Homework 4

Roly Vicaría
STAT501 Fall 2015

September 20, 2015

0.0.1 Question 1

a) True. This means that the data points were evenly distributed above and below the line.

b) False. An ideal residual plot for a valid model should display residuals randomly scattered around the residual = 0 line.

c) True. We would want to see a horizontal "band" around the residual = 0 line to show that error variances are equal.

d) False. We shold confirm the linearity and equal variance conditions first since those conditions can affect the normality.

e) False. It could be useful in determining whether or not to add another predictor to the model.

f) False. We may have independence and equal variance of error terms but still not fit the linear model.

g) False. We should question our model if any of the conditions seem in doubt. We may determine that the doubt is not significant in a given context, but that determination still needs to be made.

h) False. It is very subjective and needs to be considered in the context of the purpose for which the model was constructued.

0.0.2 Question 2

a) Based on the residuals plot, we can see that the residuals do not "bounce randomly" around the residual = 0 line. Instead it has positive values for smaller fitted values, close to 0. This calls into question the assumption that the relationship is linear. Also, the residuals don't form a "horizontal band" around the residual = 0 line, which indicates that the variances of the error terms may not be equal.

b) Yes. Adding the new predictor has improved the residual plot. It shows more of "random bounciness" around the residual = 0 line. It's not ideal, but better than before.

0.0.3 Question 3

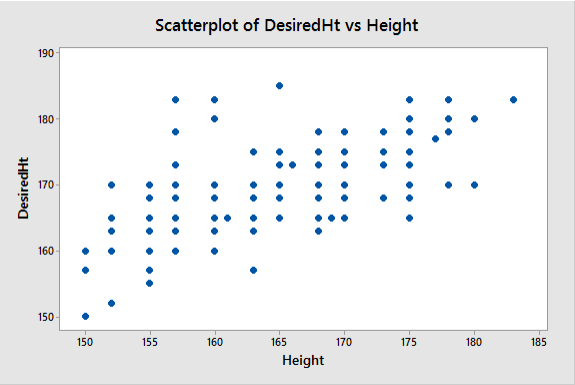
In terms of linearity, it looks like the Japan data has the most "random bouncing" of points around the residual = 0 line. The USA data seems to have negative residuals for lower fitted values, then goes positive for middle fitted values, and then scatters for the higher values. That indicates that the USA data may not be linear. The Europe data seems moderately scattered around the residual = 0 line.

In terms of the response following normal distribution, the USA data has the most normal histogram for the residuals. The Europe data looks skewed right, and the Japan data looks slightly skewed left.

In terms of the errors having equal variances, I think the USA data seems to be the most consistently distributed. The Japan data shows a little bit of a "megaphone" shape which indicates that the variance increases relative to the fitted value. The Europe data is not far off, but has some areas of greater spread than others.

0.0.4 Question 4

a) The data of height versus desired height appears to show a positive relationship between the two variables. As height increases, desired height appears to increase also.

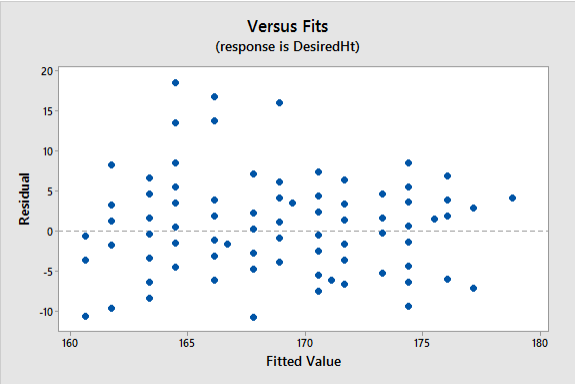


b) $DesiredHt = 77.84 + \ 0.5518\ Height$

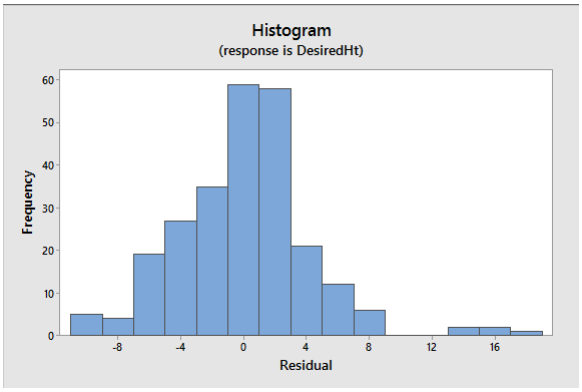
c) The R^2 value calculated by Minitab is 44.26%. This value can be interpreted as saying that about 44% of the variance in desired height is "explained" by the variation in actual height.

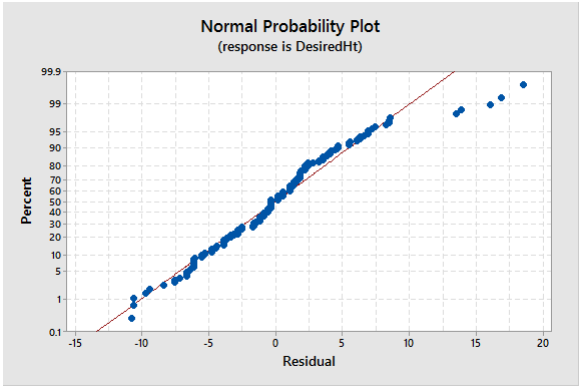
d) Some evidence supporting that the relationship is statistically significant is the t-test for b_1. The t-value is 14.06 which has a p-value of 0.000. This is strong evidence that the desired height is positively correlated with actual height with slope b_1.

e) The plot of residuals vs fitted values shows that the values appear to "randomly bounce" around the residual = 0 line. This supports the fit of the data to a linear model. The residuals also seem to distribute evenly around the residual = 0 line forming a "horizontal band" which indicates that the variance of the error terms is consistent. The only exception to the horizontal band are the appearance of a few potential outliers, 5 points with residual value greater than 10.



f) The histogram confirms what we saw in the previous plot, namely that the residuals are fairly normal with only a few potential outliers with value greater than 12. The probability plot conveys this same observation. The probability plot shows the tail to the right. It's hard to say if this is a small departure from normality or a major departure. Aside from the 5 values on the right, it is a fairly straight line which indicates normality. It may not create significant problems.





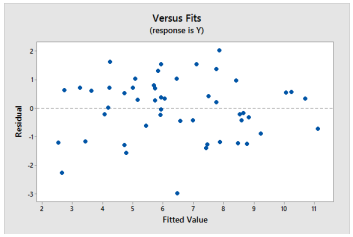
0.0.5 Question 5

a)

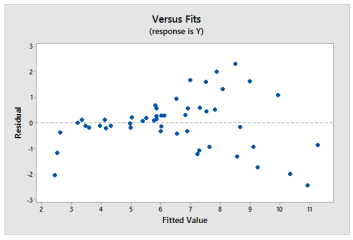
Predictor variable	S-value	R^2 value	Slope t-statistic
X_1	1.06996	80.65%	14.14
X_2	1.01855	82.46%	15.02
X_3	1.01873	82.46%	15.02
X_4	1.01878	82.45%	15.02

Based solely on these values, I would say that the models for Y vs X_2, X_3, X_4 are essentially tied for "best", followed by X_1

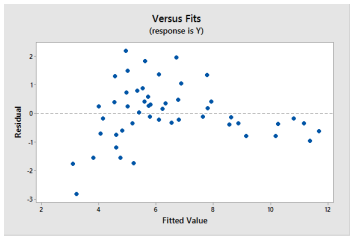
b) The residual plot for model with X_1 shows a strong "horizontal band" shape around the residual = 0 line, indicating good linearity fit. It shows a potential outlier with residual value around -3. The residuals seems to show a pretty constant variance around the residual = 0 line.



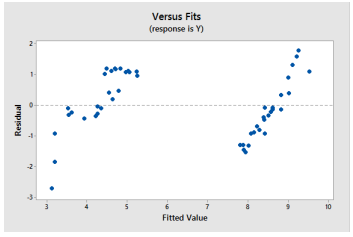
c) The residual plot for model with X_2 shows more of a "megaphone" shape aronud the residual = 0 line. It also has some negative residuals for smaller fitted values. It calls into question the linear fit of the model. It also shows that the error variance grows for larger fitted values.



d) The residual plot for the model with X_3 also shows a possible lack of linear fit for the model. It shows some negative residuals for smaller fitted values, followed by a cluster of positive values, followed by a tail of more negative values. The variance of the error terms also does not look consistent.



e) The residual plot for model with X_4 also shows lack of linear fit for the model. It shows two "clusters" of data points, the first with smaller fitted values, and another for larger fitted values. It appears that there may be another predictor value missing from the model to help explain Y.



f) Based on the residual vs fitted plots above, it would appear that the model with X_1 would be the best one. It shows the best linear fit and most consistent variance for error terms with minimal number of outliers.