengamb$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
```

Considering here X_1 explanatory variable is the sex explanatory variable, where 1 - female and 0 - male. The coefficient given is -22.11833. All other things held equal, the predicted difference gambling expenditure between male and females is 22.11833 (pounds per year).

e)

For every 1 unit increase in income (pounds per week), the predicted expenditure in gambling is expected to increase by 4.96 (pounds per year).

Problem 3)

a)

$$\text{wage}_i = \beta_0 + \beta_1 \text{education}_i + \beta_2 \text{experience}_i + \varepsilon_i$$

for $i = 1, ..., 47$.. In the data set weekly wages (wage) years of education (educ), and years of experience (exper) are abbreviated.

b)

```
library("ggplot2")
mod <- lm(uswages$wage ~ uswages$educ + uswages$exper)
mod
```

```
##
## Call:
## lm(formula = uswages$wage ~ uswages$educ + uswages$exper)
##
## Coefficients:
##   (Intercept)   uswages$educ  uswages$exper
##      -242.799         51.175          9.775
```

The coefficient for year of education is 51.175. That means, all other thing held equal, a 1 year increase of education is predicted to increase weekly wages by $51.175.

c)

```
mod2 <- lm(uswages$wage ~ uswages$exper + uswages$smsa)
ggplot(uswages, aes(x=exper, y=wage)) +
  geom_point() +
  geom_abline(intercept=mod2$coefficients[1], slope=mod2$coefficients[2], color="blue") +
  geom_abline(intercept=mod2$coefficients[1]+mod2$coefficients[3], slope=mod2$coefficients[2], color="re
```

The vertical distance between the lines is the coefficient value which is $144.2175.