Question 1: A statistician fitted several regression models to different variable pairs. For each model, she ran diagnotic plots for the residuals. For each of the following residual plots say which, if any, of the following assumptions are violated: (a) no mean trend, (b) normal distribution and (c) constant variance. Based on your answer, which models are good fits?

Model 1

Constant variance is violated because the scale-location plot is not horizontal and is more of curve.

The residual vs fitted graph is curved, but it seems to be caused by outliers, but it is okay.

Similarly, the x=y line the QQ plot is curved at the end of spectrum is caused by outliers, but the rest of it is okay.

Constant variance is violated so Model 1 is not a good fit

Model 2

Scale-location plot is not horizontal and has a curve in the middle. So, constant variance is violated.

The residual vs fitted graph is curved, so mean trend and independence is violated.

The x=y line the QQ plot is curved at the end of but the rest of it is straight, so it's okay.

Constant variance, and mean trend is violated so Model 2 is not a good fit

Model 3

The scale-location plot is more or less horizontal, so constant variance is not violated.

The residual vs fitted graph is horizontal, so mean trend is good.

The x=y line the QQ plot is curved at the end of spectrum, but it is acceptable, and can be said to be normally distributed.

There is no major violation, so Model 3 is a good fit

Model 4

Constant variance is violated because the scale-location plot is curved.

Mean trend is violated, because the residual vs fitted graph is curved.

The data is also not normally distributed as we can see in the QQ plot that it does not form a x=y line.

All three conditions are violated, so Model 4 is not a good fit

Model 5

The scale-location plot is relatively straight, i.e. it is acceptable, and constant variance is not violated.

The residual vs fitted graph is straight, so mean trend is not violated.

As we can see in the QQ plot that, the outliers cause the plot to be curved at the end, but there is no serious violations, so data is normally distributed

All three conditions are met, so Model 5 is a good fit

Model 6

The scale-location plot is straight, so constant variance is not violated.

The residual vs fitted graph is a little bit curved in the middle, but it is acceptable, so mean trend is not violated.

The QQ plot shows that there are some outliers at the beginning and at the end, but the majority of the data forms a x=y pattern, so data is normally distributed

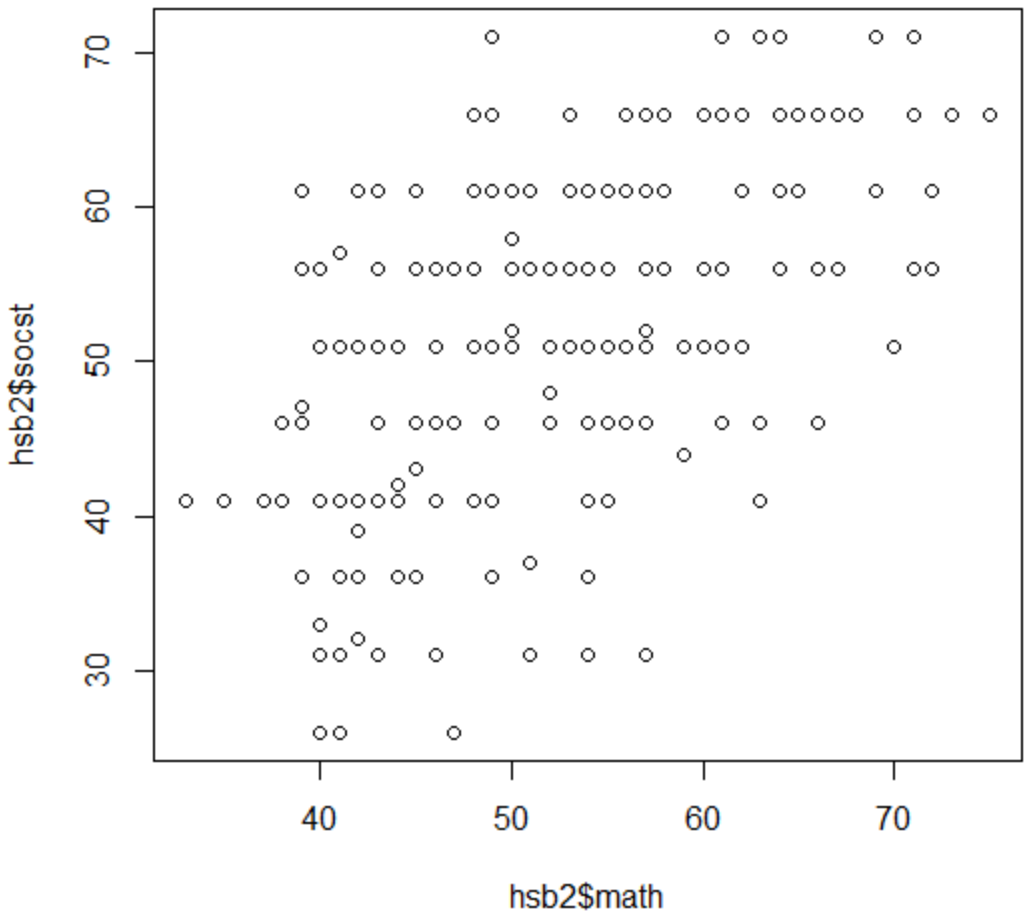
All three conditions are met, so Model 6 is a good fit

Question 2

Part 1 Do a scatter plot of social studies (socst) scores vs math scores (math). Report the correlation.

Correlation, (R) = 0.5444803

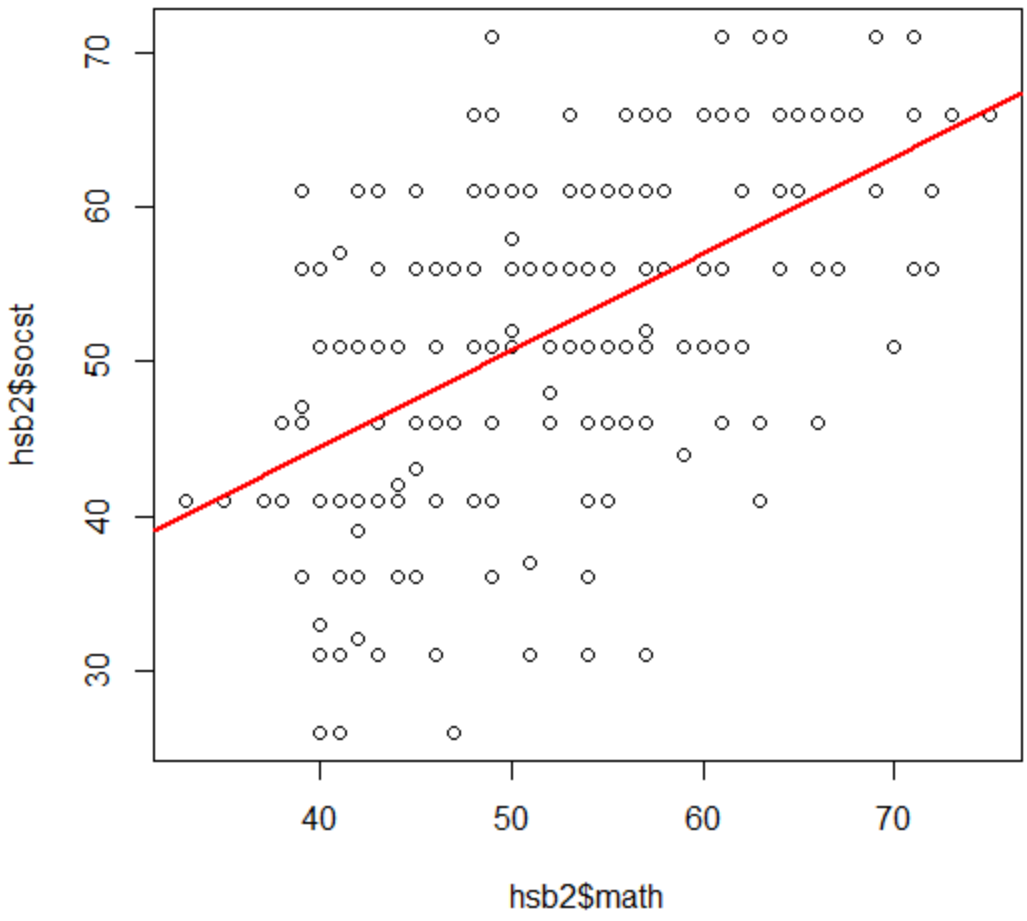
As computed in R studio




Part 2 Fit a linear regression model that can be used to predict social studies scores based on math scores. Write down the model equation that R gives you. Overlay this regression line on top of the scatterplot.

The equation is:

Y = 19.56 + 0.62*X




Part 3: Interpret the intercept and the slope. Is the intercept meaningful? Why or why not?

This would mean that for every point increase in math score, the social studies score increases by 0.62 points. The intercept would not be always useful, say in this case X = 0 then Y = 19.56. This means if the math score is 0 then the social studies score is 19.56. This would be a number which is out of range as even the minimum score of social studies is 26, and there is very little chance that someone will score a 0 in math. Therefore, extrapolation will occur as we can not know how the data will behave outside of our range.

Part 4: R reports the p-value of the slope and the intercept. What is the meaning of these p-values? What are the null and alternative hypotheses in these tests? For the p-values in your model, what can you conclude?

The p value for the intercept would be 2.36e-07, and the p value for slope is 2e-16.

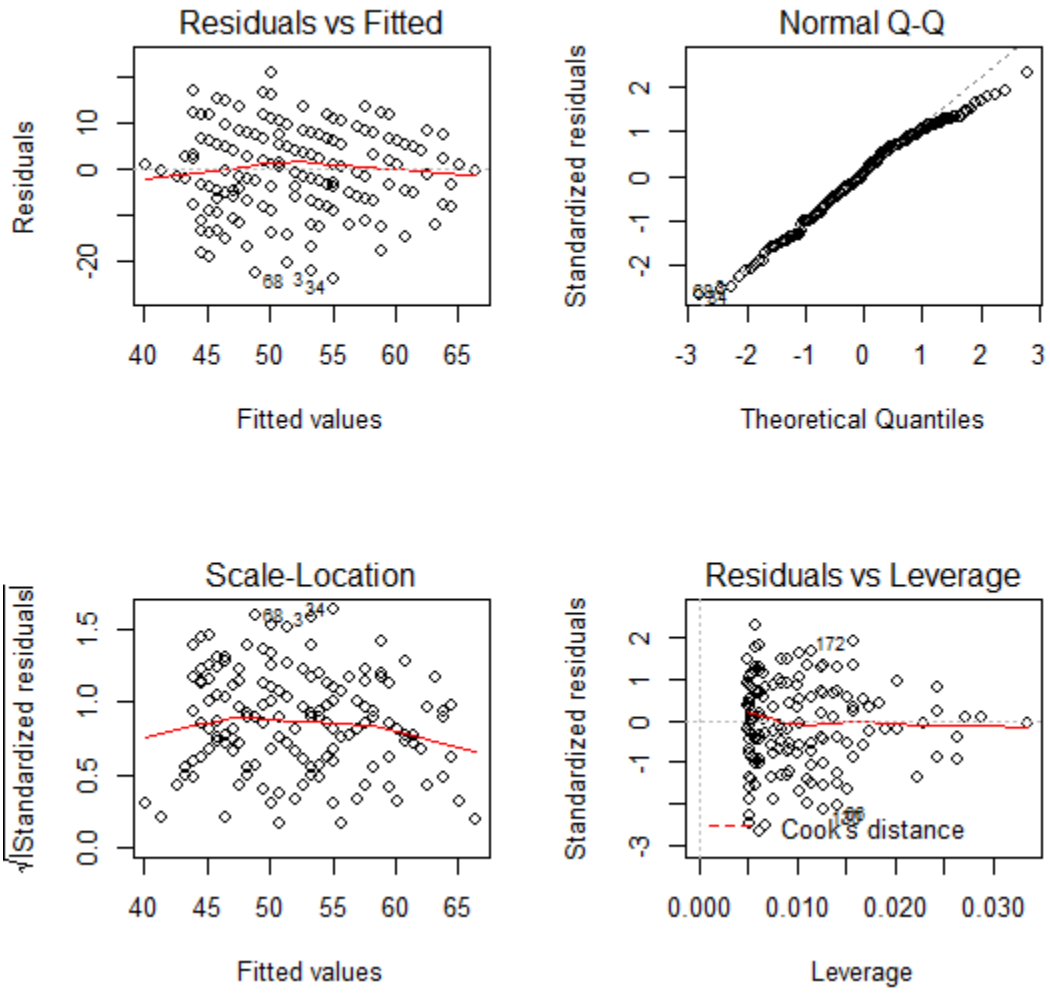
The values are the p-value for the t-test statistic under the null hypothesis.

H0: The true coefficient for increase in math scores is zero.

HA: The true coefficient for increase in math scores is not zero.

The p value reported by r is 2e-16 (two-sided) i.e. since the p value is small, we reject the null hypothesis and conclude that there is an increase in social studies score for increase in math score.

Part 5: Do a diagnostic plot for your model. Say which, if any, of the (a) no mean trend, (b) normal distribution and (c) constant variance assumptions are violated. Based on your answer, is the model a good fit? That is, is there a linear relationship between social studies scores and math scores?




The scale-location plot is horizontal so constant variance is not violated

The residual vs fitted graph is also horizontal, so mean trend is not violated.

The QQ plot shows that there are some outliers at the end, but most of the data forms a x=y pattern, so data is acceptable/normally distributed

All three conditions are met, so this model is a good fit i.e. there a linear relationship between social studies scores and math scores.

Part 6: Report the $R^2$ value of your model. What is the meaning of this value?

R2 = 0.296

The $R^2$ of a linear model describes the amount of variation in the response that is explained by the least squares line.

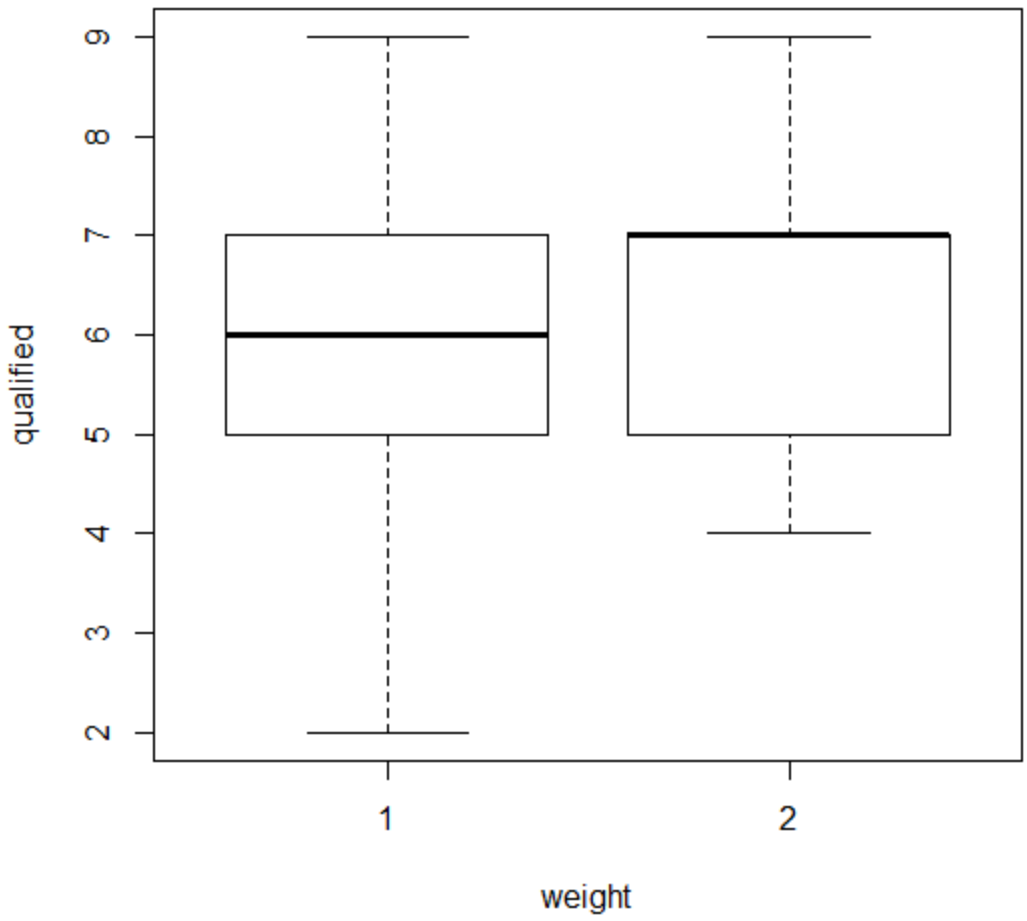
Part 7: Summarize the relationship between social studies scores and math scores in a paragraph that utilizes all of the numbers in the previous questions.

If there is an increase in math score, there is also an increase in social studies score. This can only mean that students do better in social studies than math i.e. math is harder than social studies. The residual plots, and the variance seem to confirm our previous hypothesis.

Question 3

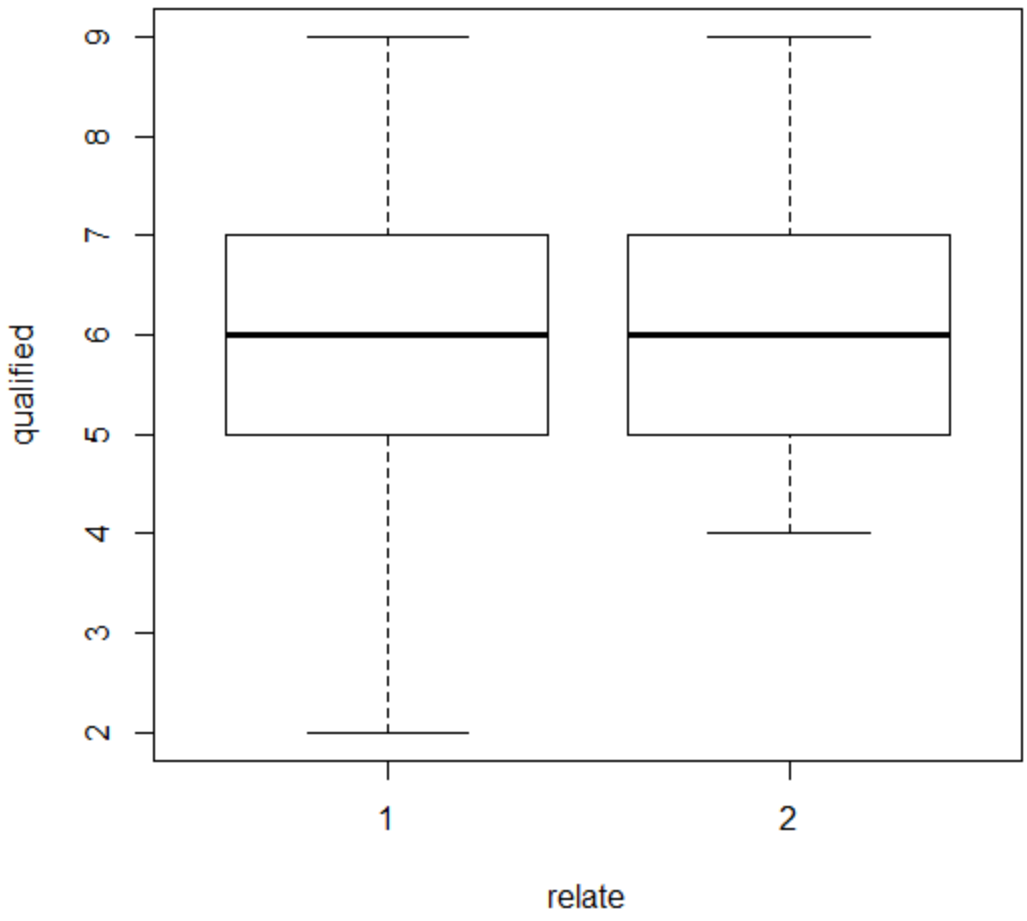
Part 1: Do a plot of qualified vs weight, qualified vs relate, and qualified vs weight:relate. For each plot, briefly describe what you see. Based on these plots alone, do you think sitting next to an obese woman has an adverse effect on the applicant's qualification ratings?

qualified vs weight



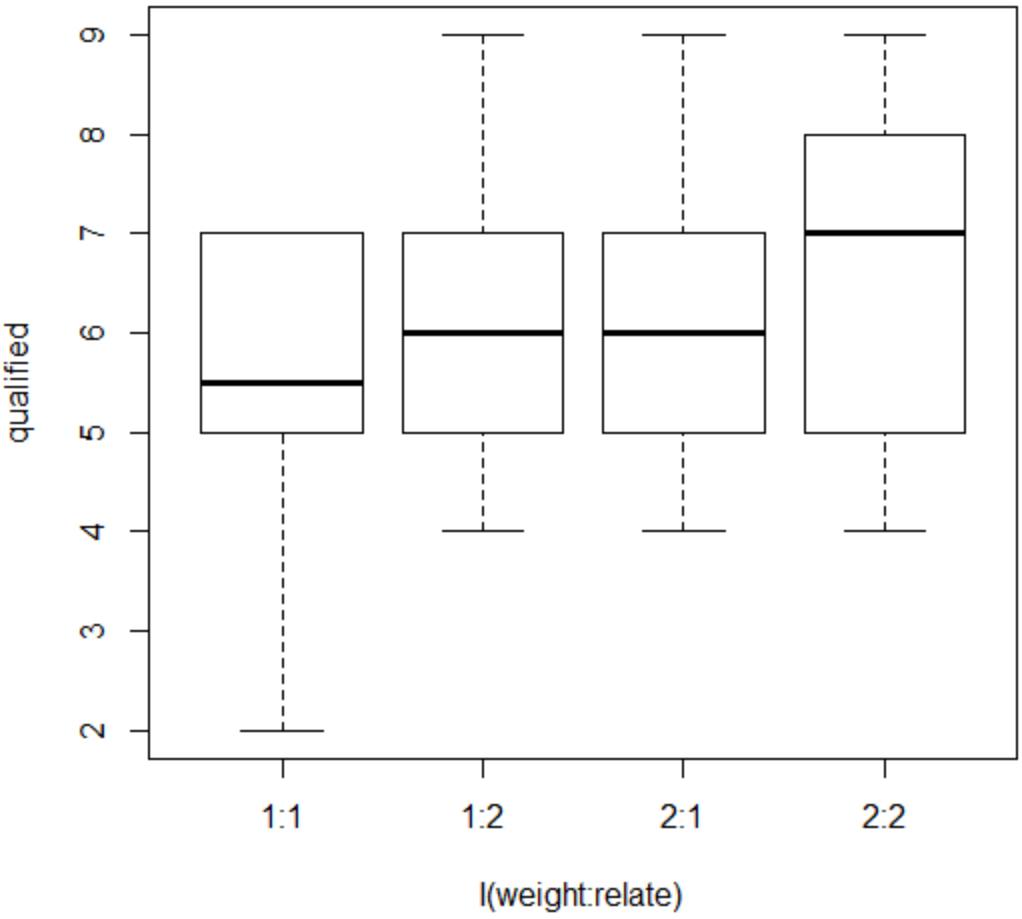

The plot suggests that average weighted women have a higher median and have a higher range of 4-9 i.e. are more likely to have a higher score than obese women. Obese women have lower median and have a range of 2-9.

qualified vs relate




Both group (girlfriend or acquaintance) have the same median score, it looks like if the candidate's partner is perceived as acquaintance then it is more likely they have a higher score, since obese women have a range of 2-9 while acquaintance women have score of 4-9.

qualified vs weight:relate




Applicant who had obese women perceived as girlfriend had the lowest median score and had the shortest qualification score range.

Applicant who had obese women perceived as acquaintance and average weighted women perceived as a girlfriend has similar outcome with both the median qualification score of 6, and range of 4-9.

Applicant who had an average weighted woman perceived as an acquaintance had the best outcome with range of 4-9 but median qualification score of 7.

Therefore, sitting next to an obese woman has an adverse effect on the applicant's qualification ratings unless she is perceived as an acquaintance.

Part 2: Run three regression models with the following definitions:

- weights.model1: qualified vs weight + relate.

- weights.model2: qualified vs weight:relate
- weights.model3: qualified vs weight + relate + weight:relate

For each model, clearly show the R command that you use, and include the R's model summary. Write down the equation that R gives you. Interpret all the coefficients and the p-values associated with the coefficients.

weights.model1: qualified vs weight + relate.

Equation: $Y = 5.67 + 0.49 * x_1 + 0.45 * x_2$

Meaning:

0.49 = average increase in qualification score of applicant, for average women compared to obese.

0.45 = average increase in qualification score of applicant, when women perceived as an acquaintance compared to women perceived as girlfriend

p-value for weight2 is not significant at 0.0137

p value for relate2: acquaintance is significant at 0.0238

This means beta1 is significantly different from 0, i.e. the data indicates that there is a significant difference in applicant's qualification score when women are average vs obese

beta2 is not significantly different from 0 (small p-value), i.e. there is significant difference in applicant's qualification score when average weight women are perceived as girlfriend vs obese girlfriend

weights.model2: qualified vs weight:relate

Equation: $Y = 6.5926 - 0.94 * x_1 – 0.4 * x_2 – 0.44 * x_3$

Meaning:

0.94 = average decrease in qualification score of applicant, when obese women perceived as acquaintance compared to girlfriend

0.4 = average decrease in qualification score of applicant, when average women perceived as a girlfriend compared to obese women perceived as acquaintance

0.44 = average decrease in qualification score of applicant, when average women perceived as acquaintance compared to average women perceived as a girlfriend

p-value for obese: acquaintance is not significant at 0.000632

p-value for average: girlfriend is significant at 0.133908

p value for average: acquaintance is significant at 0.103934

This means beta1 is not significantly different from 0, i.e. the data does not indicate strongly that there is a significant difference in applicant's qualification score when obese women are perceived as girlfriend vs acquaintance

beta2 is significantly different from 0 (small p-value), i.e. there is significant difference in applicant's qualification score when average weight women are perceived as girlfriend vs obese girlfriend

beta3 is significantly different from 0 (small p-value), i.e. there is significant difference in applicant's qualification score when average weight women are perceived as acquaintance vs obese girlfriend

weights.model3: qualified vs weight + relate + weight:relate

Equation: $Y = 5.65 + 0.54 * x_1 + 0.5 * x_2 – 0.097 * x_3$

0.54 = average increase in qualification score of applicant, for average women compared to obese.

0.50 = average increase in qualification score of applicant, when women perceived as an acquaintance compared to women perceived as girlfriend

-0.0978 = average decrease in qualification score of applicant, when obese women perceived as girlfriend compared to average weighted women perceived as an acquaintance

p-value for weight2: acquaintance is not significant at 0.0611

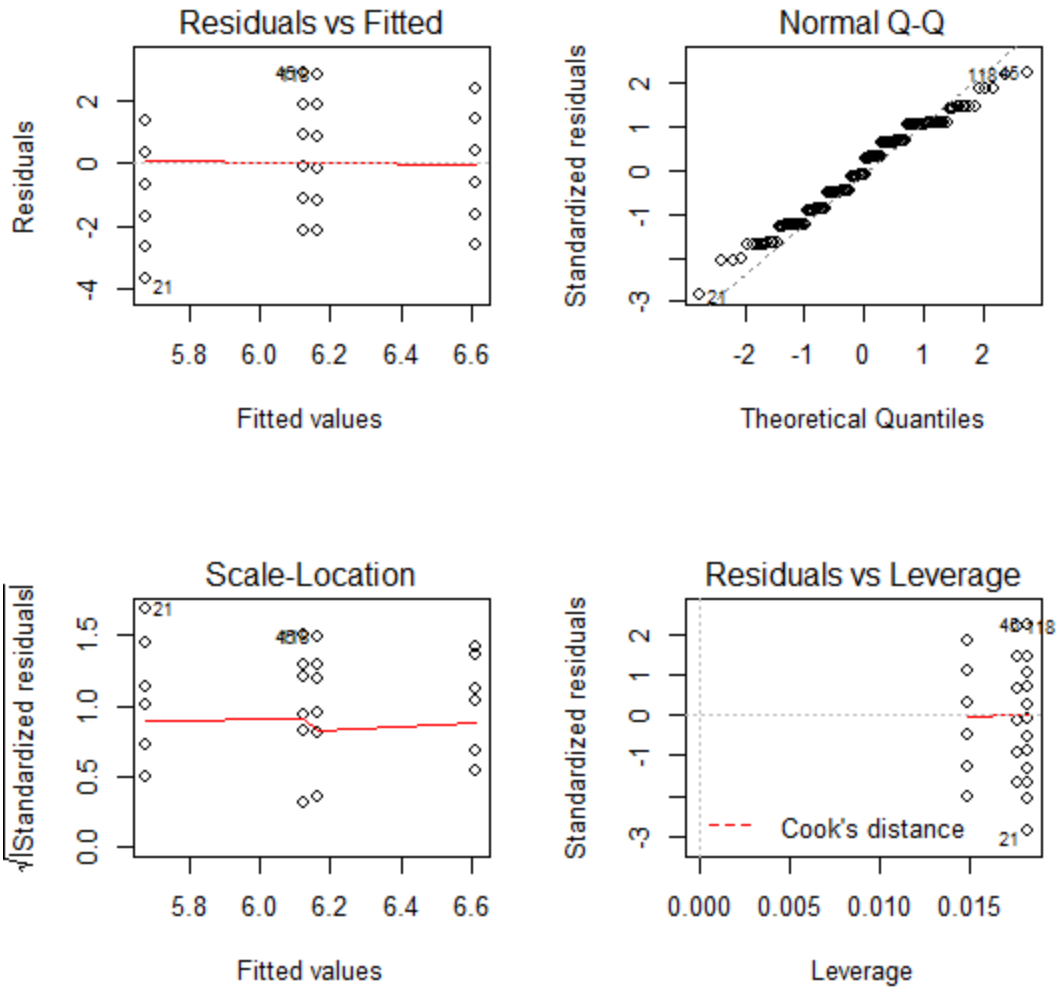
p-value for relate2: girlfriend is not significant at 0.0867

p value for weight2: relate2 is significant at 0.0.8043

Therefore beta3 is significantly different from 0 (small p-value), i.e. there is significant difference in applicant's qualification score when obese women perceived as girlfriend compared to average weighted women perceived as an acquaintance

Part 3: Do a diagnostic plot for each of your models. Say which, if any, of the (a) independence (no mean trend) (b) normal distribution and (c) constant variance assumptions are violated.

weights.model1




The scale-location plot is relatively straight, i.e. it is acceptable, and constant variance is not violated.

The residual vs fitted graph is straight, so mean trend is not violated.

The QQ plot shows that there are some outliers at the beginning and at the end, but the majority of the data forms a x=y pattern, so data is normally distributed

All three conditions are met, so Model 1 is a good fit

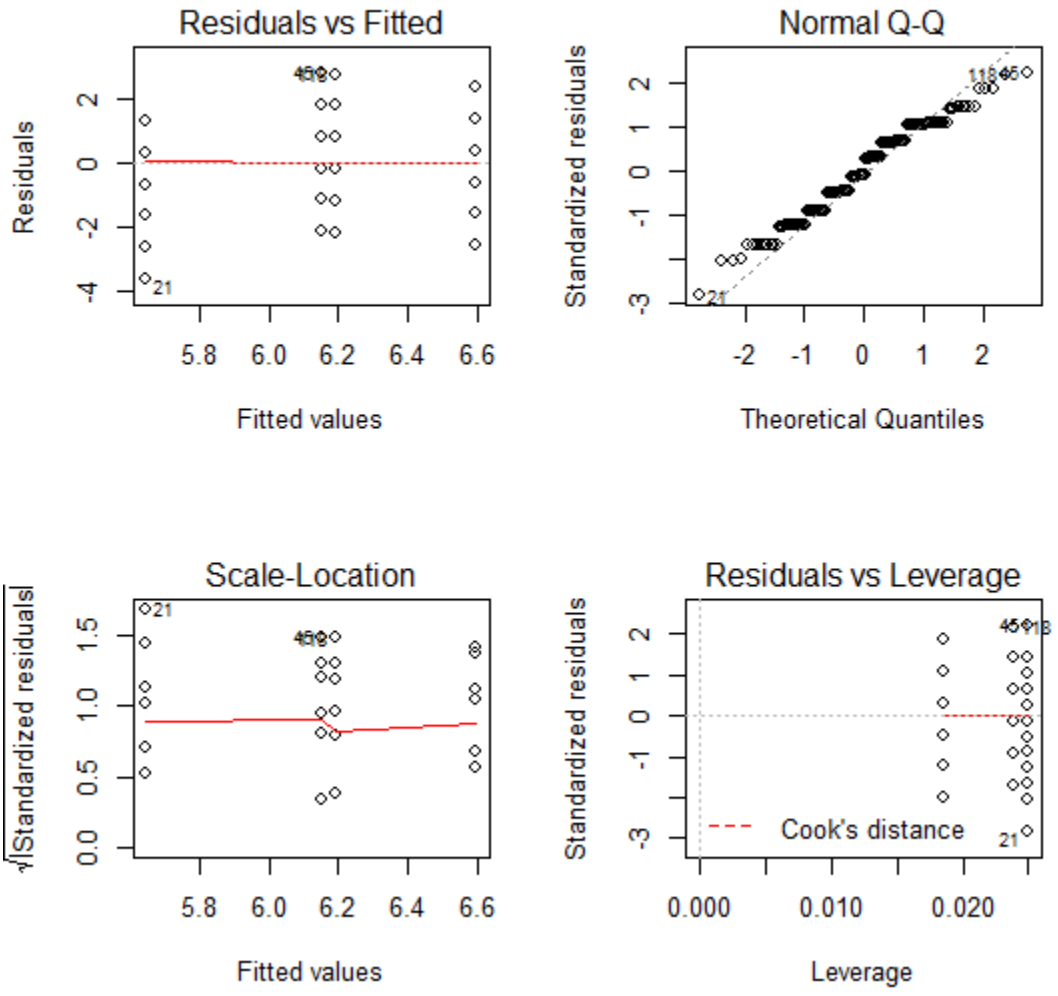
weights.model2

The scale-location plot is relatively straight, i.e. it is acceptable, and constant variance is not violated.

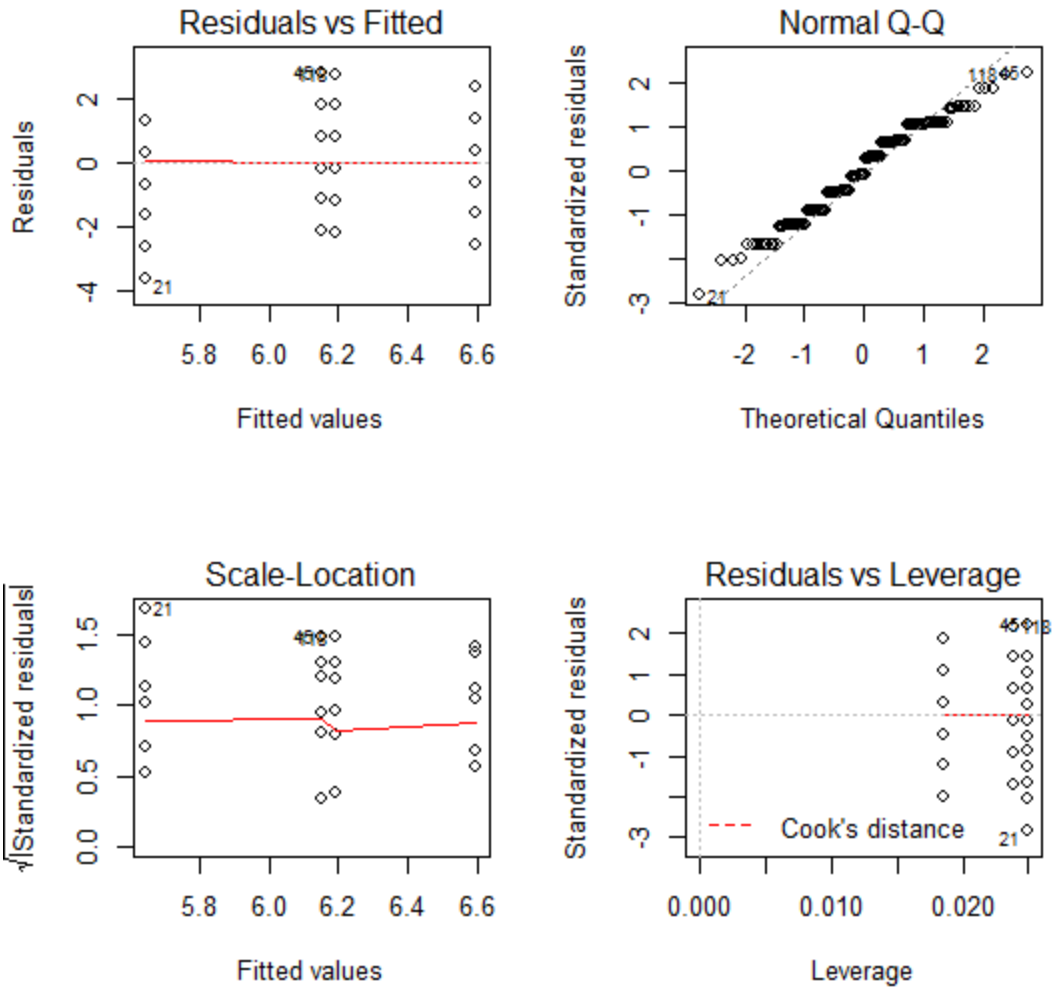
The residual vs fitted graph is straight, so mean trend is not violated.

The QQ plot shows that there are some outliers at the beginning and at the end, but the majority of the data forms a x=y pattern, so data is normally distributed

All three conditions are met, so Model 2 is a good fit




weights.model3




The scale-location plot is relatively straight, i.e. it is acceptable, and constant variance is not violated.

The residual vs fitted graph is straight, so mean trend is not violated.

The QQ plot shows that there are some outliers at the beginning and at the end, but the majority of the data forms a x=y pattern, so data is normally distributed

All three conditions are met, so Model 3 is a good fit

Part 4: Why do the coefficients estimates of weights.model3 differ from those in weights.model1 and weights.model2, even for the same variables?

The coefficients have different meanings in each graph. For example b3 in model 3 means:

average decrease in qualification score of applicant, when obese women perceived as girlfriend compared to average weighted women perceived as an acquaintance

In model 2, b3 means,

average decrease in qualification score of applicant, when average women perceived as acquaintance compared to average women perceived as a girlfriend

In model1, there is no b3.

Therefore, even though the variables are same, the models have different interpretability, hence the difference in coefficients.

Part 5: Between these three models, which one is the best? Justify your choice based on the diagnostics and the model's interpretability.

Weights.model1 would be the best. Even though all models are a good fit, model1 has a lower variance compared to both model 2 and model 3. It also has lesser variables, and a lower p-value than model 2 and model 3 which is better. Hence, it gives us the best model to predict the applicant's score between obese vs regular women, how they are perceived as girlfriends or acquaintance