Homework 7

Roly Vicaría
STAT501 Fall 2015

October 19, 2015

0.0.1 Question 1

a) FALSE. The null hypothesis in the Ryan-Joiner test is that the errors are normally distributed. So a small p-value would lead us to reject the null hypothesis and conclude they are NOT normally distributed.

b) FALSE. The confidence interval for the mean response can be valid even when the Normality requirement is not satisfied if there is a sufficiently large data set. The prediction interval, however, strongly relies on the normality requirement.

c) TRUE

d) TRUE

e) FALSE. We could also code the variable as 0, 1, 2, 3, 4.

0.0.2 Question 2

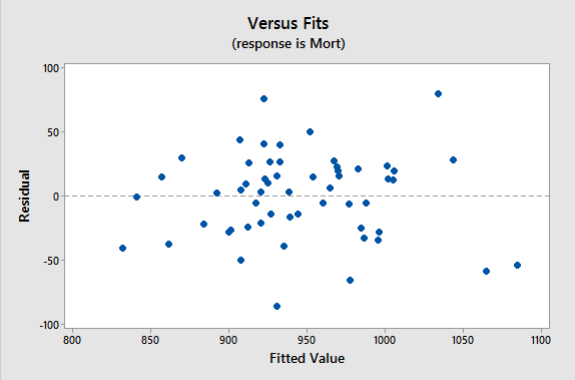
a) $Mort = 1006.2 - 15.35\ Edu + 4.214\ Nwt - 2.150\ Jant + 1.624\ Rain + 18.55\ Nox + 0.537\ Hum - 0.35\ Inc$

SSE = 60417; df = 48

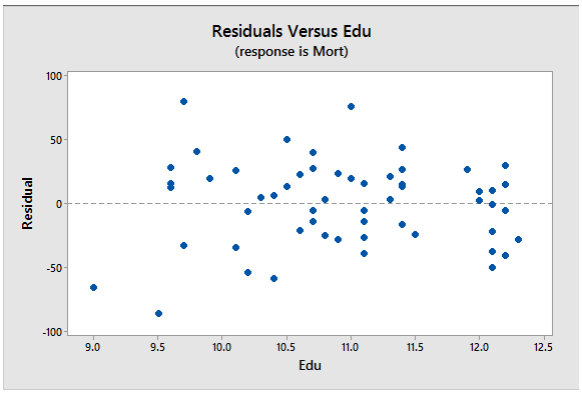
b) $F^* = \frac{SSE(Reduced) - SSE(Full)}{df_R - df_F} \div \frac{SSE(Full)}{df_F} = \frac{60948 - 60417}{50 - 48} \div \frac{60417}{48} = 0.211$

F distribution with 2 DF in numerator and 48 DF in denominator: $P(X > 0.211) = 1 - 0.189479 = 0.81$, therefore, we can conclude that the variables do not provide significant information about the response.

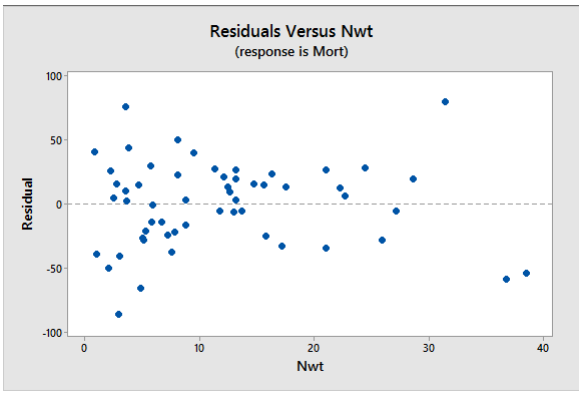
c) $Mort = 1028.2 - 15.59\ Edu + 4.181\ Nwt - 2.131\ Jant + 1.633\ Rain + 18.41\ Nox$



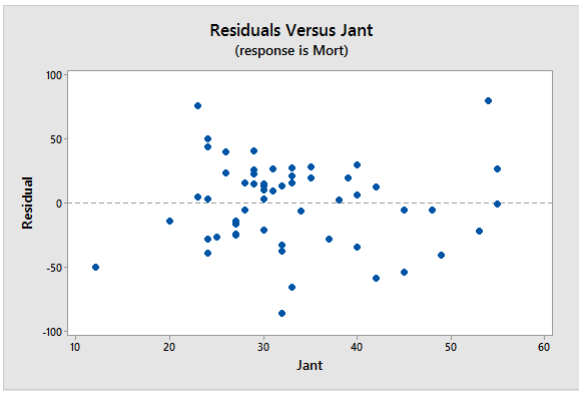
Scanning the scatterplot of residuals vs fitted values, it does appear that the vertical average remains close to 0. This affirms the **L** condition. Also, the vertical spread of the residuals remains fairly constant as we scan the plot from left to right, which affirms the **E** condition. There do not appear to be any outliers.



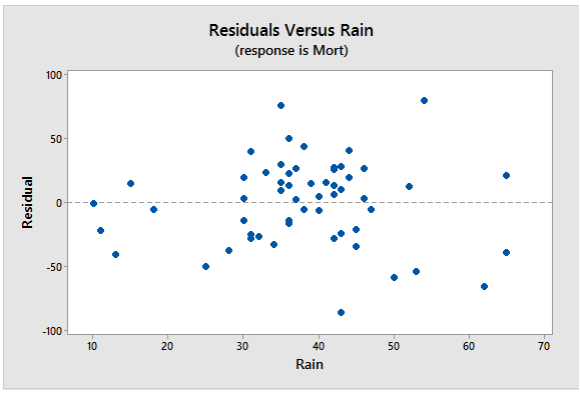
Looking at the scatterplot of residuals vs Edu, it appears that the vertical average of the residuals remains approximately 0 (affirming the **L** condition), and that the vertical spread of the residuals is fairly constant (affirming the **E** condition).



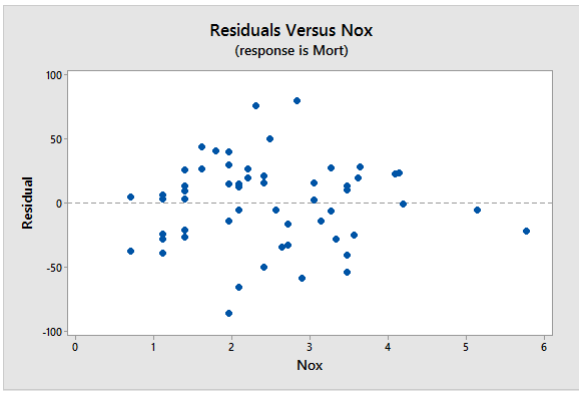
The scatterplot of residuals vs Nwt seems to affirm the **L** condition since the vertical average of the residuals appears close to 0, but the vertical spread of the residuals does not seem constant. The spread seems largest for smaller values of Nwt, less than 10.



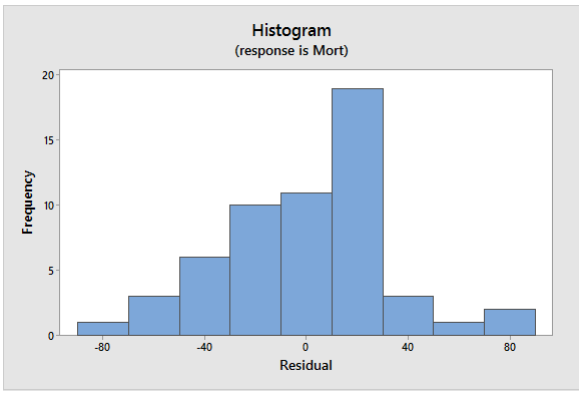
The scatterplot of residuals vs Jant looks OK. Vertical average of residuals seems close to 0 and spread seems fairly constant, affirming both **L** and **E** conditions.



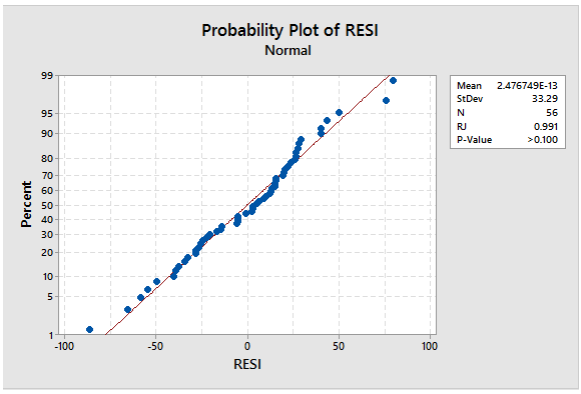
The residuals vs Rain plot also looks good. Both the **L** and **E** conditions appear satisfied.



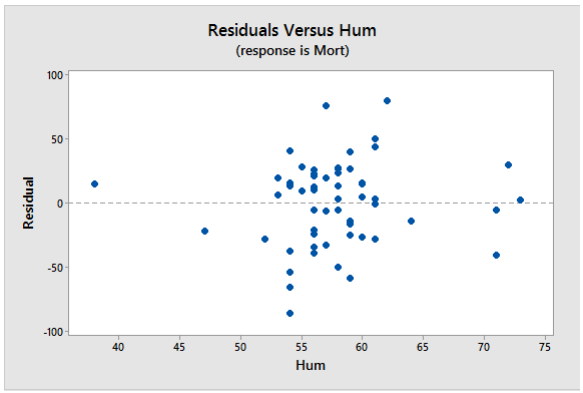
The residuals vs Nox plot looks OK. Vertical average of residuals seems close to 0 and the spreads seems fairly constant. A few more spread out values around Nox = 2, but I don't think it's enough to be suspicious.

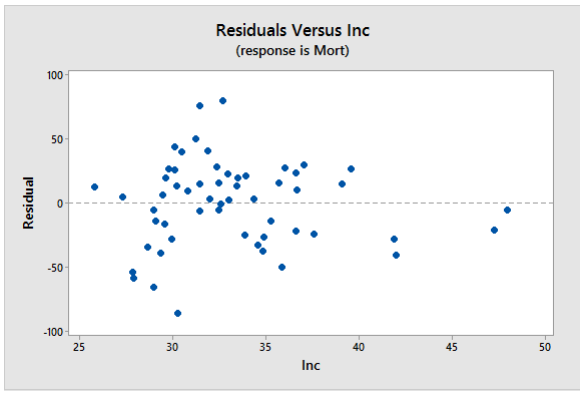


The histogram doesn't have a perfect bell shape, and looks slightly skewed, but not enough to confirm violation of normality.



The probability plot and Ryan-Joiner test support the normality condition of the errors. The test shows a p-value > 0.1 which says we fail to reject the null hypothesis and conclude that the errors are normally distributed.





The scatterplots of residuals vs Hum and Inc seem pretty unremarkable. They don't show any indication of strong linear or simple nonlinear trends. They don't appear to warrant being added to the model.

d) The 95% confidence interval for E(Mort) is (946.273, 978.945). This interval can be interpreted as saying that we can be 95% confident that the average mortality rate (age adjusted per 100,000 population) for all cities with 10 median years of education, 15 percent non-white population, 35 degrees Fahrenheit mean January temperature, 40 inches annual rainfall, and 2 units of log(nitrous oxide concentration in parts per billion), will be between 946.273 and 978.945.

e) The 95% prediction interval for Mort is (890.605, 1034.61). This interval can be interpreted as saying that we can be 95% confident that the mortality rate for a new city with 10 median years of education, 15 percent non-white population, 35 degrees Fahrenheit mean January temperature, 40 inches annual rainfall, and 2 units of log(nitrous oxide concentration in parts per billion), will be between 890.605 and 1034.61. This interval is wider than the confidence interval because it always has an extra MSE term than the confidence interval formula.

0.0.3 Question 3

The $\mathbf{X}$ matrix would be a 200 x 5 matrix where each row corresponds to a record from the data set. The columns are: 1) 1's which correspond to the β_0 constant term, 2) body weight, 3) gender (0=male,1=female), 4) age, and 5) interaction gender x age.

0.0.4 Question 4

a) At first glance, it looks like there is an interaction between degree and years of experience since it looks like the Masters line is not parallel to the other lines. However, upon further observation, I believe that the Masters data has an outlier that is a very influential point affecting the slope of that line. If we were to remove the data point at around 20 years of experience, I suspect that the 3 lines would be almost parallel. Which would indicate that there is no interaction between degree and years of experience.

b) $Salary = 40.77 + \ 2.158\ YrsExp$

i) SSE = 12412.8
ii) 61

c) $Salary = 29.27 + \ 1.841\ YrsExp + \ 28.36\ Deg1 + \ 10.71\ Deg2$

i) SSE = 4948
ii) 59

d) $F^* = \frac{SSE(Reduced) - SSE(Full)}{df_R - df_F} \div \frac{SSE(Full)}{df_F} = \frac{12412.8 - 4948}{61 - 59} \div \frac{4948}{59} = 44.505$

e) i) Df numerator=2, Df denominator=59
ii) p-value = 0.000 The conclusion we can draw from this is that the variables deg1 and deg2 provide significant information about the response.

f) i) $Salary = 29.27 + \ 1.841\ YrsExp + \ 28.36\ Deg1 + \ 10.71\ Deg2$
ii) $Salary = 29.27 + 1.841 YrsExp$ (Deg1 = 0, Deg2 = 0)
iii) $Salary = (29.27 + 10.71) YrsExp$ (Deg1 = 0, Deg2 = 1)
iv) $Salary = (29.27 + 28.36) YrsExp$ (Deg1 = 1, Deg2 = 0)

g) i) The coefficient that multiplies the variable Deg1 represents the difference in mean salary for employees with PhD's versus employees with Bachelor's degrees, assuming the same value for years of experience.
ii) The coefficient that multiplies the variable Deg2 represents the difference in mean salary for employees with Master's degrees versus employees with Bachelor's degrees, assuming the same value for the years of experience.

h) One way to do it could be to have indicator variables for managers with Bachelor's and one indicator variable for managers with Master's such that when both are equal to 0 denotes a manager with a PhD. The coefficient of interest would then be the coefficient that multiplies the variable for Master's degree managers.

i) $Salary = 26.94 + \ 2.206\ YrsExp + \ 31.26\ Deg1 + \ 13.73\ Deg2 - \ 0.427\ YrsExp * Deg1 - \ 0.471\ YrsExp * Deg2$

$H_0 : \beta_4 = \beta_5 = 0$, H_a : at least one of β_4, β_5 is not equal to 0

Test statistic: $F^* = \frac{SSE(Reduced) - SSE(Full)}{df_R - df_F} \div \frac{SSE(Full)}{df_F} = \frac{4948 - 4867.1}{59 - 57} \div \frac{4867.1}{57} = 0.474$
This test statistic has a p-value of 1 - 0.37506 = 0.625. Therefore we cannot reject the null hypothesis and conclude that there is no interaction between highest degree attained and years of experience.

0.0.5 Question 5

a) Only (iii) is true. The salary regression lines are:

$$Salary_{male} = 50 + 20GPA + .07IQ + .01GPA * IQ$$
$$Salary_{fem} = 50 + 20GPA + .07IQ + 35 + .01GPA * IQ - 10GPA$$

We can see from the equations that female salary is greater as long as GPA is less than 3.5. After 3.5, then male salaries are greater.

b) 137.1

c) I agree there is reason to doubt the interaction between GPA and IQ since it is such a small value, but we would need to run a t-test to comment further.