Module 5 Homework

Diana McSpadden

9/24/2020

Stat 6021: Homework Set 5

Name: H. Diana McSpadden

UID: hdm5s

10.02.2020

Attended Study Group With

- Caprill Wright
- Chelsea Alvarado
- Arne Newman
- Jing Huang
- Abby Bernhardt
- Brian Nam
- Loren Bushkar
- Katie Barbre

Question 1: For this first question, …

you will continue to use the dataset swiss which you also used in the last homework. 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.

Q1(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 alternative hypotheses, find the relevant test statistic, p-value, and state a conclusion in context. (For practice, try to calculate the test statistic by hand.)

```
# load the data
#?swiss
```

```
attach(swiss)
```

About Swiss

Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888.

A data frame with 47 observations on 6 variables, each of which is in percent, i.e., in [0, 100].

[,1] Fertility Ig, 'common standardized fertility measure'

[,2] Agriculture % of males involved in agriculture as occupation

[,3] Examination % draftees receiving highest mark on army examination

[,4] Education % education beyond primary school for draftees.

[,5] Catholic % 'catholic' (as opposed to 'protestant').

[,6] Infant.Mortality live births who live less than 1 year.

Question 1a Work

```
swissModel.Full <-lm(Fertility~Education + Catholic + Infant.Mortality + Agri
culture + Examination) # I put agriculture and examination on the end
swissModel.Full # view the full model coefficients

## 
## Call:
## lm(formula = Fertility ~ Education + Catholic + Infant.Mortality + 
##     Agriculture + Examination)
## 
## Coefficients:
##      (Intercept)         Education          Catholic  Infant.Mortality  
##          66.9152           -0.8709            0.1041            1.0770  
##      Agriculture       Examination  
##          -0.1721           -0.2580  

swissModel.Reduced <-lm(Fertility ~ Education + Catholic + Infant.Mortality)
swissModel.Reduced # view the reduced coefficients

## 
## Call:
## lm(formula = Fertility ~ Education + Catholic + Infant.Mortality)
## 
## Coefficients:
##      (Intercept)         Education          Catholic  Infant.Mortality  
##         48.67707          -0.75925           0.09607           1.29615  
```

I will run the anova for the reduced model vs. the full model, and also for the full model.

```
# run the anova for the reduced model vs full model
anova(swissModel.Reduced,swissModel.Full)
```

```
## Analysis of Variance Table
##
## Model 1: Fertility ~ Education + Catholic + Infant.Mortality
## Model 2: Fertility ~ Education + Catholic + Infant.Mortality + Agriculture
+
##     Examination
##   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
```

Question 1a Answer

Hypothesis Test:

H0: Coefficients for Agriculture, Examination, B4 and B5, = 0 (i.e. we use the reduced model)

Ha: at least one of the coefficients for Agriculture, Examination, B4 or B5, != 0 (i.e. support the full model)

Calculate the Test Statistic By Hand

r == number of predictors to drop == 2

n - p; n number of observations, p == number of parameters == 47 - 6 == 41

F r,n-p is what we compare to.

```
print(paste("Test Statistic: ",qf(0.95,2,41)))
```

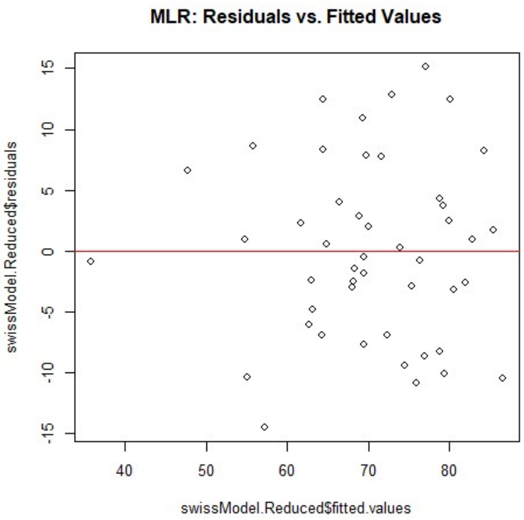
```
## [1] "Test Statistic:  3.22568384229545"
```

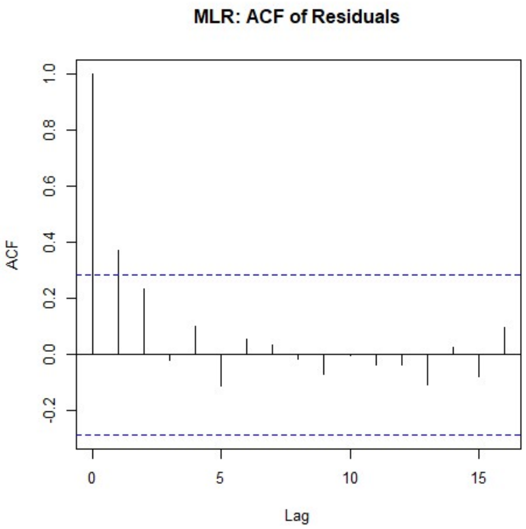
The p value is greater than 0.05, 0.05628, and the test statistic is < critical value, so I fail to reject the null hypothesis, and determine that **the reduced model is appropriate in comparison to the full model.**

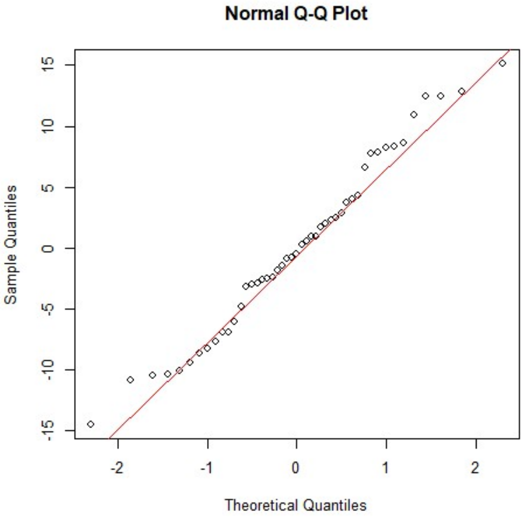
Q1(b) For the model you decide to use from part 1a, assess if the regression assumptions are met.

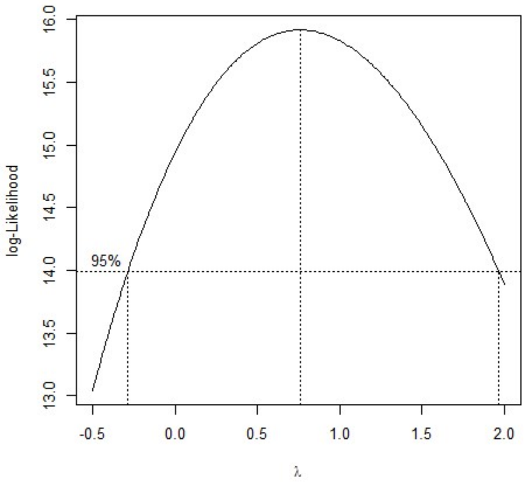
```
par(mfrow=c(2,2)) # 2 rows, 2 columns - there will be three panes to look thr
ough

# Residual plot
plot(swissModel.Reduced$fitted.values,swissModel.Reduced$residuals, main="MLR
: Residuals vs. Fitted Values")
abline(h=0,col="red")
```

```
# ACF Plot
acf(swissModel.Reduced$residuals, main="MLR: ACF of Residuals")

# QQ Plot of Residuals
qqnorm(swissModel.Reduced$residuals)
qqline(swissModel.Reduced$residuals, col="red")

## BoxCox
library(MASS)
boxcox(swissModel.Reduced, lambda = seq(-0.5, 2, 0.01))
```









Question Answer 1b

The BOXCOX contains lambda of 1, so the constant variance assumption appears to be met.

The Residual plot shows an approximate 0 mean of residuals.

The QQ plot shows two “lumps” that indicate some non-normality; however the normality of residuals is the least important of the assumptions.

The ACF plot shows that the residuals are correlated with a lag of 1, which does not meet the regression assumptions, but is most likely a representation of the data collection, and leads to the question of whether the geographic providences from which the data were collected are correlated to each other by proximity.

Question 2 (No R) The data for this question come from 113 hospitals…

The key response variable is **InfctRsk**, the risk that patients get an infection while staying at the hospital. We will look at five predictors:

- x1: Stay. Average length of stay at hospital
- x2: Cultures. Average number of bacterial cultures per day at the hospital
- x3: Age. Average age of patients at hospital
- x4: Census. The average daily number of patients
- x5: Beds. The number of beds in the hospital

R Output

```
Call: lm(formula = InfctRsk ~ Stay + Cultures + Age + Census + Beds)

Coefficients: Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.2051282 1.2075929 0.170 0.8654

Stay 0.2055252 0.0660885 3.110 0.0024 **

Cultures 0.0590369 0.0103096 5.726 9.5e-08 ***

Age 0.0173637 0.0229966 0.755 0.4519

Census 0.0010306 0.0034942 0.295 0.7686

Beds 0.0004476 0.0026781 0.167 0.8676

Residual standard error: 0.9926 on 107 degrees of freedom

Multiple R-squared: 0.4765, Adjusted R-squared: 0.4521

F-statistic: 19.48 on 5 and 107 DF, p-value: 9.424e-14
```

Analysis of Variance Table

```
Response: InfctRsk
```

```
        Df         Sum Sq         Mean Sq          F value      Pr(>F)
```

Stay 1 57.305 57.305 58.1676 1.044e-11 ***

Cultures 1 33.397 33.397 33.8995 6.154e-08 ***

Age 1 0.136 0.136 0.1376 0.71144

Census 1 5.101 5.101 5.1781 0.02487 *

Beds 1 0.028 0.028 0.0279 0.86759

Residuals 107 105.413 0.985

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?

Answer Question 2a

Based on the t-values Age, Census, and Beds appear 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 regression model. Show all steps, including your null and alternative hypotheses, the corresponding test statistic, p-value, critical value, and your conclusion in context.

Work on Question 2b

H0: the coefficients for Age, Census, and Beds = 0

Ha: at least one of the coefficients for Age, Census, or Beds != 0.

Use Following calculations r == number of predictors to drops

n - p; n number of observations, p == number of parameters

F r,n-p is what we compare to.

```
# calculate that F statistic using SSR's and SSE
F0 = ((0.136 + 5.101 + 0.028) / 3) / (105.413 / 107)
print(paste("F0 is: ", F0))

## [1] "F0 is:  1.78142164628651"

# create the critical value
criticalFValue = qf(0.95,3,107)
print(paste("The critical value is: ", criticalFValue))

## [1] "The critical value is:  2.68948977200524"
```

```
# create the pvalue
pvalue = 1-pf(F0,3,107)
print(paste("The p value for the F statistic is: ", pvalue))

## [1] "The p value for the F statistic is:  0.155092525265351"
```

Answer Question 2b The p-value for the F-statistic is 0.155 (F0 = |1.781| < |2.689|) which indicates we fail to reject H0. The simpler model is appropriate with the Stay and Cultures predictor variables.

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

- Full Model: using x1; x2; x3; x4 as the predictors for InfctRsk
- Reduced Model: using x1; x2 as the predictors for InfctRsk

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

Work on Question 2c

Because SST is not dependent on predicted y values it stays constant regardless of predictor variables being added or removed from the model. Since neither Question 2c's Model 1 or Model 2 include the regressor x5, I will recreate the ANOVA table without x5 effect on the model and add it's regression error to the total residual error. I will also need to update the degrees of freedom by adding 1 since we have 1 fewer predictor variables.

Analysis of Variance Table Without x5

Response: InfctRsk

```
        Df        Sum Sq        Mean Sq          F value      Pr(>F)
```

Stay 1 57.305 57.305 58.1676 1.044e-11 ***

Cultures 1 33.397 33.397 33.8995 6.154e-08 ***

Age 1 0.136 0.136 0.1376 0.71144

Census 1 5.101 5.101 5.1781 0.02487 *

Residuals 108 105.441 ???

Hypothesis Statements H0: The coefficients for Age and Census are == 0, i.e. use reduced model.

Ha: At least one of the coefficients for Age or Census are != 0, i.e. use the full model.

Use Following calculations r == number of predictors to drops

n - p; n number of observations, p == number of parameters

F r,n-p is what we compare to.

```
# calculate that F statistic using SSR's and SSE when dropping x3 and x4
F0 = ((0.136 + 5.101) / 2) / (105.441 / 108)
print(paste("F0 is: ", F0))

## [1] "F0 is:  2.68204967707059"

# create the critical value
criticalFValue = qf(0.95,2,108)
print(paste("The critical value is: ", criticalFValue))

## [1] "The critical value is:  3.08038686329258"

# create the pvalue
pvalue = 1-pf(F0,2,108)
print(paste("The p value for the F statistic is: ", pvalue))

## [1] "The p value for the F statistic is:  0.0729799441799894"
```

I fail to reject H0, and I will use the reduced model using Stay and Cultures as predictor variables.

Question 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. You may assume the regression assumptions are met. Some R output is given below.

Call: lm(formula = LeftArm ~ LeftFoot + RtFoot)

```
        Coefficients:    Estimate Std. Error    t value     Pr(>|t|)
```

(Intercept) 11.7104 2.5179 4.651 2.31e-05 *** LeftFoot 0.3519 0.2961 1.188 0.240 RtFoot 0.1850 0.2816 0.657 0.514 — Residual standard error: 1.796 on 52 degrees of freedom

Multiple R-squared: 0.3688, Adjusted R-squared: 0.3445

F-statistic: 15.19 on 2 and 52 DF, p-value: 6.382e-06

Explain how this output indicates the presence of multicollinearity in this regression model.

Answer Question 3 The indication of multicollinearity is that the p values for both predictors do not indicate that the predictor, when holding the other predictor constant, is a significant predictor in the model. The R output seems to say that there are no significant predictors when used simultaneously; **however** the F-statistic and corresponding p-value say that the total model is statistically significant. Together these two findings indicate that there is multicollinearity in the model.