**STAT 501 – Homework Solutions 4 – Fall 2015**

1. (**8x3 = 24 points**) Say whether the following statements about residual analysis for simple linear regression are true or false and explain your answers.
   1. If the sample mean of the residuals is zero, then this supports the linearity condition for the model.

**False, since the sample mean of the residuals is always zero, regardless of the validity of the model.**

* 1. An ideal residual plot for a valid model will display residuals with a strong positive or negative linear trend.

**False, since an ideal residual plot for a valid model will display residuals that fall within a horizontal band centered around 0, displaying no systematic tendencies to be positive or negative.**

* 1. An ideal residual plot for a valid model will display residuals with similar variation no matter the value on the horizontal axis.

**True, since the equal variance condition means that residuals should have similar variation no matter the value on the horizontal axis.**

* 1. We should only assess the linearity and equal variance conditions after first confirming the normality condition using a normal probability plot.

**False, since residuals may appear to be not normally distributed due to failure of the linearity or equal variance conditions, so these conditions should be investigated first.**

* 1. A residual plot that has a predictor variable excluded from the model on the horizontal axis provides no useful information about the adequacy of the model.

**False, since a residual plot that has a predictor variable excluded from the model on the horizontal axis can indicate whether the model could be improved materially by adding the predictor variable in question.**

* 1. We can accept the validity of our model if at least one of the four LINE conditions is supported.

**False, since we can accept the validity of our model only if all four LINE conditions are supported.**

* 1. We should question the validity of our model only if all four LINE conditions seem in doubt.

**False, since we should question the validity of our model if at least one of the four LINE conditions seems in doubt.**

* 1. Residual analysis is entirely objective so that every assumption about the regression errors can be unambiguously determined.

**False, since residual analysis requires a subjective assessment of graphics.**

1. **(2x4 = 8 points)** The dataset “Auto” gives information on 392 different cars including their mpg, weight, horsepower, displacement, year of make, country of origin etc. A car buyer is interested in regressing mpg (Y) on horsepower (X). Given below is the residual plot of the regression.



1. What departure, if any, from the simple linear regression model assumptions can be assessed from this plot?

**The residual plot indicates nonlinearity. There is distinct departure from linear model assumption with residuals on both sides being positive and negatives in the middle.**



1. The residuals are now regressed on another predictor weight. Residual plot of this regression is given below. Would you say that inclusion of weight as an additional predictor has improved the residual plot? Briefly justify your answer.

**Inclusion of weight seems to have improved the residual plot. The residual plot is not equally distributed on both sides of the 0-line and does not form a band. But the degree of departure from non-linearity is less compared to part (a)**

1. **(8 points)** “Auto” data includes cars from three different origins: 1 (USA), 2 (Europe) and 3 (Japan). The variable *acceleration* (time to accelerate) is regressed on the variable *displacement*, separately for the cars from different origins. Following are the residual plots from three separate regression equations.

|  |
| --- |
|  |
|  |
|  |

Compare the residual plots for the cars of three different origins in terms of (i) whether linearity holds (ii) whether the response follows normal distribution and (iii) whether the errors have equal variances.

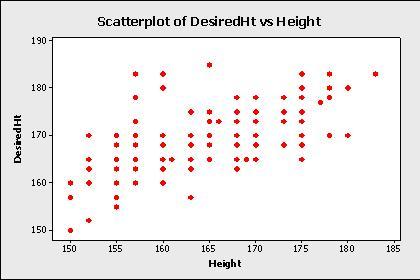
**For Origin=1 (US cars) residuals seem to follow a normal distribution. The residual histogram is more or less symmetric. Normal Q-Q plot shows mild departure for large residual values. However, the linearity assumption does not seem to hold. There does not seem to be marked departure from equal variance assumption.**

**For Origin=2 (European cars) residuals do not follow normal distribution. The residual histogram is skewed. The Q-Q plot shows marked departure from normality. However, the residuals do not show marked non-linearity, neither there is evidence for error variances not to be equal.**

**For Origin=2 (Japanese cars) residuals do not follow normal distribution. The residual histogram is negatively skewed. Moreover, the residuals show signs of inequality of variances. With larger fitted values, residuals have higher degree of dispersion.**

1. (**6x5 = 30 points**) Consider the “Desired Height” dataset. The data are from females in two introductory statistics classes at Penn State. The variables are *x* = Height (student's height in centimeters) and *y* = DesiredHt (how tall the student said they would like to be, in centimeters, if they could be any height at all). (Note: Students gave responses in inches, but we converted them to centimeters.)
   1. Draw a scatterplot of *y* = DesiredHt versus *x* = Height. Include the graph and write a brief description of the graph's main features in the context of simle linear regression. [In Minitab, the menu sequence is Graph > Scatterplot. To copy a Minitab graph, right click on it and use “Copy Graph” in the resulting menu.]

**There’s a moderately strong positive association. The plot looks like a straight line may work as a model. A few students may see a slight curvature - maybe that’s so but it definitely is slight. There may be a few outliers in the 155 to 165 cm height range. That is hard to judge from this plot.**



**Scatterplot of the Desired Height data**

* 1. Determine the estimated sample regression equation for a straight-line relationship between desired height and actual height. [In Minitab, use the menu sequence Stat > Regression > Regression > Fit Regression Model (v17, exclude last part in v16).]

**Let *y* = DesiredHt and *x* = Height. Then the equation is *ŷi* = 77.8 + 0.552*xi*.**

* 1. The output that you created for the previous part will include a value of *R*2. What is that value? Write a sentence that interprets the value in the context of this situation.

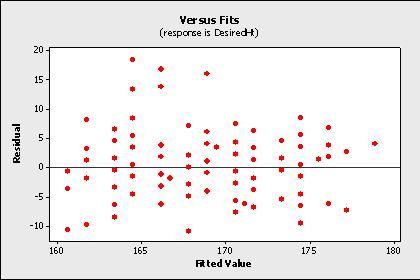
***R*2 = 44.3%, which means that Height (or, the model) “explains” (or accounts for) 44.3% of the observed variation in DesiredHt.**

* 1. In the output that you created, find evidence that the observed relationship is statistically significant. Describe the evidence that you found (for instance, the test for . . . had a *p*-value of . . ., which means . . . .)

**Two answers are possible - You can either use the *p*-value from the Height row in the report of *t*-tests for coefficients or you can use the *p*-value from the Analysis of Variance. From either, the *p-*value is 0.000, which means that the observed relationship is unlikely to be due to chance and is statistically significant.**

* 1. Create a plot of residuals versus predicted values (fits) for this situation. Include the plot and briefly discuss what the plot indicates about the validity of the model and assumptions about the errors. [In Minitab, click the Graphs button in the Regression dialog and then select “Residuals versus Fits.”]

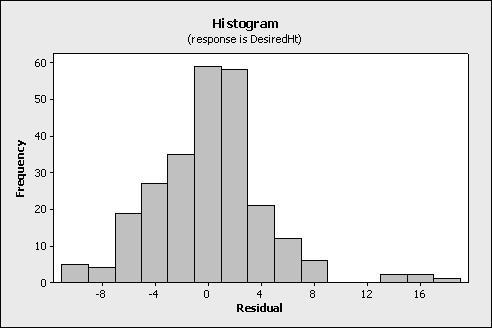
**The residual plot below generally supports the assumptions of the model. There is no curvature and the variance looks to be constant. There may be evidence of a few outliers in the data.**



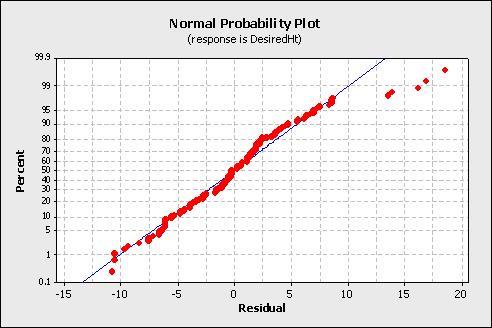
**Residual plot for the Desired Height data.**

* 1. Create a histogram of the residuals and a normal probability plot of the residuals. Include the plots and briefly discuss what they indicate about the validity of the model and assumptions about the errors. [In Minitab, click the Graphs button in the Regression dialog and then select “Histogram of residuals” and also “Normal probability plot of residuals.”]

**The bulk of the residuals seem to support the assumption of normality of errors, although there are distinct outliers on the high side (see the two plots below).**



**Histogram of residuals for the Desired Height data**

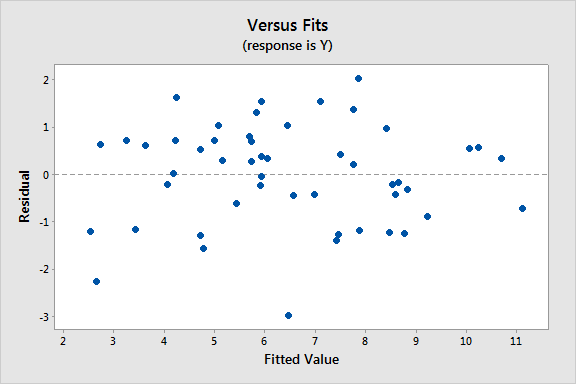


1. (**6x5 = 30 points**) Consider the “Compare” dataset. There is a single response variable, *Y*, and four potential predictor variables, *X1*, *X2*, *X3*, and *X4*. Fit four simple linear regression models, with each of the predictor variables, *X1*, *X2*, *X3*, and *X4* in turn. For each model create a residual plot with residuals on the vertical axis and fitted values on the horizontal axis.
   1. Compare the values of *S* (root mean square error), *R2*, and the slope *t*-statistic for each model (the clearest way to present this information is in a table). *Based solely on these values*, order the four models from “best” to “worst” (in terms of predicting or explaining *Y*). If two or more models appear to be essentially equivalent based on these criteria, make a note of this.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model predictor** | ***S*** | ***R2*** | ***t*-stat** |
| ***X1*** | **1.06996** | **80.65%** | **14.14** |
| ***X2*** | **1.01855** | **82.46%** | **15.02** |
| ***X3*** | **1.01873** | **82.46%** | **15.02** |
| ***X4*** | **1.01878** | **82.45%** | **15.02** |

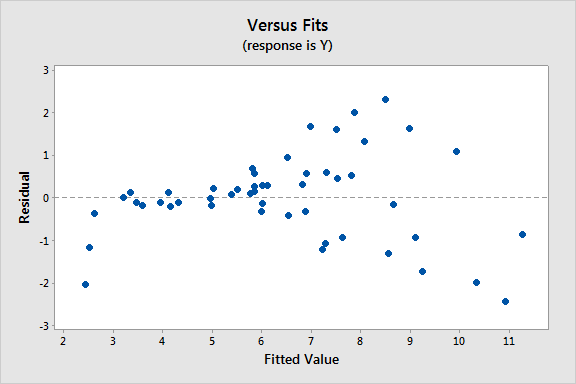
**The models with *X2*, *X3*, and *X4* all appear to be essentially equivalent based on this table, with the model with *X1* the “worst.”**

* 1. What does the residual plot for the model with *X1* as the predictor indicate about the validity of this regression model and assumptions made about the errors?



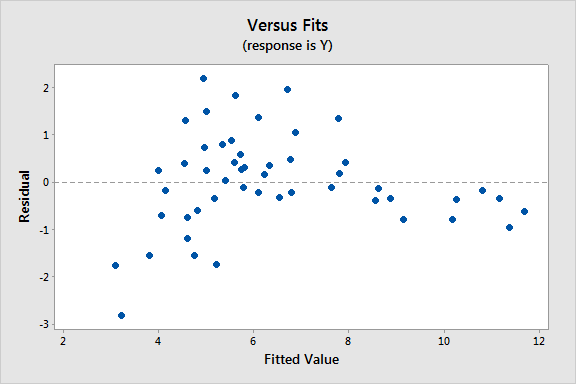
**The residual plot above generally supports the assumptions of the model. There is no obvious curvature and the variance looks to be constant. There is no evidence of outliers and the residuals appear to be independent.**

* 1. What does the residual plot for the model with *X2* as the predictor indicate about the validity of this regression model and assumptions made about the errors?



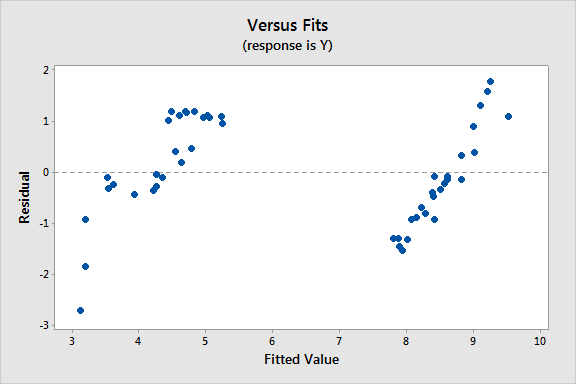
**The residual plot above does not support the assumptions of the model. Specifically, there is a strong indication of increasing variance from left to right.**

* 1. What does the residual plot for the model with *X3* as the predictor indicate about the validity of this regression model and assumptions made about the errors?



**The residual plot above does not support the assumptions of the model. Specifically, there is a strong indication of curvature.**

* 1. What does the residual plot for the model with *X4* as the predictor indicate about the validity of this regression model and assumptions made about the errors?



**The residual plot above does not support the assumptions of the model. Specifically, there is a strong indication of dependence.**

* 1. Given your answers for parts (b) to (e), which of the four models now appears to be “best?” Explain your answer.

**The “best” model is the one with *X1* since this is the only model that supports the assumptions of a simple linear regression model.**