MIDS W203 Lab 4 Final Exam

Matthew Potts

April 26, 2016

Part 1. Multiple Choice

1. D
2. B
3. E
4. B
5. D
6. C
7. E
8. A

Part 2. Test Selection

9. A
10. D
11. B
12. B
13. A
14. B

Part 3. Data Analysis

15. OLS Regression

15a. The life_quality variable measures quality of life on a 5-point scale, where 1 = ex- cellent and 5 = poor. We would prefer, however, for higher numbers to be better. Reverse the scale so that 5 = excellent and 1 = poor. What is the mean quality of life in the sample?

```
# Load libraries, load Dating data.
library(ggplot2)
library(car)
dating <- read.csv("Dating.csv")

# Look at a summary of the life_quality variable
summary(dating$life_quality)
```

```
##          1          2          3          4          5 Don't know
##        407        618        762        335        110          8
##    Refused
##         12
```

```
# The variable is a factor (we want to change to numeric) and there are a few
# "Don't know" and "Refused".  We will convert them to NAs and reverse the scale
# so 1 is poor and 5 is excellent
dating$life_quality <- recode(dating$life_quality, "1=5; 2=4; 4=2; 5=1")
summary(dating$life_quality)
```

```
##          1          2          3          4          5 Don't know
##        110        335        762        618        407          8
##    Refused
##         12
```

```
dating$life_quality <- as.numeric(as.character(dating$life_quality))
```

```
## Warning: NAs introduced by coercion
```

```
# Calculate the mean of life_quality
mean(dating$life_quality, na.rm = TRUE)
```

```
## [1] 3.392921
```

1. The mean of quality of life is 3.392921.

15b. The years_in_relationship variable measures how long a respondent has spent in their current relationship. As you recode this variable, you may find that R con- verts each text string to the wrong number. For example, the string "0" may be converted to 2 or some other number (this happens because R's as.numeric func- tion returns factor levels if they're available). If this happens, convert the variable to a character string before converting it to a numeric vector, as in the following expression:

```
# Convert a factor to a numeric
dating$years_in_relationship  <- as.numeric(as.character(dating$years_in_relationship))
```

```
## Warning: NAs introduced by coercion
```

```
# Find weird values where age minus years_in_relationship is less than
# 10 years and convert to NA.
dating$id <- as.numeric(rownames(dating))
weird <- dating$id[dating$years_in_relationship >= dating$age]
for(i in weird) {
  dating$years_in_relationship[dating$id == i] <- NA
}
```

Notice that years_in_relationship equals zero for respondents that are not cur- rently in a relationship. You should leave these values in the dataset for the pur- poses of this lab. What is the mean of years_in_relationship in the sample?

```r
#Calculate the mean of years_in_relationship in the sample
mean(dating$years_in_relationship, na.rm = TRUE)
```

```
## [1] 13.40575
```

2. The mean of years_in_relationship is 13.40575.

15c. To run a nested regression in R, your first step will be to select just the rows in your dataset that have no missing values in your final OLS model. In this case, you will want just the rows that have non-missing values for life_quality, years_in_relationship, and use_internet. How many cases does this leave you with?

```r
# Let's look at the use_internet variable since we have not before.
summary(dating$use_internet)
```

```
##            Don't know         No    Refused        Yes
##       1122          2        190          2        936
```

```r
# There are a few blanks, Don't know, and Refused. Let's set those to NA and
# recode this to binary 0 = Now and 1 = Yes.
dating$use_internet <- recode(dating$use_internet, "'No' = 0; 'Yes' = 1; else = NA")
dating$use_internet <- as.numeric(as.character(dating$use_internet))
summary(dating$use_internet)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  1.0000  1.0000  0.8313  1.0000  1.0000    1126
```

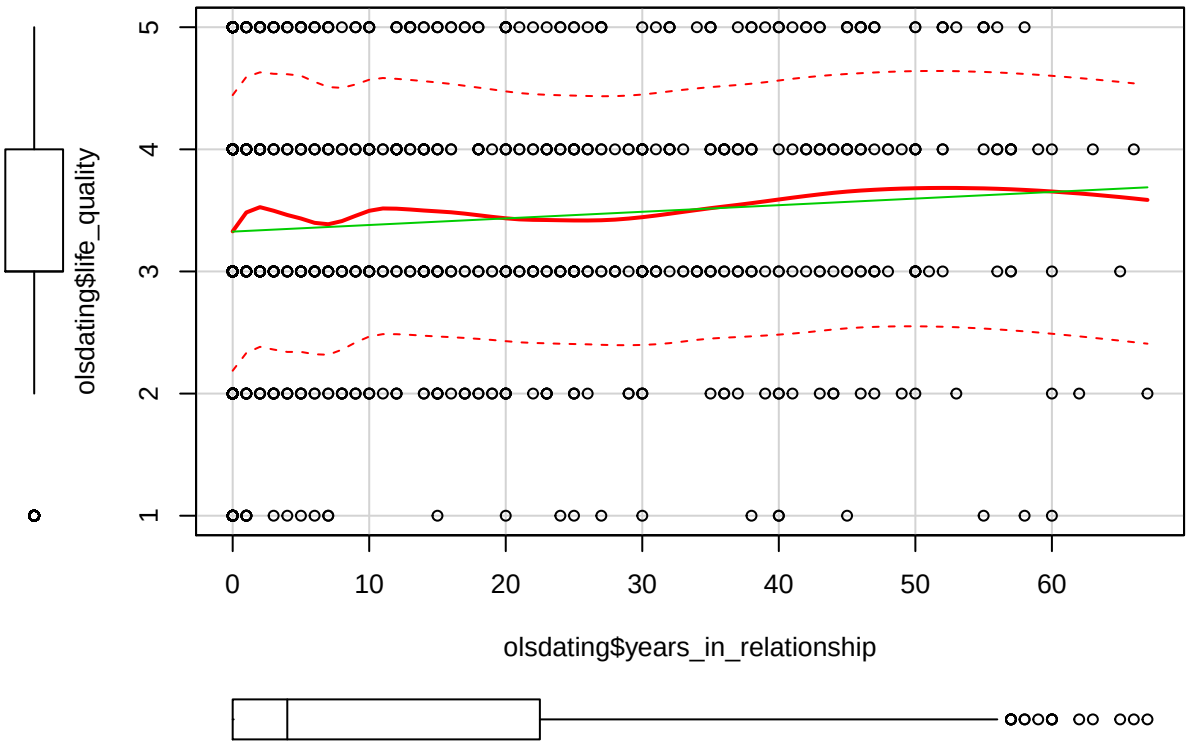
```r
# Create a dataset with the variables to want to regress and remove NAs.
olsdating <- subset(dating, select = c(life_quality, years_in_relationship, use_internet, age))
olsdating <- na.omit(olsdating)

# The number of complete cases.
nrow(olsdating)
```

```
## [1] 1088
```

3. This leaves 1,088 cases.

15d. Fit an OLS model to the data from the previous step that predicts life_quality as a linear function of years_in_relationship. What is the slope coefficient you get? Is it statistically significant? What about practically significant?

```r
# First, check the scatterplot to get a sense of the underlying
# relationship.
scatterplot(olsdating$years_in_relationship, olsdating$life_quality)
```




```
# Create the linear model, and look at a summary
model1 <- lm(life_quality ~ years_in_relationship, data = olsdating)
summary(model1)
```

```
##
## Call:
## lm(formula = life_quality ~ years_in_relationship, data = olsdating)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6509 -0.4886 -0.3263  0.6737  1.6737
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              3.326300   0.041892  79.402  < 2e-16 ***
## years_in_relationship 0.005411   0.002012   2.689  0.00728 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.093 on 1086 degrees of freedom
## Multiple R-squared:  0.006614,   Adjusted R-squared:  0.0057
## F-statistic: 7.231 on 1 and 1086 DF,  p-value: 0.007275
```

4. The slope coefficient for years_in_relationship is 0.005411. The p-value for years_in_relationship is statistically significant at an alpha of 0.05. However, the R-squared value is fairly low, indicating that years_in_relationship explains only 0.6614% in the variance of life_quality. Practically this number is very low.

15e. Now fit a second OLS model to the data. Keep life_quality as your dependent variable, but now use both years_in_relationship and use_internet as your explan- atory variables. What is the slope coefficient for use_internet? Is it statistically significant? What about practically significant?

```
# Create another linear model with use_internet as an added predictor, and
# look at a summary
model2 <- lm(life_quality ~ years_in_relationship + use_internet, data = olsdating)
summary(model2)
```

```
##
## Call:
## lm(formula = life_quality ~ years_in_relationship + use_internet,
##     data = olsdating)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.63083 -0.54154 -0.00025  0.60553  2.00763
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.992368   0.084482  35.420  < 2e-16 ***
## years_in_relationship 0.005252   0.001994   2.633  0.00857 **
## use_internet          0.402102   0.088596   4.539 6.29e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.083 on 1085 degrees of freedom
## Multiple R-squared:  0.02512,    Adjusted R-squared:  0.02333
## F-statistic: 13.98 on 2 and 1085 DF,  p-value: 1.012e-06
```

5. The slope coefficient for use_internet is 0.402102. The p-value for years_in_relationship is statistically significant at an alpha of 0.05. The R-squared value is still fairly low but much higher than the simpler model. The amount of variation being explained by the model increased from 0.66% to 2.5%. However, the practical significance is still low.

15f. Compute the F-ratio and associated p-value between your two regression models. Assess the improvement from your first model to your second.

```
# compare the model improvement with anova
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: life_quality ~ years_in_relationship
## Model 2: life_quality ~ years_in_relationship + use_internet
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1086 1297.4
## 2   1085 1273.2  1    24.173 20.599 6.294e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6. According to the test, the second model is a significant improvement over the first model. This makes sense given the difference in the amounts of variance in life_quality explained by the two models.

16. Logistic Regression

16a. What are the odds that a respondent in the sample has flirted online at some point (flirted_online)?

```
# Look at flirted_online and maybe do some cleaning
summary(dating$flirted_online)
```

```
##            Don't know         No    Refused        Yes
##        357          2       1496          6        391
```

```
dating$flirted_online <- recode(dating$flirted_online, "'No' = 0; 'Yes' = 1; else = NA")
dating$flirted_online <- as.numeric(as.character(dating$flirted_online))
summary(dating$flirted_online)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.0000  0.0000  0.2072  0.0000  1.0000     365
```

```
meanflirt <- mean(dating$flirted_online, na.rm = TRUE)

odds <- meanflirt / (1-meanflirt)
odds
```

```
## [1] 0.2613636
```

1. The odds that someone flirted online at some point is 0.26.

16b. Conduct a logistic regression to predict flirted_online as a function of where a respondent lives (usr). What Akaike Information Criterion (AIC) does your model have?

```
# Look at the usr variable to see if any cleaning needs to be done.
summary(dating$usr)
```

```
##             Rural Suburban    Urban
##        2      450     1037      763
```

```
dating$usr[dating$usr == " "] <- NA
dating$usr <- droplevels(dating$usr)
summary(dating$usr)
```

```
##    Rural Suburban    Urban     NA's
##      450     1037      763        2
```

```r
# Pull all of the complete cases for the logistic regression model.
logdating <- subset(dating, select = c(flirted_online, usr))
logdating <- na.omit(logdating)

# Run a bivariate logistic regression
model3 <- glm(flirted_online ~ usr, data = logdating, family = binomial())
summary(model3)
```

```
##
## Call:
## glm(formula = flirted_online ~ usr, family = binomial(), data = logdating)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7592  -0.7592  -0.6731  -0.5432   1.9934
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.8392     0.1554 -11.837  < 2e-16 ***
## usrSuburban   0.4697     0.1764   2.663  0.00774 **
## usrUrban      0.7427     0.1799   4.127 3.67e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1922.0  on 1884  degrees of freedom
## Residual deviance: 1903.4  on 1882  degrees of freedom
## AIC: 1909.4
##
## Number of Fisher Scoring iterations: 4
```

2. The model has an AIC of 1909.4.

16c. According to your model, how much bigger are the odds that an urban respondent has flirted online than the odds that a rural respondent has flirted online? Is this effect practically significant?

```r
exp(model3$coefficients)
```

```
## (Intercept) usrSuburban    usrUrban
##   0.1589404   1.5995763   2.1015464
```

3. The odds that an urban respondent has flirted is about 110% higher than or 2.1 times the odds for rural respondents. This seems to makes sense and has fairly large practical significance in terms of predicting which respondents have flirted online for not.