Final

Greg Ceccarelli

August 21, 2015

Part 1. Multiple Choice (32 points)

1. d
2. b
3. e
4. b
5. d
6. a
7. e
8. a

Part 2. Test Selection (24 points)

9. a
10. d
11. b
12. b
13. a
14. b

Part 3: Data Analysis (44 points)

15. OLS Regression

```
#set up working envrionment
##running on macbook air, set relevant directory
setwd('/Users/ceccarelli/MIDS/DATASCI_W203/Assignments/Labs/Final Exam/')
rm( list = ls() )

#Load relevant packages
#library(ggplot2)
library(car)
library(psych)
```

```
##
## Attaching package: 'psych'
##
## The following object is masked from 'package:car':
##
##     logit
```

```r
library(gmodels)
library(MASS)
library(plyr)

#load supplied R data file
dating <- read.csv("Dating.csv",header = TRUE)

#inspect life_quality variable
summary(dating$life_quality)
```

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

```r
levels(dating$life_quality)
```

```
## [1] "1"          "2"          "3"          "4"          "5"
## [6] "Don't know" "Refused"
```

```r
#wrapper function to easily recode factors
changelevels <- function(f, ...) {
    f <- as.factor(f)
    levels(f) <- list(...)
    f
}

dating$life_quality_cleaned <- changelevels(dating$life_quality, "5"=c("1"), "4"=c("2"), "3"=c("3"),"2"=

levels(dating$life_quality_cleaned)
```

```
## [1] "5"  "4"  "3"  "2"  "1"  "NA"
```

```r
##appears NA is a string in dating$life_quality_cleaned, recode to true NA
is.na(dating) <- dating=="NA"

# Can fix the remaining "NA" after fix by recreating the factor
dating$life_quality_cleaned <- factor(dating$life_quality_cleaned)

#check summary output
summary(dating$life_quality_cleaned)
```

```
##    5    4    3    2    1 NA's
##  407  618  762  335  110   20
```

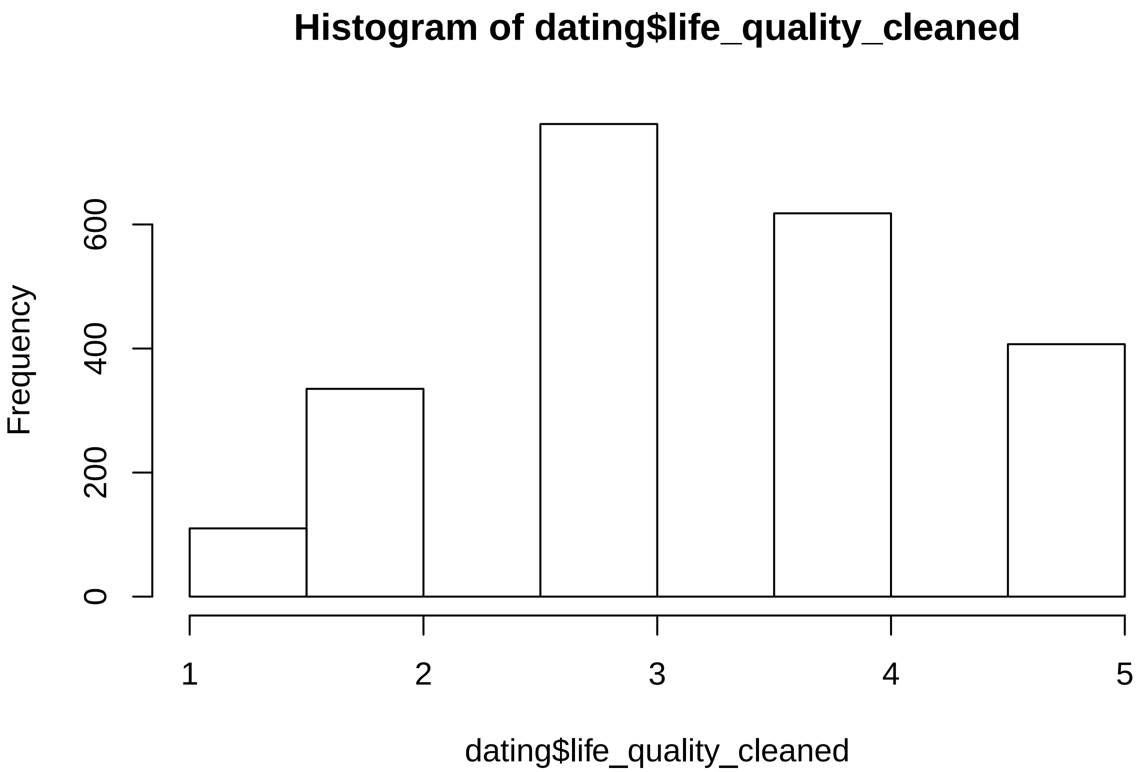
```r
#convert to numeric to use in regression using ali's method
dating$life_quality_cleaned <- as.numeric(as.character(dating$life_quality_cleaned))

#compute the mean for life quality to answer 15. a
mean(dating$life_quality_cleaned, na.rm =TRUE)
```

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

```
summary(dating$life_quality_cleaned)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   3.000   3.000   3.393   4.000   5.000      20
```

```
hist(dating$life_quality_cleaned)
```

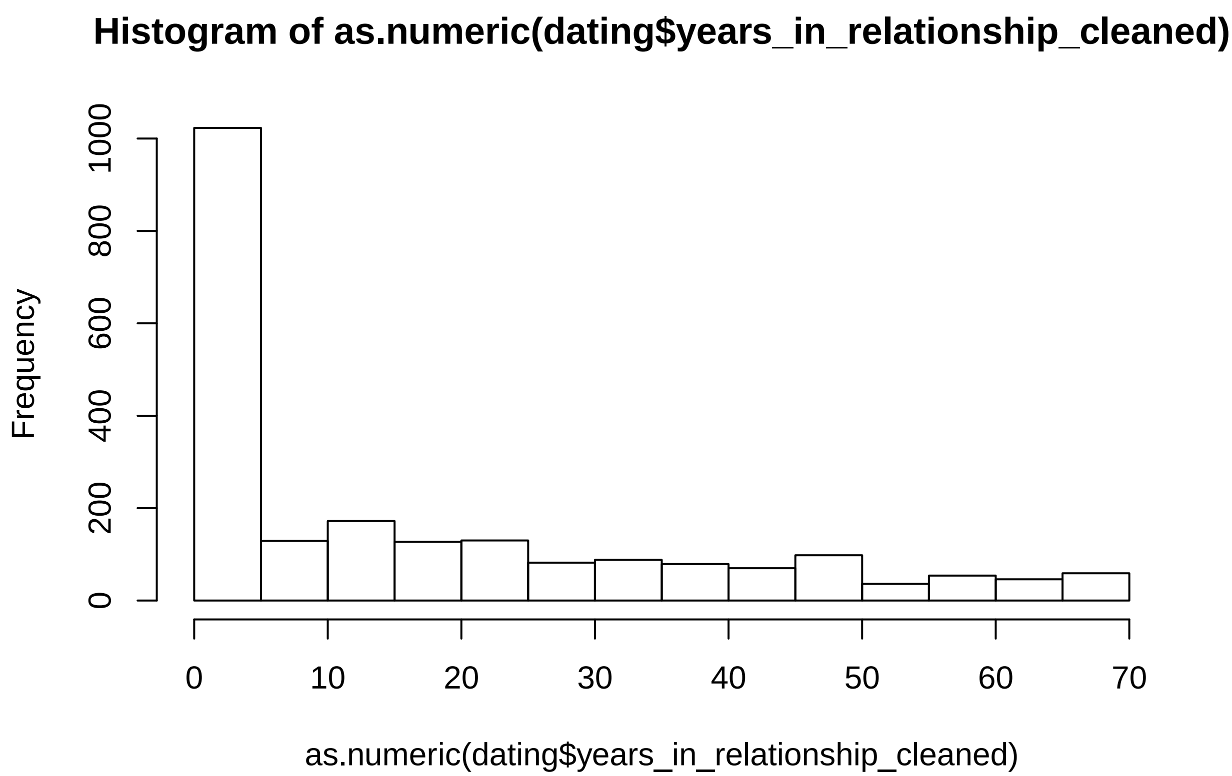



```
##review values for dating$years_in_relationship & age
counts_yrs = as.data.frame(table(dating$years_in_relationship))
#counts_yrs[with(counts_yrs,order(-Freq)),]

counts_age = as.data.frame(table(dating$age))
#counts_age[with(counts_age,order(-Freq)),]

dating$years_in_relationship_cleaned <- dating$years_in_relationship


#in this instance, use plyr instead to mapvalues
dating$years_in_relationship_cleaned <- mapvalues(dating$years_in_relationship_cleaned, from = c(" ","R

##appears NA is a string in dating$years_in_relationship_cleaned, recode to true NA
is.na(dating) <- dating=="NA"

# Can fix the remaining "NA" after fix by recreating the factor
dating$years_in_relationship_cleaned <- factor(dating$years_in_relationship_cleaned)
```

```r
#recheck counts to make sure nothing amiss
counts = as.data.frame(table(dating$years_in_relationship_cleaned))
#counts[with(counts,order(-Freq)),]

#check total counts, 59 NA's not included
sum(as.numeric(counts$Freq))
```

```
## [1] 2193
```

```r
hist(as.numeric(dating$years_in_relationship_cleaned))
```




```r
#update to be numeric following Alis Guidance
dating$years_in_relationship_cleaned <- as.numeric(as.character(dating$years_in_relationship_cleaned))

##finally, check for spurious values when years_in_relationship exceeds age
dating$years_in_relationship_cleaned.spurious <- as.numeric(as.character(dating$years_in_relationship_c

table(dating$years_in_relationship_cleaned.spurious)
```

```
##
## FALSE  TRUE
##  2192     1
```

```r
##One value where this occurs, recode to NA
#View(data.frame(dating[dating[,"years_in_relationship_cleaned.spurious"],]))

is.na(dating$years_in_relationship_cleaned) <- dating$years_in_relationship_cleaned.spurious==TRUE
```

```
dating$years_in_relationship_cleaned <- factor(dating$years_in_relationship_cleaned)

dating$years_in_relationship_cleaned <- as.numeric(as.character(dating$years_in_relationship_cleaned))
#compute the mean for life quality to answer 15. b
mean(dating$years_in_relationship_cleaned,  na.rm =TRUE)
```

```
## [1] 13.43887
```

```
##check use_internet
summary(dating$use_internet)
```

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

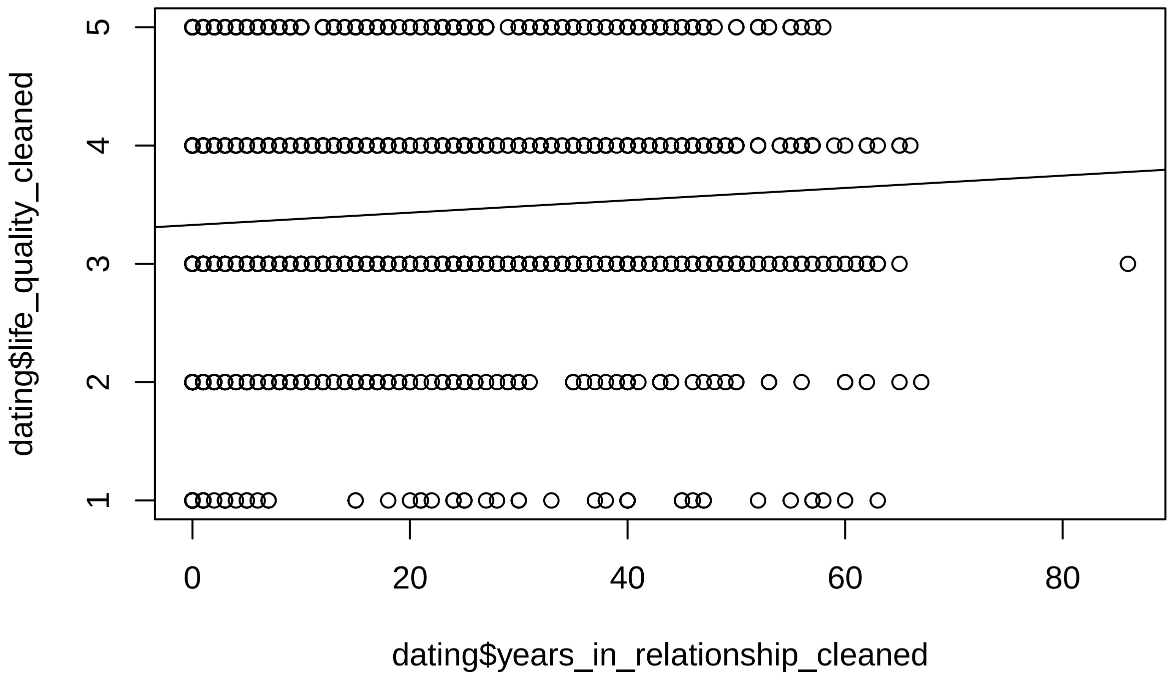
```
#remap values
dating$use_internet_cleaned <- mapvalues(dating$use_internet, from = c(" ","Don't know","Refused"), to=

## NA is a string in dating$use_internet_cleaned, recode to true NA
is.na(dating) <- dating=="NA"

# Can fix the remaining "NA" after fix by recreating the factor
dating$use_internet_cleaned<- factor(dating$use_internet_cleaned)

##check use_internet again
summary(dating$use_internet_cleaned)
```

```
##   No  Yes NA's
##  190  936 1126
```

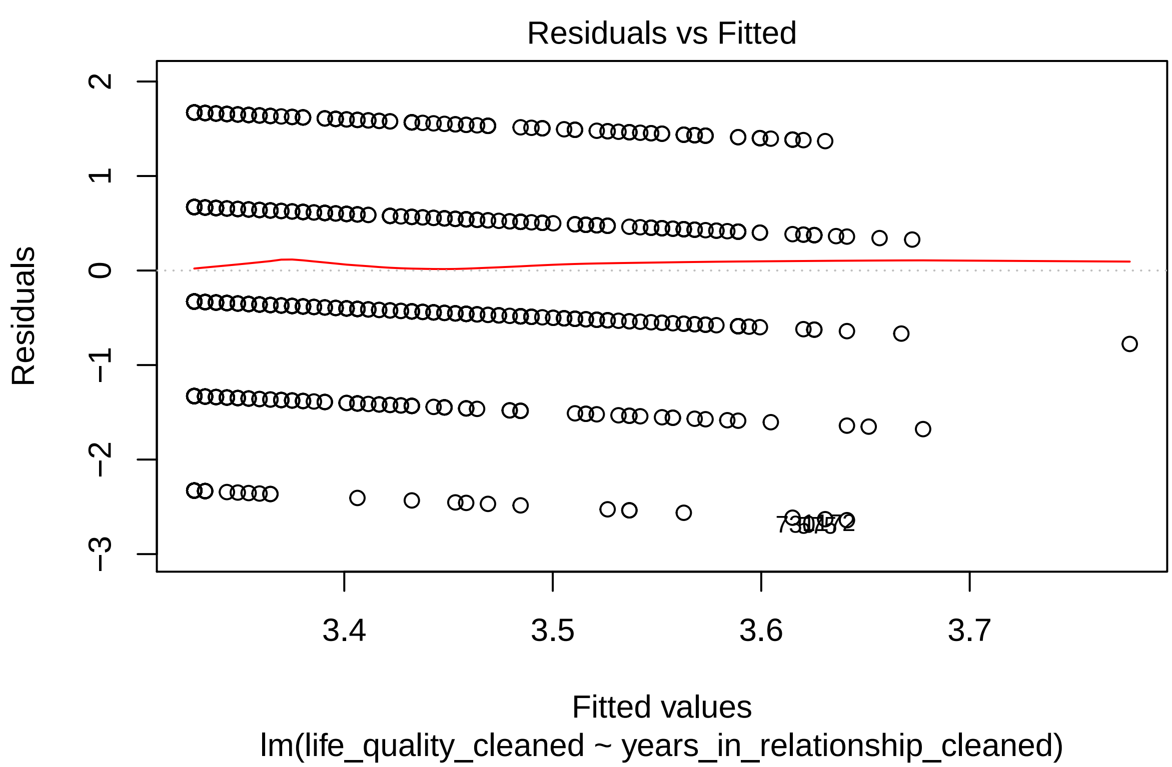
```
#compute lim_rows logical vector for the 3 variables in question
lim_rows = complete.cases(dating$life_quality_cleaned, dating$use_internet_cleaned, dating$years_in_rel

## just the complete cases, count the rows with nrow to answer 15. c
dating_lim = dating[lim_rows,]
nrow(dating_lim)
```

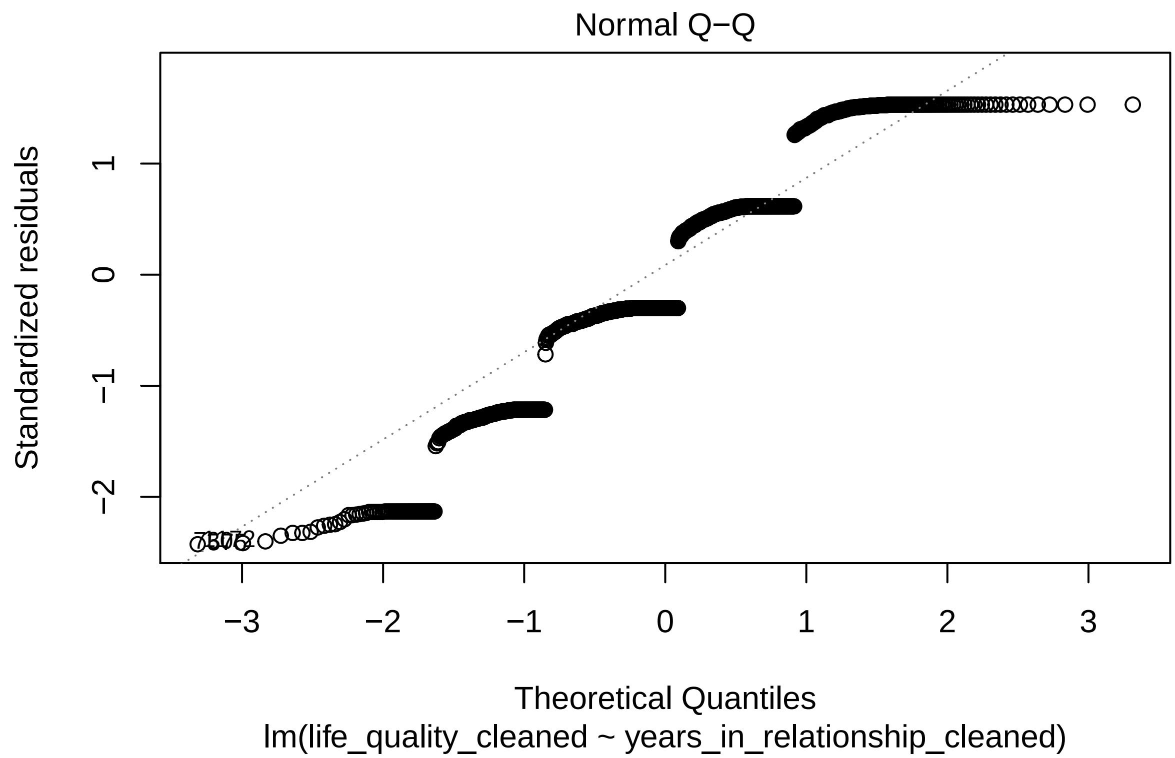
```
## [1] 1089
```

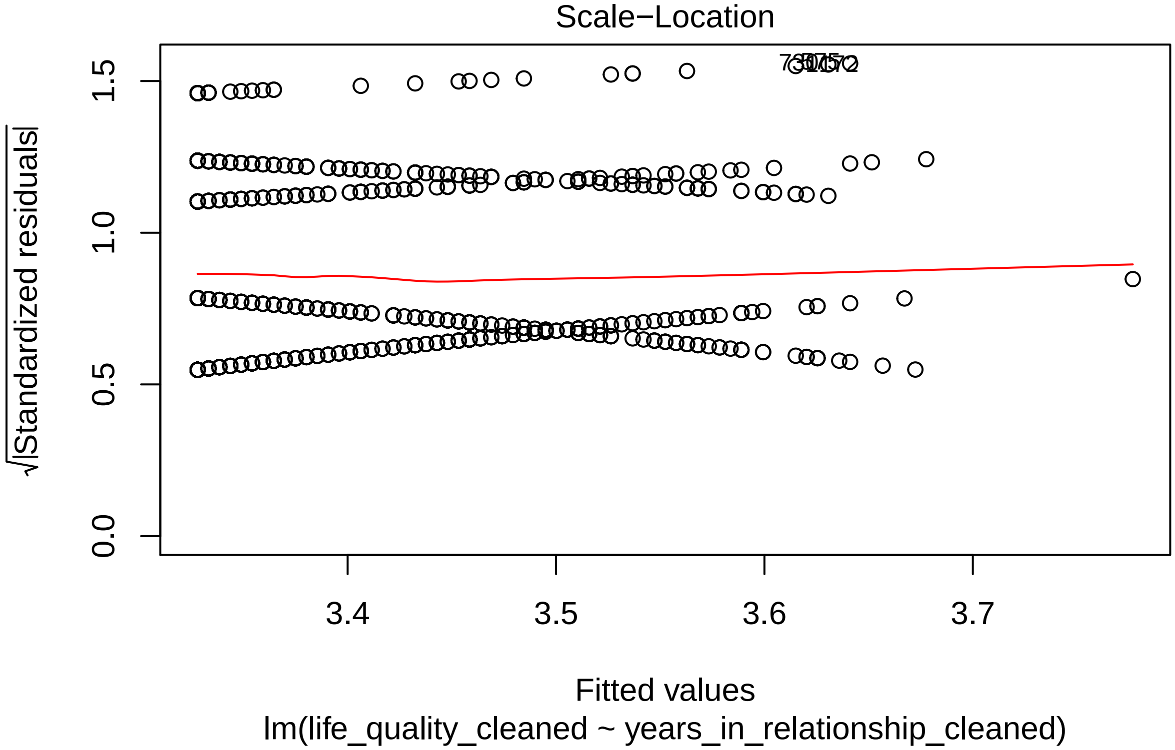
```
lmod1 <- lm(life_quality_cleaned ~ years_in_relationship_cleaned, dating_lim)
plot(dating$life_quality_cleaned ~ dating$years_in_relationship_cleaned)
abline(lmod1)
```

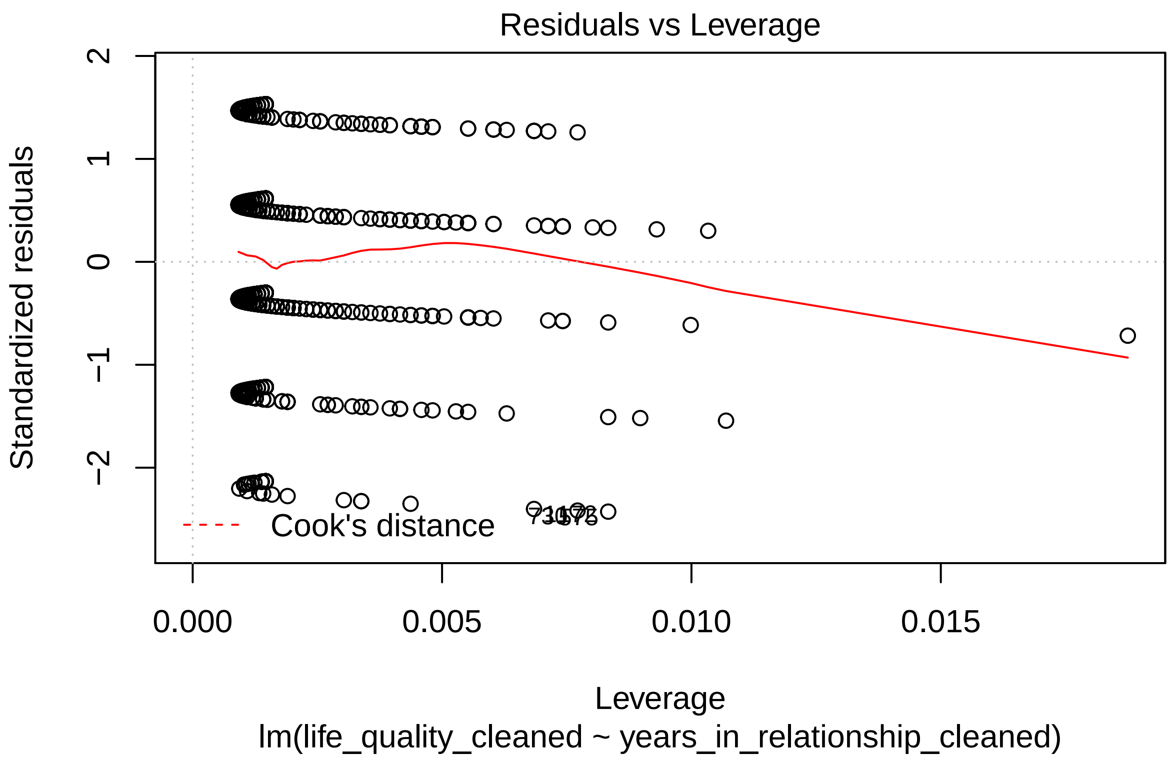



```r
##plot model to review
plot(lmod1)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(life_quality_cleaned ~ years_in_relationship_cleaned)
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
Fitted values
lm(life_quality_cleaned ~ years_in_relationship_cleaned)
0.0
0.5
1.0
1.5
3.4
3.5
3.6
3.7




```r
summary(lmod1)
```

```
##
## Call:
## lm(formula = life_quality_cleaned ~ years_in_relationship_cleaned,
##     data = dating_lim)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6411 -0.4846 -0.3280  0.6720  1.6720
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.328043   0.041812  79.595  < 2e-16 ***
## years_in_relationship_cleaned 0.005218   0.001994   2.617  0.00899 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.093 on 1087 degrees of freedom
## Multiple R-squared:  0.006262,   Adjusted R-squared:  0.005348
## F-statistic:  6.85 on 1 and 1087 DF,  p-value: 0.008987
```

```r
#first models coefficient
coef(lmod1)
```

```
##                   (Intercept) years_in_relationship_cleaned
##                   3.328042604                   0.005217756
```

```r
#fit multivariate ols linear model: life_quality outcome, use_internet
#and years_in relationship as predictors
lmod2 <- lm(life_quality_cleaned ~ years_in_relationship_cleaned + use_internet_cleaned, dating_lim)

summary(lmod2)
```

```
##
## Call:
## lm(formula = life_quality_cleaned ~ years_in_relationship_cleaned +
##     use_internet_cleaned, data = dating_lim)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62734 -0.53989 -0.00155  0.60413  2.00874
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.991259   0.084405  35.439  < 2e-16 ***
## years_in_relationship_cleaned 0.005144   0.001976   2.604  0.00935 **
## use_internet_cleanedYes       0.404609   0.088345   4.580 5.19e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.083 on 1086 degrees of freedom
## Multiple R-squared:  0.02509,    Adjusted R-squared:  0.0233
## F-statistic: 13.98 on 2 and 1086 DF,  p-value: 1.017e-06
```

```r
coef(lmod2)
```

```
##                   (Intercept) years_in_relationship_cleaned
##                   2.991258733                   0.005143746
##       use_internet_cleanedYes
##                   0.404609136
```

```r
# compare the model improvement with anova
anova(lmod1, lmod2)
```

```
## Analysis of Variance Table
##
## Model 1: life_quality_cleaned ~ years_in_relationship_cleaned
## Model 2: life_quality_cleaned ~ years_in_relationship_cleaned + use_internet_cleaned
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1087 1298.0
## 2   1086 1273.4  1    24.595 20.975 5.191e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# also check the AIC
AIC(lmod1)
```

```
## [1] 3287.664
```

```r
AIC(lmod2)
```

```
## [1] 3268.832
```

RESPONSE:

A. What is the mean quality of life in the sample? 3.392921

B. What is the mean of "years in relationship" in the sample? 13.43887

C. How many cases does this leave you with? 1089

D. Fit an OLS Model. What is the slope coefficient you get? Is it statistically significant? What about practically significant? The slope coefficient is 0.005217756, meaning that every unit increase in "years in relationship" increases "life quality" by that amount. It is statistically significant but given the extremely small R squared, not practically significant given the very low percent of variance explained.

E. Fit a second OLS model. What is the slope coefficient for use_interne ? Is it statistically significant? What about practically significant? The slope coefficient for use internet is 0.404609136, meaning that internet users on average had a rating for "life quality" that much higher. In this case it is higher statistically significant with a p value of 5.19e-06.

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

Using ANOVA, The F-ratio is 20.975 and is highly statisitically significant. Reduced AIC between model 1 and 2 indicate that we've improved predictive ability and that we have a higher quality model (with improved R squared)

16. Logistic Regression

```r
summary(dating$flirted_online)
```

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

```r
#remap values
dating$flirted_online_cleaned <- mapvalues(dating$flirted_online, from = c(" ","Don't know","Refused"),

## NA is a string in dating$flirted_online_cleaned, recode to true NA
is.na(dating) <- dating=="NA"

# Can fix the remaining "NA" after fix by recreating the factor
dating$flirted_online_cleaned<- factor(dating$flirted_online_cleaned)

summary(dating$flirted_online_cleaned)
```

```
##   No  Yes NA's
## 1496  391  365
```

```r
#create dummy variables
contrasts(dating$flirted_online_cleaned)<-contr.treatment(2,base = 1)

levels(dating$flirted_online_cleaned)
```

```
## [1] "No"  "Yes"
```

```r
summary(dating$flirted_online_cleaned)
```

```
##   No  Yes NA's
## 1496  391  365
```

```r
dating$flirted_online_cleaned_recode <- ifelse(dating$flirted_online_cleaned=="Yes", 1, 0)

#t=table(dating$flirted_online_cleaned)
#D = data.frame( matrix(t,ncol=2))
#D$num.obs = D[,1]+D[,2]
#D$srate = D[,2]/D$num.obs

##check updated values
summary(dating$flirted_online_cleaned_recode)
```

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

```r
##compute mean
mean(dating$flirted_online_cleaned_recode, na.rm=TRUE)
```

```
## [1] 0.2072072
```

```r
#compute odds for question 16. a
mean(dating$flirted_online_cleaned_recode, na.rm=TRUE)/(1-mean(dating$flirted_online_cleaned_recode, na
```

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

```r
#review usr variable
summary(dating$usr)
```

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

```r
dating$usr_cleaned <- mapvalues(dating$usr, from = c(" "), to= c("NA"))

## NA is a string in dating$flirted_online_cleaned, recode to true NA
is.na(dating) <- dating=="NA"

# Can fix the remaining "NA" after fix by recreating the factor
dating$usr_cleaned<- factor(dating$usr_cleaned)

summary(dating$usr_cleaned)
```

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

```r
# Only incorporate complete cases
#compute lim_rows logical vector for the 3 variables in question
lim_rows_glm = complete.cases(dating$flirted_online_cleaned_recode, dating$usr_cleaned)

## just the complete cases
dating_lim_glm = dating[lim_rows_glm,]
summary(dating_lim_glm$usr_cleaned)
```

```
##    Rural Suburban    Urban
##      350      888      647
```

```r
#create dummy variables, make surburban base case
#contrasts(dating$usr_cleaned)<-contr.treatment(3,base = 1)
#dating$usr_cleaned

## Begin with a bivariate logistic regression given
model1 = glm(flirted_online_cleaned_recode ~ usr_cleaned, data=dating_lim_glm, family=binomial())

#grab AIC criterion
summary(model1)
```

```
##
## Call:
## glm(formula = flirted_online_cleaned_recode ~ usr_cleaned, family = binomial(),
##     data = dating_lim_glm)
##
## 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 ***
## usr_cleanedSuburban   0.4697     0.1764   2.663  0.00774 **
## usr_cleanedUrban      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
```

```r
exp(coef(model1))
```

```
##         (Intercept) usr_cleanedSuburban    usr_cleanedUrban
##           0.1589404           1.5995763           2.1015464
```

```r
#output odds, manually create odds ratio in answer
exp(coef(model1))
```

```
##         (Intercept) usr_cleanedSuburban    usr_cleanedUrban
##           0.1589404           1.5995763           2.1015464
```

```r
##odds ratio
2.1015464/0.1589404
```

```
## [1] 13.22223
```

RESPONSE:

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

The percentage of folks who flirted online is 0.2072072, thus, the odds that someone would have flirted online is 0.2072072/(1-0.2072072) = .26136. On the otherhand, the odds that someone did not flirt online is (1-0.2072072)/0.2072072 = 3.82

B. Conduct a logistic regression to predict flirted_online as a function of where a respondent lives (usr). What Akaike Information Criterion (AI) does your modelhave?

1909.4

C. how much bigger are the odds that an urban respondent thas flirted online than the odds that a rural respondent has flirted online? Is this effect practically significant?

Based on the output, the odds are: usr_cleanedUrban: 2.1015464 (Intercept)->Rural 0.1589404 Odds Ratio: 13.22223, that is, the odds of an urban respondant having flirted online are 13 times the odds of a rural respondant. While there is no "accepted test" to calculate practical significance in this scenaro, this result certainly would seem to be, especially given the fact that each of coefficie t values are statistically significant.