Homework #3

Yefrid Cordoba

Problem 1

A regression analysis relating test scores (Y) to training hours (X) produced the following fitted equation: $\hat{y} = 25 - 0.5x$.

a.

What is the fitted value of the response variable corresponding to $ x = 7$?

R/

```
y = 25 - 0.5 * 7
y
```

```
[1] 21.5
```

The fitted value when x = 7 is y = 21.5

b.

What is the residual corresponding to the data point with = 3 and = 30?.
Is the point above or below the line? Why?

R/

```
y_hat = 25 - 0.5 * 3
y_hat
```

```
[1] 23.5
```

```r
30-y_hat
```

```
[1] 6.5
```

The residual is 6.5, it is above the line due to the positive error, because after evaluating the fitted equation with 3, the $\hat{y}$ is lower than the y corresponding to the data point.

c.

If x increases 3 units, how does $\hat{y}$ change?

R/

```r
-0.5 * 3
```

```
[1] -1.5
```

It x increase 3 units, the estimated value y will decrease 1.5 units.

d.

An additional test score is to be obtained for a new observation at $= 6$.
Would the test score for the new observation necessarily be 22? Explain.

R/
Not necessarily, this is because there is a probability associated to each point on the regression that is assumed to be normal, to determine what is the range at which we can be sure about the true future value of y when $x = 6$, we would need to estimate the prediction interval.

e.

The error sums of squares (SSE) for this model were found to be 7. If there were $n = 16$ observations, provide the best estimate for σ^2.

R/

```
SSE = 7
n = 16
p = 2 #2 parameters estimated in the model b0 and b1
s2 = SSE / (n - 2)
s2
```

```
[1] 0.5
```

The best estimate is $\sigma^2 = 0.5$.

Problem 2

The dataset "Healthy Breakfast" contains, among other variables, the Consumer Reports ratings of 77 cereals and the number of grams of sugar contained in each serving. Considering "Sugars" as the explanatory variable and "Rating" as the response variable generated the following fitted regression equation:

$$\widehat{rating} = 59.3 - 2.4 * sugars$$

The "Analysis of variance" partial portion of the R output is shown below

a.

Find the missing values

Source	DF	SS	MS	F	P
Regression	1	8654.7	8654.7	102.35	1.1537e-15
Error	75	6342.1	84.56		
Total	76	14996.8			

```
nt <- 77
SSr <- 8654.7
MSr <- 8654.7
SST <- 14996.8
dfr <- SSr / MSr
dfr
```

```
[1] 1
```

```r
SSE <- SST - SSr
SSE
```

```
[1] 6342.1
```

```r
p <- dfr + 1
p
```

```
[1] 2
```

```r
dfe <- nt-p
dfe
```

```
[1] 75
```

```r
dfT <- dfe + dfr
dfT
```

```
[1] 76
```

```r
MSE <- SSE / dfe
MSE
```

```
[1] 84.56133
```

```r
F_v <- MSr / MSE
F_v
```

```
[1] 102.3482
```

```r
P_v <- pf(F_v, dfr, dfe, lower.tail = F)
P_v
```

```
[1] 1.153683e-15
```

b.

Find R^2 value and interpret that number.

R/

```
R_sq <- SSr / SST
R_sq
```

```
[1] 0.5771031
```

Based on our R^2, 57% of the variability of the data can be explained by the model.

c.

What is the estimated value of σ^2.

R/

```
s2 = SSE / (nt - p)
s2
```

```
[1] 84.56133
```

The estimated variance is $\sigma^2 = 84.56$.

d.

Test $H_o : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$ using F - test

R/
Assuming a 5% level of significance, and given that the P-value (1.1537e-15)from the table is much lower than our level of significance, we reject the null hypothesis and have strong evidence to conclude that β_1 is different than 0.

Problem 3

Athletes are constantly seeking measures of the degree of their cardiovascular fitness prior to a major race. Athletes want to know when their training is at a level that will produce a peak performance. One such measure of fitness is the time to exhaustion from running on a treadmill at a specified angle and speed. The important question is then “Does this measure of cardiovascular fitness translate into performance in a 10-km running race?” Twenty experienced distance runners who professed to be in top condition were evaluated on the treadmill and then had their times recorded in a 10-km race.

You are provided with the following information

$$\hat{\beta}_0 = b_0 = 58.816$$

$$\hat{\beta}_1 = b_1 = -1.867$$

$$s(b_1) = 0.346$$

$$MSE = 4.417$$

$$\hat{y} = 58.816 - 1.867x$$

a.

Test the hypothesis that there is a linear relationship between the amount of time needed to run a 10-km race and the time to exhaustion on a treadmill.

R/

```r
bo <- 58.816
b1 <- -1.867
MSE <- 4.417
sb1 <- 0.346
tobs <- (b1 - 0) / sb1
tobs < -qt(0.975, 18)
```

```
[1] TRUE
```

We have to test:$H_o : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$

Assuming an $\alpha = 0.05$, and knowing that $t_{obs} < -t_{(0.975,18)}$ we reject the null hypothesis and conclude that the time of exhaustion on a treadmill has a linear relationship to the time needed to run a 10-km race.

b.

Construct a 95% confidence interval for β_1.

R/

```r
Up <- b1 + qt(0.975, 18) *sb1
Up
```

```
[1] -1.140081
```

```r
Lo <- b1 - qt(0.975, 18) *sb1
Lo
```

```
[1] -2.593919
```

We are 95% confident that the estimated mean time to run a 10-km race decreases between 1.14 and 2.59 minutes for each additional minute that an athlete times on the treadmill.

c.

Construct a 95% confidence interval for the estimated mean 10-km race when the treadmill time is 10 minutes. Given that $sd\{ \hat{y}_{x_h=10}\} = 0.7342$

R/

```r
y_hat <- bo + b1 * 10
y_hat
```

```
[1] 40.146
```

```r
sd10 <- 0.7342
Up10 <- y_hat + sd10 * qt(0.975, 18)
Up10
```

```
[1] 41.6885
```

```r
Lo10 <- y_hat - sd10 * qt(0.975, 18)
Lo10
```

```
[1] 38.6035
```

We are 95% confident that the estimated mean time when the treadmill time is 10 min for a 10-km race is between 38.6 and 41.7 minutes.

d.

Construct a 95% prediction interval for an athlete whose treadmill time is 10 minutes. $sd(\hat{y}_{\text{pred}}) = \sqrt{mse + \left(sd\{\hat{y}_{x_h=10}\}\right)^2} = 2.226$

```r
sdpred <- 2.226
Upp <- y_hat + sdpred * qt(0.975, 18)
Upp
```

```
[1] 44.82265
```

```r
Lop <- y_hat - sdpred * qt(0.975, 18)
Lop
```

```
[1] 35.46935
```

We are 95% confident that the mean time when an athlete times 10 min on the treadmill will be between 35.5 and 44.8 min for a 10-km race.

Problem 4

The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test score (X). The results of the study follow. Assume that first-order regression model $Y_i = \beta_0 + \beta_1 * X_i + \epsilon_i$ is appropriate. **(Data is provided in the Moodle Page)**

```r
gpa_data<-read.table('Homework3 Data.txt', header=T)
names(gpa_data)
```

```
[1] "GPA" "ACT"
```

```r
head(gpa_data)
```

```
    GPA ACT
1 3.897  21
2 3.885  14
3 3.778  28
4 2.540  22
5 3.028  21
6 3.865  31
```

```r
tail(gpa_data)
```

```
      GPA ACT
115 1.486  31
116 3.885  20
117 3.800  29
118 3.914  28
119 1.860  16
120 2.948  28
```

```r
attach(gpa_data)
```

a.

Compute the mean and variance of ACT test score and GPA.

R/

```r
library(tidyverse)
gpa_data |>
  summarise(across(
    where(is.numeric),
    list(mean = mean, var = var),
    na.rm = TRUE
  ))
```

```
  GPA_mean   GPA_var ACT_mean  ACT_var
1  3.07405 0.4151719   24.725 19.99937
```

b.

Compute the correlation between ACT test score and GPA. Comment on the strength and direction of the linear relationship between the variables.

R/

```r
cor(gpa_data$ACT, gpa_data$GPA, method = "pearson")
```

```
[1] 0.2694818
```

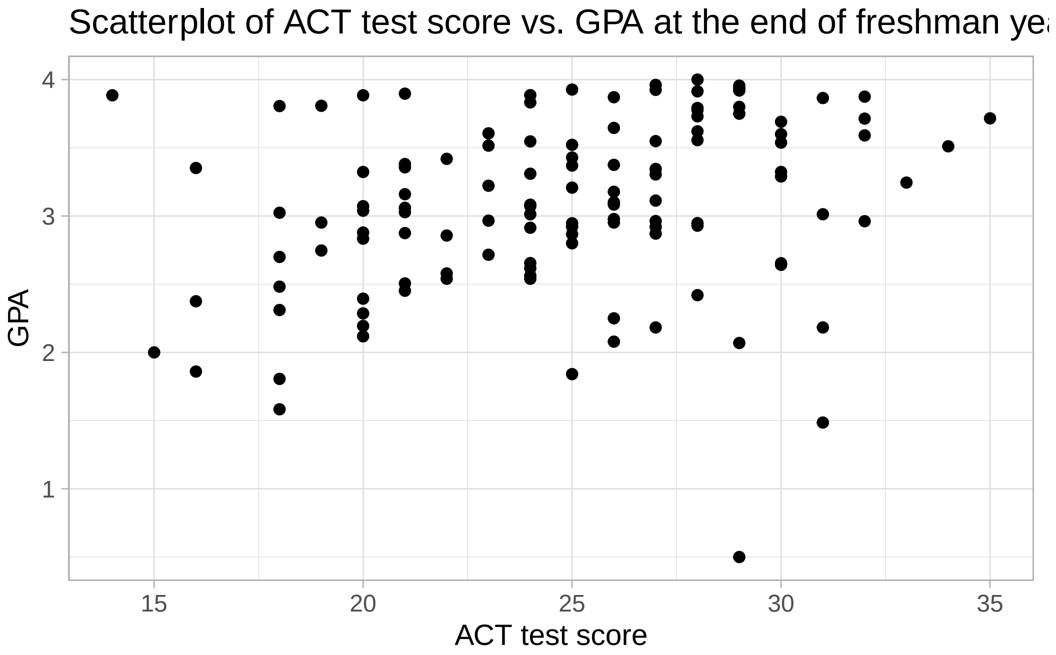
There is a positive correlation between ACT test score and GPA, that means that a higher test score is related to a higher GPA at the end of the freshman year but this correlation is weak.

c.

Construct a scatter plot. Is the relationship approximately linear?

R/

```r
gpa_data |>
  ggplot(aes(x = ACT, y = GPA)) +
  labs(x = 'ACT test score',
       y = 'GPA',
       title = 'Scatterplot of ACT test score vs. GPA at the end of freshman year')+
  theme_light() +
  geom_point()
```



The relationship for ACT test score and GPA is weakly positive and linear.

d.

Run a linear regression to predict GPA based on the ACT score. Give the regression equation.

R/

```r
model <- lm(GPA ~ ACT, gpa_data)
summary(model)
```

```
Call:
lm(formula = GPA ~ ACT, data = gpa_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.74004 -0.33827  0.04062  0.44064  1.22737 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
ACT          0.03883    0.01277   3.040  0.00292 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6231 on 118 degrees of freedom
Multiple R-squared:  0.07262,   Adjusted R-squared:  0.06476
F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917
```

The regression equation for this model is:

$$\widehat{GPA} = 2.11405 + 0.03883 * ACT_{testscore}$$

e.

What is the point estimate of the change in the mean response when the entrance test score increases by one point? And increases by 4 points?

R/
The estimated change in the mean GPA is 0.03883 points more when the ACT test score increases by one point

```
0.03883*4
```

```
[1] 0.15532
```

When the ACT test score increases by 4 points, it is estimated that the mean GPA at the end of the freshman year will be 0.15532 points higher.

f.

Based on your answer in (e), predict the GPA of a student who scored 20 on the ACT.

R/

```
GPA20 <- 2.11405 + 0.03883 * 20
GPA20
```

```
[1] 2.89065
```

The predicted mean GPA for a student who scored 20 on the ACT is 2.89.

g.

Estimate σ^2 and σ

R/
From the regression model we get $\sqrt{MSE} = 0.6231 =$ standard error
$\sigma^2 = MSE$

```
MSE <- 0.6231 ^2
MSE
```

```
[1] 0.3882536
```

Estimates:

$$\sigma^2 = 0.3883$$

$$\sigma = 0.6231$$

h.

Give a point estimate and 95% confidence interval for the slope and interpret each of these in words.

```
confint(model, 'ACT', level = 0.95)
```

```
         2.5 %     97.5 %
ACT 0.01353307 0.06412118
```

Point estimate $= 0.03883$
For each point increase in the ACT test score, it is estimated that the GPA will increase by 0.03883 points.

We are 95% confident that the true change in the GPA for each ACT test point obtained by the students is between 0.0135 and 0.0641 more.

i.

Obtain a 95% interval estimate of the mean GPA for students whose ACT test score is 28. Interpret your confidence interval.

R/

```
data <- data.frame(ACT = 28)
predict(model, data, interval = 'confidence',
        level = 0.95, se.fit = TRUE)
```

```
$fit
       fit      lwr      upr
1 3.201209 3.061384 3.341033

$se.fit
[1] 0.07060873

$df
[1] 118

$residual.scale
[1] 0.623125
```

We are 95% confident that the true mean GPA value is between 3.061 and 3.341 when the ACT test score is 28.

j.

Predict GPA using a 95% prediction interval for students whose ACT test score is 28.

R/

```
data <- data.frame(ACT = 28)
predict(model, data, interval = 'prediction',
        level = 0.95, se.fit = TRUE)
```

```
$fit
       fit      lwr      upr
1 3.201209 1.959355 4.443063
```

```
$se.fit
[1] 0.07060873

$df
[1] 118

$residual.scale
[1] 0.623125
```

For students with an ACT score of 28, a 95% prediction interval for freshman year GPA is 1.959 to 4.443, meaning that under similar conditions about 95% of those students' GPAs are expected to fall in this range.

Problem 5

A medical study was conducted to study the relationship between infants' systolic blood pressure and two explanatory variables, weight (kgm) and age (days). The portion of data for 25 infants are shown here.

```r
age <- c(3, 4, 5, 6, 3, 4, 5, 6, 3, 4, 5, 6, 3, 4, 5, 6, 3, 4, 5, 6, 3, 4, 5, 6, 6)
weight <- c(2.61, 2.67, 2.98, 3.98, 2.87, 3.41, 3.49, 4.03, 3.41, 2.81, 3.24, 3.75, 3.18, 3.
systolic_BP <- c(80, 90, 96, 102, 81, 96, 99, 110, 88, 90, 100, 102, 86, 93, 101, 103, 86, 9
# Create a data frame
data <- data.frame(age, weight, systolic_BP)
```

a.

Write a first-order multiple regression model relating Systolic BP to Age and Weight.

R/

$$SystolicBP = \beta_0 + \beta_1 * Age + \beta_2 * Weight + \epsilon_i$$

b.

Fit Multiple linear regression models to these data and obtain the estimated regression equation.

R/

```r
MLRmodel <- lm(systolic_BP ~ age + weight, data)
summary(MLRmodel)
```

```

Call:
lm(formula = systolic_BP ~ age + weight, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.1779 -1.2224  0.2005  1.5164  4.5465 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  57.2644     3.7986  15.075 4.44e-13 ***
age           5.8041     0.6415   9.048 7.22e-09 ***
weight        3.3162     1.5522   2.136    0.044 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.454 on 22 degrees of freedom
Multiple R-squared:  0.9199,	Adjusted R-squared:  0.9126 
F-statistic: 126.3 on 2 and 22 DF,  p-value: 8.696e-13
```

The estimated multiple linear regression model is:

$$\widehat{SystolicBP} = 57.2644 + 5.8041 * Age + 3.3162 * Weight$$

c.

Obtain the estimated residual standard deviation.

R/
The estimated residual standard error = estimated residual standard deviation.

$$\sigma = \sqrt{MSE} = 2.454$$

d.

Provide an interpretation of $\hat{\beta}_2$, the coefficient of weight.

R/
If the age of an infant is constant, for each additional kilogram of weight that an infant has, the mean systolic blood pressure will increase by 3.32.

e.

Can the hypothesis of no overall predictive value of the model be rejected at the $\alpha = 0.01$ level?

R/
Given the p-value(8.696e-13) obtained for the overall predictive value of the model is much lower than the significance level, it is valid to reject the hypothesis that the model has no overall predictive power on the systolic blood preassure of infants.