STAT 420 – Homework 4

1. Restaurant Wait Times (without R)

a.

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \left(X'X\right)^{-1}\left(X'Y\right) = \begin{bmatrix} 0.6 & -0.2 & -0.2 \\ -0.2 & 0.1 & 0 \\ -0.2 & 0 & 0.4 \end{bmatrix} \begin{bmatrix} 30 \\ 67 \\ 18 \end{bmatrix} = \begin{bmatrix} 1.0 \\ 0.7 \\ 1.2 \end{bmatrix}$$

b. The t-test is not an option for the multiple regression model. Use an F-test.

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

$\text{SSTotal} = SYY = \sum\left(y_i - \bar{y}\right)^2 = 13.5$; $\text{SSError} = RSS = \sum\left(y_i - \hat{y}_i\right)^2 = 5$

Next, calculate SSRegression: $\text{SSReg} = \text{SSTotal} - \text{SSError} = 13.5 - 5 = 8.5$.

Completing the ANOVA table,

Source	SS	df	MS	F
Regression	$\sum\left(\hat{y}_i - \bar{y}\right)^2 = 8.5$	$p - 1 = 2$	4.25	**5.95**
Error	$\sum\left(y_i - \hat{y}_i\right)^2 = 5$	$n - p = 7$	0.714	
Total	$\sum\left(y_i - \bar{y}\right)^2 = 13.5$	$n - 1 = 9$		

According to the F-distribution, the critical region is $F > F_{\alpha}(2,7) = F_{0.05}(2,7) =$ **4.74**. Since the test statistic does lie the critical region, we reject H_0 and conclude that the model does a significant job of predicting wait time.

c. Calculate estimated variance of the slope estimate. From the ANOVA table, we'll use MSE = 0.714 as the estimate of the variance of the residuals, and we'll pull C_{33} from $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$.

$$\hat{\text{V}}\text{ar}\left[\hat{\beta}_2\right] = \hat{\sigma} \cdot C_{33} = (0.714)(0.4) = 0.2856$$

Calculate the test statistic.

$$t = \frac{\hat{\beta}_2 - \beta_{20}}{\sqrt{\hat{\text{V}}\text{ar}\left[\hat{\beta}_2\right]}} = \frac{1.2 - 0}{\sqrt{0.2856}} = 2.245$$

There are $n - p = 7$ degrees of freedom. According to the t-distribution, the critical region is $|t| > t_{\alpha/2}(n - p) = t_{0.025}(7) = 2.365$. Since the test statistic does <u>not</u> lie the critical region (barely), we fail to reject H_0 and conclude that whether the order is for dine-in is <u>not</u> a significant predictor in the model.

d. Calculate estimated variance of the slope estimate. From the ANOVA table, we'll use MSE = 0.714 as the estimate of the variance of the residuals, and we'll pull C_{22} from $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$.

$$\hat{\text{V}}\text{ar}\left[\hat{\beta}_1\right] = \hat{\sigma} \cdot C_{22} = (0.714)(0.1) = 0.0714$$

Calculate the test statistic.

$$t = \frac{\hat{\beta}_2 - \beta_{20}}{\sqrt{\hat{\mathrm{V}}\mathrm{ar}\left[\hat{\beta}_2\right]}} = \frac{0.7-1}{\sqrt{0.0714}} = -1.12$$

There are $n - p = 7$ degrees of freedom. According to the t-distribution, the critical region is $|t| > t_{\alpha}(n-p) = t_{0.05}(7) = 1.895$. Since the (absolute value of the) test statistic does <u>not</u> lie the critical region, we fail to reject H_0 and conclude that it's more plausible that each additional item order adds at least an extra minute to wait time.

e. Calculate estimated variance of the intercept estimate. We'll use MSE = 0.714 as the estimate of the variance of the residuals, and we'll pull C_{11} from $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$.

$$\hat{\mathrm{V}}\mathrm{ar}\left[\hat{\beta}_0\right] = \hat{\sigma} \cdot C_{11} = (0.714)(0.6) = 0.4284$$

The 95% confidence interval for β_0 is

$$\hat{\beta}_0 \pm t_{\alpha/2}(n-p) \cdot \sqrt{\hat{\mathrm{V}}\mathrm{ar}\left[\hat{\beta}_0\right]} = 1.0 \pm t_{0.025}(7) \cdot \sqrt{0.4284} = 1.0 \pm 2.365 \cdot 0.6545 = 1.0 \pm 1.55 = (-0.55, 2.55)$$

f. We already calculated $\hat{\mathrm{V}}\mathrm{ar}\left[\hat{\beta}_2\right]$ back in part c. So, the 90% confidence interval for β_2 is

$$\hat{\beta}_2 \pm t_{\alpha/2}(n-p) \cdot \sqrt{\hat{\mathrm{V}}\mathrm{ar}\left[\hat{\beta}_2\right]} = 1.2 \pm t_{0.05}(7) \cdot \sqrt{0.2856} = 1.2 \pm 1.895 \cdot 0.5344 = 1.2 \pm 1.01 = (0.19, 2.21)$$

g. The vector representing these predictors is $\boldsymbol{x}_0 = \begin{bmatrix} 1 & 3 & 1 \end{bmatrix}$. The estimate for the average wait time is

$$\hat{y} = \begin{bmatrix} 1 & 3 & 1 \end{bmatrix} \begin{bmatrix} 1.0 \\ 0.7 \\ 1.2 \end{bmatrix} = 1 + 0.7(3) + 1.2(1) = 4.3.$$

To calculate the estimate for the variance of the estimate, we need

$$\boldsymbol{x}_0'(X'X)^{-1}\boldsymbol{x}_0 = \begin{bmatrix} 1 & 3 & 1 \end{bmatrix} \begin{bmatrix} 0.6 & -0.2 & -0.2 \\ -0.2 & 0.1 & 0 \\ -0.2 & 0 & 0.4 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 1 \end{bmatrix} \begin{bmatrix} -0.2 \\ 0.1 \\ 0.2 \end{bmatrix} = 0.3.$$

So, the 95% confidence interval is

$$\hat{y} \pm t_{\alpha/2}(n-p) \cdot \sqrt{\hat{\mathrm{V}}\mathrm{ar}\left[\hat{Y} \mid x\right]} = 4.3 \pm t_{0.025}(7) \cdot \sqrt{\hat{\sigma}^2 \cdot 0.3} = 4.3 \pm 2.365 \cdot \sqrt{0.714 \cdot 0.3}$$

$$= 4.3 \pm 1.09 = (3.21, 5.39)$$

We are 95% that the average wait time for a dine-in order of 3 items is between 3.21 and 5.39 minutes.

h. The vector representing these predictors is $\boldsymbol{x}_0 = \begin{bmatrix} 1 & 4 & 0 \end{bmatrix}$. The estimate for the average wait time is

$$\hat{y} = \begin{bmatrix} 1 & 4 & 0 \end{bmatrix} \begin{bmatrix} 1.0 \\ 0.7 \\ 1.2 \end{bmatrix} = 1 + 0.7(4) + 1.2(0) = 3.8 .$$

To calculate the estimate for the variance of the estimate, we need

$$\boldsymbol{x}_0'(X'X)^{-1}\boldsymbol{x}_0 = \begin{bmatrix} 1 & 4 & 0 \end{bmatrix} \begin{bmatrix} 0.6 & -0.2 & -0.2 \\ -0.2 & 0.1 & 0 \\ -0.2 & 0 & 0.4 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 4 & 0 \end{bmatrix} \begin{bmatrix} -0.2 \\ 0.2 \\ -0.2 \end{bmatrix} = 0.6 .$$

So, the 95% prediction interval is

$$\hat{y} \pm t_{\alpha/2}(n-p) \cdot \sqrt{\hat{\text{V}}\text{ar}[Y \mid x]} = 3.8 \pm t_{0.05}(7) \cdot \sqrt{\hat{\sigma}^2 \cdot (1+0.6)} = 3.8 \pm 1.895 \cdot \sqrt{0.714 \cdot 1.6}$$
$$= 3.8 \pm 2.03 = (1.77, 5.83)$$

There is a 90% chance that a person who orders 4 items for carry-out will have to wait between 1.77 and 5.83 minutes.

i. From the ANOVA table in part b, we see that $\text{SSTotal} = 13.5$ and $\text{SSReg} = 8.5$. Thus, $R^2 = \dfrac{8.5}{13.5} = 0.629$. That is, the model explains about 63% of the variation in wait time.

2. Restaurant Wait Times (with R)

a.

b.

c.

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
   Min     1Q Median     3Q    Max 
 -1.00  -0.50  -0.25   0.50   1.50 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)                 
(Intercept)   1.0000     0.6547   1.528   0.1705              # part a
x1            0.7000     0.2673   2.619   0.0345 *
x2            1.2000     0.5345   2.245   0.0596 .            # part c
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8452 on 7 degrees of freedom
Multiple R-squared:  0.6296,    Adjusted R-squared:  0.5238 
F-statistic:  5.95 on 2 and 7 DF,  p-value: 0.03092           # part b
```

d.

```
> se.beta1hat = summary(fit)$coef[2,2];   se.beta1hat
[1] 0.2673
> t = (.7-1)/se.beta1hat;                 t
[1] -1.122
> p.value = pt(t, 7);                     p.value
[1] 0.1493
```

e.
```
> confint(fit, level=.95)[1,]
 2.5 % 97.5 %
-0.548  2.548
```

f.
```
> confint(fit, level=.90)[3,]
   5 %   95 %
0.1873 2.2127
```

g.
```
> predict(fit, data.frame(x1=3, x2=1), interval=c("conf"), level=.95)
  fit   lwr   upr
1 4.3 3.205 5.395
```

h.
```
> predict(fit, data.frame(x1=4, x2=0), interval=c("pred"), level=.90)
  fit   lwr   upr
1 3.8 1.775 5.825
```

i.
```
> summary(fit)$r.squared
[1] 0.6296
```

3. Cars (without R)

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

$$\text{SSTotal} = SYY = \sum\left(y_i - \bar{y}\right)^2 = 1518.8\,; \qquad \text{SSError} = RSS = \sum\left(y_i - \hat{y}_i\right)^2 = 95.5$$

Next, calculate SSRegression: $\text{SSReg} = \text{SSTotal} - \text{SSError} = 1518.8 - 95.5 = 1423.3$.

Completing the ANOVA table,

Source	SS	df	MS	F
Regression	$\sum\left(\hat{y}_i - \bar{y}\right)^2 = 1423.3$	$p - 1 = 3$	474.4	**29.84**
Error	$\sum\left(y_i - \hat{y}_i\right)^2 = 95.5$	$n - p = 6$	15.9	
Total	$\sum\left(y_i - \bar{y}\right)^2 = 1518.8$	$n - 1 = 9$		

According to the F-distribution, the critical region is $F > F_{\alpha}(3,6) = F_{0.05}(2,7) =$ **4.76**. Since the test statistic does lie the critical region, we reject H_0 and conclude that the model does a significant job of predicting mileage.

b. Test $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$ at the 5% level of significance. State the value of the test statistic, critical value(s), and a decision.
Calculate estimated variance of the slope estimate. From the ANOVA table, we'll use MSE = 15.9 as the estimate of the variance of the residuals, and we'll pull C_{22} from $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$.

$$\hat{\text{V}}\text{ar}\left[\hat{\beta}_1\right] = \hat{\sigma} \cdot C_{22} = (15.9)(0.007) = 0.1113$$

Calculate the test statistic.

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\hat{\text{V}}\text{ar}\left[\hat{\beta}_1\right]}} = \frac{0.92 - 0}{\sqrt{0.1113}} = 2.758$$

There are $n - 1 = 9$ degrees of freedom. According to the t-distribution, the critical region is $|t| > t_{\alpha/2}(n-1) = t_{0.025}(9) = 2.262$. Since the test statistic does lie in the critical region, we reject H_0 and conclude that the engine's horsepower is a significant predictor in the model.

c. The vector representing these predictors is $\boldsymbol{x}_0 = \begin{bmatrix} 1 & 100 & 100 & 20 \end{bmatrix}$. The estimate for the average wait time is

$$\hat{y} = \begin{bmatrix} 1 & 100 & 100 & 20 \end{bmatrix} \begin{bmatrix} 338.06 \\ 0.92 \\ -2.77 \\ -3.22 \end{bmatrix} = 338.06 + 0.92(100) - 2.77(100) - 3.22(20) = 88.66.$$

To calculate the estimate for the variance of the estimate, we need

$$\begin{aligned} \boldsymbol{x}_0'(X'X)^{-1}\boldsymbol{x}_0 &= \begin{bmatrix} 1 & 100 & 100 & 20 \end{bmatrix} \begin{bmatrix} 566.23 & 1.987 & -6.048 & -3.835 \\ 1.987 & 0.007 & -0.021 & -0.014 \\ -6.048 & -0.021 & 0.065 & 0.040 \\ -3.835 & -0.014 & 0.040 & 0.032 \end{bmatrix} \begin{bmatrix} 1 \\ 100 \\ 100 \\ 20 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 100 & 100 & 20 \end{bmatrix} \begin{bmatrix} 83.429 \\ 0.307 \\ -0.848 \\ -0.595 \end{bmatrix}. \\ &= 17.43 \end{aligned}$$

Note: This value of 17.43 calculated using values of $(\mathbf{X}^T\mathbf{X})^{-1}$ rounded to the nearest thousandth is seemingly somewhat far from the more precise value of 12.55 found when using the exact values of $(\mathbf{X}^T\mathbf{X})^{-1}$.

So, the 95% prediction interval is

$$\begin{aligned} \hat{y} \pm t_{\alpha/2}(n-p) \cdot \sqrt{\hat{\mathrm{V}}\mathrm{ar}[Y \mid x]} &= 88.66 \pm t_{0.025}(6) \cdot \sqrt{\hat{\sigma}^2 \cdot (1 + 17.43)} \\ &= 88.66 \pm 2.447 \cdot \sqrt{15.9 \cdot 18.43} \\ &= 88.66 \pm 41.89 \\ &= (46.77, 130.55) \end{aligned}$$

There is a 95% chance that a vehicle with the given specs will get between 46.77 and 130.55 miles per gallon of mileage.