HW3

Jackson Connor

Packages

```
library(tidyverse)
```

Problem 1

A regression analysis relating test scores (Y) to training hours (X) produced the following fitted equation: ŷ=25-0.5x.

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

$\hat{y} = 25 - 0.5(7) = 21.5$

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

$\hat{y} = 25 - 0.5x = 25 - 0.5(3) = 23.5$

residual $= e_i = y_i - \hat{y}_i = 30 - 23.5 = 6.5$

The residual corresponding to the data point is above the line since the calculated residual is a positive value.

c. If x increases 3 units, how does ŷ change?

For each one unit increase of x, y decreases by 0.5; so, for a 3 unit increase of x, y would decrease 1.5 units.

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

The test score would not necessarily be 22 as the model is used to estimate values and built on a sample set of observations. If you used the model to estimate a value, it would be 22, however the actual value of the observation may differ slightly.

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.

$\hat{\sigma}^2 = MSE = \frac{SSE}{(n-p)} = \frac{7}{16-1} = \frac{7}{15} = 0.467$

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:

(rating)^=59.3-2.40*sugars

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

Source	DF	SS	MS	F	P
Regression	**1**	8654.7	8654.7	**102.348**	**1.11e-15**
Error	**75**	**6342.1**	**84.561**		
Total	**76**	14996.8			

a. Find the missing values.

$df_{reg} = 2 - 1 = 1$

$df_{error} = 77 - 2 = 75$

$df - total = 75 + 1 = 76$

SSE = SST - SSR = 14996.8 - 8654.7 = 6342.1

MSE = $\frac{SSE}{n-p} = \frac{6342.1}{75} = 84.561$

F = $\frac{MSR}{MSE} = \frac{8654.7}{84.561} = 102.348$

```
1-pf(102.348, 1, 75)
```

```
[1] 1.110223e-15
```

b. Find R^2 value and interpret that number.

$R^2 = \frac{SSR}{SST} = \frac{8654.7}{14996.8} = 0.577$

57.7% of variation in the ratings is due to the sugars in the cereal.

c. What is the estimated value of σ^2?

$\sigma^2 = MSE = 84.561$

d. Test $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$ using F-test

1.11e-15 < 0.05, therefore we reject the null hypothesis and have strong evidence to conclude the alternative that $\beta_1 \neq 0$, ie that there is a strong linear relationship.

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:

^_0=b_0=58.816, ^_1=b_1=-1.867,s(b_1)=0.346,MSE=4.417

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.

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

$t* = \frac{b_1 - 0}{sb_1} = \frac{-1.867}{0.346} = -5.396$

```
2*(1-pt(5.396, 18))
```

```
[1] 3.972477e-05
```

3.972e-05 < 0.05, so we reject our null that β_1 is equal to 0 and have strong evidence to conclude there is a linear relationship between the amount of time needed to run a 10km race and the time to exhaustion on a treadmill.

b. Construct a 95% confidence interval for 1.

$\hat{\beta}_1 \pm t_{1-\frac{\alpha}{2}, df_{error}} * sd(\hat{\beta}_1)$ = -1.867 $\pm$ (2.101*0.346)

```
qt(.975, 18)
```

```
[1] 2.100922
```

(-2.59, -1.14)

c. Construct a 95% confidence interval for the estimated mean 10-km race when the treadmill time is 10 minutes. Given that sd{yˆ_(x_h=10)}=0.7342

Point Estimate: $y(10) = 58.816 + (-)1.867(10) = 40.146

40.146 ± 2.101(0.7324)

(38.60, 41.69)

d. Construct a 95% prediction interval for an athlete whose treadmill time is 10 minutes. sd(y ˆ*pred*)=√(*mse*+(*sd*{*y*ˆ(x_h=10)})ˆ2)=2.226

40.146 ± 2.101(2.226)

(35.47, 44.82)

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= _0+ _1*X_i+ _i is appropriate.

```
gpa_data<-read.table(file.choose(), header=T)

names(gpa_data)
```

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

```
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
gpa_data |>
  summarise(mean_GPA = mean(GPA), var_GPA = var(GPA),
            mean_ACT = mean(ACT), var_ACT = var(ACT))
```

```
  mean_GPA   var_GPA mean_ACT  var_ACT
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
gpa_data |>
  summarize(correlation = cor(ACT, GPA))
```

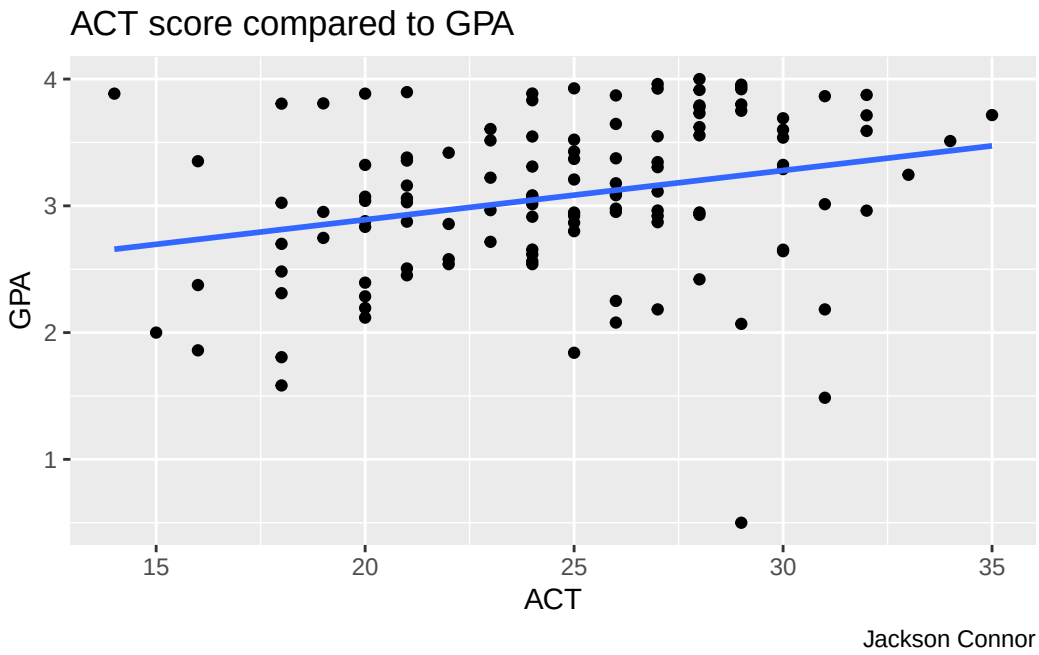
```
  correlation
1   0.2694818
```

There is a weak positive correlation between ACT score and student GPA.

c. Construct a scatter plot. Is the relationship approximately linear?

```r
gpa_data |>
  ggplot(aes(x = ACT, y = GPA)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE) +
  labs(title = "ACT score compared to GPA", caption = "Jackson Connor")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Yes, this relationship is approximately linear.

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

```
gpalm <- lm(GPA ~ ACT, data = gpa_data)
summary(gpalm)
```

```
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 is: $\hat{y} = 2.11 + 0.039$*ACT

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?

Increase by one point: $\hat{y} = 2.11 + 0.039*1 = 2.149$, with a 0.039 increase for each one increase in test score.

Increase by four points: $\hat{y} = 2.11 + 0.039*4 = 2.266$, with a .156 increase for each four point increase in test score.

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

$\hat{y}(20) = 2.11 + 0.039(20) = 2.89$

g. Estimate σ^2 and σ

$\hat{\sigma} = 0.6231$

$\hat{\sigma^2} = 0.6231^2 = 0.3883$

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

PE: 0.039, which is the GPA increase we could expect with every one increase in ACT score.

$0.039 \pm 1.980272(0.01277)$

```
qt(0.975, 118)
```

```
[1] 1.980272
```

```
confint(gpalm, "ACT", 0.95)
```

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

We would expect the true value of increase for each point increase in ACT score to be between (0.0135, 0.0641).

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

$\hat{y}(28) = 2.11 + 0.039(28) = 3.202$

$S_{xx} = (\frac{s}{se\beta_1})^2 = (\frac{0.6231}{.01277})^2 = 2380.86$

$sd(\hat{y}_i) = s\sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}} = 0.6231\sqrt{\frac{1}{120} + \frac{28+24.7}{2380.86}} = 0.0708$

$3.202 \pm 1.980272(0.0708)$

(3.06, 3.34)

We would expect that the true mean GPA for students who scored a 28 on the ACT is between 3.06 and 3.34.

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

$sd(\hat{y}_{pred}) = \sqrt{mse + (sd(\hat{y}_i))^2} = \sqrt{0.6231^2 + 0.0708^2} = 0.6271$

$3.202 \pm 1.980272(0.6271)$

(1.96, 4.44)

We would expect the true GPA of a student who scored a 28 on the ACT to be between 1.96 and 4.44.

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.

```
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, 3.13, 3.98, 4.55, 3.41, 3.35, 3.75, 3.83, 3.18, 3.52,
            3.49, 3.81, 4.03)
systolic_BP <- c(80, 90, 96, 102, 81, 96, 99, 110, 88, 90, 100, 102, 86, 93,
                 101, 103, 86, 91, 100, 105, 84, 91, 95, 104, 107)
data <- data.frame(age, weight, systolic_BP)
```

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

Systolic BP $= \beta_0 + \beta_1 * age + \beta_2 * weight$

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

```
sbp <- lm(systolic_BP ~ age + weight)
summary(sbp)
```

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

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
```

equation: $\hat{y} = 57.26 + 5.80(\text{age}) + 3.32(\text{weight})$

c. Obtain the estimated residual standard deviation.

residual standard deviation $= 2.454$

d. Provide an interpretation of β_2, the coefficient of weight.

Holding age constant, for each increase in weight, systolic blood pressure rises by 3.32.

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

$H_0 : \beta_1 = \beta_2 = 0$

H_a: at least one β is not equal to zero

p-value = 8.696e-13

8.696e-13 $<$ 0.01

Therefore, we can reject the hypothesis that there is no overall predictive value of the model and have strong evidence to conclude that the model does have overall predictive value.