Stat 6021: Homework Set 6

Tom Lever

10/13/22

1. For this first question, you will continue to use the data set `swiss`, which you also used in the last homework. Load the data. For more information about the data set, type `?swiss`. This data set encapsulates a standardized Fertility measure and socioeconomic indicators for each of 47 French-speaking provinces of Switzerland at about 1888.

```
head(swiss, n = 3)
```

```
##              Fertility Agriculture Examination Education Catholic
## Courtelary        80.2        17.0          15        12     9.96
## Delemont          83.1        45.1           6         9    84.84
## Franches-Mnt      92.5        39.7           5         5    93.40
##              Infant.Mortality
## Courtelary               22.2
## Delemont                 22.2
## Franches-Mnt             20.2
```

   (a) In the previous homework, you fit a model with the fertility measure as the response variable and used all the other variables as predictors. Now, consider a simpler model, using only the last three variables as predictors: *Education*, *Catholic*, and $Infant.Mortality$. Carry out an appropriate hypothesis test to assess which of these two models should be used. State the null and alternate hypotheses, find the relevant test statistic and p-value, and state a conclusion in context. For practice, try to calculate the test statistic by hand.

   We conduct a partial F test to investigate if the predictors *Agriculture* and *Examination* omitted from the reduced model are jointly insignificant in the context of the full multiple linear model and all predictors.

```
library(TomLeversRPackage)
full_model <- lm(
    Fertility ~ Agriculture + Examination + Education + Catholic + Infant.Mortality,
    data = swiss
)
reduced_model <- lm(
    Fertility ~ Education + Catholic + Infant.Mortality, data = swiss
)
analyze_variance_for_reduced_and_full_linear_models(reduced_model, full_model)
```

```
## Analysis of Variance Table
##
## Model 1: Fertility ~ Education + Catholic + Infant.Mortality
## Model 2: Fertility ~ Agriculture + Examination + Education + Catholic +
##     Infant.Mortality
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     43 2422.2
## 2     41 2105.0  2     317.2 3.0891 0.05628 .
```

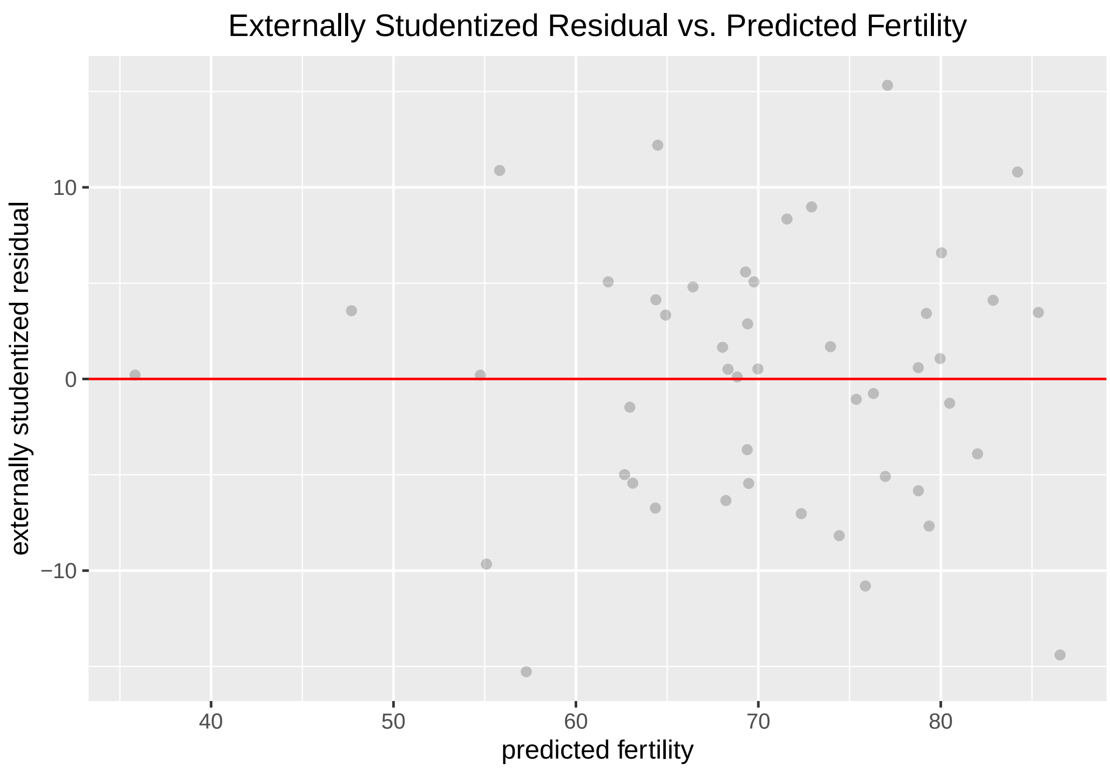
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## test statistic F0 for Partial F Test: 3.08908079753289
## Fc(alpha = 0.05, predictors_dropped = 2, DFRes(full) = 41) = 3.22568384229545
## P(F > F0) for Partial F Test: 0.0562831355250011
## significance level: 0.05
```

The test statistic for the Partial F Test $F_0 = 3.089$. Since the test statistic is less than a critical value $F_c = 3.226$, we have insufficient evidence to reject a null hypothesis that the regression coefficients for the predictors *Agriculture* and *Examination* omitted from the reduced model are 0. We have insufficient evidence to support an alternate hypothesis that a regression coefficient for an omitted predictor is not 0. The predictors *Agriculture* and *Examination* are jointly insignificant in the context of the full multiple linear model and all predictors. The reduced model should be used.

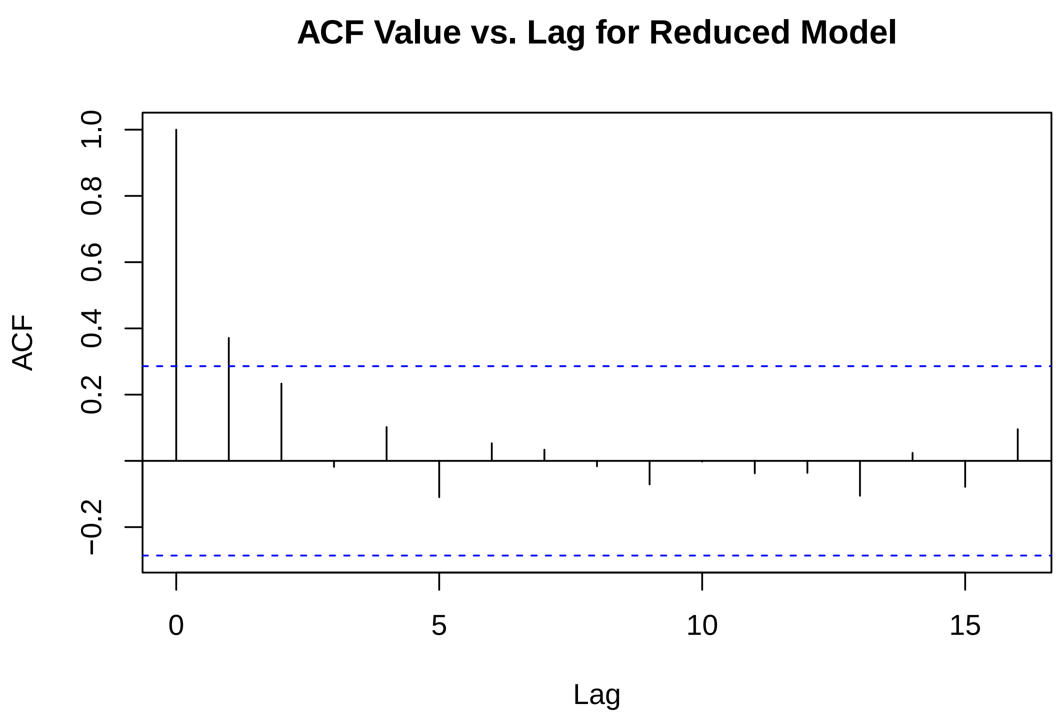
The p-value for the Partial F Test $p = 0.056$. Since this p-value is greater than a significance level $\alpha = 0.05$, we have insufficient evidence to reject a null hypothesis that the regression coefficients for the predictors *Agriculture* and *Examination* omitted from the reduced model are 0. We have insufficient evidence to support an alternate hyptohesis that a regression coefficient for an omitted predictor is not 0. The predictors *Agriculture* and *Examination* are jointly insignificant in the context of the full multiple linear model and all predictors. The reduced model should be used.

(b) For the model you decide to use from part 1a, assess if the regression assumptions are met.

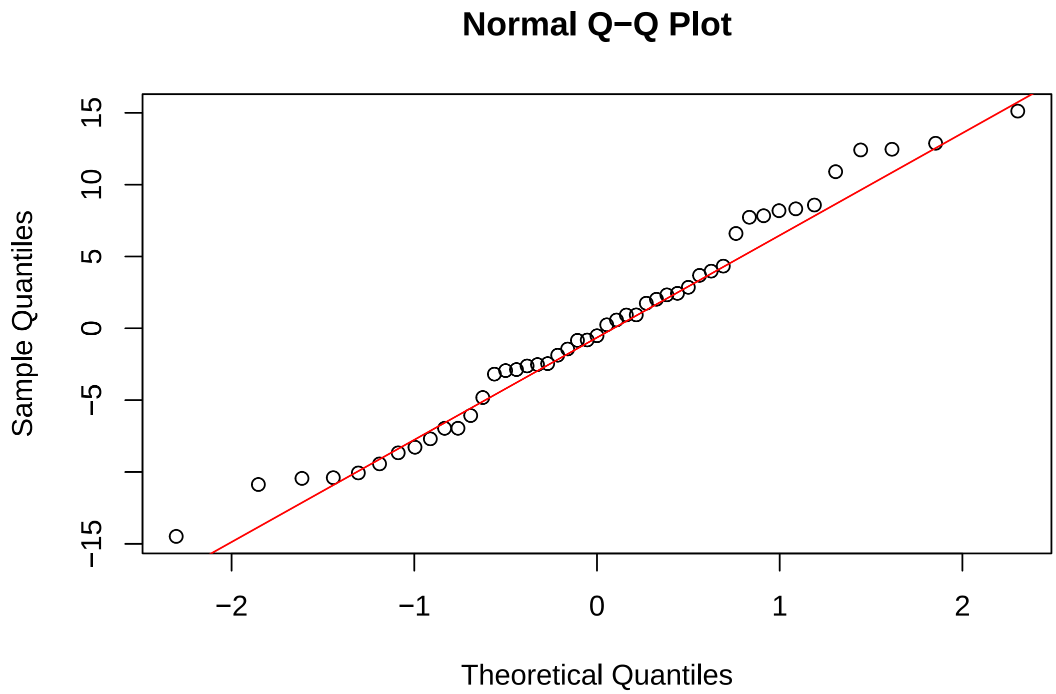
```r
library(ggplot2)
ggplot(
    data.frame(
        externally_studentized_residual = full_model$residuals,
        predicted_fertility = reduced_model$fitted.values
    ),
    aes(x = predicted_fertility, y = externally_studentized_residual)
) +
    geom_point(alpha = 0.2) +
    geom_hline(yintercept = 0, color = "red") +
    labs(
        x = "predicted fertility",
        y = "externally studentized residual",
        title = "Externally Studentized Residual vs. Predicted Fertility"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 0)
    )
```




```
acf(reduced_model$residuals, main = "ACF Value vs. Lag for Reduced Model")
```




```
qqnorm(reduced_model$residuals)
qqline(reduced_model$residuals, col = "red")
```



1. The assumption that the relationship between response / fertility and predictors is linear, at least approximately, is met cannot be addressed.
2. The assumption that the residuals of the linear model of fertility versus predictors have mean 0 is met. Residuals are evenly scattered around $e = 0$ at random.
3. The assumption that the distributions of residuals of the linear model for different predictors have constant variance is met. Residuals are evenly scattered around $e = 0$ with constant vertical variance.
4. The assumption that the residuals of the linear model are uncorrelated is not met. The ACF value for lag 0 is always 1; the correlation of the vector of residuals with itself is always 1. Since the ACF value for lag 1 is significant, we have sufficient evidence to reject a null hypothesis that the residuals of the linear model are uncorrelated. We have sufficient evidence to conclude that the residuals of the linear model are correlated. We have sufficient evidence to conclude that the assumption that the residuals of the linear model are uncorrelated is not met.
5. The assumption that the residuals of the linear model are normally distributed is met. A linear model is robust to these assumptions. Considering a plot of sample quantiles versus theoretical quantiles for the residuals of the linear model, since observations lie near the line of best fit / their theoretical values, a probability vs. externally studentized residuals plot / distribution is normal.

2. You may only use R as a simple calculator or to find p-values or critical values. The data for this question come from 113 hospitals. The key response variable is $InfRsk$, the risk that patients get an infection while staying at the hospital. We will look at five predictors:

- x_1: $Stay$: Average length of stay at hospital.
- x_2: $Cultures$: Average number of bacterial cultures per day at the hospital.
- x_3: Age: Average age of patients at hospital.
- x_4: $Census$: The average daily number of patients.
- x_5: $Beds$: The number of beds in the hospital.

Some R output in shown in the prompt for this homework. You may assume the regression assumptions are met. Only use the provided R output to answer the rest of part 2.

(a) Based on the t statistics, which predictors appear to be insignificant?

Since the R output describes an F statistic for a Partial F Test following an F distribution with $DF_R = 5$ and $DF_{Res} = 107$ degrees of freedom, a critical value t_c for each test statistic t_0 follows a Student's t distribution with $DF_{Res} = 107$ degrees of freedom. This critical value is

```
significance_level <- 0.05
number_of_confidence_intervals <- 1
residual_degrees_of_freedom <- 107
critical_value_tc <- qt(
    significance_level / (2*number_of_confidence_intervals),
    residual_degrees_of_freedom,
    lower.tail = FALSE
)
critical_value_tc
```

```
## [1] 1.982383
```

```
critical_value_tc <- calculate_critical_value_tc(
    significance_level,
    number_of_confidence_intervals,
    residual_degrees_of_freedom
)
critical_value_tc
```

```
## [1] 1.982383
```

Since test statistics $t_{0,\ Stay}$ and $t_{0,\ Cultures}$ are greater than critical value t_c, predictors $Stay$ and $Cultures$ are significant. Since corresponding p-values are less than a significance level $\alpha = 0.05$, predictors $Stay$ and $Cultures$ are significant.

Since test statistics $t_{0,\ Age}$, $t_{0,\ Census}$, and $t_{0, Beds}$ are less than critical value t_c, predictors Age, $Census$, and $Beds$ are insignificant. Since corresponding p-values are greater than significance level α, predictors Age, $Census$, and $Beds$ are insignificant.

(b) Based on your answer in part 2a, carry out the appropriate hypothesis test to see if those predictors can be dropped from the multiple linear regression model. Show all steps, including your null and alternate hypotheses; the corresponding test statistic, p-value, and critical value; and your conclusion in context.

We consider predictors $Stay$ and $Cultures$ to be kept predictors. We consider predictors Age, $Census$, and $Beds$ to be dropped predictors. The regression sum of squares for the dropped predictors given that the kept predictors are already in the model, and the sum of regression sum of squares for dropped predictors given that kept predictors are already in the model

$$SS_R\left(\boldsymbol{x}_d | \boldsymbol{x}_k\right) = SS_{R,\ Age} + SS_{R,\ Census} + SS_{R,\ Beds} = 0.136 + 5.101 + 0.028 = 5.265$$

The number of dropped predictors $d = 3$. The regression mean square for the dropped predictors given that the kept predictors are already in the model

$$MS_R\left(\boldsymbol{x}_d | \boldsymbol{x}_k\right) = \frac{SS_R\left(\boldsymbol{x}_d | \boldsymbol{x}_k\right)}{d} = \frac{5.265}{3} = 1.755$$

The residual mean square $MS_{Res} = 0.985$. The test statistic for the Partial F Test

$$F_0 = \frac{MS_R\left(\boldsymbol{x}_d | \boldsymbol{x}_k\right)}{MS_{Res}} = \frac{1.755}{0.985} = 1.782$$

Since the R output describes an F statistic for a Partial F Test following an F distribution with $DF_R = 5$ and $DF_{Res} = 107$ degrees of freedom, a critical value for the Partial F Test $F_c = 2.299$.

```
regression_degrees_of_freedom <- 5
critical_value_Fc <- qf(
    significance_level,
    regression_degrees_of_freedom,
    residual_degrees_of_freedom,
    lower.tail = FALSE
)
critical_value_Fc
```

```
## [1] 2.299234
```

```
critical_value_Fc <- calculate_critical_value_Fc(
    significance_level,
    regression_degrees_of_freedom,
    residual_degrees_of_freedom
)
critical_value_Fc
```

```
## [1] 2.299234
```

Since this test statistic F_0 is less than the critical value $F_c = 2.299$, we have insufficient evidence to reject a null hypothesis that the regression coefficients for the dropped predictors are 0. We have insufficient evidence to support an alternate hypothesis that a regression coefficient for a dropped predictor is not 0. The dropped predictors are jointly insignificant. The predictors Age, $Census$, and $Beds$ may be dropped simultaneously from the multiple linear model with the above summary and analysis of variance.

```
test_statistic_F0 = 1.782
p_value <- pf(
    test_statistic_F0,
    regression_degrees_of_freedom,
    residual_degrees_of_freedom,
    lower.tail = FALSE
)
p_value
```

```
## [1] 0.1226644
```

```
p_value <- calculate_p_value_from_F_statistic_and_regression_and_residual_degrees_of_freedom(
    test_statistic_F0, regression_degrees_of_freedom, residual_degrees_of_freedom
)
p_value
```

```
## [1] 0.1226644
```

Since the p-value $p = 0.123$ corresponding to this F statistic $F_0 = 1.782$ is greater than a significance level $\alpha = 0.05$, we have insufficient evidence to reject a null hypothesis that the regression coefficients for the dropped predictors are 0. We have insufficient evidence to support an alternate hypothesis that a regression coefficient for a dropped predictor is not 0. The dropped predictors are jointly insignificant. The predictors Age, $Census$, and $Beds$ may be dropped simultaneously from the multiple linear model with the above summary and analysis of variance.

(c) Suppose we want to decide between two potential models:

- Model 1 using x_1, x_2, x_3, and x_4 as the predictors for $InfctRsk$
- Model 2 using x_1 and x_2 as the predictors for $InfctRsk$

Carry out the appropriate hypothesis test to decide which of models 1 and 2 should be used. Be sure to show all steps in your hypothesis test.

Let n be the number of observations for the multiple linear model with the above summary and analysis. 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.

$$SS_T = \sum_{i=1}^{n}\left[(y_i - \bar{y})^2\right] = SS_R + SS_{Res}$$

Consider the full model to be a multiple linear regression model of response $InfctRsk$ and all five predictors.

The residual sum of squares of the full model when all five predictors are in the model

$$SS_{Res,\ full} = 105.413$$

The regression sum of squares of the full model when all five predictors are in the model

$$SS_{R,\ full} = SS_{R,\ Stay} + SS_{R,\ Cultures} + SS_{R,\ Age} + SS_{R,\ Census} + SS_{R,\ Beds}$$

$$SS_{R,\ full} = 57.305 + 33.397 + 0.136 + 5.101 + 0.028$$

$$SS_{R,\ full} = 95.967$$

The total sum of squares

$$SS_T = SS_{R,\ full} + SS_{Res,\ full} = 95.967 + 105.413 = 201.38$$

Let the number of predictors $k = 5$. Let

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

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

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

$$\delta = SS_R \left(\left[\hat{\beta}_k \right] \middle| \hat{\boldsymbol{\beta}}_{w/o \ \beta_k} \right) = \sum_{i=1}^{n} \left[\beta_k \ x_{k,i} \ y_i \right] = \beta_k \ \boldsymbol{x_k}^T \ \boldsymbol{y}$$

SS_R decreases by δ and SS_{Res} increases by δ.

We remove predictor x_5 / *Beds* from our multiple linear model.

$SS_{R, \ w/o \ x_5}$ is 95.939; $SS_{Res, \ w/o \ x_5}$ is 105.441.

We consider predictors *Stay* and *Cultures* to be kept predictors. We consider predictors *Age* and *Census* to be dropped predictors. The regression sum of squares for the dropped predictors given that the kept predictors are already in the model, and the sum of regression sum of squares for dropped predictors given that kept predictors are already in the model

$$SS_{R, \ w/o \ x_5} \left(\boldsymbol{x}_{d, \ w/o \ x_5} | \boldsymbol{x}_k \right) = SS_{R, \ Age} + SS_{R, \ Census} = 0.136 + 5.101 = 5.237$$

The number of dropped predictors $d_{w/o \ x_5} = 2$. The regression mean square for the dropped predictors given that the kept predictors are already in the model

$$MS_{R, \ w/o \ x_5} \left(\boldsymbol{x}_{d, \ w/o \ x_5} | \boldsymbol{x}_k \right) = \frac{SS_{R, \ w/o \ x_5} \left(\boldsymbol{x}_{d, \ w/o \ x_5} | \boldsymbol{x}_k \right)}{d_{w/o \ x_5}} = \frac{5.237}{2} = 2.6185$$

The regression degrees of freedom $DF_{R, \ w/o \ x_5} = 4$.

The residual degrees of freedom $DF_{Res, \ w/o \ x_5} = n - p_{w/o \ x_5} = n - (p-1) = n - p + 1 = DF_{Res} + 1 = 108$.

The residual mean square

$$MS_{Res, \ w/o \ x_5} = \frac{SS_{Res, \ w/o \ x_5}}{DF_{Res}} = \frac{105.441}{108} = 0.976$$

The test statistic for the Partial F Test

$$F_{0, \ w/o \ x_5} = \frac{MS_{R, \ w/o \ x_5} \left(\boldsymbol{x}_{d, \ w/o \ x_5} | \boldsymbol{x}_k \right)}{MS_{Res, \ w/o \ x_5}} = \frac{2.6185}{0.976} = 2.683$$

```r
regression_degrees_of_freedom_without_x5 <- 4
residual_degrees_of_freedom_without_x5 <- 108
critical_value_Fc_without_x5 <- calculate_critical_value_Fc(
    significance_level,
    regression_degrees_of_freedom_without_x5,
    residual_degrees_of_freedom_without_x5
)
critical_value_Fc_without_x5
```

```
## [1] 2.455767
```

Since the test statistic $F_{0, \ w/o \ x_5} = 2.683$ is greater than the critical value $F_{c, \ w/o \ x_5} = 2.456$, we have sufficient evidence to reject a null hypothesis that the regression coefficients for the dropped predictors are 0. We have sufficient evidence to support an alternate hypothesis that a regression coefficient for a dropped predictor is not 0. The dropped predictors are jointly significant. The predictors Age and $Census$ cannot be dropped simultaneously from the multiple linear model without the predictor $Beds$.

```
test_statistic_F0_without_x5 <- 2.683
p_value_without_x5 <- calculate_p_value_from_F_statistic_and_regression_and_residual_degrees_o
    test_statistic_F0_without_x5,
    regression_degrees_of_freedom_without_x5,
    residual_degrees_of_freedom_without_x5
)
p_value_without_x5
```

```
## [1] 0.03529164
```

Since the p-value $p_{w/o\ x_5} = 0.0353$ is less than the significance level $\alpha = 0.05$, we have sufficient evidence to reject a null hypothesis that the regression coefficients for the dropped predictors are 0. We have sufficient evidence to support an alternate hypothesis that a regression coefficient for a dropped predictor is not 0. The dropped predictors are jointly significant. The predictors Age and $Census$ cannot be dropped simultaneously from the multiple linear model without the predictor $Beds$.

3. This question is based on a data set seen in Homework Set 4. Data from 55 college students are used to estimate a multiple regression model with response variable $LeftArm$ and predictors $LeftFoot$ and $RtFoot$. All variables were measured in centimeters. You may assume the regression assumptions are met. Some R output is given in the prompt for this homework. Explain how this output indicates the presence of multicollinearity in this regression model.

Per section 9.4.4: Multicollinearity: Multicollinearity Diagnostics: Other Diagnostics in *Introduction to Linear Regression Analysis* (Sixth Edition) by Douglas C. Montgomery et al., "if the overall F statistic is significant but the individual t statistics are all nonsignificant, multicollinearity is present".

Since the R output describes an F statistic for a Partial F Test following an F distribution with $DF_R = 2$ and $DF_{Res} = 52$ degrees of freedom, a critical value F_c for the overall F statistic $F_0 = 15.19$ follows a F distribution with $DF_R = 2$ and $DF_{Res} = 52$ degrees of freedom. This critical value is

```
regression_degrees_of_freedom <- 2
residual_degrees_of_freedom <- 52
critical_value_Fc <- calculate_critical_value_Fc(
    significance_level,
    regression_degrees_of_freedom,
    residual_degrees_of_freedom
)
critical_value_Fc
```

```
## [1] 3.175141
```

Since F_0 is greater than F_c, the overall F statistic is significant.

Since the R output describes an F statistic for a Partial F Test following an F distribution with $DF_R = 2$ and $DF_{Res} = 52$ degrees of freedom, a critical value t_c for each predictor t statistic t_0 follows a Student's t distribution with $DF_{Res} = 52$ degrees of freedom. This critical value is

```
critical_value_tc <- calculate_critical_value_tc(
    significance_level,
    number_of_confidence_intervals,
    residual_degrees_of_freedom
)
critical_value_tc
```

```
## [1] 2.006647
```

Since the test statistic for predictor $LeftFoot$ is less than critical value t_c, the test statistic for predictor $LeftFoot$ is non-significant. Since the corresponding p-value is greater than significance level $\alpha = 0.05$, the test statistic for predictor $LeftFoot$ is non-significant. Since the test statistic for predictor $RtFoot$ is less than critical value t_c, the test statistic for predictor $RtFoot is non-significant. Since the corresponding p-value is greater than significance level $\alpha = 0.05$, the test statistic for predictor $RtFoot$ is non-significant.

Since the overall F statistic is significant but the test statistics for individual predictors are all non-significant, multicollinearity is present.