IST772 Problem Set 8

Shivani Sanjay Mahaddalkar

The homework for week 8 is based on exercises 2-8 on pages 181-182 but with changes as noted in this notebook (i.e., follow the problems as given in this document and not the textbook).

Attribution statement: (choose only one) 2. I did this homework with help from the book and the professor and these Internet sources:
https://en.wikipedia.org/wiki/Variance_inflation_factor

Chapter 8, Exercise 2

The data sets package in R contains a small data set called swiss that contains n = 47 observations of socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888. Use "?swiss" to display help about the data set.

Create and interpret a bivariate correlation matrix using cor(swiss) keeping in mind the idea that you will be trying to predict the Fertility variable. Which other variable might be the single best predictor? (1 pt)

```
cor(swiss)
```

```
##                   Fertility Agriculture Examination   Education   Catholic
## Fertility         1.0000000  0.35307918  -0.6458827 -0.66378886  0.4636847
## Agriculture       0.3530792  1.00000000  -0.6865422 -0.63952252  0.4010951
## Examination      -0.6458827 -0.68654221   1.0000000  0.69841530 -0.5727418
## Education        -0.6637889 -0.63952252   0.6984153  1.00000000 -0.1538589
## Catholic          0.4636847  0.40109505  -0.5727418 -0.15385892  1.0000000
## Infant.Mortality  0.4165560 -0.06085861  -0.1140216 -0.09932185  0.1754959
##                  Infant.Mortality
## Fertility             0.41655603
## Agriculture          -0.06085861
## Examination          -0.11402160
## Education            -0.09932185
## Catholic              0.17549591
## Infant.Mortality      1.00000000
```

The cor(swiss) command gives the correlation matrix of correlation coefficients between all the variables in the data frame. Fertility has the highest correlation with the variable Education. It has a negative correlation of -0.6637, which means that as the percent education of draftees beyond primary school goes up, fertility will go down.

Chapter 8, Exercise 3

Run a multiple regression analysis on the swiss data with lm(), using Fertility as the dependent variable and Education and Agriculture as the predictors. (1 pt) Check the diagnostics. (1 pt) Say whether or not the overall R-squared was significant. If it was significant, report the value and say in your own words whether it seems like a strong result or not. Review the significance tests on the coefficients (B-weights). For each one that was significant, report its value and say in your own words whether it seems like a strong result or not. (1 pt)

```
fertilitylm <-lm(Fertility~Education+Agriculture, data= swiss)
summary(fertilitylm)

##
## Call:
## lm(formula = Fertility ~ Education + Agriculture, data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.3072  -6.6157  -0.9443   8.7028  20.5291
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 84.08005    5.78180  14.542  < 2e-16 ***
## Education   -0.96276    0.18906  -5.092  7.1e-06 ***
## Agriculture -0.06648    0.08005  -0.830    0.411
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.479 on 44 degrees of freedom
## Multiple R-squared:  0.4492, Adjusted R-squared:  0.4242
## F-statistic: 17.95 on 2 and 44 DF,  p-value: 2e-06
```

The overall adjusted R-squared is 0.4242 with a p-value of less than 0.05, which means the overall adjusted R-squared value is significant. It means that the variables account for 42.42% of variability in the Fertility variable. 42.42% does not seem like a strong enough result. On inspecting the significance of the independent variables individually, we can see that the weight for Education is significant, whereas for Agriculture the pr value is greater than 0.05, therefore the weight for Agriculture is insignificant. For Education and Intercept both significant values the weights are -0.96276 and 84.08005. It means that for every increase in percent education, the fertility is predicted to go down by 0.927 units. # Chapter 8, Exercise 4

Using the results of the analysis from Exercise 2, construct a prediction equation for Fertility using all three of the coefficients from the analysis (the intercept along with the two B-weights). If you observed a province with scores for Education of 8 and Agriculture of 35, what would you predict the Fertility rate to be (you can use your equation or the predict function)? Show your calculation and the resulting value of Fertility. (1 pt)

```
#The prediction equation would be:
# Fertility = 84.08005 - 0.96276 * Education - 0.06648 * Agriculture
Fertility <- 84.08005 - (0.96276 * 8) - (0.06648 * 35)
Fertility #Fertility is predicted to be 74.05117 for Education 8 and
Agriculture 35.
```

```
## [1] 74.05117
```

Chapter 8, Exercise 5

Run a multiple regression analysis on the swiss data with lmBF(), using Fertility as the dependent variable and Education and Agriculture as the predictors. (1 pt) Interpret the resulting Bayes factor in terms of the odds in favor of the alternative hypothesis (be sure to note the hypotheses being compared). Do these results strengthen or weaken your conclusion to exercise 2? (1 pt)

```
library(BayesFactor)
```

```
## Warning: package 'BayesFactor' was built under R version 4.0.4
```

```
## Loading required package: coda
```

```
## Loading required package: Matrix
```

```
## ************
## Welcome to BayesFactor 0.9.12-4.2. If you have questions, please contact
Richard Morey (richarddmorey@gmail.com).
##
## Type BFManual() to open the manual.
## ************
```

```
swiss.mcmc <-lmBF(Fertility~Education+Agriculture, data=swiss)
summary(swiss.mcmc)
```

```
## Bayes factor analysis
## --------------
## [1] Education + Agriculture : 8927.474 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

This shows that the odds are in favour of the null hypothesis, which means that the weights of the variables Education and Agriculture are not zero and they contribute in predicting the fertility variable. This strengthens the conclusion to the previous exercise which stated that the variability of fertility was 42.42% due to these factors.

Chapter 8, Exercise 6

Run lmBF() with the same model as for Exercise 4, but with the options posterior=TRUE and iterations=10000. (1 pt) Interpret the resulting information about the coefficients. (1 pt)

```
swiss.mcmc1 <-lmBF(Fertility~Education+Agriculture, data=swiss,
posterior=TRUE, iterations=10000)
summary(swiss.mcmc1)
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                 Mean       SD  Naive SE Time-series SE
## mu          70.15144  1.43230 0.0143230      0.0143230
## Education   -0.88663  0.19888 0.0019888      0.0021959
## Agriculture -0.06094  0.08031 0.0008031      0.0008031
## sig2        95.83204 22.03809 0.2203809      0.2483591
## g            0.80147  3.93760 0.0393760      0.0451146
##
## 2. Quantiles for each variable:
##
##                 2.5%     25%      50%        75%     97.5%
## mu          67.34618 69.1870 70.16771  71.132600  72.91073
## Education   -1.27278 -1.0198 -0.88877  -0.753073  -0.48840
## Agriculture -0.21788 -0.1140 -0.06116  -0.008231   0.09631
## sig2        62.72060 80.3723 92.97372 107.574885 146.51622
## g            0.06768  0.1779  0.32551   0.663132   3.78365
```

The mean values of the intercept and the coefficients are close to the values we got from the lm function. The HDIs for the intercept and Education do not include zero. However, the HDI for Agriculture contains zero, and from the lm model we saw that Agriculture was not significant. The Bayes model strengthens that conclusion.

Chapter 8, Exercise 7

Run install.packages() and library() for the "car" package. The car package is "companion to applied regression" rather than more data about automobiles. Read the help file for the vif() procedure and then look up more information online about how to interpret the results. Then write down in your own words a "rule of thumb" for interpreting vif. (1 pt)

```
library(car)
```

```
## Loading required package: carData
```

vif() function calculates the variance inflation factor, which means that it calculates the severity of multi-collinearity. It square root of vif indicates how much larger the standard error increases compared to if that variable had 0 correlation to the other predictor variables in the model. If the vif() is greater than 10, it means that the two factors have strong correlation with each other and it would be better to drop one of the factors.

Chapter 8, Exercise 8

Run vif() on the results of the model from Exercise 3. (1 pt) Interpret the results. Then run a model that predicts Fertility from all five of the predictors in swiss. Run vif() on those results and interpret what you find. (1 pt)

```
vif(fertilitylm)
```

```
##   Education Agriculture
##    1.692016    1.692016
```

Both the factors seem like they do not have a high multicollinearity

```
fertilitylm1 <- lm(Fertility~., data= swiss)
vif(fertilitylm1)
```

```
##      Agriculture      Examination        Education         Catholic
##         2.284129         3.675420         2.774943         1.937160
## Infant.Mortality
##         1.107542
```

None of the factors have numbers even higher than 5. It can be said that multi collinearity is highly unlikely to be an issue.