STA 100 Homework 5 Solution

Yidong Zhou

1. (a) H_0 : Death is independent of treatment (Surgery, Watchful Waiting).
   H_A: Death is dependent of treatment (Surgery, Watchful Waiting).

   (b) See the following table for the row, column, and grand totals.

	Surgery	WW	Total
Died	83	106	189
Alive	264	242	506
Total	347	348	695

The expected frequencies are

$$e_1 = \frac{347 \times 189}{695} = 94.364, \quad e_2 = \frac{348 \times 189}{695} = 94.636,$$

$$e_3 = \frac{347 \times 506}{695} = 252.636, \quad e_4 = \frac{348 \times 506}{695} = 253.364.$$

Therefore, the test statistic is

$$\begin{aligned} T &= \sum_{i=1}^{4} \frac{(o_i - e_i)^2}{e_i} \\ &= \frac{(83 - 94.364)^2}{94.364} + \frac{(106 - 94.636)^2}{94.636} + \frac{(264 - 252.636)^2}{252.636} + \frac{(242 - 253.364)^2}{253.364} \\ &= 3.754. \end{aligned}$$

The null distribution for the test statistic is χ_1^2. The critical value for $\alpha = 0.05$ is thus $\chi_1^2(0.05) = 3.84$.

   (c) From chi-square Table with df $= 1$, we find that $P(\chi_1^2 > 2.71) = 0.10$ and $P(\chi_1^2 > 3.84) = 0.05$. The range of p-value is thus $(0.05, 0.10)$.

   (d) Since p-value $> \alpha = 0.05$, we fail to reject the null at the 0.05 level of significance.

   (e) We support the claim that death is independent of treatment at the 0.05 level of significance.

2. (a) H_0 : Pain is independent of treatment (Angioplasty, Bypass Surgery).
   H_A: Pain is dependent of treatment (Angioplasty, Bypass Surgery).

   (b) See the following table for the row, column, and grand totals.

	A	B	Total
Pain	111	74	185
No pain	402	441	843
Total	513	515	1028

The expected frequencies are

$$e_1 = \frac{513 \times 185}{1028} = 92.32, \quad e_2 = \frac{515 \times 185}{1028} = 92.68,$$

$$e_3 = \frac{513 \times 843}{1028} = 420.68, \quad e_4 = \frac{515 \times 843}{1028} = 422.32.$$

Therefore, the test statistic is

$$
\begin{aligned}
T &= \sum_{i=1}^{4} \frac{(o_i - e_i)^2}{e_i} \\
&= \frac{(111 - 92.32)^2}{92.32} + \frac{(74 - 92.68)^2}{92.68} + \frac{(402 - 420.68)^2}{420.68} + \frac{(441 - 422.32)^2}{422.32} \\
&= 9.20.
\end{aligned}
$$

The null distribution for the test statistic is χ_1^2. The critical value for $\alpha = 0.01$ is thus $\chi_1^2(0.01) = 6.63$.

(c) From chi-square Table with df $= 1$, we find that $P(\chi_1^2 > 6.63) = 0.01$ and $P(\chi_1^2 > 10.83) = 0.001$. The range of p-value is thus $(0.001, 0.01)$.

(d) Since p-value $< \alpha = 0.01$, we reject the null at the 0.01 level of significance.

(e) We cannot support the claim that pain is independent of treatment at the 0.01 level of significance.

3. (a) Here we have $n_1 = 84 + 87 = 171, n_2 = 2916 + 4913 = 7829$. The Wilson-adjusted sample proportions are

$$
\tilde{p}_1 = \frac{84 + 1}{171 + 2} = 0.4913, \quad \tilde{p}_2 = \frac{2916 + 1}{7829 + 2} = 0.3725.
$$

The standard error for $\tilde{p}_1 - \tilde{p}_2$ is

$$
\mathrm{SE}_{\tilde{p}_1 - \tilde{p}_2} = \sqrt{\frac{0.4913 \times (1 - 0.4913)}{171 + 2} + \frac{0.3725 \times (1 - 0.3725)}{7829 + 2}} = 0.0384.
$$

The 95% confidence interval for $p_1 - p_2$ is thus

$$
(0.4913 - 0.3725) \pm 1.96 \times 0.0384
$$

or $(0.0435, 0.1941)$.

(b) We are 95% confident that the difference in the proportion of people who develop CHD between smokers and nonsmokers falls within the range of 0.0435 to 0.1941, with smokers having a higher proportion.

(c) Yes, it suggests a dependence on CHD and smoking since the confidence interval does not contain 0.

(d) No, it does not support the claim since 0.20 is not included in the interval.

4. (a) Here we have $I = 4$, $n = \sum_{i=1}^{4} n_i = 40$, and

$$
\begin{aligned}
\bar{Y} &= \frac{\sum_{i=1}^{4} n_i \bar{Y}_i}{n} \\
&= \frac{10 \times 3.22 + 10 \times 3.57 + 10 \times 2.87 + 10 \times 2.98}{40} \\
&= 3.16.
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\mathrm{SSB} &= \sum_{i=1}^{4} n_i (\bar{Y}_i - \bar{Y})^2 \\
&= 10 \times (3.22 - 3.16)^2 + 10 \times (3.57 - 3.16)^2 + 10 \times (2.87 - 3.16)^2 + 10 \times (2.98 - 3.16)^2 \\
&= 2.882.
\end{aligned}
$$

and

$$\begin{aligned}
\text{SSW} &= \sum_{i=1}^{4}(n_i - 1)s_i^2 \\
&= (10-1)\times 0.54^2 + (10-1)\times 0.35^2 + (10-1)\times 0.21^2 + (10-1)\times 0.23^2 \\
&= 4.600.
\end{aligned}$$

Therefore,

$$\text{SSTO} = \text{SSB} + \text{SSW} = 2.882 + 4.600 = 7.482.$$

If follows that

$$\text{MSB} = \frac{\text{SSB}}{4-1} = 0.9607, \quad \text{MSW} = \frac{\text{SSW}}{40-4} = 0.1278.$$

The ANOVA table is as follows.

Source	**df**	**SS**	**MS**
Between groups	3	2.882	0.9607
Within groups	36	4.600	0.1278
Total	39	7.482	

(b) H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ v.s. H_A: The μ_i's are not all equal.

(c) The test statistic is

$$T = \frac{\text{MSB}}{\text{MSW}} = \frac{0.9607}{0.1278} = 7.5172.$$

The null distribution of the test statistic is $F_{3,36}$. The critical value for $\alpha = 0.01$ is thus $F_{3,30}(0.01) = 4.51$.

(d) From F Table with numerator df = 3 and denominator df = 30, we find that $P(F_{3,30} > 7.05) = 0.001$ and $P(F_{3,30} > 9.99) = 0.0001$. The range of p-value is thus $(0.0001, 0.001)$.

(e) Since the p-value $< \alpha = 0.01$, we reject the null at the 0.01 level of significance.

(f) We conclude at the 0.01 level of significance that at least two of the average GPA's of the four sororities are different.

(g) We could falsely reject the null and thus possibly made a Type I error.

(h) To construct family-wise 99% confidence intervals for $\mu_2-\mu_1, \mu_2-\mu_3$, and $\mu_2-\mu_4$. The individual coverage probability for each confidence interval is $1-\alpha/3$ where $\alpha = 0.01$. The $1-\alpha/3$ confidence interval for $\mu_i - \mu_j$ is given by

$$\left(\bar{Y}_i - \bar{Y}_j\right) \pm t_{n-I}(\alpha/(2\times 3)) \times \text{SE}_{\bar{Y}_i - \bar{Y}_j}.$$

Running `qt(p = 1 - 0.01 / (2 * 3), df = 40 - 4)` in R, we know that $t_{36}(0.01/6) = 3.143858$. The family-wise 99% confidence intervals for $\mu_2 - \mu_1, \mu_2 - \mu_3$, and $\mu_2 - \mu_4$ are thus as follows.

$$\begin{aligned}
&(\bar{Y}_2 - \bar{Y}_1) \pm t_{36}(0.01/6) \times \sqrt{\text{MSW} \times \left(\frac{1}{n_2} + \frac{1}{n_1}\right)} \\
&= (3.57 - 3.22) \pm 3.143858 \times \sqrt{0.1278 \times \left(\frac{1}{10} + \frac{1}{10}\right)} \\
&= (-0.1526, 0.8526),
\end{aligned}$$

$$\begin{aligned}
&(\bar{Y}_2 - \bar{Y}_3) \pm t_{36}(0.01/6) \times \sqrt{\text{MSW} \times \left(\frac{1}{n_2} + \frac{1}{n_3}\right)} \\
&= (3.57 - 2.87) \pm 3.143858 \times \sqrt{0.1278 \times \left(\frac{1}{10} + \frac{1}{10}\right)} \\
&= (-0.1974, 1.2026),
\end{aligned}$$

$$
\begin{aligned}
&(\bar{Y}_2 - \bar{Y}_4) \pm t_{36}(0.01/6) \times \sqrt{\text{MSW} \times \left(\frac{1}{n_2} + \frac{1}{n_4}\right)} \\
&= (3.57 - 2.98) \pm 3.143858 \times \sqrt{0.1278 \times \left(\frac{1}{10} + \frac{1}{10}\right)} \\
&= (0.0874, 1.0926).
\end{aligned}
$$

(i) The confidence interval for $\mu_2 - \mu_4$ suggest a significant difference in the means as it does not include zero.

5. (a) Here we have $I = 3, n = \sum_{i=1}^{3} n_i = 21$, and

$$
\begin{aligned}
\bar{Y} &= \frac{\sum_{i=1}^{3} n_i \bar{Y}_i}{n} \\
&= \frac{7 \times 2.57 + 7 \times 3.71 + 7 \times 4.29}{21} \\
&= 3.5233.
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\text{SSB} &= \sum_{i=1}^{3} n_i (\bar{Y}_i - \bar{Y})^2 \\
&= 7 \times (2.57 - 3.5233)^2 + 7 \times (3.71 - 3.5233)^2 + 7 \times (4.29 - 3.5233)^2 \\
&= 10.72027
\end{aligned}
$$

and

$$
\begin{aligned}
\text{SSW} &= \sum_{i=1}^{3} (n_i - 1) s_i^2 \\
&= (7 - 1) \times (0.98^2 + 1.11^2 + 1.38^2) \\
&= 24.5814.
\end{aligned}
$$

Therefore,

$$\text{SSTO} = \text{SSB} + \text{SSW} = 10.7203 + 24.5814 = 35.3017.$$

If follows that

$$\text{MSB} = \frac{\text{SSB}}{3 - 1} = 5.3601, \quad \text{MSW} = \frac{\text{SSW}}{21 - 3} = 1.3656.$$

The ANOVA table is as follows.

Source	**df**	**SS**	**MS**
Between groups	2	10.7202	5.3601
Within groups	18	24.5814	1.3656
Total	20	35.3017	

(b) H_0: $\mu_1 = \mu_2 = \mu_3$ v.s. H_A: The μ_i's are not all equal.

(c) The test statistic is

$$T = \frac{\text{MSB}}{\text{MSW}} = \frac{5.3601}{1.3656} = 3.9251.$$

The null distribution of the test statistic is $F_{2,18}$. The critical value for $\alpha = 0.01$ is thus $F_{2,18}(0.01) = 6.01$.

(d) From F Table with numerator df = 2 and denominator df = 18, we find that $P(F_{2,18} > 3.55) = 0.05$ and $P(F_{2,18} > 4.90) = 0.02$. The range of p-value is thus $(0.02, 0.05)$.

(e) Since the p-value $> \alpha = 0.01$, we fail to reject the null at the 0.01 level of significance.

(f) There is insufficient evidence to conclude that there is any difference among the average number of calls in the morning, afternoon and night shifts, at the 0.01 level of significance.

(g) We could falsely fail to reject the null and thus possibly made a Type II error.

(h) To construct family-wise 99% confidence intervals for $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, and $\mu_2 - \mu_3$. The individual coverage probability for each confidence interval is $1 - \alpha/3$ where $\alpha = 0.01$. The $1 - \alpha/3$ confidence interval for $\mu_i - \mu_j$ is given by

$$\left(\bar{Y}_i - \bar{Y}_j\right) \pm t_{n-I}(\alpha/(2 \times 3)) \times \mathrm{SE}_{\bar{Y}_i - \bar{Y}_j}.$$

Running `qt(p = 1 - 0.01 / (2 * 3), df = 21 - 3)` in R, we know that $t_{18}(0.01/6) = 3.380362$. The family-wise 99% confidence intervals for $\mu_1 - \mu_2, \mu_1 - \mu_3$, and $\mu_2 - \mu_3$ are thus as follows.

$$\begin{aligned}
&(\bar{Y}_1 - \bar{Y}_2) \pm t_{18}(0.01/6) \times \sqrt{\mathrm{MSW} \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \\
&= (2.57 - 3.71) \pm 3.380362 \times \sqrt{1.3656 \times \left(\frac{1}{7} + \frac{1}{7}\right)} \\
&= (-3.2515, 0.9715),
\end{aligned}$$

$$\begin{aligned}
&(\bar{Y}_1 - \bar{Y}_3) \pm t_{18}(0.01/6) \times \sqrt{\mathrm{MSW} \times \left(\frac{1}{n_1} + \frac{1}{n_3}\right)} \\
&= (2.57 - 4.29) \pm 3.380362 \times \sqrt{1.3656 \times \left(\frac{1}{7} + \frac{1}{7}\right)} \\
&= (-3.8315, 0.3915),
\end{aligned}$$

$$\begin{aligned}
&(\bar{Y}_2 - \bar{Y}_3) \pm t_{18}(0.01/6) \times \sqrt{\mathrm{MSW} \times \left(\frac{1}{n_2} + \frac{1}{n_3}\right)} \\
&= (3.71 - 4.29) \pm 3.380362 \times \sqrt{1.3656 \times \left(\frac{1}{7} + \frac{1}{7}\right)} \\
&= (-2.6915, 1.5315).
\end{aligned}$$

(i) Yes, the confidence intervals are consistent with the conclusion in (f) since all of them include zero.

6. (a) The response variable is the peak flow, and the explanatory variable is the height.

(b) The slope is

$$b_1 = r\frac{s_Y}{s_X} = 0.32725 \times \frac{117.9952}{8.5591} = 4.5114.$$

The intercept is

$$b_0 = \bar{Y} - b_1\bar{X} = 660 - 4.5114 \times 180.4118 = -153.9098.$$

(c) Slope: When height of male increases by 1 cm, we expect peak flow to increase by 4.5114 liters/min, on average.
Intercept: No practical meaning because height can never equal to 0.

(d) The prediction based on the fitted regression line is

$$\hat{Y} = -153.9098 + 4.5114 \times 174 = 631.0738.$$

(e) Their average difference in peak flow would be

$$10 \times b_1 = 45.114.$$

7. (a) $H_0 : \rho = 0$ v.s. $H_A : \rho \neq 0$ or $H_0 : \beta_1 = 0$ v.s. $H_A : \beta_1 \neq 0$.

(b) The test statistic is

$$T = r\sqrt{\frac{n-2}{1-r^2}} = 0.32725 \times \sqrt{\frac{17-2}{1-0.32725^2}} = 1.3413.$$

The null distribution for the test statistic is t_{15}. The critical value for $\alpha = 0.05$ is thus $t_{15}(0.05/2) = 2.131$.

(c) From t Table with df = 15, we find that $P(t_{15} > 1.341) = 0.10$ and $P(t_{15} > 1.753) = 0.05$. The range of p-value is thus $(0.10, 0.20)$.

(d) Since the p-value $> \alpha = 0.05$, we fail to reject the null at the 0.05 level of significance.

(e) There is insufficient evidence to conclude that that peak flow is linearly related with height.

(f) The residual standard deviation is

$$s_e = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{198909.3}{17-2}} = 115.1548.$$

The standard error for b_1 is

$$\text{SE}_{b_1} = \frac{s_e}{s_X\sqrt{n-1}} = \frac{115.1548}{8.5591 \times \sqrt{17-1}} = 3.3635.$$

The 95% confidence interval for the slope is thus

$$4.5114 \pm 2.131 \times 3.3635$$

or $(-2.6562, 11.6790)$. The confidence interval is consistent with the conclusion in (e) since it includes zero.

(g) We are 95% confident that the when the height increases by 1 cm, we expect the peak flow to increase by between -2.6562 liters/min and 11.6790 liters/min on average.

Problem 8

```
blood = read.csv("blood.csv", header = T)
head(blood)
```

```
##   Type Disease
## 1    O     yes
## 2    O     yes
## 3    O     yes
## 4    O     yes
## 5    O     yes
## 6    O     yes
```

(a) – (b)

```
# create the observaion table, group by Type and Disease
O = xtabs(~ Type + Disease, data = blood)
test_result = chisq.test(O)
```

```
## Warning in chisq.test(O): Chi-squared approximation may be incorrect
```

```
test_result
```

```
##
##  Pearson's Chi-squared test
##
## data:  O
## X-squared = 10.654, df = 3, p-value = 0.01375
```

According to the R result, the test-statistic is $T = 10.654$, and p-value is 0.01375.

(c) Since p-value $> \alpha = 0.01$, we fail to reject the null at the 0.01 level of significance. We conclude that blood type and whether to develop a disease are independent.

```
test_result$observed
```

```
##     Disease
## Type no yes
##   A  12  15
##   AB  7   2
##   B   8  17
##   O   9  30
```

```
test_result$expected
```

```
##     Disease
## Type    no   yes
##   A   9.72 17.28
##   AB  3.24  5.76
##   B   9.00 16.00
##   O  14.04 24.96
```

(d) The observed frequency for blood type A is 15, while the expected frequency is 17.28. Blood type A is thus less likely to have the disease than what we expected if the null was true.

(e) The observed frequency for blood type A is 30, while the expected frequency is 24.96. Blood type A is thus more likely to have the disease than what we expected if the null was true.

(f)

```r
(test_result$observed - test_result$expected)^2/test_result$expected
```

```
##     Disease
## Type        no       yes
##   A  0.5348148 0.3008333
##   AB 4.3634568 2.4544444
##   B  0.1111111 0.0625000
##   O  1.8092308 1.0176923
```

The group blood type "AB" and no disease contributes most to the test statistic.

Problem 9

```r
IQ = read.csv("IQ.csv")

head(IQ)
```

```
##   group iq
## 1     A 44
## 2     A 40
## 3     A 44
## 4     A 39
## 5     A 25
## 6     A 37
```

(a)

```r
anova = aov(iq ~ group, data = IQ)
summary(anova)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## group        2   1529   764.7   20.02 7.84e-07 ***
## Residuals   42   1604    38.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

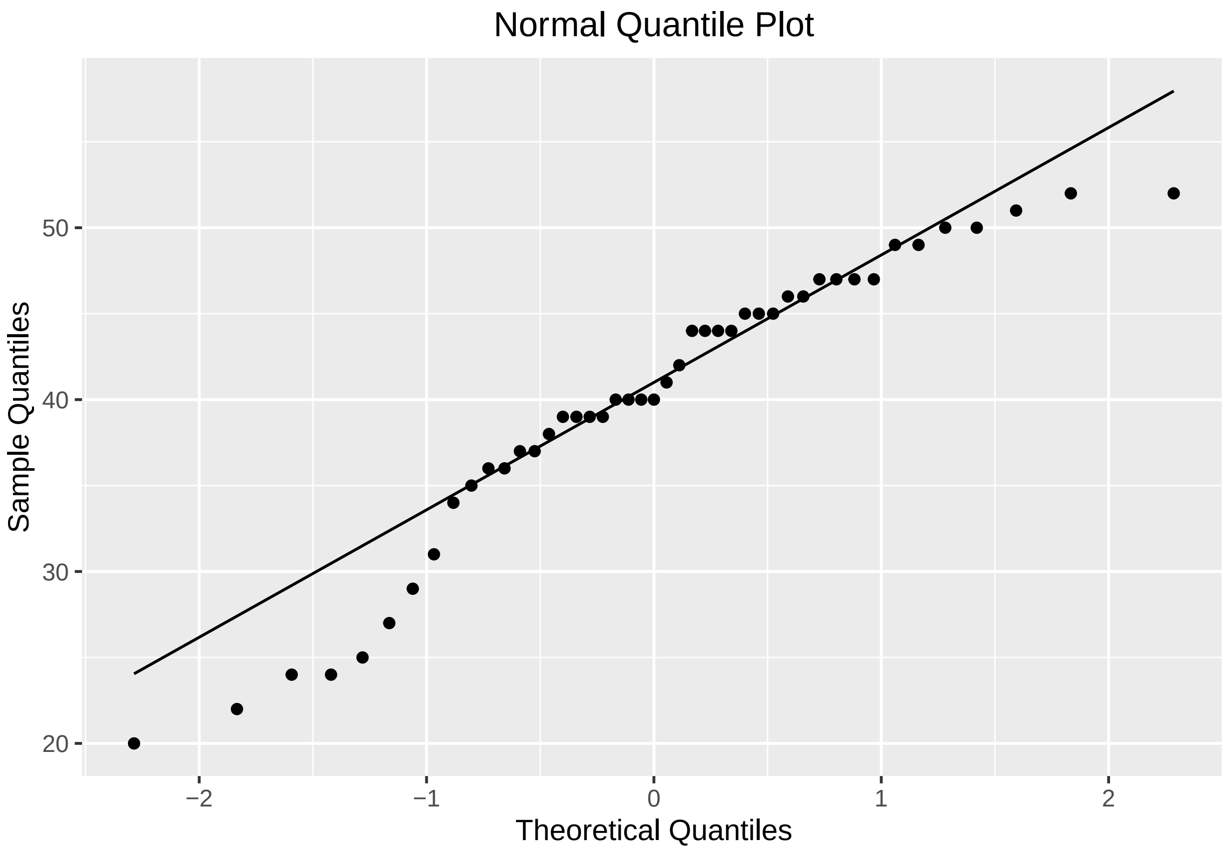
(b) The test statistic is 20.02 and the p-value is 7.84×10^{-7}.

(c) Since p-value $< \alpha = 0.05$, we fail to reject the null at the 0.05 level of significance.

(d) There is significant difference for the mean IQ of students among the three majors.

(e)

```r
library(ggplot2)
ggplot(IQ,
       aes(sample = iq)) +
  stat_qq() +
  stat_qq_line() +
  labs(x = "Theoretical Quantiles",
       y = "Sample Quantiles",
       title = "Normal Quantile Plot") +
  theme(plot.title = element_text(hjust = 0.5))
```



This data do not appear to be approximately normally distributed.

(f)

```
library(asbio)
```

```
## Loading required package: tcltk
```

```
bonfCI(y = IQ$iq, x = factor(IQ$group), conf.level = 0.95)
```

```
##
## 95% Bonferroni confidence intervals
##
##                 Diff     Lower    Upper  Decision Adj. p-value
## muA-muB  -0.06667  -5.69471  5.56138    FTR H0            1
## muA-muC     -12.4 -18.02805 -6.77195 Reject H0        6e-06
## muB-muC -12.33333 -17.96138 -6.70529 Reject H0        7e-06
```

(g) The confidence intervals for $\mu_A - \mu_C$ and $\mu_B - \mu_C$ suggest a significant difference in the means.

Problem 10

```
fitness = read.csv("fitness.csv")
head(fitness)
```

```
##   Tread  Run
## 1   7.5 43.5
## 2   7.8 45.2
## 3   7.9 44.9
```

```
## 4   8.1 41.1
## 5   8.3 43.8
## 6   8.7 44.4
```

(a)

```r
reg = lm(Run ~ Tread, data = fitness)
summary(reg)
```

```
##
## Call:
## lm(formula = Run ~ Tread, data = fitness)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9440 -1.5788  0.1860  0.7863  4.5603
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  59.9211     3.1166  19.226 1.90e-13 ***
## Tread        -1.9601     0.3164  -6.194 7.59e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.921 on 18 degrees of freedom
## Multiple R-squared:  0.6807, Adjusted R-squared:  0.6629
## F-statistic: 38.37 on 1 and 18 DF,  p-value: 7.589e-06
```

The slope and intercept of the fitted regression line are -1.9601 and 59.9211.

(b)

```r
confint(reg, 'Tread', level = 0.95)
```

```
##           2.5 %    97.5 %
## Tread -2.624957 -1.295313
```

(c) From the summary table, we find that $s_e = 1.921$.

(d) From the summary table, we find that $r^2 = 0.6807$.

(e) Yes, the interval suggests a significant linear relationship.