Audrey Leung

Part 1: Multiple Choice (28 points)

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

Part 2: Test Selection (24 points)

1. **A**
   Categorical + Categorical = Chi-Squared
2. **D**
   Polytomous + Metric = ANOVA
3. **B**
   Binary Categorical (assume 'flirted_online' is binary) + Metric = t-test
4. **B**
   Polytomous + Metric = ANOVA
5. **A**
   Continuous nominal + Metric = Pearson
6. **B**
   Wilcoxon Rank-Sum Test – for small sample size < 30

Part 3: Short Answer (8 points)

1. What are the null and alternative hypotheses for the test in Question 2.1 (relating marital_status to use_reddit)?

The null hypothesis is that there is no relationship between the two variables, marital status and use of reddit; they are independent.The alternative hypothesis is that the two variables, marital status and use of reddit, are not independent.

2. Imagine that you have a dataset that has standardized GRE test scores for every US state. A colleague tells you that he conducted a single t-test and found that California has a statistically significantly higher test score than New York. What statistical test would you use to evaluate whether test scores differ by state? How would you evaluate all of the individual (pairwise) differences between states?

I would use ANOVA to evaluate whether test scores differ by state. I would evaluate all the pairwise differences between states by using an independent samples t-test.

Part 4: Data Analysis (40 points) 1. OLS Regression

a. **Data preparation**: In this data, missing values are very obvious because they are strings such as "Refused". Properly deal with any missing values for *life_quality*, *years_in_relationship*, and *use_internet*.

I assumed that "Refused", "Don't know", and " " were missing values.

b. After you reverse the coding, what is the mean quality of life in the sample?

```
> mean(Dating$life_qual_num, na.rm=TRUE)
[1] 3.392921
```

The mean quality of life in the sample is 3.39.

c. What is the mean of *years_in_relationship* in the sample?

```
> mean(Dating$yrsinrel_num, na.rm=TRUE)
[1] 13.47697
```

The sample mean number of years in a relationship is 13.48**.**

d. 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?

```
> sum(lim_rows)
[1] 1090
```

This leaves 1,090 cases of rows with non-missing values.

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

```
> summary(dating_model1)

Call:
lm(formula = life_qual_num ~ yrsinrel_num, data = Dating_lim)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6296 -0.4799 -0.3302  0.6698  1.6698

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.33022    0.04170  79.853   <2e-16 ***
yrsinrel_num  0.00499    0.00197   2.533   0.0115 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.093 on 1088 degrees of freedom
Multiple R-squared:  0.005861,  Adjusted R-squared:  0.004947
F-statistic: 6.414 on 1 and 1088 DF,  p-value: 0.01146


> sqrt(summary(dating_model1)$r.squared)
[1] 0.07655665
```

Model: life quality = 3.33 + (.005 x years in a relationship)

The slope coefficient (b_1) for years in a relationship is 0.005. This means that for a one unit increase in years in relationship (predictor variable), then our model predicts a 0.005 unit increase in quality of life (outcome variable). Since the p-value is 0.0115, < 0.05, the slope coefficient or the relationship between life quality and years in a relationship is statistically

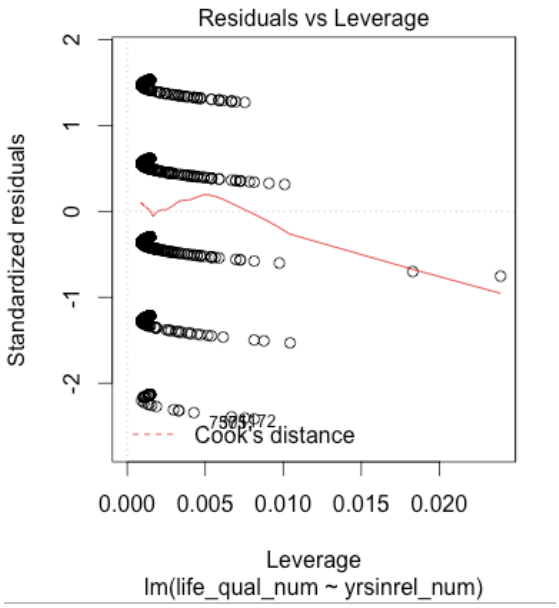
significant. This means that the probability of these t-values (or larger) occurring if the values of b in the population were 0 is less than 0.05. Therefore, this model is (weakly) statistically significant at the 0.05 level.

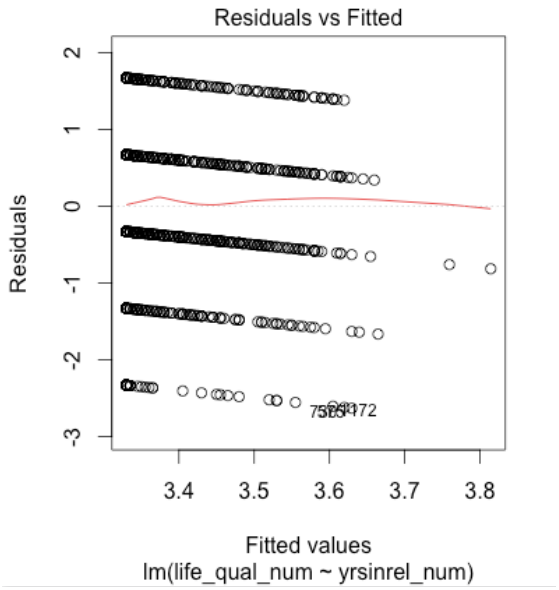
The F-ratio is 6.414, which is not significant at $p = 0.01$; $p > 0.05$. This tells us that there is less than a 1% chance that an F-ratio this large would happen if the null hypothesis were true. Therefore, we cannot conclude that the regression model results in significantly better prediction of life quality than if we used the mean value of life quality; the regression model overall does not predict life quality very well.

The Mutliple R^2 value of 0.00586 tells us that years in relationship can account for .586% of the variation in quality of life. 99.5% of the variation cannot be explained by years in a relationship alone. There must be other variables that also have influence, and the relationship is not practically significant.

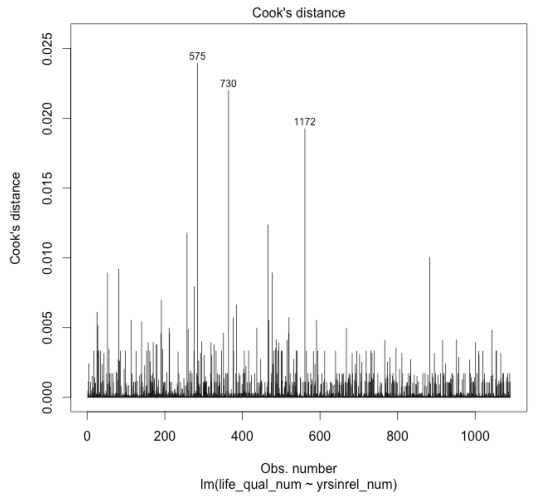
Since the dependent variable is ordinal, I could also calculate Spearman's rho to assess practical significance. Spearman's rho is 0.082, which is < 0.1 and shows very low effect size or practical significance[1]. Since the p-value is <0.05, the null hypothesis is rejected and the true rho is statistically not equal to zero. Thus, rank on years in a relationship tells us little about rank on life quality.

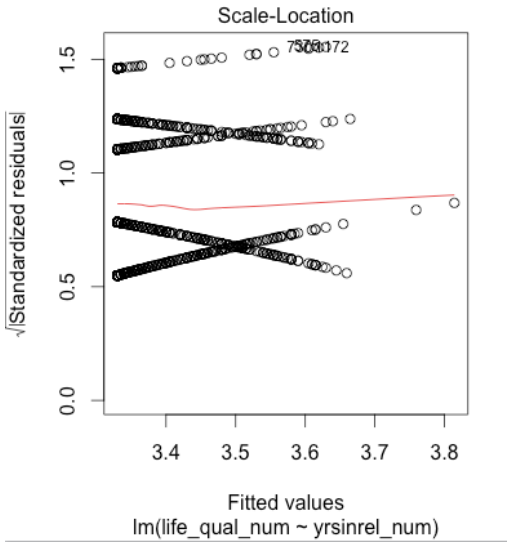
f. Evaluate the regression diagnostics from your model.

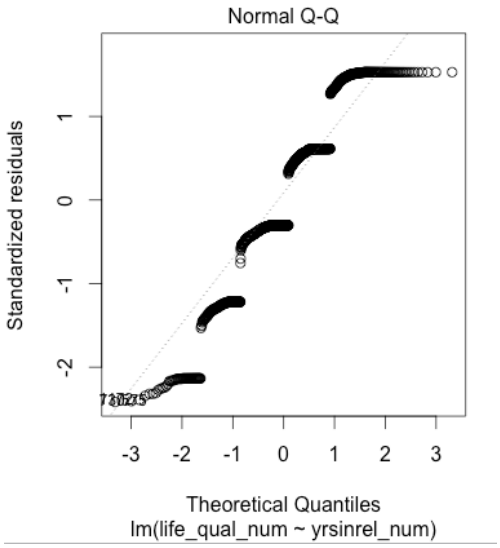


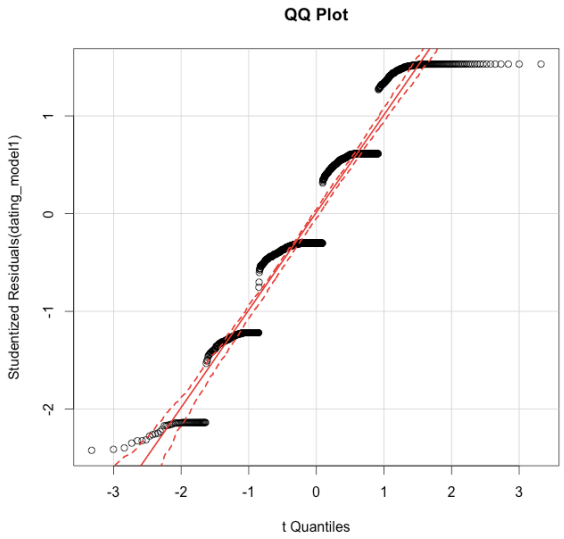


[1] Rules of thumb on magnitude of effect sizes http://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/effectSize









Since the graph for residuals vs. fitted values shows the residuals as fairly randomly distributed around the horizontal line representing a residual error of zero, this is indicative that the assumptions of linearity, randomness, and homoscedasticity have been met.

The residuals vs. leverage plot is a measure of each point's influence on the regression model. Cook's distance is superimposed, which is another measure of the importance of each observation to the regression. Cook's distance is a function of leverage and residuals, and it is a measure of how much influence a single observation has on the model. Smaller distances means that removing the observation has little affect on the regression results. Distances larger than 1 are suspicious and suggest the presence of a possible outlier or a poor model.

The Cook's distance plot showed three observations that were outliers. However, there were no studentized residuals with Bonferroni $p < 0.05$.

There is no obvious trend in the scale-location plot. The scale-location plot shows the square

root of the standardized residuals (sort of a square root of relative error) as a function of the fitted values.

The Q-Q plot shows that the residual errors in the model are random, normally distributed variables since they follow the diagonal line. However, at the extremes, the points become more distant from the line, which could indicate a skewed distribution.

The Durbin-Watson value is 1.99, which is so close to 2 that the assumption of independence has probably been met.

The assumption of no multicollinearity is met, because there is only one independent variable.

g. 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 explanatory variables. What is the slope coefficient for use_internet? Is it statistically significant? What about practically significant?

```
> summary(dating_model2)

Call:
lm(formula = life_qual_num ~ yrsinrel_num + use_internet_num,
    data = Dating_lim)

Residuals:
     Min       1Q   Median       3Q      Max
-2.61852 -0.53523 -0.01881  0.60195  2.00568

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.994316   0.084309  35.516  < 2e-16 ***
yrsinrel_num      0.004899   0.001952   2.509   0.0122 *
use_internet_numYes 0.403738   0.088325   4.571 5.41e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.083 on 1087 degrees of freedom
Multiple R-squared:  0.02461,   Adjusted R-squared:  0.02282
F-statistic: 13.71 on 2 and 1087 DF,  p-value: 1.314e-06
```

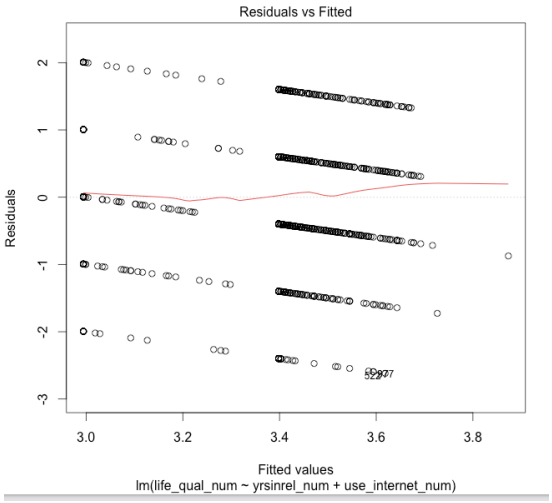
Model: life quality = 2.99 + (.0049 x years in a relationship) + (0.404 x use internet)

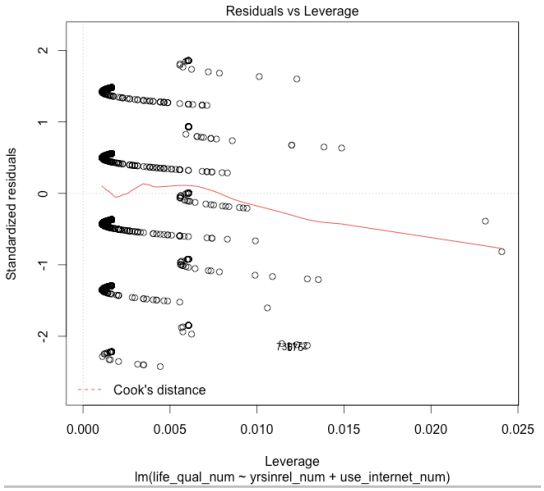
The slope coefficient(b_2) for use_internet is 0.404. For a one unit increase in use_internet (predictor variable), then our model predicts a 0.404 unit increase in quality of life (outcome variable). Since the p-value is 5.41e-06 and $p < 0.05$, the slope coefficient or the relationship between life quality and using internet is statistically significant. This means that the probability of these t-values (or larger) occurring if the values of b in the population were 0 is less than 0.05., This model is highly statistically significant at the 0.05 level.

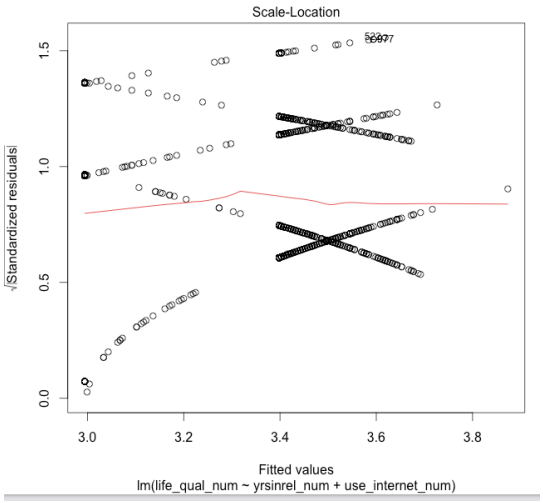
The Multiple R^2 value of 0.02461 tells us that using the internet can account for 2.46% of the variation in quality of life. 97.54% of the variation cannot be explained by using the internet and years in a relationship alone. The R^2 around 0.02 means that the results have a small effect or

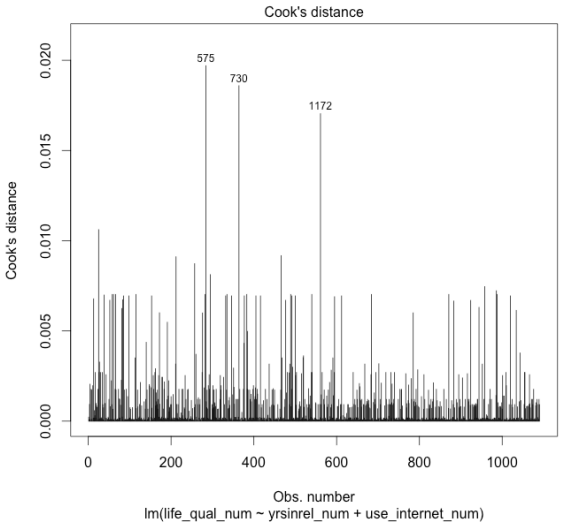
small practical significance[2]. Use of internet and years in a relationship has a small association with the quality of life.

h. Evaluate the regression diagnostics from your second model.

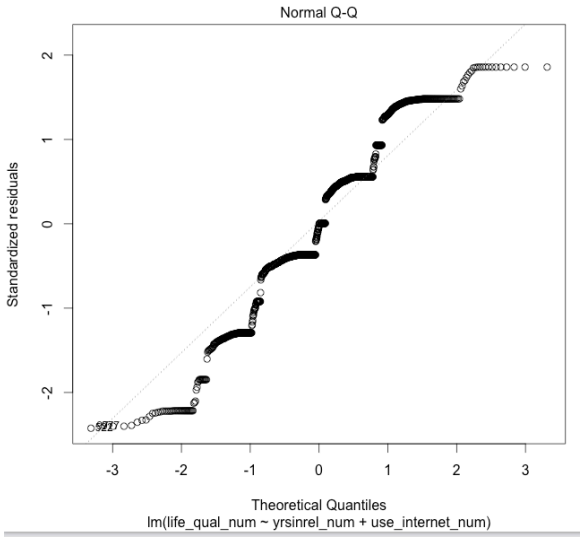


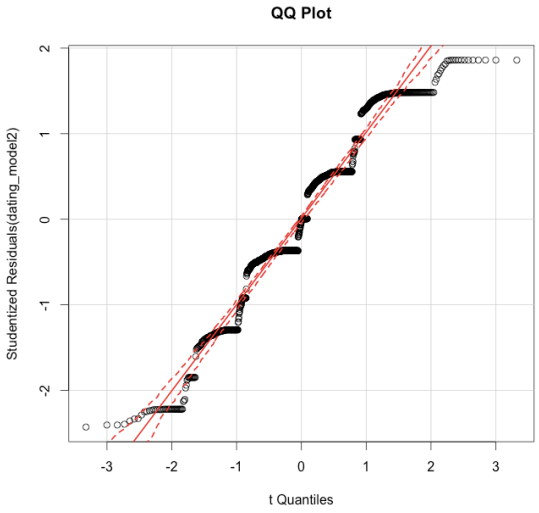






[2] Rules of thumb on magnitude of effect sizes http://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/effectSize





Since the graph for residuals vs. fitted values shows the residuals as fairly randomly distributed around the horizontal line representing a residual error of zero, this is indicative that the assumptions of linearity, randomness, and homoscedasticity have been met.

The residuals vs. leverage plot is a measure of each point's influence on the regression model. Cook's distance is superimposed, which is another measure of the importance of each observation to the regression. Cook's distance is a function of leverage and residuals, and it is a measure of how much influence a single observation has on the model. Smaller distances means that removing the observation has little affect on the regression results. Distances larger than 1 are suspicious and suggest the presence of a possible outlier or a poor model.

The Cook's distance plot showed three observations that were outliers. However, there were no studentized residuals with Bonferroni $p < 0.05$.

There is no obvious trend in the scale-location plot. The scale-location plot shows the square root of the standardized residuals (sort of a square root of relative error) as a function of the fitted values.

The Q-Q plot shows that the residual errors in the model are random, normally distributed variables since they follow the diagonal line. However, at the extremes, the points become more distant from the line, which could indicate a skewed distribution.

The Durbin-Watson value is 1.993, which is so close to 2 that the assumption of independence has probably been met.

The assumption that there is no multicollinearity is met, because since the average VIF is not substantially greater than 1, the regression is probably not biased.

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

```
> anova(dating_model1, dating_model2)
Analysis of Variance Table

Model 1: life_qual_num ~ yrsinrel_num
Model 2: life_qual_num ~ yrsinrel_num + use_internet_num
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1   1088 1298.7
2   1087 1274.2  1    24.493 20.894 5.41e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The F-ratio between the two regression models is 20.89. The p value is 5.41e-06, which is less than 0.05, so we can say that the second OLS regression significantly improved the fit of the model to the data compared to the first regression model. $F(1, 1087) = 20.89$, $p<0.001$.