STAT 420 – Homework 3

1. Time Use (without R)

a. We're already given two of the necessary values for the ANOVA table.

$$\text{SSTotal} = SYY = \sum\left(y_i - \bar{y}\right)^2 = 2170; \qquad \text{SSError} = RSS = \sum\left(y_i - \hat{y}_i\right)^2 = 442$$

Next, calculate SSRegression.

$$\text{SSReg} = \text{SSTotal} - \text{SSError} = 2170 - 442 = 1728 \text{, or}$$

$$\text{SSReg} = \hat{\beta}_1^2 \cdot SXX = \left(-4\right)^2 \cdot 108 = 1728$$

Completing the ANOVA table,

Source	SS	df	MS	F
Regression	$\sum\left(\hat{y}_i - \bar{y}\right)^2 = 1728$	$p - 1 = 1$	1728	**19.55**
Error	$\sum\left(y_i - \hat{y}_i\right)^2 = 442$	$n - p = 5$	88.4	
Total	$\sum\left(y_i - \bar{y}\right)^2 = 2170$	$n - 1 = 6$		

According to the F-distribution, the critical region is $F > F_{\alpha}(1,5) = F_{0.05}(1,5) =$ **6.61**. Since the test statistic does lie the critical region, we reject H_0 and conclude that the model does a significant job of predicting physical activity hours.

b. Calculate the t-test statistic.

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{s_e / \sqrt{SXX}} = \frac{\left(-4\right) - 0}{9.402 / \sqrt{108}} = -4.421$$

There are $n - 2 = 5$ degrees of freedom. According to the t-distribution, the critical region is $|t| > t_{\alpha/2}(5) = t_{0.025}(5) = 2.571$. Since the test statistic does lie the critical region, we reject H_0 and conclude that the model does a significant job of predicting physical activity hours.

Note: This is the same decision as in part a, and $(t)^2 = (-4.421)^2 = 19.55 = F$.

c. We are 90% confident that the average change in TV viewing due to a one hour increase in physical activity is between –5.8 and –2.2 hours.

$$\hat{\beta}_1 \pm t_{\alpha/2} \cdot \frac{s_e}{\sqrt{SXX}} = -4 \pm t_{0.05} \cdot \frac{s_e}{\sqrt{SXX}} = -4 \pm 2.015 \cdot \frac{9.402}{\sqrt{108}} = -4 \pm 1.823 = \left(-5.8, -2.2\right)$$

d. Calculate the t-test statistic.

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{s_e / \sqrt{SXX}} = \frac{\left(-4\right) - \left(-2\right)}{9.402 / \sqrt{108}} = -2.211$$

There are $n - 2 = 5$ degrees of freedom. According to the t-distribution, the critical region is $|t| > t_{\alpha}(5) = t_{0.10}(5) = 1.476$. Since the test statistic does lie the critical region, we reject H_0 and conclude that each additional hour of physical activity would result in at least two fewer hours of TV viewing.

e. Calculate the t-test statistic using the standard deviation of $\hat{\beta}_0$.

$$t = \frac{\hat{\beta}_0 - \beta_{00}}{s_e \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SXX}}} = \frac{104-100}{9.402 \cdot \sqrt{\frac{1}{7} + \frac{18^2}{108}}} = 0.240$$

There are $n - 2 = 5$ degrees of freedom. According to the t-distribution, the critical region is $|t| > t_\alpha(5) = t_{0.05}(5) = 2.015$. Since the test statistic does not lie the critical region, we fail to reject H_0 and conclude that there's not enough evidence to support the fitness guru's claim.

f. We are 90% confident that the mean number of TV viewing hours in a week when a person engages in 20 hours of physical activity is between 16 and 32 hours.

$$\hat{y} \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}} = (104 - 4 \cdot 20) \pm t_{0.05}(5) \cdot 9.402 \cdot \sqrt{\frac{1}{7} + \frac{(20-18)^2}{108}}$$

$$= 24 \pm 2.015 \cdot 9.402 \cdot \sqrt{\frac{1}{7} + \frac{(20-18)^2}{108}} = 24 \pm 8 = (16, 32)$$

g. There's a 90% probability that the number of TV viewing hours in a week when a person engages in 20 hours of physical activity will be between 3.4 and 44.6 hours.

$$\hat{y} \pm t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}} = (104 - 4 \cdot 20) \pm t_{0.05}(5) \cdot 9.402 \cdot \sqrt{1 + \frac{1}{7} + \frac{(20-18)^2}{108}}$$

$$= 24 \pm 2.015 \cdot 9.402 \cdot \sqrt{1 + \frac{1}{7} + \frac{(20-18)^2}{108}} = 24 \pm 20.6 = (3.4, 44.6)$$

h. Calculate the t-test statistic using the standard deviation of $\mathrm{E}[Y \mid x = 14] = \mu_{Y|x=14}$ which is the same as when calculating a confidence interval for $\mu_{Y|x}$ as in part f.

$$t = \frac{\hat{y} - \mu_0}{s_e \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}}} = \frac{(104 - 4 \cdot 14) - 40}{9.402 \cdot \sqrt{\frac{1}{7} + \frac{(14-18)^2}{108}}} = 1.577$$

There are $n - 2 = 5$ degrees of freedom. According to the t-distribution, the critical region is $|t| > t_\alpha(5) = t_{0.05}(5) = 2.015$. Since the test statistic does not lie the critical region, we fail to reject H_0. We conclude that there's not enough evidence to suggest that a person who engages in only 2 hours of physical activity per day (14 hours per week) will watch more than 40 hours of TV in that week.

2. Time Use (with R)

a.

b.

```
> TVfit <- lm(y~x)
> summary(TVfit)

Call:
lm(formula = y ~ x)

Residuals:
  1   2   3   4   5   6   7 
-10  -4   3   4 -11  12   6 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) 104.0000    16.6682   6.239  0.00155 **
x            -4.0000     0.9047  -4.421  0.00688 **              ## part b
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.402 on 5 degrees of freedom
Multiple R-squared:  0.7963,     Adjusted R-squared:  0.7556 
F-statistic: 19.55 on 1 and 5 DF,  p-value: 0.006884              ## part a
```

c.

```
> confint(TVfit, level=.90)
               5 %    95 %
(Intercept) 70.413 137.587
x           -5.823  -2.177
```

d.

```
> sig = summary(TVfit)$sigma;    sig
[1] 9.402
> t = (-4- -2)/(sig/sqrt(108));  t
[1] -2.211
> p.value = pt(t, 5);            p.value
[1] 0.03902
```

Just for some extra knowledge, here's how to do it just with `lm`. Note that the *p*-value is for a two-sided alternative, so divide it by 2 to match our one-sided test.

```
> fit.d=lm(y~x+offset(-2*x))
> summary(fit.d)
...

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  104.000     16.668    6.24   0.0015 **
x             -2.000      0.905   -2.21   0.0780 . 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

e.

```
> se = summary(TVfit)$coef[1,2]; se   ## SE of beta0hat from table
[1] 16.67
> t = (104-100)/(se);            t
[1] 0.24
> p.value = 1 - pt(t, 5);        p.value
[1] 0.4099
```

f.

```
> predict(TVfit, data.frame(x=20), interval=c("conf"), level=.90)
  fit   lwr   upr
1  24 15.96 32.04
```

g.

```
> predict(TVfit, data.frame(x=20), interval=c("pred"), level=.90)
  fit   lwr   upr
1  24 3.421 44.58
```

h. You can calculate the SE of the estimate for $\mu_{Y|x}$ using R like a calculator.

```
> t = (104-4*14 - 40)/(9.402*sqrt(1/7+(18-14)**2/108));  t
[1] 1.577
> p.value = 1 - pt(t, 5);                                p.value
[1] 0.08777
```

Or use `predict.lm`. The option `se.fit` will include the calculation in the output for you.

```
> SE =
predict(TVfit,data.frame(x=14),interval=c("conf"),level=.90,se.fit=T)$se.f
it; SE
[1] 5.072
> t = (104-4*14 - 40)/SE;     t
[1] 1.577
> p.value = 1 - pt(t, 5);     p.value
[1] 0.08778
```

3. Cars (with R)

a.

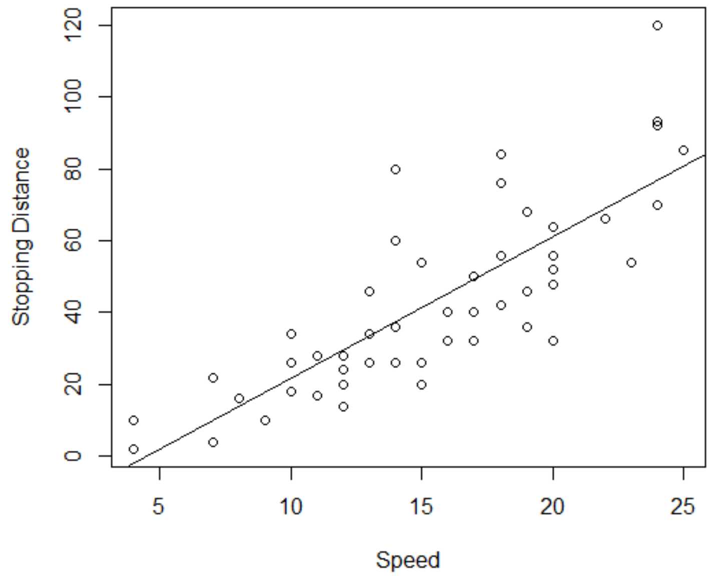
```
> car.fit <- lm(dist ~ speed, data=cars)
> car.fit

Call:
lm(formula = dist ~ speed, data = cars)

Coefficients:
(Intercept)        speed
     -17.58         3.93
```

b.

```
> plot(cars$speed, cars$dist)
> abline(car.fit)
```



c.

```
> confint(car.fit, level=.90)
                 5 %    95 %
(Intercept) -28.915 -6.244
speed         3.236  4.629
```

d. Viewing either *t*-test on β_1 or the equivalent *F*-test on the model, we see that the *p*-value of nearly zero (1.5×10^{-12}) is enough evidence to suggest that the slope (and thus the model) is significant.

```
> summary(car.fit)

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
   Min      1Q Median      3Q     Max
-29.07  -9.53  -2.27    9.21   43.20

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.579      6.758   -2.60    0.012 *
speed          3.932      0.416    9.46  1.5e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.4 on 48 degrees of freedom
Multiple R-squared:  0.651,     Adjusted R-squared:  0.644
F-statistic: 89.6 on 1 and 48 DF,  p-value: 1.49e-12
```