HW Questions

This homework concerns a study on obesity and qualification.

Reference: Hebl, M. R., & Mannix, L. M. (2003). The weight of obesity in evaluating others: A mere proximity effect. Personality and Social Psychology Bulletin, 29, 28-38.

People who are obese face a great deal of prejudice and discrimination. But what about people who are somehow associated with an obese person? In the study above, participants had to rate how qualified a particular job applicant was. This applicant was sitting by a woman. The researchers manipulated the following two variables: the weight of the woman and the relationship between the woman and the applicant. The woman was either obese or of average weight. This woman was also portrayed as being the applicant’s girlfriend or a woman simply waiting to participate in a different experiment.

Data: weight.txt

Variables descriptions:

• Weight: The weight of the woman sitting next to the job applicant. 1 = obese, 2 = average weight

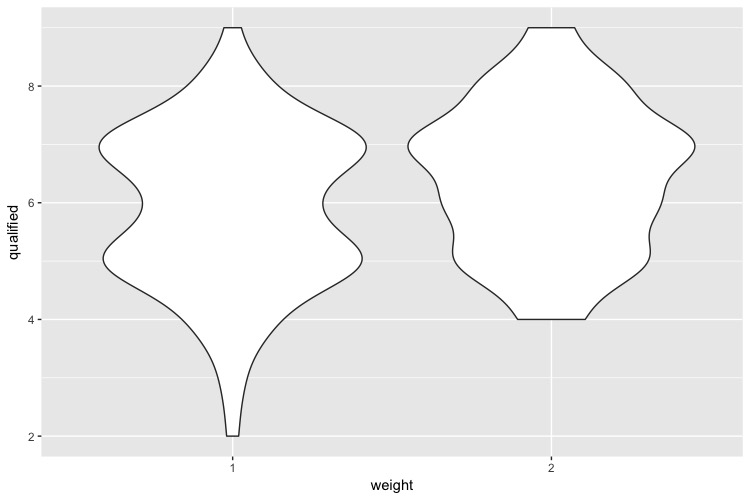
• Relate: type of relationship between the job application and the woman seated next to him

1 = girlfriend, 2 = (waiting for another experiment)

• Qualified: a numerical. Larger numbers represent higher professional qualification ratings

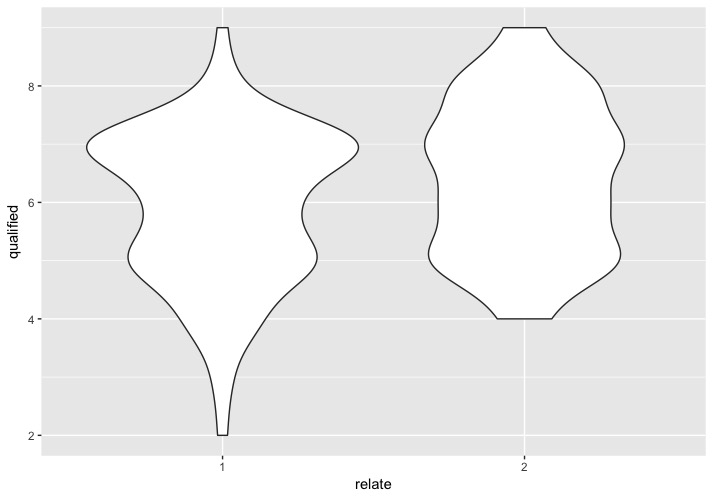
1. **Do a plot of qualified vs weight, qualified vs relate, and qualified vs weight:relate. For each plot, briefly describe what you see. Based on these plots alone, do you think sitting next to an obese woman has an adverse effect on the applicant’s qualification ratings?**

Qualified vs Weight Plot



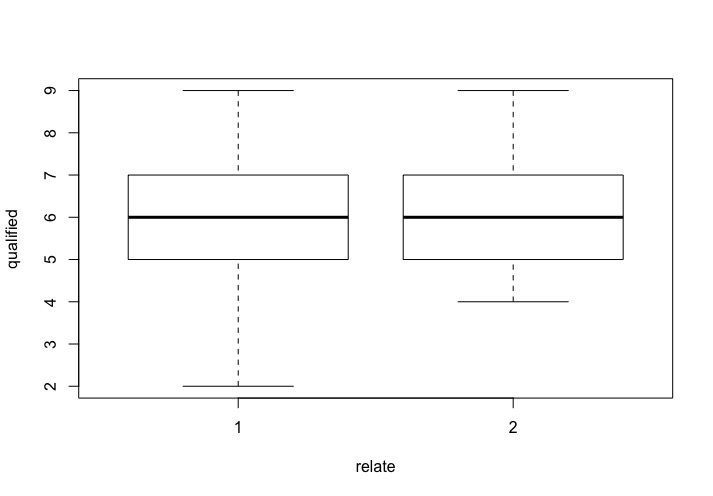
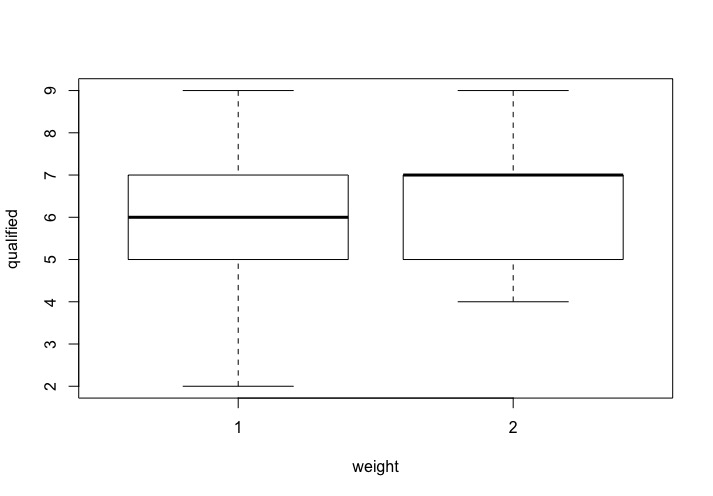
*According to the variable description, if the women sitting beside the job applicant is obese, then the professional qualification rating could be anywhere from two to nine. But it is less possible to be on the extreme ratings but more likely to be around five or seven, where are two peaks for this situation. Then if the women next to the applicants is in average weight, then their professional qualification rating would be possibly higher, and a rating of seven is where most applicants happen to be.*

Qualified vs Relate Plot



*According to the variable definition, in this graph shows that when the women seated who seated next to the job application is his girlfriend, then the professional qualification varies in a large range with a peak appears at rating around seven. If the women is just an acquaintance, then none of the job applicants have a professional qualification ratings lower than four, and the ratings among them spreads out pretty evenly.*

Qualified vs weight and relate interacted Plot



*Two graphs do not show large differences and almost have the same median value for qualified rating no matter how the applicants and the women related or what is the weight of the women except for the median when weight of the women is average weight is the same as the third quartile.*

1. **Run a linear regression model, called weights.model1, of qualified vs weight and relate. Clearly show the R command that you use, and include the R’s model summary. Write down the equation that R gives you.**

*I used the R command: weights.model1 <- lm(qualified ~ weight + relate, data = wei)*

*summary(weights.model1)*

*The summary given by R is:*

*Call: lm(formula = qualified ~ weight + relate, data = wei)*

*Residuals:*

*Min 1Q Median 3Q Max*

*-3.6765 -1.1235 -0.1235 0.8765 2.8765*

*Coefficients: Estimate Std. Error t value Pr(>|t|)*

*(Intercept) 5.6765 0.1748 32.482 <2e-16 \*\*\**

*weight2 0.4887 0.1963 2.489 0.0137 \**

*relate2 0.4470 0.1960 2.281 0.0238 \**

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Residual standard error: 1.294 on 173 degrees of freedom*

*Multiple R-squared: 0.06565, Adjusted R-squared: 0.05485*

*F-statistic: 6.078 on 2 and 173 DF, p-value: 0.002811*

*From this summary, we know the function is:*

1. **Interpret all the coefficients and the p-values associated with the coefficients.**

*Intercept means the value of professional qualification rating when weight is in category 1 and relate is also in category 1. The p-value here is <2e-16, which is extremely small. This p-value means that the coefficient value 5.6765 is significantly different from zero, which means that putting the intercept here would be significantly different from having nothing here.*

*The coefficient for variable weight is 0.4887. This means the difference of qualification rating if we change the weight input from category one to two while keeping the relate at category one. The p-value is 0.0137, which is bigger than a conservative alpha that equal to 0.01. Therefore, having a coefficient for weight with such value 0.4887 is not significantly different from not having any.*

*The coefficient for variable relate is 0.4470. Similarly, this means if we change the variable relate from category one to two, the qualification rating would increase 0.4470. The p-value is 0.0238 > alpha = 0.01. Then the coefficient of the relate variable is not significant from zero.*

1. **Report the R2 and adjusted R2 of your model. What are the meaning of these values?**

*From the summary by R, the R-squared value is 0.06565, and the adjusted R-squared is 0.05485. Firstly, the values of both r-squared are around 0.06 = 6%, which is very close to zero. Thus, the linear function we generated explain little about the trend of the data, and points are averagely very far from the linear line we generated. Secondly, since the adjusted r-squared value is even small, this means that the variables we choose are also not effective enough. We probably need more representative variables and need to reselect what variables to put into the model.*

1. **Run another linear regression model, called weights.model2, of qualified vs weight:relate alone. Write down the equation that R gives you. Interpret all the coefficients and the p-values associated with the coefficients. Report the R2 and adjusted R2.**

*From the summary given by R, we know the regression model formula is:*

*For the interaction terms, there are totally four possibilities for the women:*

1. *She is weight = 1 and relate = 1 (She is obese and the girlfriend of the applicant.)*
2. *She is weight = 2 and relate = 1 (She is around the average weight and the girlfriend of the applicant.)*
3. *She is weight = 1 and relate = 2 (She is obese and waiting for another experiment.)*
4. *She is weight = 2 and relate = 2 (She is around the average weight and waiting for another experiment.)*

*Since R select the fourth one as the reference/baseline, the intercept value means the value of the qualification rating of the applicant if the women seating beside him is around the average weight and waiting for another experiment. The p-value is smaller than 2e-16 < alpha = 0.01. Thus it is significant to include this coefficient.*

*Since the women can only be in one of the category among four and each category only have two possibility, if she is in 1), the value of the variables will be: , , Hence all other interacting variable terms will zero except*

*The first variable is with coefficient value -0.9426, which means that when the women seatting beside the applicant is obese and his girlfriend, then the average applicant’s qualification rating is 6.5926 – 0.9426 = 5.6500. The p-value for this variable coefficient is 0.000632 < alpha = 0.01. Hence having this coefficient is significant different from not having it (coefficient equals to zero).*

*The second variable is with coefficient value -0.4021. This indicate that if the women is around average weight and girlfriend, then the mean of qualified value is 6.5926 - 0.4021 = 6.1905. The p-value for it is 0.133908 > alpha = 0.01. Hence it is not significant to include this variable.*

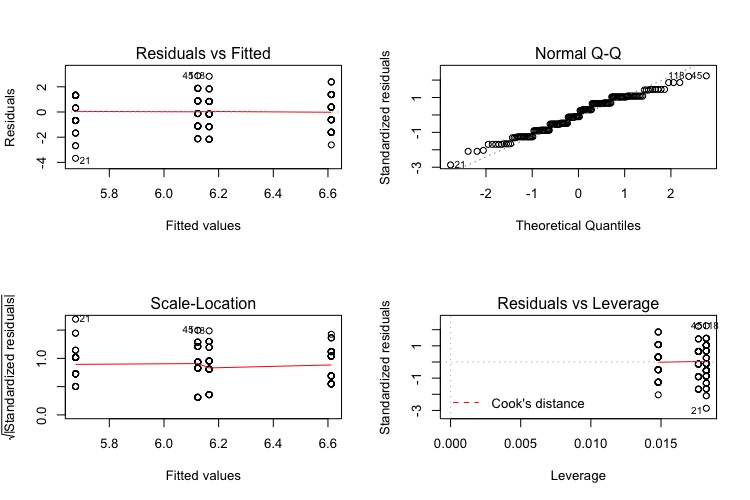
*The third one is with coefficient value -0.4426. This tell us that if the women is obese and waiting for another experiment, the mean of the applicant’s professional qualification rating is 6.5926 - 0.4426 = 6.1500. The p-value for it is 0.103934 > alpha = 0.01, which means that it is not significant to include this interacting variable term.*

*Since R set as the baseline, which shows the value by the coefficient of intercept, so all values are shown as NA (Not applicable or Not available).*

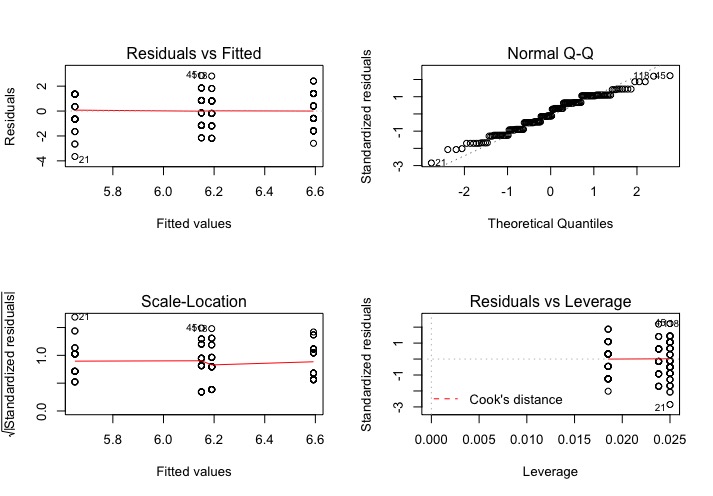
*The value of R-squared is 0.06599 and the adjusted R-squared is 0.0497. We know that when the r-squared value is close to zero, it means that the distance from each point to the mean is very large. We know this from the graphs at first question that the maximum value, 3rd-quartile, 1st-quartile, and the minimum value are all located far away from the median point, so we could image they are all far away from the mean value, too. The difference between the r-squared value and the adjusted r-squared value means that the variables in the equations are not very necessary and we need to eliminate some and possibly change them to more effective variables that could represent the data more efficiently.*

1. **Do a diagnostic plot for your two models. Say which, if any, of the (a) independence (no mean trend) (b) normal distribution and (c) constant variance assumptions are violated.**

**Diagnostic plot for weights.model1**

****

**Diagnostic plot for weights.model2**



*The diagnostic graphs for both models are very similar and does not violate any assumptions.*

*This apply to both graphs: Both residuals vs Fitted graph and Scale-Location graph have an average trend line approximately horizontal with dots distribute evenly above the line and below the line. Since the points are in four categories mentioned in the previous question, so it is reasonable that the points do not distribute randomly all over the graph but form four vertical lines. Thus, the first and the third assumption are hold. Then the graph Normal Q-Q forms a approximate diagonal line, where points are all follow the trend of the line with little variances. And the last graph shows that there’s no outliers for this data set in either models.*

1. **Between the two models you have run, which one is better? Why?**

*Weights.model1 is better because each coefficient of variables has smaller p-value. Also the model has a larger adjusted R-squared value since it contains less variables.*

1. **Run another linear regression model, called weights.model3, which includes an interaction term between weight and relate. Write down the equation that R gives you. Interpret all the coefficients and the p-values associated with the coefficients.**

*Equation R gives is:*

*The reference category is weight = 1 and relate = 1. Thus, the intercept means that when the women seatting beside the applicant is obese and his girlfriend, the qualification rating of the applicant is averagely 5.65. It is significant to maintain the intercept because the p-value for it is smaller than 2e-16 < alpha = 0.01.*

*Then if the women is around average weight (in weight category two), then its value of qualification rating become 5.65 + 0.54048 = 6.19048 ~= 6.1905, which is also the coefficient of in model 2 meaning the average rating of applicant who seated beside a women who waited to take another experiment and in average weight. The value of its p-value is 0.0611 > alpha = 0.01, which means that having is variable is not significantly different from not having it.*

*Similarly, if the applicant seated beside the women who is his girlfriend and obese, the qualification rating will be 0.5 higher than who seat beside the women who is not related to him. Notice the value of the average rating is again same as the value for coefficient of in model 2. The p-value is 0.0867 > alpha = 0.01, so it is not significant to maintain this variable.*

*The last interacting variable operate as an adjustment for the value of rating when the women is his girlfriend and obese. Because does not equal to the mean of qualification rating in such situation where. The p-value is 0.8043 >> alpha = 0.01, which means that adding such adjustment does not significantly change the result value.*

1. **What effect did the interaction term have on the model, and why do you think this is the case?**

*Interaction adjust the value of the professional qualification rating when the women satisfy both category two of weight and relate. Because from model two we know that the mean of qualified value is 6.5926. Without this adjustment, the equation become*

*This result is close but does not directly equal to the average, so adding the interaction term makes an exact fit for the data. However, it leads to an over-fit, too.*

1. **Bonus question: In what way is weights.model2 equivalent to weights.model3? In what way is it different? Between these two models, which one is better, and why?**

*As mentioned in question 8, that both model fit the data exactly but include different variables in the model. The result of fitting the data exactly is that plugging in the number by each situation would give us the exact same number, which is the mean of all data collected in such situation.*

*If considering the r-square, they are exactly the same, so it does not show which one is better.*

*From other values, model 3 is better in my point of because the p-value for each variable’s coefficient is smaller. In addition, although both model have three variables, the first two variables in model 3 is simpler with the same effect on the result.*