Homework 6

Hyun Suk (Max) Ryoo (hr2ee)

10/11/2021

1) For this first question, you will use the dataset swiss which is part of the datasets package. Load the data. For more information about the data set, type ?swiss. The goal of the data set was to assess how fertility rates in the Swiss (French-speaking) provinces relate to a number of demographic variables.

Set up

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.2

## -- Attaching packages ------------------------------------------------------------------------------

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## Warning: package 'ggplot2' was built under R version 4.0.2

## Warning: package 'tidyr' was built under R version 4.0.2

## Warning: package 'readr' was built under R version 4.0.2

## Warning: package 'dplyr' was built under R version 4.0.2

## Warning: package 'stringr' was built under R version 4.0.2

## Warning: package 'forcats' was built under R version 4.0.2

## -- Conflicts ------------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

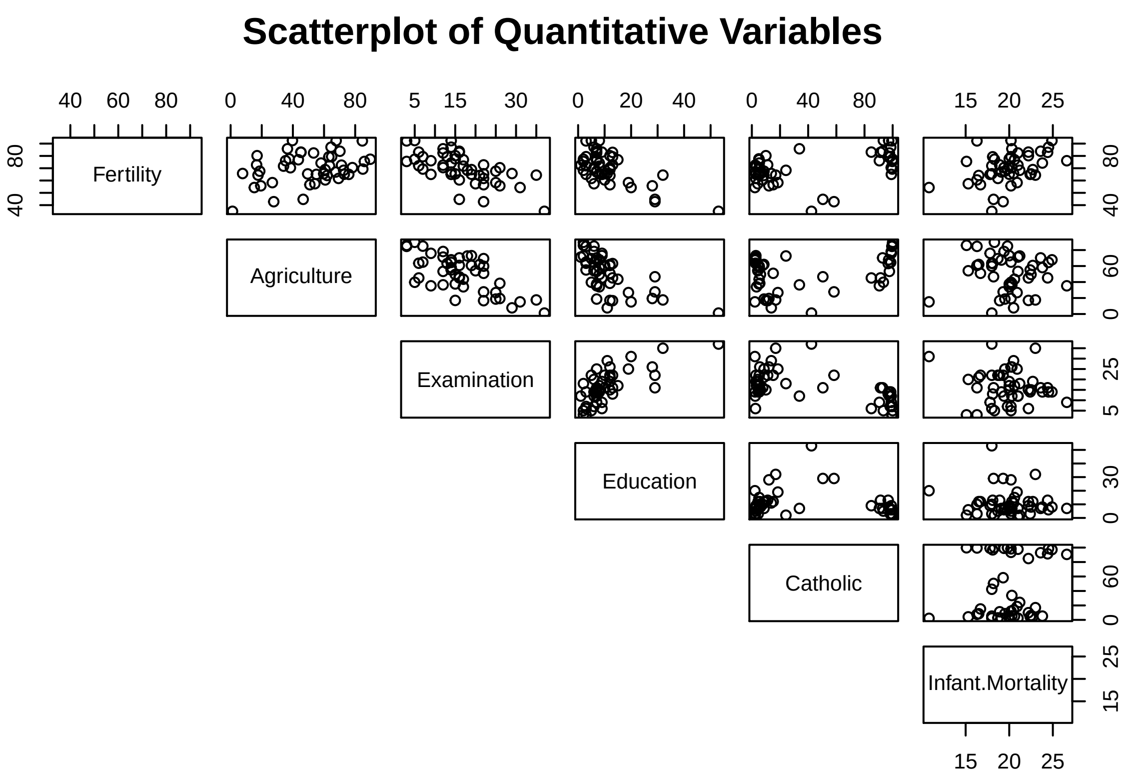
```
library(datasets)
data <- swiss
head(data)
```

```
##              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
## Moutier           85.8        36.5          12         7    33.77
## Neuveville        76.9        43.5          17        15     5.16
```

```
## Porrentruy        76.1        35.3           9          7     90.57
##                Infant.Mortality
## Courtelary                 22.2
## Delemont                   22.2
## Franches-Mnt               20.2
## Moutier                    20.3
## Neuveville                 20.6
## Porrentruy                 26.6
```

A) Create a scatterplot matrix and find the correlation between all pairs of variables for this data set. Answer the following questions based on the output:

```r
pairs(data, lower.panel = NULL, main="Scatterplot of Quantitative Variables")
```




```r
round(cor(data),3)
```

```
##                  Fertility Agriculture Examination Education Catholic
## Fertility            1.000       0.353      -0.646    -0.664    0.464
## Agriculture          0.353       1.000      -0.687    -0.640    0.401
## Examination         -0.646      -0.687       1.000     0.698   -0.573
## Education           -0.664      -0.640       0.698     1.000   -0.154
## Catholic             0.464       0.401      -0.573    -0.154    1.000
## Infant.Mortality     0.417      -0.061      -0.114    -0.099    0.175
##                  Infant.Mortality
## Fertility                   0.417
## Agriculture                -0.061
## Examination                -0.114
## Education                  -0.099
## Catholic                    0.175
## Infant.Mortality            1.000
```

A-I) Which predictors appear to be linearly related to the fertility measure?

From the correlation matrix, we are able to see that the Examination and Education predictors have a strong negative correlation with Fertility. From the scatter plot we can also check that it does seem to be a strong negative linear correlation.

A-II) Do you notice if any of the predictors are highly correlated with one another? If so, which ones?

With looking at the scatter plot and correlation matrix we can see many predictors are correlated with one another. (Agriculture, Examination), (Agriculture, Education), (Examination, Education), and (Examination, Catholic) seem to be pairs that are correlated with one another

B) Fit a multiple linear regression with the fertiliy measure as the response variable and all the other variables as predictors. Use the summary() function to obtain the estimated coefficients and results from the various hypothesis tests for this model.

```
result<-lm(Fertility~., data=data)
summary(result)
```

```
##
## Call:
## lm(formula = Fertility ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2743  -5.2617   0.5032   4.1198  15.3213
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
## Agriculture      -0.17211    0.07030  -2.448  0.01873 *
## Examination      -0.25801    0.25388  -1.016  0.31546
## Education        -0.87094    0.18303  -4.758 2.43e-05 ***
## Catholic          0.10412    0.03526   2.953  0.00519 **
## Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.165 on 41 degrees of freedom
## Multiple R-squared:  0.7067, Adjusted R-squared:  0.671
## F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

B-I) What is being tested by the ANOVA F statistic? What is the relevant conclusion in context?

The idea being tested is whether our multiple linear regression model is useful in predicting the response variable, in this case the fertility variable. The anova F statistic outputs a very small p-value, which means that our multiple linear regression model is useful in predicting the response vairable of fertility. The Anova F test has a null hypothesis that all coeffiencts for the predictors are zero. From the output we can reject the null hypothesis and say there is at least one coeffient for the predictors that is not zero.

B-II) Look at the numerical values of the estimated slopes as well as their p-values. Do they seem to agree with or contradict with what you had written in your answer to part 1a? Briefly explain what do you think is going on here.

From part A-I, it was stated that Education and Examination have a high negative correlation with the fertility variable. However, our summary shows that the examination variable is actually insiginificant and doesn't play a big role in determining the response vairable of fertility. From part A-II we saw that the examination variable is correlated with many other variables not only with fertility. When looking at the model as a whole since examination is correlated with the other variables in the grand scheme of the model, it doesn't need to be added and is insignificant. Also, we can see a contradicting statement between the agriculture variable and fertility variable. The Agriculture and Fertifilty variables were negatively correalted (negative slope), but in the multiuple linear regression summary we can see that the slope is positive, which is contradicting to what we saw before.

2) Data from n = 113 hospitals are used to evaluate factors related to the risk that patients get an infection while in the hospital.....You may assume the regression assumptions are met.

2-A) What is the value of the estimated coefficient of the variable Stay? Write a sentence that interprets this value.

0.237029 is the estiamted coefficient of the variable stay. The percentage of patients who get an infection while hospitalized increases by 0.2307029 with every one unit increase of average length of stay while the other variables are constant.

2-B)

Derive the test statistic, p-value, and critical value for the variable Age. What null and alternative hypotheses are being evaluated with this test statistic? What conclusion should we make about the variable Age?

$H_0 : \beta_{age} = 0$ $H_A : \beta_{age} \neq 0$

The t-statistic is will be the computation of $\frac{\text{Estimate}}{\text{Std. Error}} \rightarrow \frac{-0.014701}{0.022708} \rightarrow -0.6196495$

The p-value was computed using the pt function.

```
pt(-0.6196495, 108) * 2
```

```
## [1] 0.5367937
```

We can see that the p-value is abhove 0.05, which means we fail to reject the null hypothesis. Another way to test without p-value will be to find the critical value as computed below.

```
qt(0.975, 108)
```

```
## [1] 1.982173
```

We can see that the t statistic is not greater than the critical value computed, which is another reason we fail to reject the null hypothesis. The conclusion is that the age predictor is not significant for the multiple linear regression model that utilizes all the predictors given in the dataset.

2-C) A classmate states: "The variable Age is not linearly related to the predicted infection risk." Do you agree with your classmate's statement? Briefly explain.

I would disagree. Simply stating the above statment from the output shown is misleading. It could be that when looking at the predictor age and response infecection risk seperately, they could be linearly related to one another.

2-D) Using the Bonferroni method, construct 95% joint confidence intervals for β_1, β_2, and β_3.

The bonferroni confidence interavls are defined as the following. $\hat{\beta}_j \pm t_{\frac{\alpha}{2*pred},n-p} se(\hat{\beta}_j)$ The t-value used was computed as follows

```
qt(1 - 0.05/6, 108)
```

```
## [1] 2.431841
```

$$\hat{\beta}_1 \pm t_{1-\frac{\alpha}{2*pred},n-p} se(\hat{\beta}_1)$$
$$0.237209 \pm t_{1-\frac{0.05}{2*3},108} * 0.060957$$
$$0.237209 \pm 2.431841 * 0.060957$$
$$(0.0889713, 0.3854467)$$

$$\hat{\beta}_2 \pm t_{1-\frac{\alpha}{2*pred},n-p} se(\hat{\beta}_2)$$
$$-0.014071 \pm t_{1-\frac{0.05}{2*3},108} * 0.022708$$
$$-0.014071 \pm 2.431841 * 0.022708$$
$$(-0.06929325, 0.04115125)$$

$$\hat{\beta}_3 \pm t_{1-\frac{\alpha}{2*pred},n-p} se(\hat{\beta}_3)$$
$$0.020383 \pm t_{1-\frac{0.05}{2*3},108} * 0.005524$$
$$0.020383 \pm 2.431841 * 0.005524$$
$$(0.00694951, 0.03381649)$$

2-E) Fill in the values for the ANOVA table for this regression model

Source of Variation	DF	SS	MS
Regression	4	$s^2 * F * DF = 1.04^2 * 19.56 * 4 = 84.62438$	$s^2 * F = 1.04^2 * 19.56 = 21.1561$
Error	108	$DF * s^2 = 108 * 1.04^2 = 116.8128$	$s^2 = 1.04^2 = 1.0816$
Total	112	$SSR + SSE = 84.62438 + 116.8128 = 201.4372$	*

2-F) What is the R^2 for this model? Write a sentence that interprets this value in context.

$R^2 = \frac{SSR}{SST} = \frac{84.82438}{201.4372} = 0.4210959$

The R^2 value statest that around 42% of the variation of thre reponse variables can be explaine dby the predictors in the the model provided.

2-G) What is the R^2_{adj} for the model?

$R^2_{adj} = 1 - \frac{SS_{res}/(N-k)}{SS_{Total}/(N-1)} = 1 - \frac{116.8128/108}{201.4372/(113-1)} = 0.3986253$

3) Data from 55 college students are used to estimate a multiple regression model with response variable LeftArm, with predictors LeftFoot and RtFoot. All variables were measured in centimeters. Some R output is given below. A classmate points out that there appears to be a contradiction in the R output, namely, while the ANOVA F statistic is significant, the t statistics for both predictors are insignificant. Is your classmate's concern warranted? Briefly explain.

From the anova tests states that our model is useful when predicting the response response variable. However, looking at the predictors indivudually we can see that they show to be insignificant. This means that when one predictor is in the model the other is not is not siginificant. It is critical to look at the predictors one by one. It could be that we don't need both of the variables for the linear model and in fact one or the other will get the job done in prediction analysis.

4 Show that H is idempotent i.e HH = H

$$\begin{aligned}
H &= HH \\
&= (X(X'X)^{-1}X')(X(X'X)^{-1}X') \\
&= X(X'X)^{-1}(X'X)(X'X)^{-1}X' \\
&= XI(X'X)^{-1}X' \\
&= X(X'X)^{-1}X' \\
&= H
\end{aligned}$$