ST512-SP26-Homework-3

Jason Menjivar

Problem 1

a

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

b

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

e = y - $\bar{y}$ = 30 - 23.5 = 6.5

The residual has a value of 6.5. Given that the residual is positive, the point will lie above the regression line. This is because our observed value of 30 is greater than the predicted value of 23.5.

c

In our model or fitted equation, the slope is -0.5. Thus, for every 1 unit increase in x, $\hat{y}$ will decrease by 0.5. If x is increasing by 3 units, this would then mean that $\hat{y}$ will decrease by 1.5 units.

-0.5(3) = -1.5

d

$\hat{y}$ = 25 - 0.5(6) = 22

Based on the fitted equation, the predicted value is 22. However, this is only a **predicted** value, not an actual observed value. There is random error associated with the model, thus the actual observed test score could differ from 22.

e

SSE = 7

n = 16

P = 2 (2 parameters, slope and intercept)

$\sigma^2 = \frac{SSE}{n-P} = \frac{7}{14} = 0.5$

Problem 2

a

Source	DF	SS	MS	F	P
Regression	1	8654.7	8654.7	102.35	1.11e-15
Error	75	6342.1	84.56	–	–
Total	76	14996.8	–	–	–

b

$R^2 = \frac{SSR}{SST} = \frac{8654.7}{6342.1} = 0.577$

Approximately 57.7% of variation in rating can be explained by the sugar content (in g).

c

$\sigma^2 = 84.56$

d

Based on the calculated value in our ANOVA table, the p-value is very small <0.001. If we use a standard significance level of 0.05, we can reject the null hypothesis. Therefore, concluding that there is evidence that sugar content is a significant predictor of cereal ratings.

Problem 3

a

t.value = $\frac{\hat{\beta_1}}{sd(\hat{\beta_1})} = \frac{-1.867}{0.346}$ = -5.396

p.value = 2 * p(t > |t.value|) = .00004

Using a standard significance level of 0.05, our estimated p.value is much smaller. Therefore, reject the null hypothesis, there is a significant relationship between treadmill time and race time.

b

CI: $\hat{\beta}_1 \pm t_{1-\alpha/2, df_{error}} * sd(\hat{\beta}_1)$

t.stat = 2.101

CI: -1.867 $\pm$ 2.101 * 0.346

95% CI: (-2.594, -1.140)

c

58.816 -1.867(10) = 40.146 = point estimate

CI: 40.146 $\pm$ 2.101 * 0.7342

CI: (38.58, 41.71)

d

CI: 40.146 $\pm$ 2.101 * 2.226

CI: (35.47, 44.82)

Problem 4

a

```
  mean_ACT mean_GPA  var_ACT   var_GPA
1   24.725  3.07405 19.99937 0.4151719
```

b

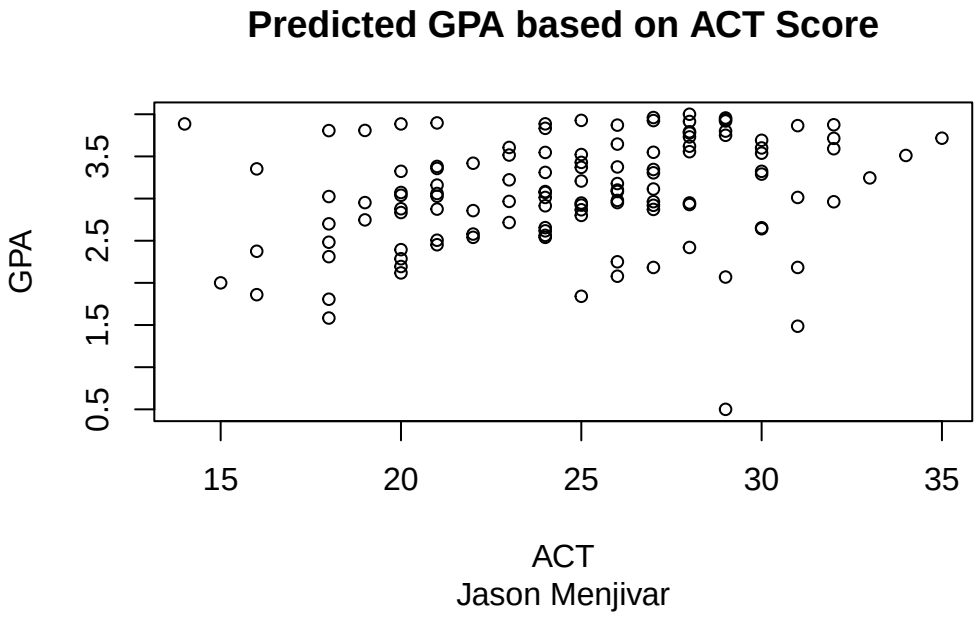
```
(cov(ACT,GPA))/(sd(ACT)*sd(GPA))
```

```
[1] 0.2694818
```

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

c

```
plot(ACT, GPA, cex= 0.8,
     main="Predicted GPA based on ACT Score",
     sub="Jason Menjivar")
```



d

```
regmodel <- lm(GPA~ACT)
summary(regmodel)
```

```
Call:
lm(formula = GPA ~ ACT)

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
```

$\hat{Y} = 2.11405 + 0.03883 * (ACT) + \epsilon$

$\hat{Y}$ = GPA

e

For a 1 point increase in ACT score, the point estimate of the change in mean GPA will be 0.03883. For a 4 point increase in ACT score, the point estimate of the change in the mean GPA will be 0.15532.

f

ACT = 20

GPA = 2.11405 + 0.03883(20) = 2.89

g

Based on regression model from part d:

$\sigma = 0.6231$

$\sigma^2 = (0.6231)^2 = 0.3883$

h

point estimate $(\hat{\beta}_1) = 0.03883$

For every 1 point increase in ACT score, the mean GPA will increase by 0.03883 points.

0.03883 $\pm$ 1.980 * 0.01277

(0.0135, 0.0641)

i

point estimate: 3.201

$s(\hat{Y}) = \sqrt{0.38825(\frac{1}{120} + \frac{3.275^2}{2379.925})} = 0.0706$

3.201 $\pm$ 1.980 * 0.0706

CI: (3.06, 3.34)

j

SE = $\sqrt{0.38825 + 0.00499} = 0.627$

3.201 $\pm$ 1.980 * 0.627

(1.96, 4.44)

Problem 5

a

$Y = \beta_o + \beta_1(age) + \beta_2(weight) + \epsilon$

b

MLR model at the end of doc.

$\hat{Y} = 57.2644 + 5.8041 * (age) + 3.3162 * (weight) + \epsilon$

c

$\hat{e} = 2.454$

d

For every 1 unit increase in weight, the systolic BP will increase by 3.3162 units.

e

Yes, the hypothesis can be rejected. Based on our MLR model, the p-value is 8.69e-13. This is much smaller than $\alpha = 0.01$, thus we would reject the null hypothesis.

R code

problem 1

```
1-pf(102.35, 1, 75)
```

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

problem 3

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

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

```
qt(0.975, 18)
```

```
[1] 2.100922
```

problem 4

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

```
  mean_ACT mean_GPA  var_ACT   var_GPA
1   24.725  3.07405 19.99937 0.4151719
```

problem 5

```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, 3.13, 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)
# Create a data frame
data <- data.frame(age, weight, systolic_BP)
```

```r
summary(lm(systolic_BP ~ age + weight))
```

```
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
```