Isabelle Rennenberg

Report STA512

Due: 18Apr23 5:45 PM

**Introduction**

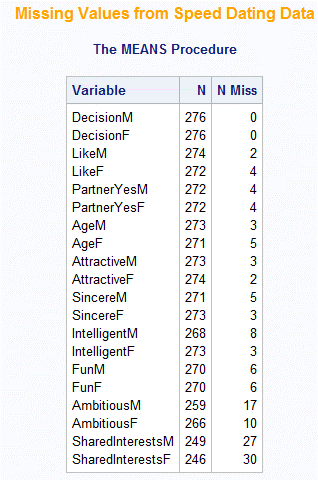
We are looking to create an equation that can predict a dater’s opinion of the person they are dating based on certain variables.. The given variables were attractiveness, sincerity, intelligence, fun, ambitiousness, and shared interests of their partner.

It is important to the client whether or not it matters that the daters are of the same race.

How many couples had the same race and how many did not? Does it matter if the partners are close in age? How many are close in age and how many are not?

Make sure to give a good idea of how close your estimated like ratings were to actual like ratings.

**Descriptive Statistics**

****Our experimental unit is one couple, each row is a couple’s date. Each value itself is the ranking of one individual. The data was collected from 276 heterosexual couples with 22 different variables of expression of the evaluation of the date. There are two character variables: race male and female. The rest are numerical, including age, and either are represented as 1 or 0 or on a scale of 1 to 10.

1=Yes 0=No

I am thinking that it is difficult to predict a way for us to understand if we do not separate male and females- we want to know like, but we have two different like variables, one for males and one for females.

First note- it looks like there are a lot of missing values. Proc means will naturally not count them, but this is important for my code going forward.

**Figure 1. Missing Values**

A screen shot of a graph

Description automatically generated with medium confidenceA screen shot of a graph

Description automatically generated with medium confidenceI decided to divide the data into two models, one with male and one with female like as compared to that gender’s rating of attractiveness, sincerity, intelligence, fun, ambitiousness, and shared interests for the other gender. What these two scatterplots show are jackknife vs. the predicted values. This is a way to assess the stability of the regression model. “Jackknife” in this instance is a type of residual, which measures difference between the observed and predicted values. Essentially, it measures the “goodness” of a fit for a regression model. All this to say, when we plot these residuals with our predicted values, we are looking for the plot to not look like it has any obvious patterns. Such patterns include funneling, sinusoidal curve, or more. These patterns would indicate that there is an issue with the data, whether that was homoscedasticity (the variance, or change, or all the residuals should be equally scattered, so no pattern!), independence or linearity. All this to say: there should be no pattern! From the data generated, **there is no pronounced funneling,** so we are good in terms of homoscedasticity. Variance seems to be homogenous throughout this plot. There is no sinusoidal pattern, so **independence is satisfied**. As you can note, there are parallel diagonal lines, likely due to gaps in the data (refer to figure 1). This also can be due to discrete ways of measuring things. There are no pronounced, systematic patterns. Consistent will what we would expect, assuming we took random samples across males and females. This pattern can also arise from data in which the possible values are 0 and 1 (as they are for your data!). Parallel diagonal lines in this case are not indicative of a pattern that would raise any alarms. Looks good so far!

**Figure 3. Scatterplot of Female Like**

**Figure 2. Scatterplot of Male Like**

Now, it is good to look into the actual values for jackknife residuals. Jackknife residuals that are greater than the absolute value of 2 generally raise some concern. These values generally have a strong pull on the regression and can be indicative of outliers or errors in the data. Here are the jackknife values of concern for both the male and female regressions.

A screenshot of a computer

Description automatically generated with medium confidenceLooks like there is an observation of concern in the female area.

A screenshot of a graph

Description automatically generated with medium confidenceMale PlotsA screenshot of a graph

Description automatically generated with low confidence-3.8 is an extreme observation. I would look into eliminating it.

A picture containing text, screenshot, font, number

Description automatically generatedA screenshot of a table with numbers

Description automatically generated with low confidence

A screenshot of a graph

Description automatically generated with medium confidenceFemale plotsA screenshot of a graph

Description automatically generated with low confidence-3.3 and 4.1 are extreme observations. I would look into eliminating them. From what I can see looking into the observations, I don’t see any jarring reasons as to why they should be eliminated.

A screenshot of a computer screen

Description automatically generated with low confidenceA screenshot of a table with numbers

Description automatically generated with low confidence

2(1+6)/276=0.051, any leverage greater than that are of concern. For Cook’s distance, anytime this value exceeds 1. 4.188, observation 255, does not have any concern over the cook’s, but there is slight concern with the leverage as it is over 0.05. Other values are slightly over but none are over for cooks.

How many variables does the optimal model have for this regression? Here we will consider polynomial terms. For Like, our options are to square or cube this term. But if we think about it logically, the only options are 0 or 1. Squaring and cubing these terms does nothing. Our other variables, however, are on the scale of 1-10, so they could benefit from squaring or cubing.

**Selection of the Models and Type of Analysis Employed**

**A screenshot of a computer

Description automatically generated with medium confidence**

**A screenshot of a computer

Description automatically generated with low confidence**

**A picture containing text, screenshot, number, font

Description automatically generated**

**Diagnostics**

**Summary of Findings**