Stat 6021: Final Exam

Tom Lever

11/29/22

1. The data for this question comes from a study examining the link between mortality and air pollution in 60 American cities from 1960. The response variables is *mortality*, the total age-adjusted mortality rate per 100,000 people. At first, the study wished to examine the relationship between *mortality* and a city's relative sulfur dioxide pollution potential, denoted as $sulfur$.

   A simple linear regression model is fitted, with no transformations on any variable.

   $$mortality_i = \beta_0 + \beta_1 sulfur_i + \epsilon_i$$

   where errors $epsilon_i$ are independently, identically, and normally distributed with mean $\mu = 0$ and constant variance $Var(\epsilon_i) = \sigma^2$. You may assume the regression assumptions are met. The output from R is shown in the prompt for this exam.

   a) Report the estimated regression equation. What is the interpretation of the estimated slope in context?

   $$\widehat{mortality} = \hat{\beta}_0 + \hat{\beta}_1 sulfur$$

   $$\widehat{mortality} = \left(917.8870 \ \frac{deaths}{y(100,000 \ people)}\right) + \left(0.4179 \ \frac{deaths}{y(100,000 \ people)}\right) sulfur$$

   For an increase in a city's relative sulfur dioxide pollution potential of 1 unit, the estimated age adjusted mortality rate in deaths per year 100,000 people increases by 0.4179.

   b) What is the value of the estimated variance for the error terms $\widehat{Var(\epsilon)} = \hat{\sigma}^2$?

   $$\widehat{Var(\epsilon)} = \hat{\sigma}^2 = MS_{Res} = \frac{SS_{Res}}{df_{Res}} = {SE_{Res}}^2$$

   $$\widehat{Var(\epsilon)} = {SE_{Res}}^2 = \left(56.77 \ \frac{deaths}{y(100,000 \ people)}\right)^2 = 3,222.833 \ \left(\frac{deaths}{y(100,000 \ people)}\right)^2$$

   c) Based on the output, construct the corresponding ANOVA table for this model. Be sure to show all relevant calculations.

	DF	SS	MS	F0	p
Regression	1	41,413.403	41,413.403	12.85	?
Residual	58	186,924.308	3,222.833	*	*
Total	59	228,337.711	*	*	*

$$df_R = k = 1$$

$$df_{Res} = n - p = 60 - 2 = 58$$

$$df_T = df_R + df_{Res} = 1 + 58 = 59$$

$$MS_{Res} = \widehat{Var(\epsilon)} = 3,222.833 \left(\frac{deaths}{y(100,000\ people)}\right)^2$$

$$F_0 = \frac{MS_R}{MS_{Res}} = 12.85$$

$$MS_R = F_0\, MS_{Res} = (12.85)\left[3,222.833 \left(\frac{deaths}{y(100,000\ people)}\right)^2\right] = 41,413.403 \left(\frac{deaths}{y(100,000\ people)}\right)^2$$

$$MS_R = \frac{SS_R}{df_R}$$

$$SS_R = MS_R\, df_R = \left[41,413.403 \left(\frac{deaths}{y(100,000\ people)}\right)^2\right](1) = 41,413.403 \left(\frac{deaths}{y(100,000\ people)}\right)^2$$

$$MS_{Res} = \frac{SS_{Res}}{df_{Res}}$$

$$SS_{Res} = MS_{Res}\, df_{Res} = \left[3,222.833 \left(\frac{deaths}{y(100,000\ people)}\right)^2\right](58) = 186,924.308 \left(\frac{deaths}{y(100,000\ people)}\right)^2$$

$$SS_T = SS_R + SS_{Res} = 41,413.403 \left(\frac{deaths}{y(100,000\ people)}\right)^2 + 186,924.308 \left(\frac{deaths}{y(100,000\ people)}\right)^2$$

$$SS_T = 228,337.711 \left(\frac{deaths}{y(100,000\ people)}\right)^2$$

d) One member of the study believes the total age adjusted mortality rate per 100,000 people increases, on average, by more than 0.35, per unit increase in a city's relative sulfur dioxide pollution potential. Carry out the corresponding hypothesis test. Be sure to state the null and alternative hypotheses, calculate the test statistic, and state your conclusion in context.

Given a significance level $\alpha = 0.05$, we test a null hypothesis $H_0 : \beta_{1,0} \leq 0.35$ that the slope of a linear model of age adjusted mortality rate per 100,000 people vs. a city's relative sulfur dioxide pollution potential is less than or equal to 0.35. If we have sufficient evidence to reject the null hypothesis, we have sufficient evidence to support an alternative hypothesis $H_1 : \beta_{1,0} > 0.35$ that the slope of the linear model is greater than 0.35. Since the alternate hypothesis involves ">", we have sufficient evidence to reject the null hypothesis if the magnitude $|t_0|$ of a test statistic t_0 is greater than a critical value $t_c = t_{\alpha, df_{Res}}$.

$$t_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE\left(\hat{\beta}_1\right)} = \frac{0.4179 - 0.35}{0.1166} = 0.582$$

$$|t_0| = 0.582$$

$$t_c = t_{\alpha, df_{Res}} = t_{0.05,58} = 1.672$$

```
library(TomLeversRPackage)
calculate_critical_value_tc(
    significance_level = 0.05,
    number_of_confidence_intervals = 1,
    residual_degrees_of_freedom = 58,
    hypothesis_test_is_two_tailed = FALSE
)
```

```
## [1] 1.671553
```

```
qt(0.05, 58, lower.tail = FALSE)
```

```
## [1] 1.671553
```

Since $|t_0| < t_c$, we have insufficient evidence to reject the null hypothesis that the slope of a linear model of age adjusted mortality rate per 100,000 people vs. a city's relative sulfur dioxide pollution potential is less than or equal to 0.35. We have insufficient evidence to support a claim that the slope of a linear model age adjusted mortality rate per 100,000 people vs. a city's relative sulfur dioxide pollution potential is greater than 0.35.

e) Based on the above simple linear regression model, the 95 percent confidence interval for the average total age adjusted mortality rate per 100,000 people among cities with relative sulfur dioxide pollution potential of 60 is (928.3882, 957.5338). Compute the corresponding 95 percent prediction interval for a city's total age adjusted mortality rate per 100,000 people when its relative sulfur dioxide pollution potential is 60.

```
library(TomLeversRPackage)
calculate_critical_value_tc(
    significance_level = 0.05,
    number_of_confidence_intervals = 1,
    residual_degrees_of_freedom = 58,
    hypothesis_test_is_two_tailed = TRUE
)
```

```
## [1] 2.001717
```

```
qt(0.05 / 2, 58, lower.tail = FALSE)
```

```
## [1] 2.001717
```

$$t_c = t_{\alpha/2, df_{Res}} = 2.002$$

$$E\left(mortality \mid sulfur_0\right) = \hat{\beta}_0 + \hat{\beta}_1 sulfur_0 = 917.8870 + (0.4179)(60) = 942.961$$

$$L_{CIMR} = E\left(\widehat{mortality \mid sulfur_0}\right) - t_c \; SE\left[E\left(\widehat{mortality \mid sulfur_0}\right)\right] = 928.3882$$

$$U_{CIMR} = E\left(\widehat{mortality \mid sulfur_0}\right) + t_c \; SE\left[E\left(\widehat{mortality \mid sulfur_0}\right)\right] = 957.5338$$

$$(L_{CIMR}, U_{CIMR}) = (928.3882, 957.5338)$$

$$t_c \; SE\left[E\left(\widehat{mortality \mid sulfur_0}\right)\right] = U - E\left(\widehat{mortality \mid sulfur_0}\right) = 957.5338 - 942.961 = 14.5728$$

$$SE\left[E\left(\widehat{mortality \mid sulfur_0}\right)\right] = \frac{U - E\left(\widehat{mortality \mid sulfur_0}\right)}{t_c} = \frac{14.5728}{2.002} = 7.2864$$

$$SE\left[E\left(\widehat{mortality \mid sulfur_0}\right)\right] = \sqrt{MS_{Res}\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]} = 7.2864$$

$$MS_{Res}\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right] = \left\{SE\left[E\left(\widehat{mortality \mid sulfur_0}\right)\right]\right\}^2 = 53.092$$

$$\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} = \frac{\left\{SE\left[E\left(\widehat{mortality \mid sulfur_0}\right)\right]\right\}^2}{MS_{Res}} = \frac{53.092}{3,222.833} = 0.0165$$

$$1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} = \frac{\left\{SE\left[E\left(\widehat{mortality \mid sulfur_0}\right)\right]\right\}^2}{MS_{Res}} = \frac{53.092}{3,222.833} = 1 + 0.0165 = 1.0165$$

$$MS_{Res}\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]=(3222.833)(1.0165)=3276.010$$

$$SE\left[\widehat{mortality}\,(sulfur_0)\right]=\sqrt{MS_{Res}\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}=57.236$$

$$\widehat{mortality}\,(sulfur_0)=E\,(mortality\mid sulfur_0)=942.961$$

$$L_{PI}=\widehat{mortality}\,(sulfur_0)-t_c\;SE\left[\widehat{mortality}\,(sulfur_0)\right]=942.961-(2.002)(57.236)=828.375$$

$$U_{PI}=\widehat{mortality}\,(sulfur_0)+t_c\;SE\left[\widehat{mortality}\,(sulfur_0)\right]=942.961-(2.002)(57.236)=1057.547$$

The 95 percent prediction interval for a city's total age adjusted mortality rate per $100,000$ people when its relative sulfur dioxide pollution potential is 60 is as follows.

$$(L_{PI},U_{PI})=(828.375,1057.547)$$

2. This question is an extension of question 1. Suppose the group is dissatisfied with the above simple linear regression model and decides to consider adding additional predictors:

- *precipitation*: Average annual precipitation in inches
- *jantemp*: Average January temperature in degrees Fahrenheit
- *popden*: Population per square mile in urbanized areas in 1960
- *nonwhite*: Percentage of non-white population in urbanized areas in 1960
- *hydrocarbons*: Relative hydrocarbon pollution potential
- *oxides*: Relative nitric oxide pollution potential

A first order additive multiple linear regression model is fitted.

$$\begin{aligned}mortality_i=\;&\beta_0\\&+\beta_1\;sulfur_i\\&+\beta_2\;hydrocarbons_i\\&+\beta_3\;oxides_i\\&+\beta_4\;precipitation_i\\&+\beta_5\;jantemp_i\\&+\beta_6\;popden_i\\&+\beta_7\;nonwhite_i\\&+\epsilon_i\end{aligned}$$

You may assume the regression assumptions are met. The output from R is shown in the prompt for this exam.

a) What is the p value for testing $H_0:\beta_2=0$ versus $H_a:\beta_2\neq 0$ in the above multiple linear regression model? You classmate says that based on the result of the test, *mortality* is not linearly associated with *hydrocarbons*. Do you agree? If not please briefly explain why.

The p value for testing $H_0:\beta_2=0$ versus $H_a:\beta_2\neq 0$ in the above multiple linear regression model is 0.25086.

I disagree. *hydrocarbons* is insignificant in the context of the multiple linear regression relationship and all the predictors in the multiple linear regression model. *mortality* may be linearly associated with *hydrocarbons* in the context of a simple linear regression relationship between *mortality* and *hydrocarbons* and a simple linear regression model of *mortality* vs. *hydrocarbons*.

b) Conduct an appropriate hypothesis test to decide between the simple linear regression model above and the multiple linear regression model above. Be sure to state the null and alternative hypotheses, calculate the test statistic, and state your conclusion in context.

The Partial F Test is used if we are to consider dropping a subset of the predictors from a full multiple linear regression model / whether the increase in the residual sum of squares SS_R is significant with the addition of predictors.

Let $\boldsymbol{\beta}_d$ denote a vector of regression coefficients for predictors to drop.

$$\boldsymbol{\beta}_d = \begin{bmatrix} hydrocarbons \\ oxides \\ precipitation \\ jantemp \\ popden \\ nonwhite \end{bmatrix}$$

$$d = |\boldsymbol{\beta}_d| = 6$$

Let $\boldsymbol{\beta}_k$ denote a vector of regression coefficients for predictors to keep.

$$\boldsymbol{\beta}_k = \begin{bmatrix} sulfur \end{bmatrix}$$

We test a null hypothesis $H_0 : \boldsymbol{\beta}_d = \mathbf{0}$ that all regression coefficients in $\boldsymbol{\beta}_d$ are 0 and that we should favor the reduced simple linear regression model with only predictor $sulfur$ corresponding to the coefficient in $\boldsymbol{\beta}_k$. If we have sufficient evidence to reject the null hypothesis, we have sufficient evidence to support an alternative hypothesis $H_a : \boldsymbol{\beta}_d \neq \mathbf{0}$ that at least one regression coefficient in $\boldsymbol{\beta}_d$ corresponds to a predictor that is significant in the context of the full multiple linear regression relationship and all the predictors in the full multiple linear regression model, and that we should consider other subsets of predictors to drop from the full multiple linear regression model or should favor the full multiple linear regression model with all of the predictors corresponding to the coefficients in $\boldsymbol{\beta}_d$.

Our test statistic F_0 follows an F distribution with $df_R = 7$ and $df_{Res} = 52$ degrees of freedom.

$$F_0 = \frac{MS_R\left(\boldsymbol{\beta}_d|\boldsymbol{\beta}_k\right)}{MS_{Res}\left(\boldsymbol{\beta}_k, \boldsymbol{\beta}_d\right)} = \frac{MS_R\left(\boldsymbol{\beta}_d|\boldsymbol{\beta}_k\right)}{MS_{Res,full}} = \frac{SS_R\left(\boldsymbol{\beta}_d|\boldsymbol{\beta}_k\right)/d}{SS_{Res}\left(\boldsymbol{\beta}_k, \boldsymbol{\beta}_d\right)/df_{Res}} = \frac{SS_R\left(\boldsymbol{\beta}_d|\boldsymbol{\beta}_k\right)/d}{SS_{Res,full}/df_{Res}}$$

We reject H_0 if F_0 is greater than a critical value $F_c = F_{\alpha,d,df_{Res}} = 2.192$.

```
calculate_critical_value_Fc(
    significance_level = 0.05,
    regression_degrees_of_freedom = 7,
    residual_degrees_of_freedom = 52
)
```

```
## [1] 2.191626
```

```
qf(0.05, 7, 52, lower.tail = FALSE)
```

```
## [1] 2.191626
```

F_0 measures the change in the regression sum of squares SS_R and the residual sum of squares SS_{Res} with removal of predictors. F_0 measures how much improvement there is in model fit when adding predictors in $\boldsymbol{\beta}_d$ to a multiple linear regression model with $\boldsymbol{\beta}_k$.

$$SS_R\left(\boldsymbol{\beta}_k, \boldsymbol{\beta}_d\right) = SS_{R,full} = \sum_{j=1}^{7} SS_{R,j} = 41,413+21,952+14,429+37,881+50+1,372+46,632 = 163,729$$

$$SS_R\left(\boldsymbol{\beta}_k\right) = SS_{R,red} = \sum_{j=1}^{1} SS_{R,j} = 41,413$$

$$SS_R\left(\boldsymbol{\beta}_d|\boldsymbol{\beta}_k\right) = SS_R\left(\boldsymbol{\beta}_k, \boldsymbol{\beta}_d\right) - SS_R\left(\boldsymbol{\beta}_k\right) = SS_{R,full} - SS_{R,red} = 163,729 - 41,413 = 122,316$$

$$F_0 = \frac{SS_R\left(\boldsymbol{\beta}_d|\boldsymbol{\beta}_k\right)/d}{SS_{Res,full}/df_{Res}} = \frac{122,316/6}{64,581/52} = 16.415$$

```
test_statistic <- 16.415
calculate_p_value_from_F_statistic_and_regression_and_residual_degrees_of_freedom(
    F_statistic = test_statistic,
    regression_degrees_of_freedom = 7,
    residual_degrees_of_freedom = 52
)
```

```
## [1] 3.396911e-11
```

```
pf(test_statistic, 7, 52, lower.tail = FALSE)
```

```
## [1] 3.396911e-11
```

Because the test statistic F_0 is greater than our critical value F_c and the p value is less than a significance level $\alpha = 0.05$, we reject our null hypothesis. We have sufficient evidence to support an alternative hypothesis $H_a : \boldsymbol{\beta}_d \neq \mathbf{0}$ that at least one regression coefficient in $\boldsymbol{\beta}_d$ corresponds to a predictor that is significant in the context of the full multiple linear regression relationship and all the predictors in the full multiple linear regression model, and that we should consider other subsets of predictors to drop from the full multiple linear regression model or should favor the full multiple linear regression model with all of the predictors corresponding to the coefficients in $\boldsymbol{\beta}_d$.