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
```

```r
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.

```r
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
```

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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

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.

We use the Partial F Test 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.

$$\boldsymbol{\beta}_d = \begin{bmatrix} \beta_{hydrocarbons} \\ \beta_{oxides} \\ \beta_{precipitation} \\ \beta_{jantemp} \\ \beta_{popden} \\ \beta_{nonwhite} \end{bmatrix} = \begin{bmatrix} \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \end{bmatrix}$$

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

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

$$\boldsymbol{\beta}_k = \left[\beta_{sulfur}\right] == \left[\beta_1\right]$$

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

$$\hat{\boldsymbol{\beta}}_d = \begin{bmatrix} \hat{\beta}_{hydrocarbons} \\ \hat{\beta}_{oxides} \\ \hat{\beta}_{precipitation} \\ \hat{\beta}_{jantemp} \\ \hat{\beta}_{popden} \\ \hat{\beta}_{nonwhite} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \\ \hat{\beta}_5 \\ \hat{\beta}_6 \\ \hat{\beta}_7 \end{bmatrix}$$

$$d = \left|\hat{\boldsymbol{\beta}}_d\right| = 6$$

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

$$\hat{\boldsymbol{\beta}}_k = \left[\hat{\beta}_{sulfur}\right] = \left[\hat{\beta}_1\right]$$

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.

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

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 with regression coefficients in $\hat{\boldsymbol{\beta}}_d$ to a multiple linear regression model with regression coefficients in $\hat{\boldsymbol{\beta}}_k$.

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

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

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

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

```
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 $\boldsymbol{\beta}_d$.

c) How would you test $\beta_2 = \beta_3 = \beta_4 = 0$ under the assumption that β_5, β_6, and β_7 are all 0? Be sure to state the null and alternative hypotheses, calculate the test statistic, and state your conclusion in context.

Let n be the number of observations for the full multiple linear model with the above summary and analysis of variance. Let $\boldsymbol{y}$ be the vector of a response values. Let y_i be the ith response value. Let $\bar{y}$ be the mean response value. Let

$$z = \frac{\left(\sum_{i=1}^{n}[y_i]\right)^2}{n}$$

Per section 3.3.1: “Test for Significance of Regression” in *Introduction to Linear Regression Analysis* (Sixth Edition) by Douglas C. Montgomery et al.,

$$SS_R = \hat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{y} - z$$

$$SS_{Res} = \boldsymbol{y}^T \boldsymbol{y} - \hat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{y}$$

$$SS_T = \boldsymbol{y}^T \boldsymbol{y} - z$$

Because the total sum of squares only depends on $\boldsymbol{y}$, the total sum of squares is constant as long as the same response values are used.

Consider the full multiple linear regression model to be the multiple linear regression model as above with all 7 predictors.

The residual sum of squares of the full multiple linear regression model when all 7 predictors are in the model

$$SS_{Res,full} = 64,581$$

The regression sum of squares of the full multiple linear regression model when all 7 predictors are in the model

$$S_{R,full} = 163,729$$

The total sum of squares

$$SS_T = SS_{R,full} + SS_{Res,full} = 163,729 + 64,581 = 228,310$$

Let the number of predictors in the full multiple linear regression model $k = 7$. Let

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,0} & x_{2,0} & \cdots & x_{k,0} \\ 1 & x_{1,1} & x_{2,1} & \cdots & x_{k,1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & x_{2,n} & \cdots & x_{k,n} \end{bmatrix}$$

If predictor x_k is removed from the full model, β_k is removed from $\hat{\boldsymbol{\beta}}$ and the right-most column of $\boldsymbol{X}$ is removed. Each element i of the n elements in $\hat{\boldsymbol{\beta}}^T \boldsymbol{X}^T$ decreases by $\beta_k \; x_{k,i}$. $\hat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{y}$ decreases by the regression sum of squares for predictor x_k given that all other predictors have been added to the multiple linear model

$$\delta_k = SS_R\left(\hat{\beta}_k \middle| \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_{k-1}\right) = \sum_{i=1}^{n} [\beta_k \; x_{k,i} \; y_i] = \beta_k \; \boldsymbol{x_k}^T \; \boldsymbol{y}$$

$SS_{R,full}$ decreases by δ_k to $SS_{R,k-1}$ and $SS_{Res,full}$ increases by δ_k to $SS_{Res,k-1}$. If instead we remove d predictors x_k, x_{k-1}, and x_{k-d+1} from the full multiple linear regression model, $SS_{R,full}$ decreases by δ to $SS_{R,k-d}$ and $SS_{Res,full}$ increases by δ to $SS_{Res,k-d}$, where

$$\delta = SS_R\left(\hat{\beta}_{k-d+1}, ..., \hat{\beta}_{k-1}, \hat{\beta}_k \middle| \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_{k-d}\right) = \sum_{i=k-d+1}^{k} \delta_i$$

We consider dropping *jantemp*, *popden*, and *nonwhite* from a full multiple linear regression model, resulting in a 4-predictor multiple linear regression model.

$$d = 3$$

$$\delta = \sum_{i=7-3+1}^{7} \delta_i = \sum_{i=5}^{7} \delta_i = 50 + 1,372 + 46,632 = 48,054$$

$SS_{R,full}$ decreases by δ to $SS_{R,4}$ and $SS_{Res,full}$ increases by δ_k to $SS_{Res,4}$.

$$SS_{R,4} = SS_{R,full} - \delta = 163,729 - 48,054 = 115,675$$

$$SS_{Res,4} = SS_{Res,full} + \delta = 64,581 + 48,054 = 112,635$$

We use the Partial F Test to consider dropping a subset of the predictors, $\{$*hydrocarbons*, *oxides*, *precipitation*$\}$, from the 4-predictor multiple linear regression model / whether the increase in the residual sum of squares $SS_{R,1}$ for the reduced simple linear regression model is significant with the addition of the predictors $\{$*hydrocarbons*, *oxides*, *precipitation*$\}$.

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

$$\boldsymbol{\beta}_d = \begin{bmatrix} \beta_{hydrocarbons} \\ \beta_{oxides} \\ \beta_{precipitation} \end{bmatrix} = \begin{bmatrix} \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}$$

$$d = |\boldsymbol{\beta}_d| = 3$$

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

$$\boldsymbol{\beta}_k = \begin{bmatrix} \beta_{sulfur} \end{bmatrix} = \begin{bmatrix} \beta_1 \end{bmatrix}$$

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

$$\hat{\boldsymbol{\beta}}_d = \begin{bmatrix} \hat{\beta}_{hydrocarbons} \\ \hat{\beta}_{oxides} \\ \hat{\beta}_{precipitation} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \end{bmatrix}$$

$$d = \left| \hat{\boldsymbol{\beta}}_d \right| = 3$$

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

$$\hat{\boldsymbol{\beta}}_k = \begin{bmatrix} \hat{\beta}_{sulfur} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_1 \end{bmatrix}$$

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 4-predictor multiple linear regression relationship and all the predictors in the 4-predictor multiple linear regression model, and that we should consider other subsets of predictors to drop from the 4-predictor multiple linear regression model or should favor the 4-predictor 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 $df_{R,4} = df_{R,full} - d = 7 - 3 = 4$ and $df_{Res,4} = df_{Res,full} + d = 52 + 3 = 55$ degrees of freedom.

$$F_0 = \frac{MS_R\left(\hat{\boldsymbol{\beta}}_d | \hat{\boldsymbol{\beta}}_k\right)}{MS_{Res}\left(\hat{\boldsymbol{\beta}}_k, \hat{\boldsymbol{\beta}}_d\right)} = \frac{MS_R\left(\hat{\boldsymbol{\beta}}_d | \hat{\boldsymbol{\beta}}_k\right)}{MS_{Res,4}} = \frac{SS_R\left(\hat{\boldsymbol{\beta}}_d | \hat{\boldsymbol{\beta}}_k\right)/d}{SS_{Res}\left(\hat{\boldsymbol{\beta}}_k, \hat{\boldsymbol{\beta}}_d\right)/df_{Res,4}} = \frac{SS_R\left(\hat{\boldsymbol{\beta}}_d | \hat{\boldsymbol{\beta}}_k\right)/d}{SS_{Res,4}/df_{Res,4}}$$

We reject H_0 if F_0 is greater than a critical value $F_c = F_{\alpha,d,df_{Res,4}} = 2.540$.

```
calculate_critical_value_Fc(
    significance_level = 0.05,
    regression_degrees_of_freedom = 4,
    residual_degrees_of_freedom = 55
)
```

```
## [1] 2.539689
```

```
qf(0.05, 4, 55, lower.tail = FALSE)
```

```
## [1] 2.539689
```

F_0 measures the change in the regression sum of squares $SS_{R,4}$ and the residual sum of squares $SS_{Res,4}$ with removal of predictors. F_0 measures how much improvement there is in model fit when adding predictors with regression coefficients in $\hat{\boldsymbol{\beta}}_d$ to a multiple linear regression model with regression coefficients in $\hat{\boldsymbol{\beta}}_k$.

$$SS_R\left(\hat{\boldsymbol{\beta}}_k, \hat{\boldsymbol{\beta}}_d\right) = SS_{R,4} = \sum_{j=1}^{4} SS_{R,j} = 41,413 + 21,952 + 14,429 + 37,881 = 115,675$$

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

$$SS_R\left(\hat{\boldsymbol{\beta}}_d | \hat{\boldsymbol{\beta}}_k\right) = SS_R\left(\hat{\boldsymbol{\beta}}_k, \hat{\boldsymbol{\beta}}_d\right) - SS_R\left(\hat{\boldsymbol{\beta}}_k\right) = SS_{R,4} - SS_{R,red} = 115,675 - 41,413 = 74,262$$

$$F_0 = \frac{SS_R\left(\hat{\boldsymbol{\beta}}_d | \hat{\boldsymbol{\beta}}_k\right)/d}{SS_{Res,4}/df_{Res}} = \frac{74,262/3}{112,635/55} = 12.087$$

```
test_statistic <- 12.0875
calculate_p_value_from_F_statistic_and_regression_and_residual_degrees_of_freedom(
    F_statistic = test_statistic,
    regression_degrees_of_freedom = 4,
    residual_degrees_of_freedom = 55
)
```

```
## [1] 4.058484e-07
```

```
pf(test_statistic, 4, 55, lower.tail = FALSE)
```

```
## [1] 4.058484e-07
```

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 4-predictor multiple linear regression relationship and all the predictors in the 4-predictor multiple linear regression model, and that we should consider other subsets of predictors to drop from the 4-predictor multiple linear regression model or should favor the 4-predictor multiple linear regression model with all of the predictors corresponding to the coefficients in $\boldsymbol{\beta}_d$.

3. The data for this question comes from a sample of 200 high school students. The response variable is y / *write*, the student's score on a standardized writing test, and the predictors are:

   - *race*, which is categorical with the 4 levels *Hispanic*, *Asian*, *African American*, and *Caucasian*
   - *female*, which is an indicator that is coded 1 if the student is female and 0 if the student is male
   - x_1 / read, the student's score on a standardized reading test

   The regression equation we are fitting is

$$E(y) = \beta_0 + \beta_1 I_1 + \beta_2 I_2 + \beta_3 I_3 + \beta_4 F + \beta_5 x_1$$

   where $I_1 = 1$ if the student's race is Asian, $I_2 = 1$ if the student's race is African American, $I_3 = 1$ if the student's race is Caucasian, and $F = 1$ if the student is female. The output from R is shown in the prompt for this exam. Assume that all the assumptions for fitting a multiple linear regression model are met.

   a) For students who are Hispanic and female, what is the estimated regression equation for *write* and *read*?

$$\widehat{E(y)} = \hat{\beta}_0 + \hat{\beta}_1 I_1 + \hat{\beta}_2 I_2 + \hat{\beta}_3 I_3 + \hat{\beta}_4 F + \hat{\beta}_5 x_1$$

$$\widehat{E(y|I_1 = I_2 = I_3 = 0, F = 1)} = \hat{\beta}_0 + \hat{\beta}_1(0) + \hat{\beta}_2(0) + \hat{\beta}_3(0) + \hat{\beta}_4(1) + \hat{\beta}_5 x_1$$

   The estimated regression equation for *write* and *read* is as follows.

$$\widehat{E(y|I_1 = I_2 = I_3 = 0, F = 1)} = \left(\hat{\beta}_0 + \hat{\beta}_4\right) + \hat{\beta}_5 x_1$$

   b) What is the value of $\hat{\beta}_4$? Interpret this value in context.

   $\hat{\beta}_4$ represents the estimated regression coefficient for the variable F in our multiple linear regression model indicating that a student is *female* is $F = 1$ and a student is *male* if $F = 0$.

   c) Suppose that there are no significant interaction terms in this regression. Briefly explain what this means in terms of the relationship between *write* and *read*.

   By assuming that there are no significant interaction terms in this regression, we assume that a student's score on a standardized writing test y / *write* has a linear relationship with a predictor I_1 that indicates that a student is *Asian* if $I_1 = 1$; a predictor I_2 that indicates that a student is *African American* if $I_2 = 1$; a predictor I_3 that indicates that a student is *Caucasian* if $I_3 = 1$; predictors I_1, I_2, and I_3 that indicate that a student is *Hispanic* if $I_1 = I_2 = I_3 = 0$; and a student's score on a standardized reading test x_1. We assume that relationship between y / *write* and any predictor does not depend on the value of other predictor and that each predictor is independent of any other predictor.

   d) Compute a 95 percent confidence interval for the difference in mean scores on the standardized writing test between students who are *Asian* and *African American*, for a given level of gender and a given value of *read*. Based on this confidence interval, write an appropriate conclusion.

   For students who are *Asian*,

$$\widehat{E(y|I_1 = 1)} = \hat{\beta}_0 + \hat{\beta}_1 I_1 + \hat{\beta}_2 I_2 + \hat{\beta}_3 I_3 + \hat{\beta}_4 F + \hat{\beta}_5 x_1$$

$$\widehat{E(y|I_1 = 1)} = \hat{\beta}_0 + \hat{\beta}_1(1) + \hat{\beta}_2(0) + \hat{\beta}_3(0) + \hat{\beta}_4 F + \hat{\beta}_5 x_1$$

$$\widehat{E(y|I_1 = 1)} = \left(\hat{\beta}_0 + \hat{\beta}_1\right) + \hat{\beta}_4 F + \hat{\beta}_5 x_1$$

   For students who are *African American*,

$$\widehat{E(y|I_2 = 1)} = \hat{\beta}_0 + \hat{\beta}_1(0) + \hat{\beta}_2(1) + \hat{\beta}_3(0) + \hat{\beta}_4 F + \hat{\beta}_5 x_1$$

$$E(\widehat{y|I_2 = 1}) = \left(\hat{\beta}_0 + \hat{\beta}_2\right) + \hat{\beta}_4 F + \hat{\beta}_5 x_1$$

$$d_{1,2} = E(\widehat{y|I_1 = 1}) - E(\widehat{y|I_2 = 1}) = \left[\left(\hat{\beta}_0 + \hat{\beta}_1\right) + \hat{\beta}_4 F + \hat{\beta}_5 x_1\right] - \left[\left(\hat{\beta}_0 + \hat{\beta}_2\right) + \hat{\beta}_4 F + \hat{\beta}_5 x_1\right] = \hat{\beta}_1 - \hat{\beta}_2$$

$$d_{1,2} = \hat{\beta}_1 - \hat{\beta}_2 = 7.34591 - 0.66265 = 6.68326$$

$$df_{Res} = n - p = n - (k+1) = 200 - 6 = 194$$

$$t_c = t_{alpha/2, df_{Res}} = t_{0.05/2,194} = 1.972$$

```r
calculate_critical_value_tc(
    significance_level = 0.05,
    number_of_confidence_intervals = 1,
    residual_degrees_of_freedom = 194,
    hypothesis_test_is_two_tailed = TRUE
)
```

```
## [1] 1.972268
```

```r
qt(0.05 / 2, 194, lower.tail = FALSE)
```

```
## [1] 1.972268
```

$$\widehat{Var\left(\beta_1\right)} = 6.643$$

$$\widehat{Var\left(\beta_2\right)} = 4.526$$

$$\widehat{Cov\left(\beta_1, \beta_2\right)} = 2.097$$

$$\widehat{Var\left(\beta_1 - \beta_2\right)} = \widehat{Var\left(\beta_1\right)} + \widehat{Var\left(\beta_2\right)} - 2\widehat{Cov\left(\beta_1, \beta_2\right)} = 6.643 + 4.526 - 2(2.097) = 6.975$$

$$\widehat{SE\left(d_{1,2}\right)} = \widehat{SE\left(\beta_1 - \beta_2\right)} = \sqrt{\widehat{Var\left(\beta_1 - \beta_2\right)}} = \sqrt{6.975} = 2.641$$

$\widehat{ME\left(d_{1,2}\right)} = 5.208\$$

$$L = d_{1,2} - \widehat{ME\left(d_{1,2}\right)} = 6.68326 - 5.208 = 1.475$$

$$U = d_{1,2} + \widehat{ME\left(d_{1,2}\right)} = 6.68326 + 5.208 = 11.891$$

A 95 percent confidence interval for the difference in mean scores on the standardized writing test between students who are $Asian$ and $African\ American$, for a given level of $female$ and a given value of $read$, is as follows.

$$(L, U) = (1.475, 11.891)$$

If we obtain many random samples of the same sample size / number of students $n = 200$ and construct similar confidence intervals for the difference between mean scores with confidence level $C = 95$ for each sample, $C = 95$ percent of samples will have a confidence interval that contains the population difference in mean scores.

Suppose we obtain many random samples of the same sample size / number of students $n = 200$ and construct a confidence interval with confidence level $C = 95$ based on each sample. The difference between the sample difference between mean scores and the population difference in mean scores in 95 percent of samples will be no greater than the margin of error $\widehat{ME\left(d_{1,2}\right)} = 5.208$.

Because this confidence interval does not contain 0, we have sufficient evidence to reject a null hypothesis that the difference in mean scores is 0. We have sufficient evidence to support an alternate hypothesis that the difference in mean scores is greater than 0 and that there is a significant difference between the mean scores on the standardized writing test between students who are $Asian$ and $African\ American$, for the same gender and score on a standardized reading test.

4. For the following statements, state whether they are true or false. If false, briefly explain why. If the statement is ambiguous, clearly explain why the statement is ambiguous and how the ambiguity is affecting your answer.

   a) When one is comparing linear regression models with different number of parameters, multiple R^2 is a better criterion for model selection than adjusted R^2.

      This statement is ambiguous.

      We follow 10.1.3: Criteria for Evaluating Subset Regression Models$ and 10.2.1: All Possible Regressions in *Introduction to Linear Regression Analysis* (Sixth Edition) by Douglas C. Montgomery et al. In determining a preferred multiple linear regression model for the Hald Cement Data, the authors consider both the multiple R^2 and the adjusted R^2. The authors use the multiple R^2 to suggest two-predictor multiple linear regression models and give context to comparing models with high adjusted R^2 and low residual mean squares. The authors seem to prefer a two-predictor model suggested by multiple R^2 with low but not minimal residual mean square, which also minimizes a criterion called Mallow's C_p. The authors consider the multiple R^2, the adjusted R^2, Mallow's C_p, and other model-selection criteria in determining a preferred multiple linear regression model.

      Let ${R^2}_p$ denote the coefficient of multiple determination for a multiple linear regression model with p variables and p terms including k predictors and intercept term $\hat{\beta}_0$. ${R^2}_p$ is a measure of the adequacy of a multiple linear regression model.

$$
{R^2}_p = \frac{SS_{R,p}}{SS_T} = 1 - \frac{SS_{Res,p}}{SS_T}
$$

      where $SS_{R,p}$ and $SS_{Res,p}$ denote the regression sum of squares and the residual sum of squares for a p-term multiple linear regression model. ${R^2}_p$ increases as p increases until we consider a full multiple linear regression model. We use this criterion in model selection by adding predictors to a model up to the point where an additional variable is not useful in that it only provides a small increase in ${R^2}_p$. Let n, P, and K be the number of observations, terms, and predictors in a full multiple linear regression model, respectively.

$$
{R^2}_0 = 1 - \left(1 - {R^2}_P\right)\left(1 + d_{\alpha,n,K}\right)
$$

$$
d_{\alpha,n,K} = \frac{KF_{\alpha,df_{R,full},df_{Res,full}}}{df_{Res,full}}
$$

      Any subset of predictors producing an ${R^2}_p$ greater than ${R^2}_0$ is an R^2 adequate subset with ${R^2}_p$ not significantly different from the coefficient of determination ${R^2}_P$ for a full multiple linear regression model. In model selection, models having large ${R^2}_p$ near ${R^2}_0$ may be preferred. Models having ${R^2}_p$ near a "knee" in a graph of ${R^2}_p$ vs. p may be preferred. Using ${R^2}_{adj,p}$ as a model selection criterion may be preferred as we cannot find an optimum value of ${R^2}_p$ for a multiple linear regression model.

      Let ${R^2}_{adj,p}$ denote the adjusted coefficient of multiple determination for a multiple linear regression model with p variables and p terms including k predictors and intercept term $\hat{\beta}_0$.

$$
{R^2}_{adj,p} = 1 - \frac{df_T}{df_{Res}}\left(1 - {R^2}_p\right)
$$

      In model selection, the model maximizing ${R^2}_{adj,p}$ may be preferred. ${R^2}_{adj,p}$ does not necessarily increase as additional predictors are introduced into a multiple linear regression model. ${R^2}_{adj,p}$ initially increases, then stabilizes, and eventually may decrease. The eventual decrease in ${R^2}_{adj,p}$ occurs when the reduction in $SS_{Res,p}$ from adding a predictor to a multiple linear regression model is not sufficient to compensate for the loss of one residual degree of freedom.

   b) When using automated search procedures such as backward elimination, the model selected by the algorithm will be the most optimal model that we can get.

This statement is false in general.

We consider "automated search procedures such as backward elimination" to be equivalent to "stepwise-type procedures... (1) forward selection, (2) backward elimination, and (3) stepwise regression... for evaluating only a small number of subset regression models by either adding or deleting regressors one at a time".

Backward elimination is often a very good variable selection procedure. It is particularly favored by analysts who like to see the effect of including all the candidate regressors, just so that nothing "obvious" will be missed.

None of the procedures generally guarantees that the best subset regresswion model of any size will be identified. Since the procedures terminate with one final equation, inexperienced analysts may conclude that they have found a model that is in some sense optimal. It is likely, not that there is one best subset model, but that there are several equally good ones. The forward-selection procedure has a general problem that once a regressor has ben added, it cannot be removed at a later step. The procedures do not necessarily lead to the same choice of final model. Forward selection tends to agree with all possible regressions for small subset sizes but not for large ones, while backward elimination tends to agree with all possible regressions for large subset sizes but not for small ones. The procedures should be used with caution. The authors prefer stepwise regression followed by backward elimination. The backward-elimination procedure is often less adversely affected by the correlative structure of the regressors than is forward selection. The choice of values for cutoffs is largely a matter of the personal preference of the analyst, and considerable latitude is often taken in regard to cutoffs.

R implementations of the procedures select first-order models without interactions or higher-order terms without introduction of iteraction or higher-order columns into a data set. A model with interaction terms or higher-order terms may be optimal. Regression assumptions for models selected by the procedures may not be met. A model with transformed respond and/or predictors may be optimal. A model selected by the procedures may not be the best model of the relationship between predictors and/or response, or for prediction. Another model of all possible models may be optimal. In forward selection, predictors are added only to a multiple linear regression model; predictors may become insignificant in the context of the multiple linear regression model and all predictors. Performing a procedure using an F statistic instead of an Akaike or Schwartz Bayesian Information Criterion (AIC or BIC) may yield models different from a preferred model. Forward selection, backward selection, and bidirectional selection may yield different models from each other and a preferred model considering all possible regressions.

c) A co-worker fits a simple linear regression model and produces the corresponding residual plot, shown in the prompt for this exam. Based on this figure, your coworker suggests transforming the predictor variable.

This statement is ambiguous first because it doesn't involve a claim that can be assessed as true or false.

This statement is ambiguous first because it doesn't suggest a goal. If our goal is to present an example of a simple linear regression model where a simple linear regression assumption that the variance of errors is constant is not met, and the variance of errors increases with fitted values, we can present this graph and a statement that a simple linear regression assumption that the variance of errors is constant is not met. If our goal is fit a simple linear regression model where all simple linear regression assumptions are met, we can present this graph and a statement that because the variance of errors increases with fitted values, we may consider applying a transformation to the response, call it y, like $y' = ln(y)$. If we fit a simple linear regression model of y' vs. predictor x, produce a residual plot, and the residual plot seems to exhibit a curve indicating a nonlinear relationship between y' and x, or a mean residual greater than 0, we may consider applying a transformation to the predictor, like $x' = ln(x)$. If we fit a simple linear regression model of y' vs. x', produce a residual plot, and the residual plot seems to exhibit a horizontal band vertically centered at $e = 0$, we may wish to produce a plot of AutoCorrelation Function values versus lag to assess errors are uncorrelated and independent, and a QQ plot to assess whether errors are normally distributed.

This statement is ambiguous because there are many ways to take suggestions and many ways to transform response and/or predictors of a simple linear regression model. As above, we may wish to apply a transformation to the predictor x like $x' = ln(x)$, but only after attempting to align the simple linear regression model with the simple linear regression assumption that the variance of the errors is constant.

5. At the end of your exam, write out the honor pledge. “On my honor, I pledge that I have neither given nor received help on this assignment”. Also, write a statement acknowledging that you have read, understood, and followed the instructions on page 1 of the prompt for this exam, as well as the guidelines listed in the Exam Guidelines document on Collab. Failure to follow these guidelines may result in points being lost.

   On my honor, I pledge that I have neither given nor received help on this assignment.

   I acknowledge that I have read, understood, and followed the instructions on page 1 on the prompt for this exam, as well as the guidelines listed in the Exam Guidelines document on Collab.