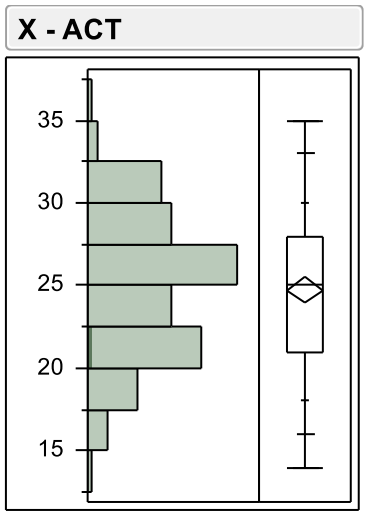
STAT 696
Winter 2011
HW #3
Major Jason Freels

3.3) Refer to GRADE POINT AVERAGE problem 1.19.

a) Prepare a box plot for the ACT scores X_i. Are there any noteworthy features in this plot?



This plot shows that the median value represents a score of 25 and the middle 50% of the observations fall between 28 and 21. Additionally, from visual inspection of the histogram the data appears to be follow a normal distribution.

b) Prepare a stem and leaf plot for the residuals. What information does this plot provide?

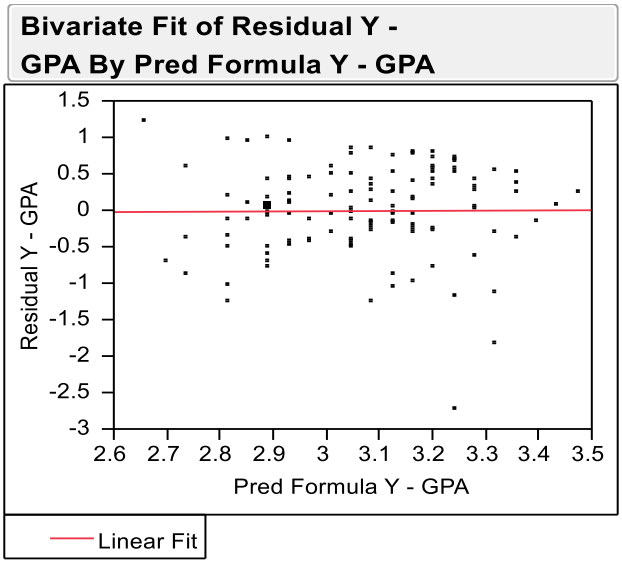
Stem and Leaf

Stem	Leaf	Count
1	0002	4
0	55555566666677777888889	24
0	11111112222223333334444444444	31
-0	444444433333332222211111111110000000000	38
-0	99887766655555	14
-1	2221000	7
-1	8	1
-2		
-2	7	1
-3		

-2|7 represents -2.7

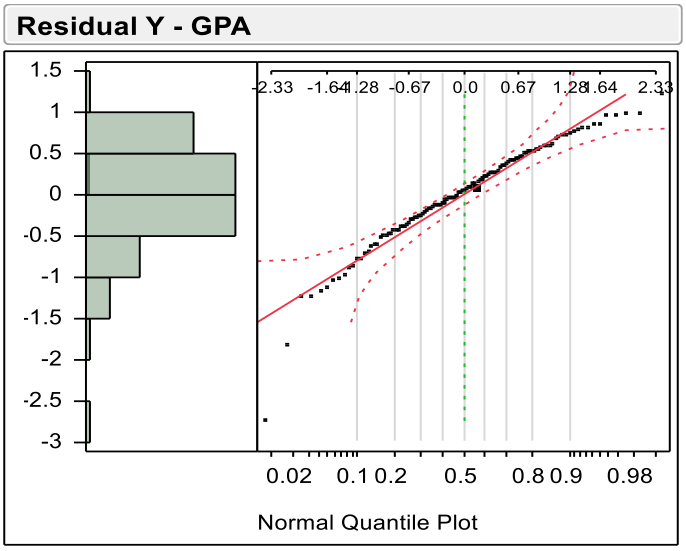
- This plot shows that the majority the residuals lie between 10 and -10, and that their distribution is slightly skewed to the right. The skewness may be the result of an outlier (-2.7). Additional plots with help make this determination.

c) Plot the residual e_i against the fitted values Y-hat$_i$. What departures from the regression model (2.1) can be studied from this plot? What are your findings?



- From this plot we can determine if any non-linearity exists within the regression function, if the variance of the error terms is consistent and if outliers exist
- The majority of the residuals lie above the predicted Y-hat values. Additionally, the variance in the residuals above the predicted line appears to be reduced when compared to the level of variance among those observations lying below the prediction line. The -2.7 residual value once again appears to be an outlier

d) Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Test the reasonableness of the normality assumption here using Table B.6 and $\alpha = .05$. What do you conclude?

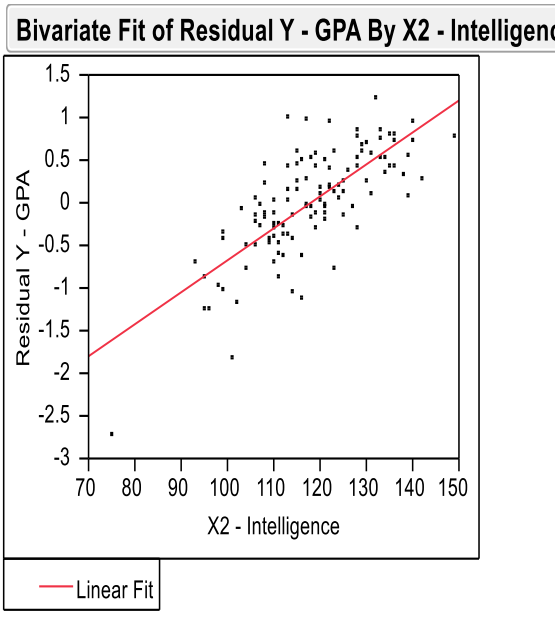


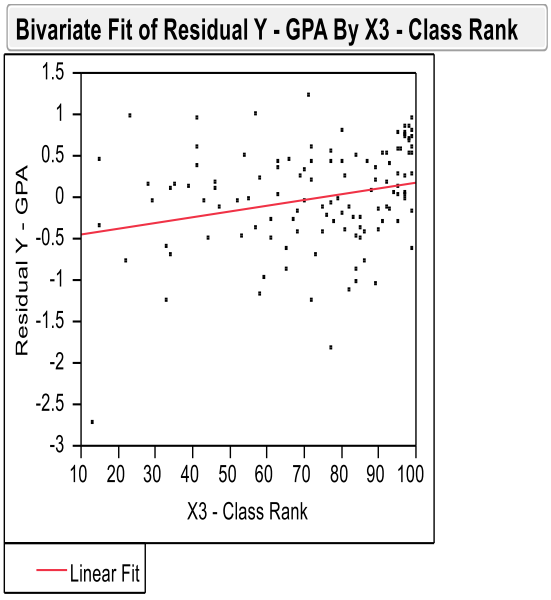
- R^2 calculated to be .9720, value obtained from Table B.6 = .987. Since .9720 < .987 we reject H_o and conclude that the data is not normally distributed. But since the calculated R^2 is high and is close to the Table B.6 and because this test is not as deterministic as some others we may be able to continue under the assumption that the residuals are normally distributed.

e) Conduct the Brown Forsythe test to determine whether or not the error variance varies with the level of X. Divide the data into the two groups, X < 26, X ≥ 26, and use α =.01. State the decision rule and conclusion. Does your conclusion support your preliminary findings in part (c)?

- The data set includes 65 entries with ACT scores below 26 and 55 entries with scores equal to or greater than 26.
- $s^2 = \frac{7.218+13.399}{118} = .174 \quad so \quad s = .418 \qquad and\ t^*_{BF} = \frac{.4380-.507}{.418\sqrt{\frac{1}{65}+\frac{1}{55}}} = 0.897$
- From Table B.2 using a two sided test with α =.01 gives t(.995, 120) = 2.617
- If ABS(t*) ≤ 2.617: conclude the error variance is constant
- If ABS(t*) > 2.617: conclude the error variance is not constant
- Since ABS(t*) is calculated at .897 < 2.617, we conclude that the error variance is constant

f) Information is given below for each student on two variables not included in the model, namely, intelligence test score (X_2) and high school class rand percentile (X_3). (Note that larger class rank percentiles indicate higher standing in the class, e.g., 1% is near the bottom of the class and 99% is near the top of the class.) Plot the residuals against X_2 and X_3 on separate graphs to ascertain whether the model can be improved by including either of these variables. What do you conclude?

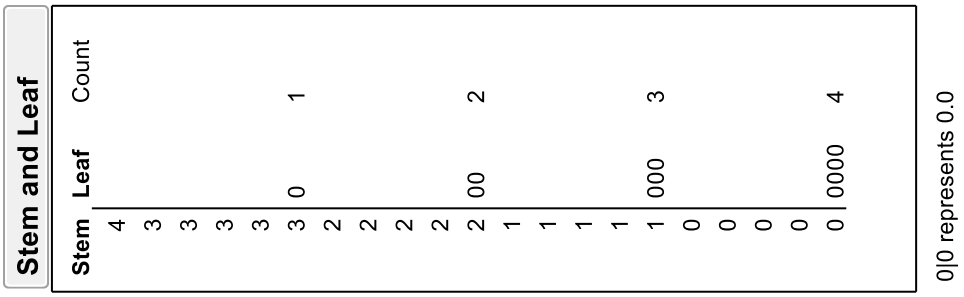




The slopes of the linear fit lines in the plots above show that intelligence has an effect on Freshman GPA and should be incorporated into the regression model, and class rank does not impact on GPA and should not be incorporated into our model. These results are as expected based on our knowledge of reality.

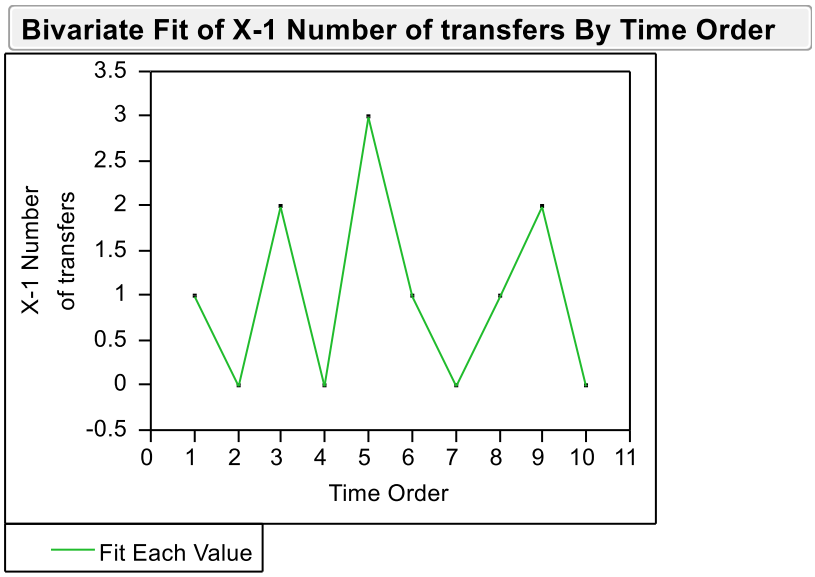
3.5) Refer to AIRCRAFT BREAKAGE problem 1.21

a) Prepare a stem and leaf plot for the number of tranfers X_i. Does the distribution of number of transfers appear to asymmetrical?



- Yes the data shows a skew toward no breakages and thus the distribution is asymmetrical.

b) The cases are given in time order. Prepare a time plot for the number of transfers. Is any systematic pattern evident in your plot? Discuss.



- The plot does not show any pattern with respect to time, with this limited amount of data.

c) Obtain the residuals e_i and prepare a stem and leaf plot of the residuals. What information is provided by your plot?

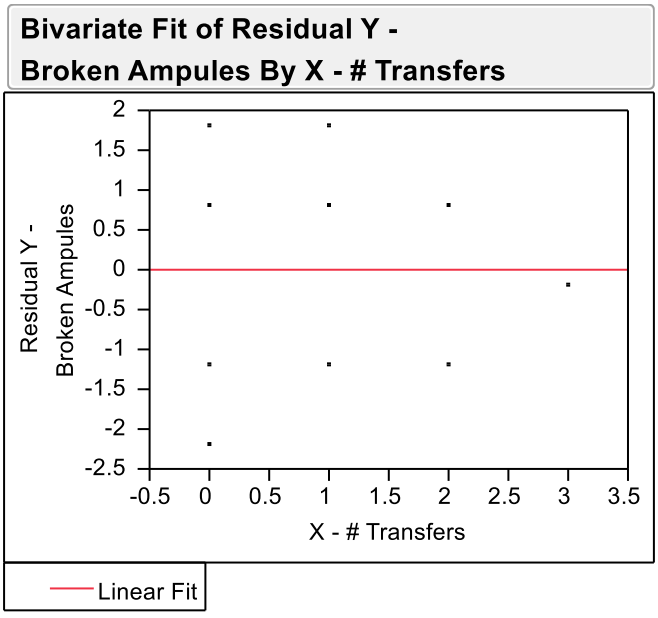
Stem and Leaf

Stem	Leaf	Count
1	88	2
1		
0	888	3
0		
-0	2	1
-0		
-1	222	3
-1		
-2	2	1
-2		

-2|2 represents -2.2

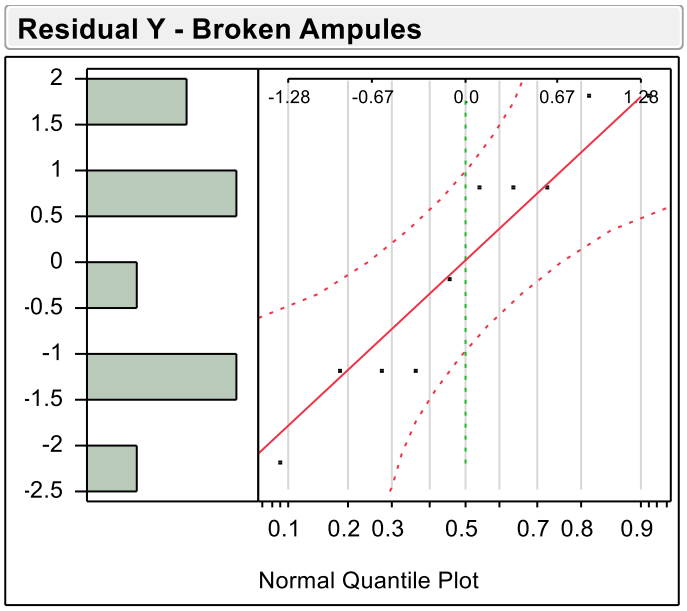
- While the data appears to be symmetric, it does not appear to be normally distributed due to the valley near the mean 0.4

d) Plot the residuals e_i against X_i to ascertain whether any departures from the regression model (2.1) are evident. What is your conclusion?



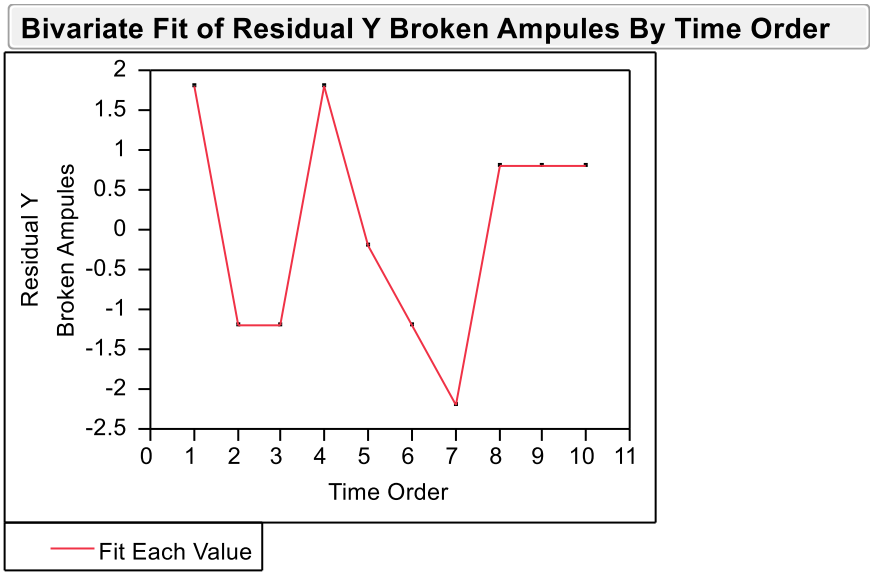
- There is less variation in the residuals as the # of transfers increase, although with such a small data set it is difficult to make such a claim conclusively.

e) Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality to ascertain whether the normal assumption is reasonable here. Use Table B.6 and $\alpha = .01$. What do you conclude?



- The calculated R^2 = .9829 which is greater than .879, the value obtained from Table B.6, thus we cannot reject H_o and thus conclude that the data is normally distributed.

f) Prepare a time plot of the residuals. What information is provided by your plot?



- The time plot does not show a time related pattern, thus the number of broken ampules is independent of time

g) Assume that (3.10) is applicable and conduct the Breusch-Pagan test to determine whether or not the error variance varies with the level of X. Use α = .10. State the alternatives, decision rule and conclusion. Does your conclusion support your preliminary findings in part (d)?

- H_o: The error variance is constant wrt X, if $X^2_{BP} < \chi^2$ (.90, 1) = 2.71, conclude H_o
- H_a: The error variance is not constant wrt X, if $X^2_{BP} > \chi^2$ (.90, 1) = 2.71, conclude H_a
- Our calculated value SSR* = 6.4, SSE = 17.6 thus $X^2_{BP} = \frac{SSR^*}{2} \div \left(\frac{SSE}{n}\right)^2 = 1.033$
- Since 1.033 < 2.71 We conclude that the error term is constant
- In part we could not conclude on the variance wrt X due to a small sample size, the fact that we now have this result give us some leverage to say that the error term is constant.

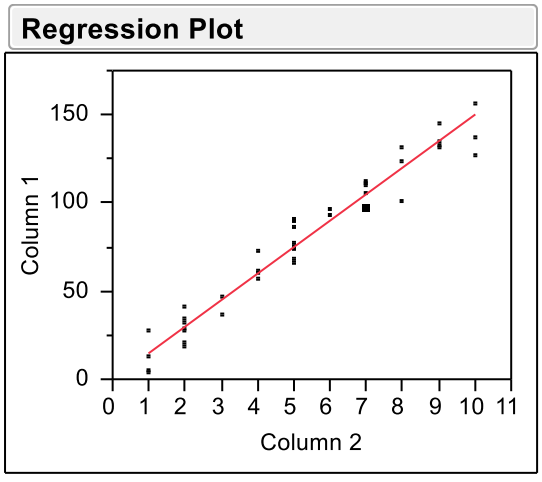
3.12) A student does not understand why the sum of squares defined in (3.16) is called pure error sum of squares "since the formula looks like the one for ordinary sum of squares." Explain.

SSPE represents the sum of squares error of the Full model in the lack of fit testing when replicates at each value of X are taken into account. It represents the actual error that cannot be attributed to the full model and it literally represents pure error. Pure error identifies the variability associated with only the error term.

3.13) Refer to COPIER MAINTENANCE problem 1.20.

a) What are the alternative conclusions when testing for lack of fit of a linear regression function?
   - H_o: $E[Y] = \beta_o + \beta_1(X)$ i.e., the reduced model is appropriate & μ_j is linearly related to X_j
   - H_a: $E[Y] = \beta_o + \beta_1(X)$ i.e., the reduced model is not appropriate & μ_j is not linearly related to X_j

b) Perform the test indicated in part (a). Control the risk of Type I error at .05. State the decision rule and conclusion.
   - If $F^* \leq F(.95, 8, 35) \approx F(.95,8,30) = 2.27$, conclude $H_{0.}$
   - If $F^* > F(.95, 8, 35) \approx F(.95,8,30) = 2.27$, conclude $H_{a.}$
   - Calculated F* = .96 ≤ Critical F of 2.27, so we conclude the null. This reduced model is appropriate



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	76960.423	76960.4	968.6572
Error	43	3416.377	79.5	Prob > F
C. Total	44	80376.800		<.0001*

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	8	618.7187	77.3398	0.9676
Pure Error	35	2797.6583	79.9331	Prob > F
Total Error	43	3416.3770		0.4766
				Max RSq
				0.9652

c) Does the test in part (b) detect other departures from regression model (2.1), such as lack of constant variance of lack of normality in the error terms? Could the results of the test of lack of fit be affected by such departures? Discuss.

- Since the lack of fit test assumes that the observations Y for a given X are independent, normally distributed, and distributions of Y have the same variance, the lack of fit test will not directly detect these departures. However, departures from normality or lack of constant error term variance errors could contribute to a false lack of fit test In other words, tests for normality, variance, and independence should have been accomplished prior to conducting the lack of fit test.

3.19) A student fitted a linear regression function for a class assignment. The student plotted the residuals e_i against Y_i and found a positive relation. When the residuals were plotted against the fitted Y-hat$_i$, the student found no relation. How could this difference arise? Which is the more meaningful plot?

- One would expect to find correlation between the residuals and the responses from which Y they were derived. A positive relation may signal the presence of an outlier at a particular level of Y_i. The fit against expected Y is meaningful and allows conclusions to be drawn on non-linearity, error term variance, and outlier influence. The expected Y versus residual plot is more meaningful.

3.20) If the error terms in a regression model are independent N(0, σ^2), what can be said about the error terms after transformation X' = 1/X is used? Is the situation the same after transformation Y' = 1/Y is used?

- Based on the eq 2.1 exchanging X for 1/X would not impact the error terms, the intercept of the regression line or the slope of the line, what would be impacted is the spread in the data wrt to the x values since an original sample range over 100 units would now be reduced to 1/100th of a unit.
- For a transformation of Y to 1/Y the data should remain unchanged since this simply involves taking the Y response to the -1 power. The result would be an across the board reduction of variance, although the relative variance would remain unchanged. In essence this transformation just scales the entire model down.