Module04HW

Diana McSpadden

9/22/2020

Assignment: Stat 6021: Homework Set 4

Name: H. Diana McSpadden

UID: hdm5s

Attended Study Group With: Caprill Wright, Chelsea Alvarado, Jen Leopold, Melanie Sattler, Loren Bushkar, JIng Hung

References:

https://pdfs.semanticscholar.org/390d/8bd14f3ae307ebb600b8658c34b7ec7505ee.pdf

Question 1: 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.

```
# load the data
?swiss

## starting httpd help server ... done

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.

```
head(swiss)
```

```
##              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
```

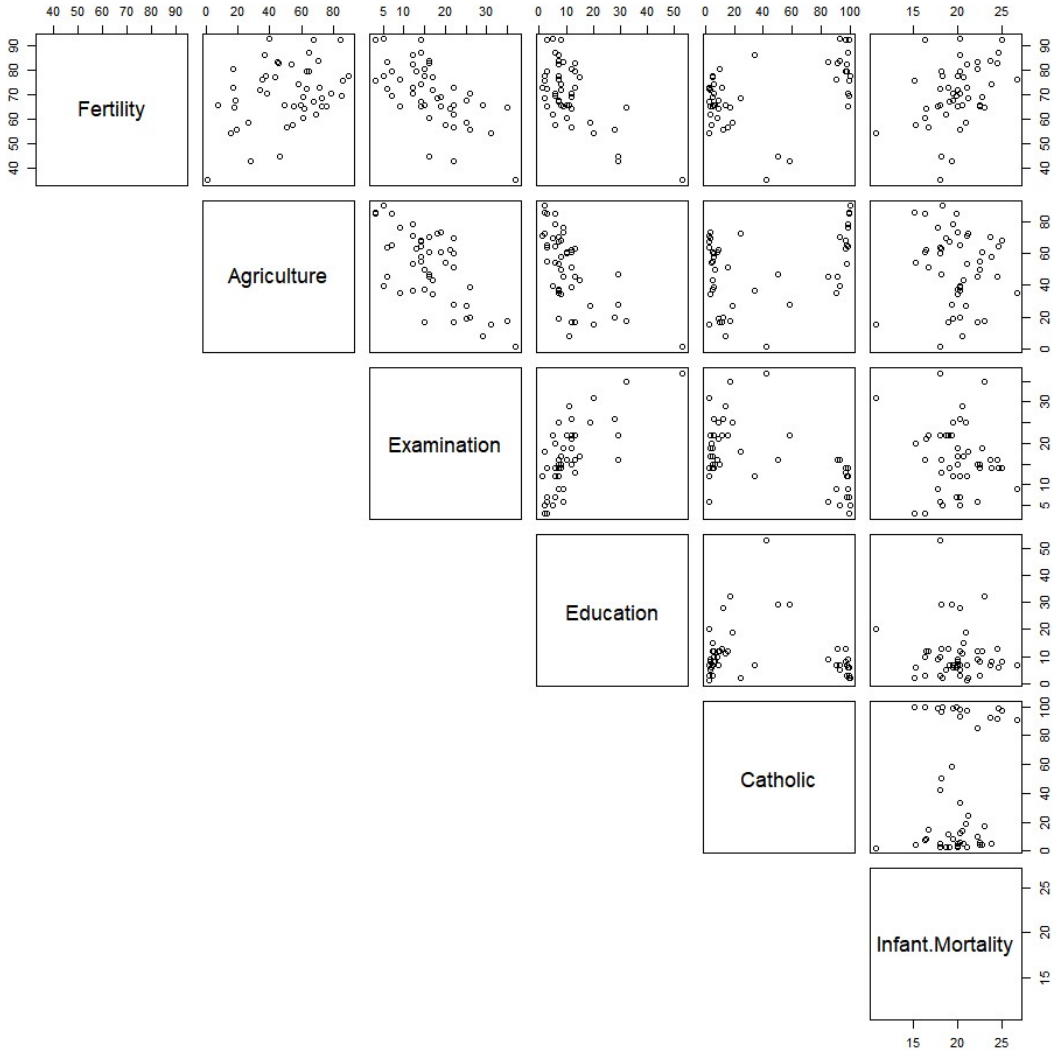
```
summary(swiss)
```

```
##    Fertility      Agriculture     Examination      Education
##  Min.   :35.00   Min.   : 1.20   Min.   : 3.00   Min.   : 1.00
##  1st Qu.:64.70   1st Qu.:35.90   1st Qu.:12.00   1st Qu.: 6.00
##  Median :70.40   Median :54.10   Median :16.00   Median : 8.00
##  Mean   :70.14   Mean   :50.66   Mean   :16.49   Mean   :10.98
##  3rd Qu.:78.45   3rd Qu.:67.65   3rd Qu.:22.00   3rd Qu.:12.00
##  Max.   :92.50   Max.   :89.70   Max.   :37.00   Max.   :53.00
##     Catholic       Infant.Mortality
##  Min.   :  2.150   Min.   :10.80
##  1st Qu.:  5.195   1st Qu.:18.15
##  Median : 15.140   Median :20.00
##  Mean   : 41.144   Mean   :19.94
##  3rd Qu.: 93.125   3rd Qu.:21.70
##  Max.   :100.000   Max.   :26.60
```

(a) Create a scatterplot matrix

and find the correlation between all pairs of variables for this data set.

```
pairs(swiss,lower.panel = NULL)
```



Answer the following questions based on the output:

i. Which predictors appear to be linearly related to the fertility measure?

```
swiss.cor <- cor(swiss)

threshold <- 0.6 # set a correlation threshold, is there a standard? A: Not r
eally.

corWorking <- swiss.cor

diag(corWorking) <- 0 # set diagonal to 0 so 1's of unrelated don't get caugh
t in threshold

swiss.cor.related <- apply(abs(corWorking) >= threshold, 1, any) # apply the
filter to the absolute value of the correlations between variables.
```

```
swiss.cor[swiss.cor.related, swiss.cor.related]
```

```
##              Fertility Agriculture Examination  Education
## Fertility    1.0000000   0.3530792  -0.6458827 -0.6637889
## Agriculture  0.3530792   1.0000000  -0.6865422 -0.6395225
## Examination -0.6458827  -0.6865422   1.0000000  0.6984153
## Education   -0.6637889  -0.6395225   0.6984153  1.0000000
```

Answer 1ai:

Visually, Fertility appears correlated to Agriculture.

From swiss.cor, Fertility appears to be linearly related to: Examination (-), Education (-).

Visually, Catholic has two or three distinct clusters that appear linearly correlated with Fertility,but a constant variance issue.

Visually, Infant.Mortality appears linearly correlated with Fertility, but also with a constant variance issue, the plot is cone shaped with decreasing variance.

ii. Do you notice if any of the predictors are highly correlated with one another? If so, which ones?

Answer aaii:

- Agriculture appears corrleated with examination(-) and education(-)
- Examination appears correlated with education (+)

(b) Fit a multiple linear regression

with the fertiliy measure as the response variable and all the other variables as predictors.

```
swissModel <-lm(Fertility ~ Agriculture + Examination + Education + Catholic
+ Infant.Mortality)

swissModel
```

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

Use the summary() function to obtain the estimated coefficients and results from the various hypothesis tests for this model.

```
summary(swissModel)

##
## Call:
## lm(formula = Fertility ~ Agriculture + Examination + Education +
##     Catholic + Infant.Mortality)
##
## 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
```

i. What is being tested by the ANOVA F statistic? What is the relevant conclusion in context?

Answer 1bi:

ANOVA F statistic is calculated by (MS-explained by model) / (MS-residual errors in model). Fewer unexplained errors, compared to fewer explained errors result in a larger F0.

F0 explains the adequacy of our model in comparison to a normal distribution of errors when B1 … Bj are all == 0.

F0 describes the of our model, incuding all predictor variables, in predicting fertility.

The F0 statistic says that at least one of our predictor variables is statistically significant.

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.

Answer 1bii:

When Agriculture, Examination, Education, Catholic, and Infant.Mortality are all used as predictor variables for Swiss fertility in 1888, Education, Catholic, and Infant.Mortality are have the most predictive power based on their B p-values.

Some aspects of this finding agree and some disagree:

Agree with my initial reading of the plots:

Because Education was correlated with both Agriculture and Examination these B p-values indicate that when all three are included in the model, Education has the most predictive value.

Disagree with my initial visual reading of the 4 plots:

I had seen too much variance in the Infant.Mortality plot to consider it a good linear relationship.

I had not considered Catholic a good linearly relationship because of the two clusters. It appeared more of a piece-wise linear relationship.

Conclusion

With MLR all the predictor variable not only influence the response variable, but influence the influence of each other. Catholic appears to have predictive value in the center of the residual plot, and Infant.Mortality appears to have predictive power at the + and - extremes of the residual plot. I believe that the combination of the two variables together creates the higher predictive power when they are used together in the model. The summary table does not contradict a visual reading after considering the interactivity if the predictors.

2. (No R required) Data from n = 113 hospitals are used …

to evaluate factors related to the risk that patients get an infection while in the hospital. The response variable is **InfctRsk**, the percentage of patients who get an infection while hospitalized.

The predictors are

- **Stay**, the average length of stay,
- **Age**, the average patient age,
- **Xrays**, a measure of how many Xrays are done in the hospital, and
- **Services**, a measure ofhow many different services the hospital orders.

We consider the following multiple regression equation:

E(InfctRsk) = B0 + B1Stay + B2Age + B3Xrays + B4Services.

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

Knowns from chart

B0: 0.170874

B1 (Stay): 0.237209

B2 (Age): -0.014071

B3 (Xrays): 0.020383

B4 (Services): 0.022718

Est. Standard Error, t value, p-value of t B1 (Stay): 0.0609571 3.891 0.00173***

Est. Standard Error, t value, p-value of t B2 (Age): 0.022708 ____ _____

Est. Standard Error, t value, p-value of t B3 (Xrays): 0.005524 3.690 0.000354***

Est. Standard Error, t value, p-value of t B4 (Services): 0.006970 3.260 0.001493**

Residual Standard Error: 1.04 on 108 degrees freedom

Multiple R squared: ___

Adjusted R squared: ___

F-stat: 19.56 on 4 and 108 degrees freedom

p-value of F0 = 3.96e-12

(a) What is the value of the estimated coeffcient of the variable Stay?**

Write a sentence that interprets this value.

Answer 2a:

The estimated coefficient of Stay, B1, is **0.237209**

An interpretation of this coefficient is: In the presence of all of the model's predictor variables, each unit increase in Stay produces a 0.237209 increase in the percentage of patients who get a HAI (healthcare acquired infection) in the hospital.

My model predicts an increase of 0.237 percent infections for each unit increase in the average length of hospital stay when controlling for the effect of other predictors.

(b) Derive the test statistic, p-value, and critical value for the variable Age.

t0 = Bj-hat / se(Bj-hat)

```
# calculate the B2 test statistics
B2.teststatistic = -0.014071 / 0.022708 # se(Bj-hat) is given
print(paste("The B2 test statistic is: ", B2.teststatistic))

## [1] "The B2 test statistic is:  -0.619649462744407"

# calculate the B2 p-value of the test statistic
B2.pvalue = 2*pt(-abs(B2.teststatistic), df=(113 - 4 - 1))  # calc p Value -
two-sided is the 2 *
print(paste("The B2 p-value is: ", B2.pvalue))

## [1] "The B2 p-value is:  0.536793700887277"
```

```r
# calculate the B2 critical value to compare the test statistic to
criticalValue = qt(0.975, (113 - 4 - 1))
print(paste("The critical value is: ", criticalValue))

## [1] "The critical value is:  1.98217348330773"
```

Answer 2b:

The B2 test statistic is: **-0.6196**

The B2 p-value is: **0.5368**

The critical value is: **1.9822**

What null and alternative hypotheses are being evaluated with this test statistic? **

H0: B2 = 0

Ha: B2 != 0 in the presence of other Bi's where i != 2.

What conclusion should we make about the variable Age?**

We can eliminate Age from the model predicting % of patients infected, when we are testing with a model with the other predictors.

(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.

Answer: 2c

No, I do not agree. The classmate's statement is not complete because it does not include the phrase "when controlling for the effect of other predictors." If one wanted to test whether Age is related to infection risk one should run a simple linear regression.

(d) Using the Bonferroni method, construct 95% joint confidence intervals for B1, B2,and B3.

** 2d Work**

I am using B1 = Stay, B2 = Age, and B3 = Xray. I have not removed Age from the model, so Age remains B2.

delta multiplier = t((alpha/2)/g,n-k-1)

where,

g = 3 = number of intervals we are constructing

n = number of observations

k = number of regressors

alpha/2p

CI for Bj-hat = Bj-hat +- (delta * se(Bj-hat))

```
B1 <- 0.237209
se.B1 <- 0.060957

B2 <- -0.014071
se.B2 <- 0.022708

B3 <- 0.020383
se.B3 <- 0.005524

delta <- qt((1 - (.05/6)), (113 - 4 - 1))

print(paste("tvalue: ",delta))
```

```
## [1] "tvalue:  2.43184113858753"
```

```
B1.CI.low <- B1 - (delta * se.B1)
B1.CI.high <- B1 + (delta * se.B1)
print(paste("The simultaneous 95% CI for B1 is: ", B1.CI.low, " - ", B1.CI.hi
gh))
```

```
## [1] "The simultaneous 95% CI for B1 is:  0.0889712597151198  -  0.38544674
028488"
```

```
B2.CI.low <- B2 - (delta * se.B2)
B2.CI.high <- B2 + (delta * se.B2)
print(paste("The simultaneous 95% CI for B2 is: ", B2.CI.low, " - ", B2.CI.hi
gh))
```

```
## [1] "The simultaneous 95% CI for B2 is:  -0.0692932485750457  -  0.0411512
485750457"
```

```
B3.CI.low <- B3 - (delta * se.B3)
B3.CI.high <- B3 + (delta * se.B3)
print(paste("The simultaneous 95% CI for B3 is: ", B3.CI.low, " - ", B3.CI.hi
gh))
```

```
## [1] "The simultaneous 95% CI for B3 is:  0.00694950955044247  -  0.0338164
904495575"
```

Answer 2d:

The simultaneous 95% CI for B1 is: 0.08897 - 0.38545

The simultaneous 95% CI for B2 is: -0.06929 - 0.04115

The simultaneous 95% CI for B3 is: 0.00694 - 0.0338

(e) Fill in the values for the ANOVA table for this regression model.

2e Work

```
n = 113 # number of observations
k = 4 # number regressor variables
p = k + 1

MSE = 1.04^2

SSE = MSE * (n-p)

F0 = 19.56

MSR = MSE * F0

SSR = MSR * k

SST = SSE + SSR

R2 = SSR / SST

print(paste("Regression df: ",k))

## [1] "Regression df:  4"

print(paste("Error df: ",n - p))

## [1] "Error df:  108"

print(paste("Total df: ",n - 1))

## [1] "Total df:  112"

print(paste("MSE: ", MSE))

## [1] "MSE:  1.0816"

print(paste("SSE: ", SSE))

## [1] "SSE:  116.8128"

print(paste("F0: ", F0))

## [1] "F0:  19.56"

print(paste("MSR: ", MSR))

## [1] "MSR:  21.156096"

print(paste("SSR: ", SSR))

## [1] "SSR:  84.624384"
```

```
print(paste("SST: ", SST))
```

```
## [1] "SST:  201.437184"
```

```
print(paste("R2: ", R2))
```

```
## [1] "R2:  0.420103092783505"
```

2e Answer:

Source of Variation	DF	SS	MS
Regression	4	**84.624384**	**21.156096**
Error	113-4-1 = **108**	1.0816 * 108=**116.8128**	**1.0816**
Total	113-1 = **112**	**201.437184**	***

(f) What is the R2 for this model? Write a sentence that interprets this value in context.

2f Work

R2 = SSR / SST

```
print(paste("R2 for the model is: ", R2))
```

```
## [1] "R2 for the model is:  0.420103092783505"
```

2f Answer:

R2 for the model is: 0.42. Interpreted as: 42% of the variability in percentage of patients infected is predicted by the model; HOWEVER, because our model contains four predictor variables, this interpretation is suspect because, by definition, as each predictor variable is added R2 cannot decrease.

(g) What is the R2 adj for this model?

2g Work

R2-adj = 1 - ((SSE / (n - p)) / (SST / (n-1)))

```
R2.adj <- 1 - ((SSE / (n - p)) / (SST / (n-1)))

print(paste("R2 Adjusted for the model is: ", R2.adj))
```

```
## [1] "R2 Adjusted for the model is:  0.398625429553265"
```

2f Answer

R2 Adjusted for the model is: **0.3986**

Question 3 (No R required) 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.

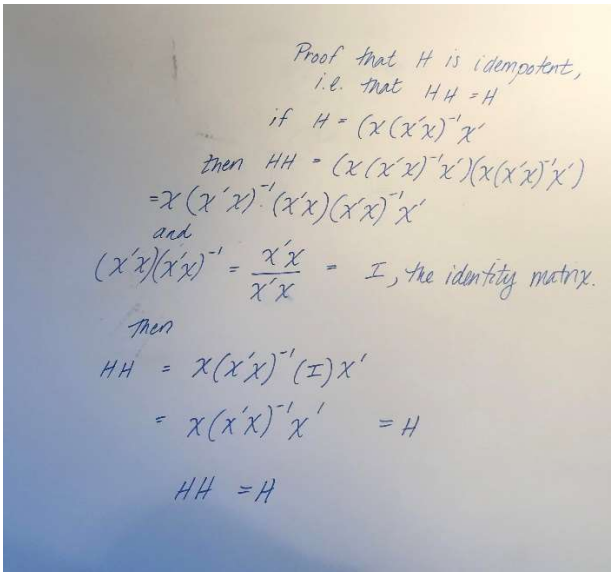
Answer Question 3

This seems very strange, but it is possible that even though the two regressors are not significant when the other is held constant (as shown by the t statistics), but that colinearity is present, and that the "bad" t statistics with a "good" F statistic means that the regressors are correlated and we don't know which one we can remove. We do know that one of the predictors has a slope != 0.

Question 4 (No R required) Recall in matrix notation, …

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

Question 4 Work and Answer



Question 5 Please remember …

to complete the Module 4 Guided Question Set Participation Self and Peer-Evaluation Questions via Test & Quizzes on Collab. **Done**