Smith SCHOOL OF BUSINESS Queen's University

Global Master of Management Analytics

GMMA 860
Acquisition and Management of Data

Professor Alexander Scott

Assignment #1
April 5, 2020

William Chak Lim Chan

Order of files:

Filename	Pages	Comments and/or Instructions

Additional Comments:

Apologies for the super late submission. I had been furloughed from Air Canada on March 31, then signed my offer with Merck, all while trying to figure out EI/CERB. Here is my assignment, finally.

Question 1

a. Explain in language that your manager is likely to understand how multiple imputation deals with missing values.

Multiple imputation seeks to replace missing values in a variable with a linear regression. An example would be trying to predict (hockey) goals scored (player) with time on ice, games played, plus/minus, total assists, but you were missing some time on ice data because the camera tracking technology was faulty during certain periods. You would estimate time on ice with a regression using the other variables and then impute the average of those values back into your original regression model. It's important to note that the number of imputations we'll use will be equal to the percentage of missing observations.

Original Model: Goals Scored = B0 + B1TOI + B2GP + B3PM + B4TA + e

Imputation Model: TOI = B0 + B1GS + B2GP + B3PM + B4TA + e (run this x times based on the % missing)

b. Under what condition(s) could multiple imputation be used reliably to deal with missing values. Provide an original example (i.e. not ones that I have provided or that we have discussed in class) to illustrate when multiple imputation could reliably be used – if there are more than one condition, be sure to illustrate them all and describe how they apply in your example.

The hockey example above I provided is a prime candidate for multiple imputation as it would contain MCAR data, as long as the camera tracking technology wasn't intentionally programmed to systematically omit ice time values.

c. If the conditions you describe above were not met, what else could you do? What might some problems / concerns be with such an approach?

If the data was not MCAR, we would have to omit the player from the specific game which would provide inaccurate results (i.e. we would underpredict the number of goals since the player has one less game recorded). That would be the worst-case scenario. A step above this would be to use subject matter expertise and blindly impute an average. We would of course need to add heavy disclaimers to our results.

d. Estimate the model y = B0 + B1 X1 + B2 X2 + … B5 X5 using multiple imputation to correct for missing values.

```
> summary(pool(reg1))
         term   estimate  std.error statistic       df      p.value
1 (Intercept) 982.771628 504.550902 1.9478146 91.18673 0.0545129892
2          X1   6.228724   3.507875 1.7756401 91.00753 0.0791349596
3          X2  14.444035   7.158360 2.0177856 82.49564 0.0468641579
4          X3  15.385880   4.470965 3.4412887 91.74031 0.0008734881
5          X4  11.005507   5.418197 2.0312121 88.72609 0.0452259437
6          X5   4.177562   4.379467 0.9538974 89.54697 0.3427035249
> View(imputed)
> View(imputed)
```

e. According to your results, does X2 belong in that model? Explain why / why not.

Based on the model, X2's p-value is 0.04 < 0.05, therefore it belongs in the model. I would remove X1, X5.

f. Suppose your non-technical manager sees your R output (by the way – never let this happen!) She notices that there are several estimates for the B1 term if you summarize the regression – explain to her why this is the case.

"Yeah, so the first B1 estimate isn't exactly accurate because the first regression was run with a data set that had missing values. Ignore that. We ran another regression using a data set that had imputed values which generated a more improved estimate and model. If you're asking me which one, I trust, or should be kept, it would be the latter one."

Question 2

a. Present your final model and the estimated parameters. This can be cut and paste from R or other sources if that is appropriate. What steps did you go through to develop this model?

My final model is:
Rating = B0 + B1Price + B2+Alcohol + B3Sulphates + B4Canada + B5US + error.

```
#ASSIGNMENT 2 - QUESTION 2 - PROBLEM A

missing <- read_excel("OneDrive - Queen's University/Global Master of Management Analytics/GMMA 860 - Acquisition and Manage

head(wine)
summary(wine)

wine$Canada <- ifelse(wine$Country == "Canada", 1, 0)
wine$US <- ifelse(wine$Country == "US", 1, 0)
wine$France <- ifelse(wine$Country == "France", 1, 0)
wine$Italy <- ifelse(wine$Country == "Italy", 1, 0)

view(wine)

#First Regression Model
reg_wine <- lm(Rating ~ Price + Alcohol + Residual_Sugar + Sulphates + pH + Canada + US + Italy + France + US, wine)

summary(reg_wine)
plot(reg_wine)

#Second Reg Model - FINAL MODEL
reg_wine1 <- lm(Rating ~ Price + Alcohol + Sulphates + Canada + US, wine)
summary(reg_wine1)
plot(reg_wine1)
```

34:1 (Top Level) R Script

Console Terminal Jobs

```
Residuals:
     Min       1Q   Median       3Q      Max
-18.6544  -3.5835   0.1769   4.3040  15.6913

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.07262    5.67223   0.189     0.85
Price         0.93506    0.05132  18.219  < 2e-16 ***
Alcohol       3.65097    0.43591   8.375 5.17e-13 ***
Sulphates   -13.52199    1.10329 -12.256  < 2e-16 ***
Canada       -9.90520    1.54513  -6.411 5.75e-09 ***
US          -11.20268    1.59074  -7.042 3.08e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.159 on 94 degrees of freedom
Multiple R-squared:  0.8718,    Adjusted R-squared:  0.8649
F-statistic: 127.8 on 5 and 94 DF,  p-value: < 2.2e-16
```

The following steps were taken to build the model:

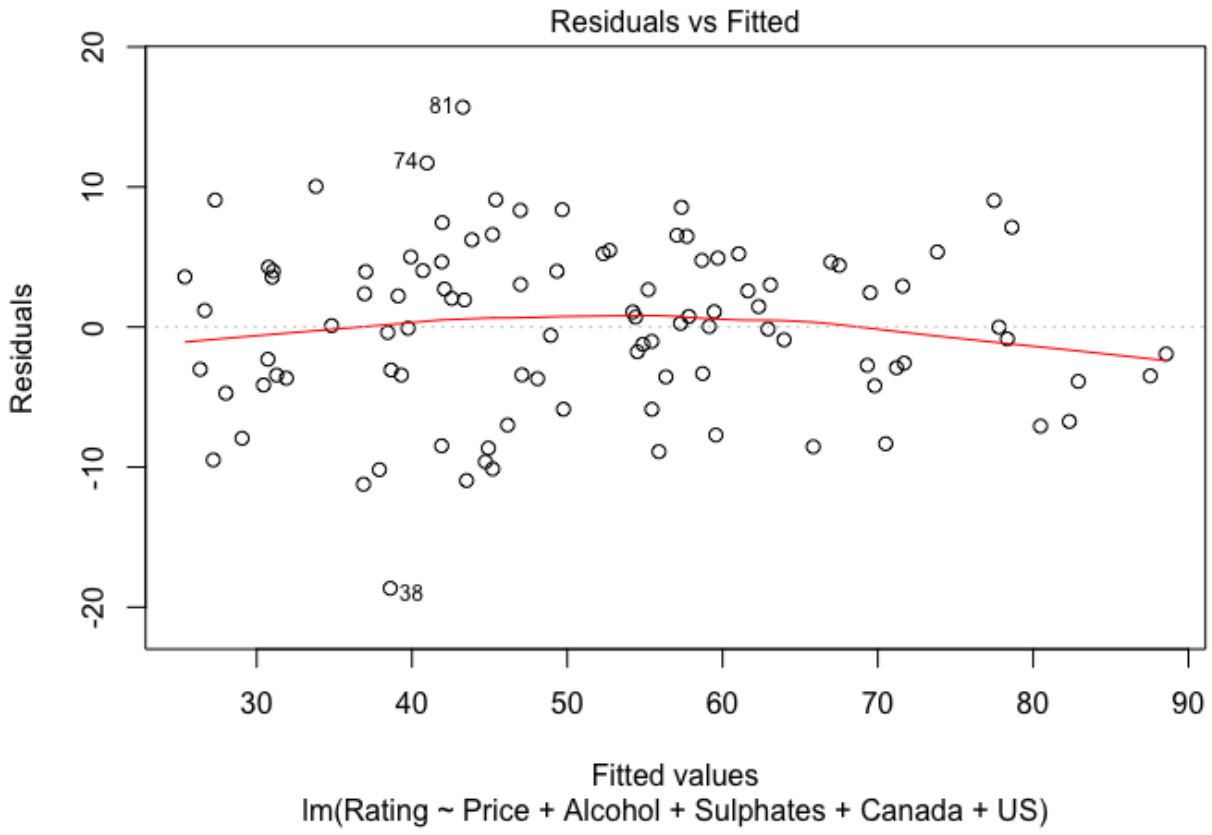
1. Import and Review Data
2. Create Dummy Variables from each country
3. Run regression with all variables
4. Review significance of each variable
5. Review plots
6. Remove Residual_Sugar, pH, Italy, France from model
7. Re-run model
8. Review plots

This was the original model:

```
Call:
lm(formula = Rating ~ Price + Alcohol + Residual_Sugar + Sulphates +
    pH + Canada + US + Italy + France + US, data = wine)

Residuals:
     Min       1Q   Median       3Q      Max
-19.1229  -3.7267   0.7814   4.0237  15.3828

Coefficients: (1 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      8.15141    7.15969   1.139    0.258
Price            0.93568    0.05218  17.931  < 2e-16 ***
Alcohol          3.42869    0.47999   7.143 2.18e-10 ***
Residual_Sugar   0.14392    0.66945   0.215    0.830
Sulphates      -13.86710    1.11938 -12.388  < 2e-16 ***
pH              -0.87922    1.11652  -0.787    0.433
Canada         -11.95777    2.01500  -5.934 5.28e-08 ***
US             -12.99630    2.01693  -6.444 5.45e-09 ***
Italy           -3.17215    1.95270  -1.624    0.108
France                NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.15 on 91 degrees of freedom
Multiple R-squared:  0.8762,    Adjusted R-squared:  0.8653
F-statistic: 80.52 on 8 and 91 DF,  p-value: < 2.2e-16
```

b. Does the data appear to be heteroskedastic? Why or why not? Show evidence.

The plot does not appear to be heteroskedastic. There is not obvious cone shaped plot.



c. Assuming there are no data problems, what would a wine be rated if it comes from France, has a price of $39.99, and alcohol content of 13.9%, sulphates of 0.5, and residual sugar of 1.96?

Rating = -4.66233 + 1.025 (39.99) + 3.29479 (.139) -12.9989 (.5) + 9.01079
The wine's rating would be 39.29. Residual Sugar does not belong in the model as it was proven to be non-significant in a previous run.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.66233    6.51241  -0.716    0.476
Price         1.02500    0.05835  17.567  < 2e-16 ***
Alcohol       3.29479    0.52556   6.269 1.06e-08 ***
Sulphates   -12.99899    1.30085  -9.993  < 2e-16 ***
France        9.01079    2.02684   4.446 2.37e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.248 on 95 degrees of freedom
Multiple R-squared:  0.8205,    Adjusted R-squared:  0.8129
F-statistic: 108.6 on 4 and 95 DF,  p-value: < 2.2e-16
```

d. Would increasing the price of a wine increase its expert rating? Be sure to clearly explain your thinking. (HINT: consider what a model can actually tell you about a relationship)

No, absolutely not. Changing the price of the wine would not necessarily lead to a high rating. This is not how we interpret the model. There is no direct causation involved, only that these variables are correlated. Further analysis must be run to prove that changing the price has a direct causation to higher wine ratings.

R Script

```
#ASSIGNMENT 2 - QUESTION 1

#PROBLEM D

#IMPORT LIBRARIES
library(readxl)
library(tidyverse)
library(mice)

#IMPORT DATA
missing <- read_excel("OneDrive - Queen's University/Global Master of Management Analytics/GMMA 860 - Acquisition and Management of Data/Assignment #2/GMMA860_Assignment2_Data.xlsx", sheet = "Missing")

#VIEW SUMMARIES
head(missing)
summary(missing)

#CREATE THE REGRESSION MODEL
reg <- lm(Y ~ X1 + X2 + X3 + X4 + X5, missing)

#VIEW THE SUMMARY
summary(reg)

#IMPUTE DATA
imputed <- mice(missing, m=5, maxit=30, meth='pmm', seed=1)

#CREATE REGRESSION WITH IMPUTED DATA
reg1 <- with(imputed, lm(Y ~ X1 + X2 + X3 + X4 + X5))
summary(reg1)
summary(pool(reg1))

#ASSIGNMENT 2 - QUESTION 2

#PROBLEM A

#IMPORT DATA
missing <- read_excel("OneDrive - Queen's University/Global Master of Management Analytics/GMMA 860 - Acquisition and Management of Data/Assignment #2/GMMA860_Assignment2_Data.xlsx", sheet = "Wine")

#VIEW SUMMARY
head(wine)
summary(wine)

#DUMMY THE VARIABLES
wine$Canada <- ifelse(wine$Country == "Canada", 1, 0)
wine$US <- ifelse(wine$Country == "US", 1, 0)
wine$France <- ifelse(wine$Country == "France", 1, 0)
```

```
wine$Italy <- ifelse(wine$Country == "Italy", 1, 0)

#VIEW AND CONFIRM DATA
view(wine)

#RUN FIRST REGRESSION MODEL
reg_wine <- lm(Rating ~ Price + Alcohol + Residual_Sugar + Sulphates + pH + Canada + US + Italy + France +
US, wine)

#ANALYZE PLOTS
summary(reg_wine)
plot(reg_wine)

#Second Reg Model - FINAL MODEL
reg_wine1 <- lm(Rating ~ Price + Alcohol + Sulphates + Canada + US, wine)
summary(reg_wine1)
plot(reg_wine1)

#PROBLEM C

#RUN MODEL WITH 4 VARIABLES
reg_wine2 <- lm(Rating ~ Price + Alcohol + Sulphates + France, wine)
summary(reg_wine2)
plot(reg_wine2)
```