271B
Fall 2020
Final Exam

There are 3 parts to this exam (100 points total).

- Part 1: Multiple Choice (35 points)
- Part 2: Test Selection (25 points)
- Part 3: Data Analysis (40 points)

You must work individually on all Parts (I, II and III) of this exam. This exam is open-book, but you may not discuss your answers with anyone else.

As done with the other labs, please write up a lab report with your answers and upload the report, Rmd script and html file to bCourses. Please upload the following:

- Lab report: FinalExam_YourLastName.pdf
- Your Rmd script: FinalExam_YourLastName.R
- Html output from Rmd file, named FinalExam_YourLastName.html

The exam is due by **11:59pm on Tuesday, December 15th**. We will *not* accept late exams.

Part 1: Multiple Choice (35 points)

For the following questions, please choose the best answer and **bold** the correct answer in your response. Only one correct option per question.

For Q1, 2 & 3:

Suppose you are a dog and you want to run an ordinary least squares regression to predict how many treats you will be given today. You have studied the human carefully and want to try modeling his behavior more formally. Your dependent variable, Y, = number of treats you will get today. You hypothesize that the number of times you fetch your favorite stick (X1) and the number of times you cuddle with the human (X2) are associated with how many treats you will get. Your hypothesized model looks like this:

$$Y = b0 + b1 * X1 + b2 * X2 + \varepsilon$$

1. What statistical test should you use to test the significance of your overall model?
    a. T-tests with correction for multiple comparisons
    b. T-test
    c. Ordinary Least Squares
    d. ANOVA
    e. Chi square
    f. Pearson Correlation

2. What is the null hypothesis for this test?
    a. At least one coefficient is equal to zero.
    b. b0 = 0
    c. b1 = 0
    d. b2 = 0
    e. b0 = b1 = b2
    f. All coefficients for each independent variable equal zero.
    g. Except for the y-intercept, at least one coefficient is not equal to zero.

3. Suppose you rejected the null hypothesis for your test in Q1. What can you conclude?
    a. There are statistically significant relationships between each independent variable and the dependent variable.
    b. b0 is not equal to 0
    c. b1 is not equal to 0
    d. b2 is not equal to 0
    e. b, c, and d
    f. c and d
    g. At least one of statements b, c, or d is true.
    h. None of the above

4. You are analyzing data from an experiment to see if "video configuration" (measured as three nominal types: X, Y, and Z) is significantly associated with "zoom meeting productivity" (measured as a unidimensional metric scale from 0-50). What is the appropriate order (from first to last) that you should use to conduct your analyses of statistical significance?
    a. ANOVA between configurations and productivity; if significant, follow with a Chi-square between configurations and productivity, and then check the standardized residuals in each cell to see which ones are significant.
    b. Chi-square between configurations and productivity; if significant, follow with an ANOVA between configurations and productivity, and then t-tests between each configuration and productivity.
    c. Individual t-tests between each configuration and productivity; if significant, follow with an ANOVA between configurations and productivity, and then Bonferroni corrections for the f-test.

    d. **ANOVA between configurations and productivity; if significant, follow with Bonferroni-corrected t-tests between each configuration and productivity.**

5. Imagine that you select a random sample of tech workers (Sample A: N=200) from the San Francisco bay area and find that their mean number of blockchain references per day is 20. The population mean of blockchain references among bay area tech workers is known to be 12 references per day with a standard deviation of 2.9. Now, imagine that you select another random sample of tech workers (Sample B: N = 40) from the same population. Compared to Sample A, would you be equally likely, more likely, or less likely to find a sample mean of 20 in Sample B?
    a. **More likely**
    b. Less likely
    c. Equally likely
    d. Cannot be determined because each random sample is different

6. Which of the following would provide the best *operational definition* of “rabbit happiness”?
    a. A unidimensional scale of “rabbit happiness”
    b. A unidimensional scale of “rabbit happiness” with sufficient precision, validity, and reliability
    c. An expert assessment of “rabbit happiness”
    d. A statement that describes precisely how to determine statistical and practical significance of the concept of “rabbit happiness”
    e. **A statement that describes precisely what “rabbit happiness” is and exactly how it will be measured.**
    f. None of the above.

7. Suppose you have a random sample of new Reddit.com users (N=500). What is the most appropriate way to tell if there is a difference in the average number of articles that that a person reads in their first three months versus their second three months?
    a. OLS Regression
    b. Bonferroni-corrected t-test comparisons
    c. Chi-square test of independence
    d. Independent samples t-test
    e. **Paired samples t-test**
    f. Effect size calculation
    g. None of the above

Part 2: Test Selection (25 points)

The Pew Internet and American Life Project collects survey data on a variety of topics related to online behavior. You will be working with a subset of data (Dating.csv) from a 2013 survey on online dating.

In this section, there are several questions that apply to the "Dating.csv" dataset included with this exam.

Recall that surveys are generally weighted in order to compensate for over- or under-representation of subgroups. These weights appear in the "weight" and "standwt" columns of the Pew dataset. For the sake of simplicity, however, we will not need to use these weight values because we will not be conducting any generalizable analyses.

For each question below, *select the most appropriate statistical procedure from the provided choices*. You should examine the variables in Dating.csv and the provided codebook (Abridged Codebook.pdf) to determine whether you should treat each variable as metric, binary or category variables, and to determine which test to run.

You should assume that you would appropriately deal with missing values, if applicable. **For the purpose of these tests, answers like "Refused" and "Don't Know" should be considered as missing/NA.**

Please note that in the questions below, we will treat ordinal Likert-style questions such as "life quality" as a metric variable.

For the following questions, please **bold** the correct letter in your response. Only one correct option per question.

1. Is an individual's age (*age*) related to the region of the United States in which he or she lives (*region*)?
    a. OLS Regression
    b. Independent Samples t-test
    c. ANOVA
    d. F-test
    e. Chi square test of independence

2. Is an individual's race (*race*) related to the state (*state*) in which he or she lives?
    a. ANOVA
    b. Chi square test of independence
    c. OLS Regression
    d. Independent Samples t-test
    e. Pearson Correlation

3. Are individuals who have used a dating site older than individuals who have not used a dating site (*used_dating_site*, *age*)?
    a. Pearson Correlation
    b. Independent Samples t-test
    c. ANOVA
    d. Chi square test of independence
    e. Levene's Test

4. Is lower perceived quality of life (*life_quality, treated as a metric variable*) related to age (*age*)?
    a. Independent Samples T-test
    b. ANOVA
    c. Pearson Correlation
    d. Wilcoxon Signed-Rank test
    e. Dependent Samples t-test

5. Do 30-year-old men (*age*, *sex*) have fewer children (children0_5 + children6_11 + children12_17) than 30-year-old women (*age*, *sex*)?
    a. ANOVA
    b. Dependent Samples t-test
    c. Wilcoxon Rank-Sum Test
    d. Independent Samples t-test
    e. Pearson Correlation

Part 3: Data Analysis (40 points)
In this section, you will work on the Dating.csv dataset (bCourses -> Files -> Labs -> Final Exam -> Dating.csv). An abridged codebook is provided (bCourses -> Files -> Labs -> Final Exam -> Abridged Codebook.pdf).

Depending on the options you've specified when you read in your csv file, your variables may be factor, numeric or strings, etc. In each of the questions below, check that your variable is in the correct format with class(variable) before running your analysis.

NOTE: In this section, values like "Refused" and "Don't Know" should be set to missing (NA).

In your lab report, you should clearly answer each **bold question** below:

a. The *life_quality* variable measures one's perception of quality of life for themselves and their family on a 5-point scale, where 1 = excellent and 5 = poor. In our regression, we're going to reverse-code this to a new variable.

   The following summarizes how to correctly convert your *life_quality* variable to a numeric *good_life_quality* variable.

   First, check the class of your *life_quality* variable using *class().* This will differ depending on how you've loaded your data. If it is a factor variable, make sure to convert the variable to a character string before converting it to a numeric vector, as in the following expression:

   ```
   Dating$numeric_life_quality = as.numeric(as.character(df$life_quality))
   ```

   Make sure "Don't know" and "Refused" categories are set to NA.

   Next, create a new *good_life_quality* variable which recodes the numeric values so that higher values correspond to higher levels of life quality (e.g. 1 = poor… 5 = excellent). Check your conversion using table(Dating$good_life_quality, Dating$numeric_life_quality) to make sure your 1s have been correctly coded to 5s, 2s to 4s, etc.

   Include the output of "summary(Dating$good_life_quality)" in your lab report, after making the above changes.

b. We'll be using the following variables to run a series of nested regressions with good_life_quality as our outcome variable.

   - *married* (created from *marital_status*)
   - *income_50K* (created from *income*)
   - *age*

Using the *marital_status* variable, create a binary integer *married* dummy variable with Married = 1 and 0 for all other categories except “Refused.” Set rows with “Refused” to NA for the *married* variable.

Using the *income* variable, create a binary integer *income_50K* dummy variable with incomes under $50,000 to 0, and $50,000 and above to 1. Set rows with categories corresponding to “Refused” and “Don’t know” to NA for the *income_50K* variable.

Finally, make sure your *age* variable is numeric. Set rows corresponding to “Refused” and “Don’t know” to NA for the *age* variable. Keep any values that are grouped above some number (i.e. “97 or more” should be left in the data as value 97).

Create a new dataframe, Dating_lim, that contains only rows with non-missing values for *good_life_quality*, *married, income_50K*, and *age.* Make sure Dating_lim also only contains the above four columns.

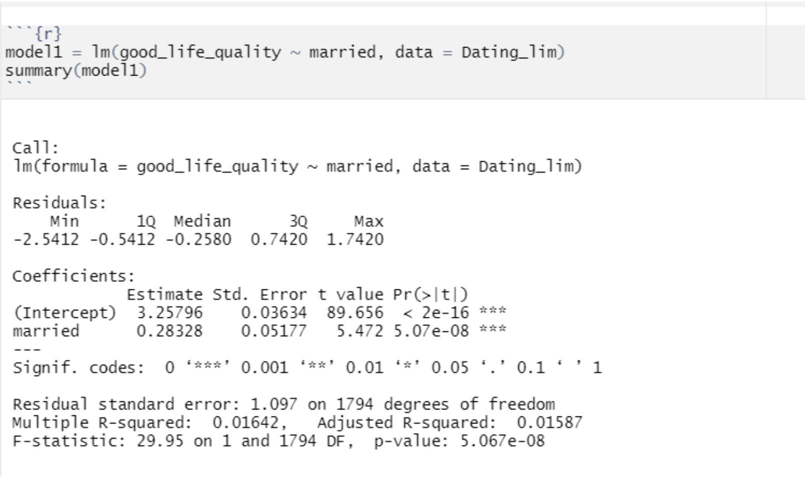
Thus, our final regression with all variables (Model 3) will be:

good_life_quality = Intercept + b1*married + b2*income_50K + b3*age + ε

Do not transform any of the variables for parts b-f (we will ask a question about transformations in part g).

How many cases are left in Dating_lim? Include the output of summary(Dating_lim) and dim(Dating_lim) below.

c. Model 1: Fit an OLS model to Dating_lim from the previous step that predicts *good_life_quality* (dependent variable) as a function of *married* (independent variable).



~~~
```{r}
model1 = lm(good_life_quality ~ married, data = Dating_lim)
summary(model1)
```
~~~

```
Call:
lm(formula = good_life_quality ~ married, data = Dating_lim)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5412 -0.5412 -0.2580  0.7420  1.7420 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.25796    0.03634  89.656  < 2e-16 ***
married      0.28328    0.05177   5.472 5.07e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.097 on 1794 degrees of freedom
Multiple R-squared:  0.01642,	Adjusted R-squared:  0.01587 
F-statistic: 29.95 on 1 and 1794 DF,  p-value: 5.067e-08
```

Is the regression significant? How do you know?

- It shows statistically significant, since the overall p-value (5.067e-08) is smaller than .05.

What is the slope coefficient? Is it statistically significant?

- The slope coefficient is around 0.28, which indicates a positive slope.
- Yes, it is statistically significant. Because the p-value (5.07e-08) for the married coefficient is less than .05.

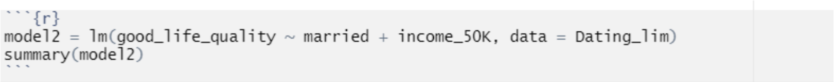
If the slope is statistically significant, provide a precise interpretation of the coefficient of *married* as it relates to the dependent variable in Model 1.

- There is a positive relationship between the dependent variable (good_life_quality) and independent variable (married), and it is statistically significant.
- Since married is coded as a binary variable, which means being married has a slightly positive effect (0.28-unit increase) on the good life quality comparing to the ones who are not being married.

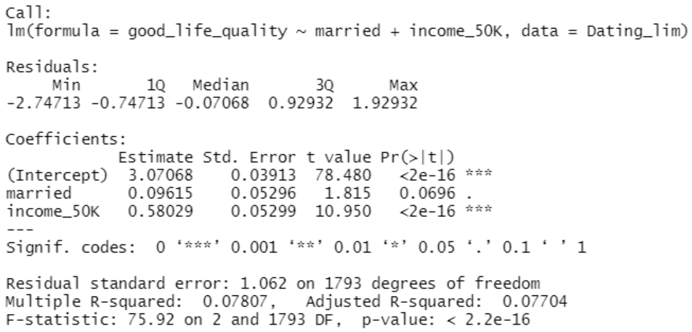
Commented [A1]: Need to double check the language.

Commented [A2R1]:

d. Model 2: Now fit a second OLS model to Dating_lim. Keep *good_life_quality* as your dependent variable, but now use both *married* and *income_50K* as your explanatory variables.



~~~
```{r}
model2 = lm(good_life_quality ~ married + income_50K, data = Dating_lim)
summary(model2)
```
~~~



```
Call:
lm(formula = good_life_quality ~ married + income_50K, data = Dating_lim)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.74713 -0.74713 -0.07068  0.92932  1.92932 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.07068    0.03913  78.480   <2e-16 ***
married      0.09615    0.05296   1.815   0.0696 .  
income_50K   0.58029    0.05299  10.950   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.062 on 1793 degrees of freedom
Multiple R-squared:  0.07807,	Adjusted R-squared:  0.07704 
F-statistic: 75.92 on 2 and 1793 DF,  p-value: < 2.2e-16
```

Is the regression significant? How do you know?

- It is statistically significant since the overall p-value (<2.2e-16) is less than .05.

What is the slope coefficient for *income_50K*? Is it statistically significant?

- The slope is 0.58, it is statistically significant since the coefficient for income_50K has a p-value (<2e-16) less than .05.

If it is statistically significant, provide a precise interpretation of the coefficient of *income_50K* as it relates to the dependent variable in Model 2.

- Since the income_50K variable is a binary variable, which means while controlling the married variable, yearly income more than 50K has a positive effect (0.58-unit increase) on the good life quality comparing to the yearly income is less than 50K,.

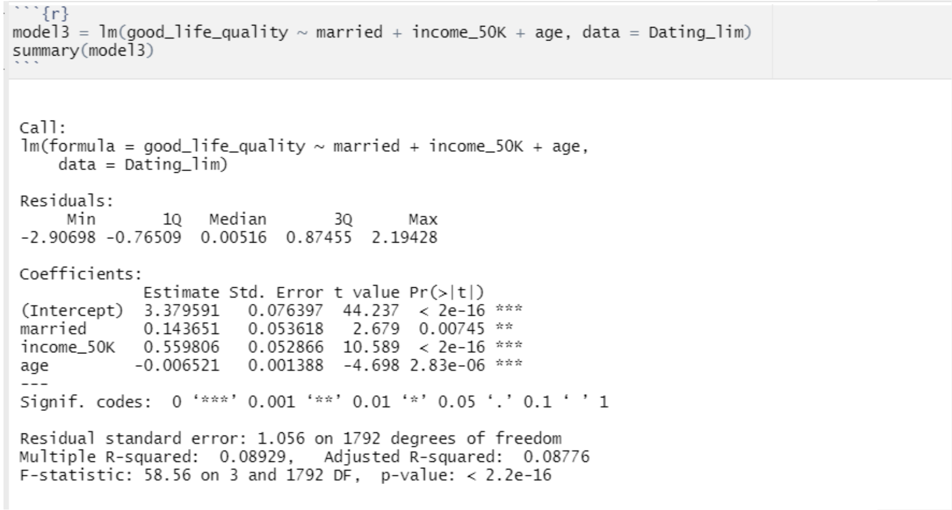
Commented [A3]: Need to double check the language.

Commented [A4R3]:

Discuss any meaningful changes for the *married* variable between Model 1 and Model 2 (i.e. after including the *income_50K* variable).

- After adding the income_50k variable, the married variable is no longer statistically significant to the overall regression. The positive direction stays the same.

e. Model 3: Time to fit a third OLS model to Dating_lim. Keeping *good_life_quality* as your dependent variable, and *married* and *income_50K* as your explanatory variables, add *age* as your third explanatory variable.



~~~
```{r}
model3 = lm(good_life_quality ~ married + income_50K + age, data = Dating_lim)
summary(model3)
```
~~~

```
Call:
lm(formula = good_life_quality ~ married + income_50K + age,
    data = Dating_lim)

Residuals:
     Min       1Q   Median       3Q      Max
-2.90698 -0.76509  0.00516  0.87455  2.19428

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.379591   0.076397  44.237  < 2e-16 ***
married      0.143651   0.053618   2.679  0.00745 **
income_50K   0.559806   0.052866  10.589  < 2e-16 ***
age         -0.006521   0.001388  -4.698 2.83e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.056 on 1792 degrees of freedom
Multiple R-squared:  0.08929,	Adjusted R-squared:  0.08776
F-statistic: 58.56 on 3 and 1792 DF,  p-value: < 2.2e-16
```

Is the regression significant? How do you know?

- Yes. The model 3 is statistically significant since the overall p-value (<2.2e-16) is less than .05.

What is the slope coefficient for *age*? Is it statistically significant?

- The slop for age is -0.0065, it is statistically significant because the p-value (2.83e-06) for the age coefficient is less than .05.

If it is statistically significant, provide a precise interpretation of the coefficient of *age* as it relates to the dependent variable in Model 3.

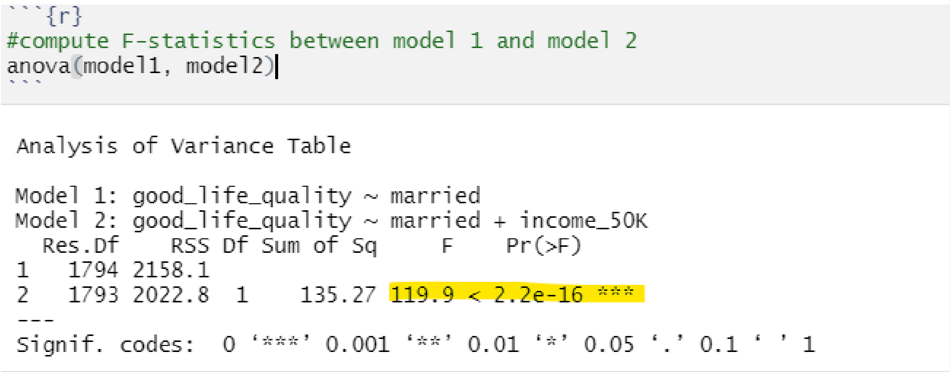
- Controlling the marriage status and income variable, the age variable has a slightly negative relationship (0.0065 unit of decrease) with the life quality.

Discuss any meaningful changes for the *married* and *income_50K* variable between Model 2 and Model 3 (i.e. after including the *age* variable).

- After adding the age variable, the married and income variable both show statistically significance; the positive direction stays the same.

f. Compute the F-statistics and associated p-values ***between*** your three regression models.

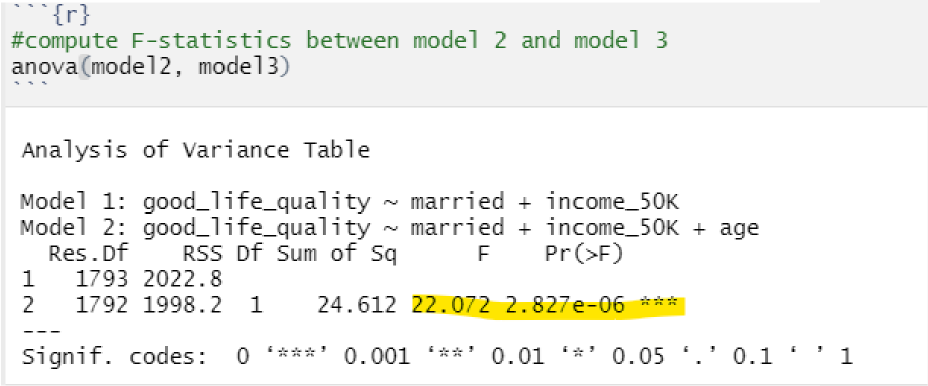
Is there a statistically significant improvement from Model 1 to Model 2 & from Model 2 to Model 3? Include the F-statistics and associated p-values in your answer.



~~~
```{r}
#compute F-statistics between model 1 and model 2
anova(model1, model2)
```
~~~

```
Analysis of Variance Table

Model 1: good_life_quality ~ married
Model 2: good_life_quality ~ married + income_50K
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)    
1   1794 2158.1                                 
2   1793 2022.8  1    135.27 119.9 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- The p-value (2.2e-16) for the F-statistic model is statistically significant, which means the model2(controlling married variable, adding the income variable) has improved the overall regression from model 1.

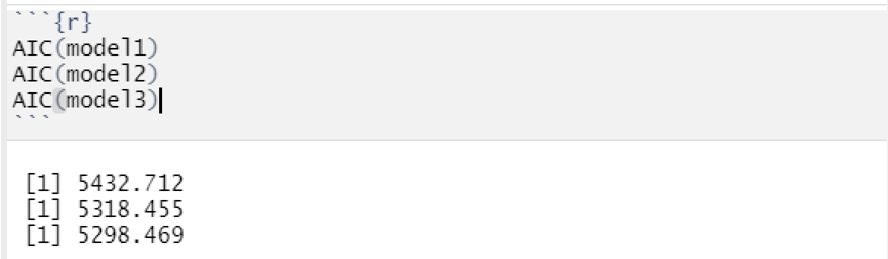


~~~
```{r}
#compute F-statistics between model 2 and model 3
anova(model2, model3)
```
~~~

```
Analysis of Variance Table

Model 1: good_life_quality ~ married + income_50K
Model 2: good_life_quality ~ married + income_50K + age
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1   1793 2022.8                                  
2   1792 1998.2  1    24.612 22.072 2.827e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- The p-value (2.827e-06) for the F-statistic model is statistically significant, which means controlling married, income variable, and adding the age variable model3 improved the overall regression model from model 2.

Next, compute and state the AIC values for Model 1, 2 and 3. What do they tell you as you compare between the three models?

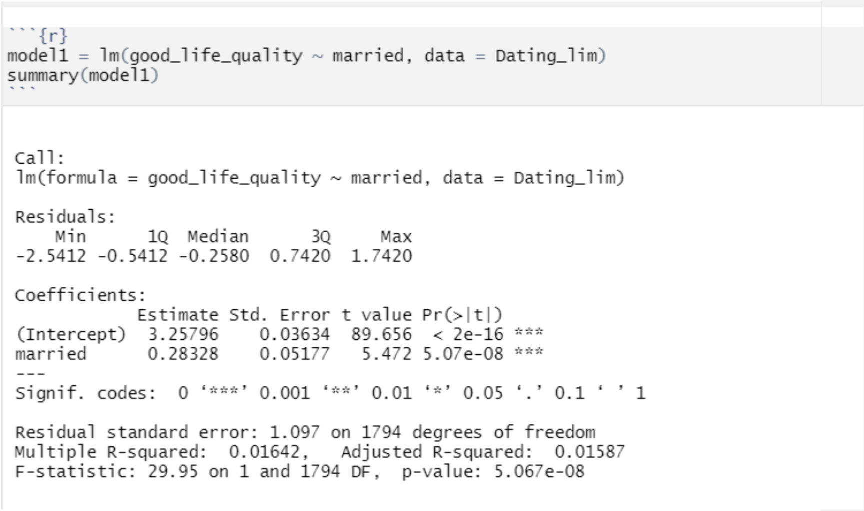


~~~
```{r}
AIC(model1)
AIC(model2)
AIC(model3)
```
~~~

```
[1] 5432.712
[1] 5318.455
[1] 5298.469
```

- Since the AIC **decreased** from model1 to model2, as well as from model2 to model3, therefore the models are improving upon one another.

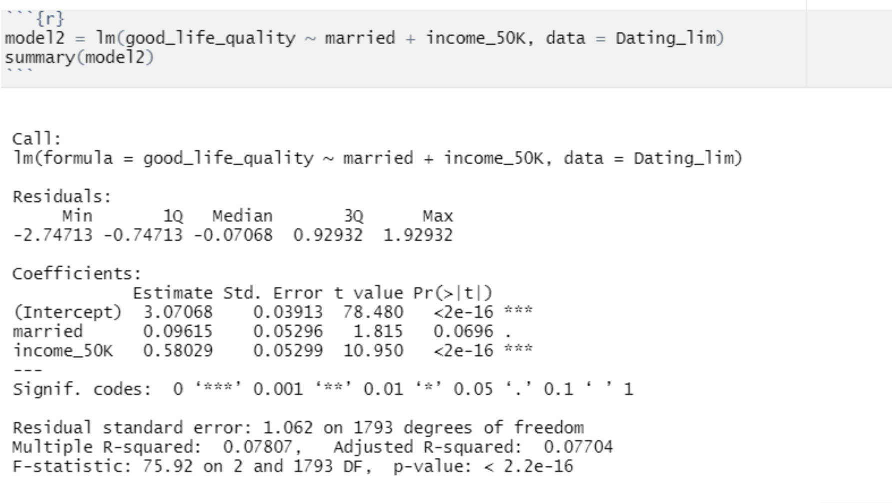
State the adjusted R^2 values for Model 1, 2 and 3. What do they tell you as you compare between the three models?



~~~
```{r}
model1 = lm(good_life_quality ~ married, data = Dating_lim)
summary(model1)
```
~~~

```
Call:
lm(formula = good_life_quality ~ married, data = Dating_lim)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5412 -0.5412 -0.2580  0.7420  1.7420 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.25796    0.03634  89.656  < 2e-16 ***
married      0.28328    0.05177   5.472 5.07e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.097 on 1794 degrees of freedom
Multiple R-squared:  0.01642,	Adjusted R-squared:  0.01587 
F-statistic: 29.95 on 1 and 1794 DF,  p-value: 5.067e-08
```

- R-squared with married variable: 0.01587; the model1 which has the married variable explains 1.587% of the data

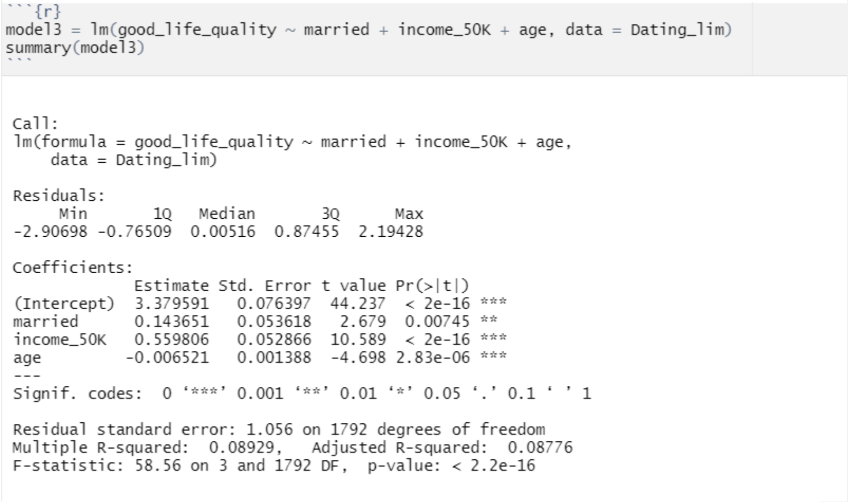


~~~
```{r}
model2 = lm(good_life_quality ~ married + income_50K, data = Dating_lim)
summary(model2)
```
~~~

```
Call:
lm(formula = good_life_quality ~ married + income_50K, data = Dating_lim)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.74713 -0.74713 -0.07068  0.92932  1.92932 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.07068    0.03913  78.480   <2e-16 ***
married      0.09615    0.05296   1.815   0.0696 .  
income_50K   0.58029    0.05299  10.950   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.062 on 1793 degrees of freedom
Multiple R-squared:  0.07807,	Adjusted R-squared:  0.07704 
F-statistic: 75.92 on 2 and 1793 DF,  p-value: < 2.2e-16
```

- R-squared adding income variable: 0.077; The model2 which has the married and income variable explains 7.7% of the data

- Model 2 shows a statistically significant improvement from model 1, because the anova between p-value is statistically significant, and the AIC decreased from Model1.



~~~
```{r}
model3 = lm(good_life_quality ~ married + income_50K + age, data = Dating_lim)
summary(model3)
```
~~~

```
Call:
lm(formula = good_life_quality ~ married + income_50K + age,
    data = Dating_lim)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.90698 -0.76509  0.00516  0.87455  2.19428 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.379591   0.076397  44.237  < 2e-16 ***
married      0.143651   0.053618   2.679  0.00745 ** 
income_50K   0.559806   0.052866  10.589  < 2e-16 ***
age         -0.006521   0.001388  -4.698 2.83e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.056 on 1792 degrees of freedom
Multiple R-squared:  0.08929,	Adjusted R-squared:  0.08776 
F-statistic: 58.56 on 3 and 1792 DF,  p-value: < 2.2e-16
```

- 8.8% of the variation in the depend variable is accountable for the variation of the model3

- Model 3 shows a statistically significant improvement from model 2, because the anova between p-value is statistically significant, and the AIC decreased from Model2.

- The r square between models is increasing as we are adding more variable (0.01589 – 0.077 – 0.087), from the ANOVA the p-values are also statistically significance, the AIC is decreasing as we are adding more variables, which means adding more variables in each model better explains the data.

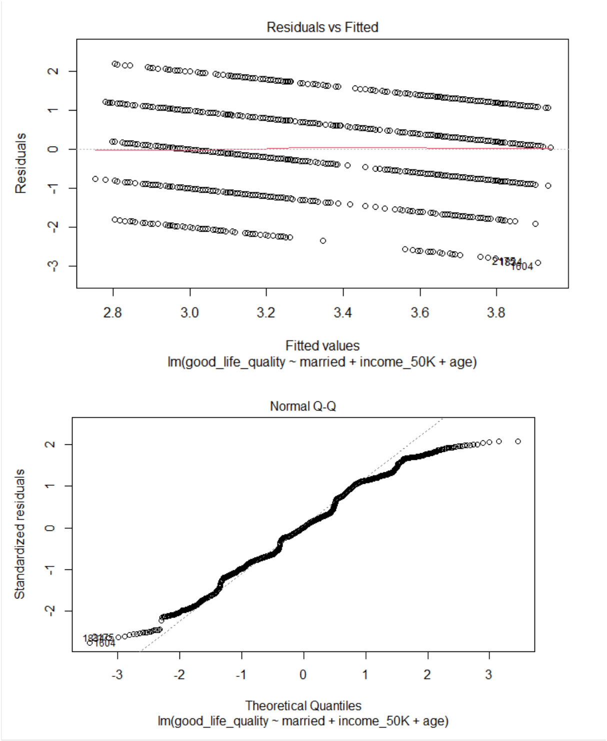
What is your “best” model, and how do you know?

- Model 3, since it has the largest r square, as well as a statistically significant p-value comparing the F-statistic with model 2, the AIC is also decreased.
- The R-squared increased the most from model1 to model2, which means adding the income variable improve the model a lot, while form model2 to model3 (adding age) only shows a slightly improvement.

g. Now we will test for heteroskedasticity of residuals for our last regression, Model 3. Please note that if we did find a problem, we would probably re-run all of our regressions with an appropriate linear transformation. However, for our purposes we are only going to check to see if we should re-run our regressions or not.

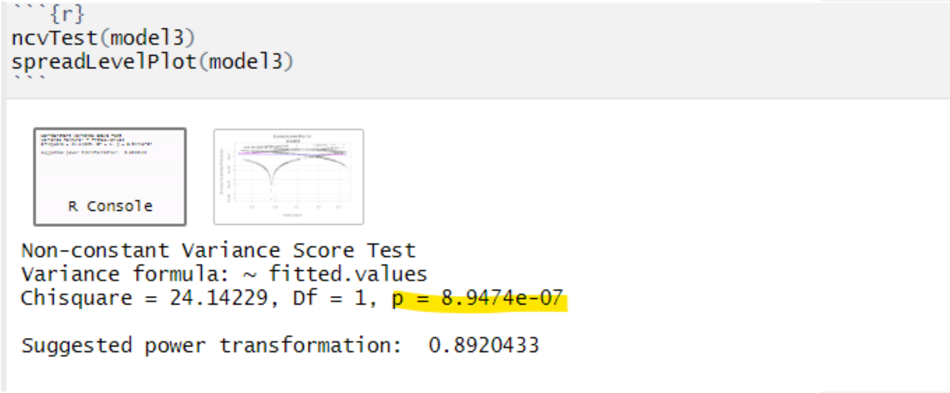
Examine the diagnostic plots for your 3rd regression. Look at the first two plots (Residuals vs Fitted plot and Normal Q-Q plot). Why do you see five distinct 'bands' in the Residuals vs Fitted plot? Does the Q-Q plot indicate normality?

- Because the question is a Likert-style question, therefore the Residual is in bands shape.



- The QQ plot is semi- normal, except for the 2 ends.

What is your p-value for the heteroskedasticity test, and is it significant?



~~~
```{r}
ncvTest(model3)
spreadLevelPlot(model3)
```
~~~

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 24.14229, Df = 1, p = 8.9474e-07

Suggested power transformation:  0.8920433
```

- The p-value shows statistically significant, which means the variance are not equal and needed transformation to reach normality.

If the heteroskedasticity test is not significant, you can stop here. If it is significant, state the power transformation you would need to apply to your dependent variable to help address the problem of heteroskedasticity. Would this transformation be enough to solve the problem of heteroskedasticity in your model (and if not, what is at least one thing you could do to address the problem)?

- Suggested power transformation: 0.892. No, after the transformation, the p-value is still statistically significant, we can try to use a robust regression.

That is all—Have a wonderful winter break! May you rest like a rabbit/dog!