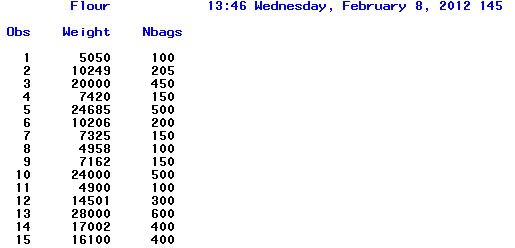
[To Projects](http://facweb.cs.depaul.edu/sjost/csc423/projects.htm" \t "_top)

**CSC 423/324 -- Project 3**

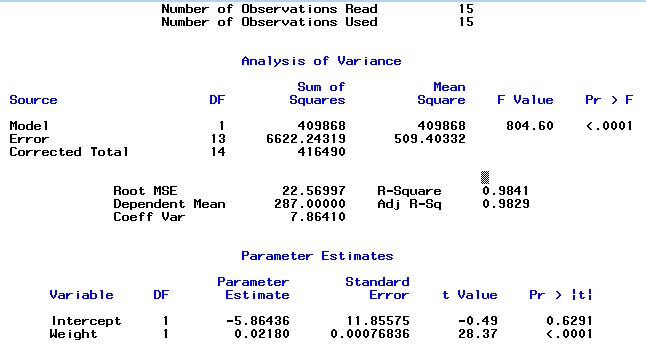
**Part A. Flour Dataset (30 pts.)**

Use the [flour dataset](http://facweb.cs.depaul.edu/sjost/csc423/projects/flour.txt) to do these problems:

1. Create and print a SAS dataset or R dataframe named Flour.



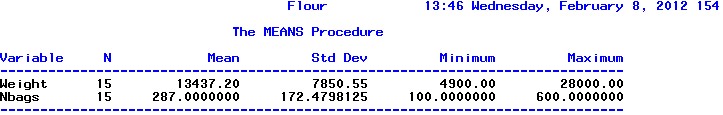
1. Use SAS or R to find the simple linear regression model for predicting NBags from Weight. Include the relevant output in your Word file.



Answer:

Nbags = -5.8644 + 0.0218 Weight

1. Use proc means to compute the means and standard deviations for Weight and NBags.



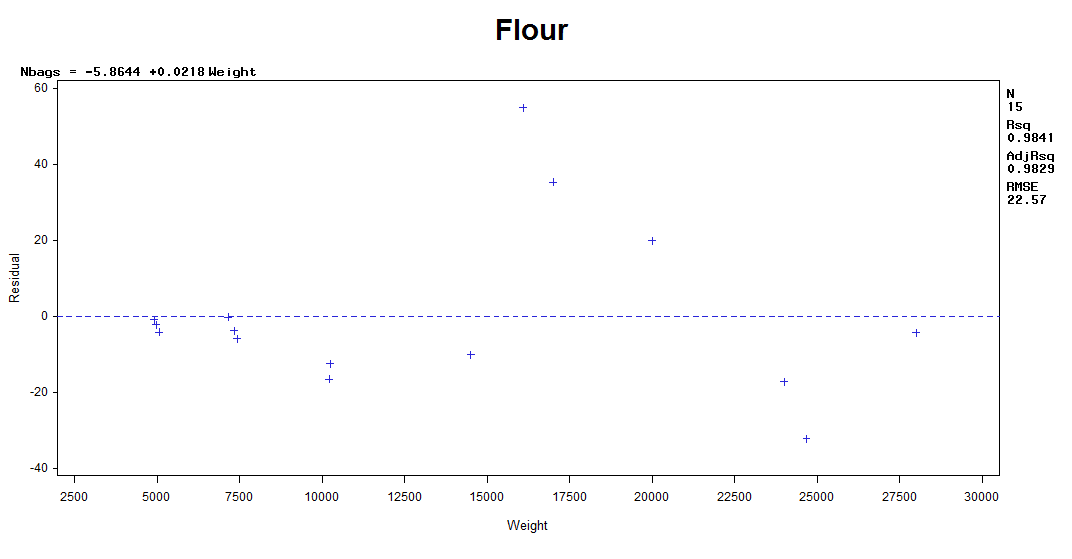
1. Compute the regression model by hand using the formula

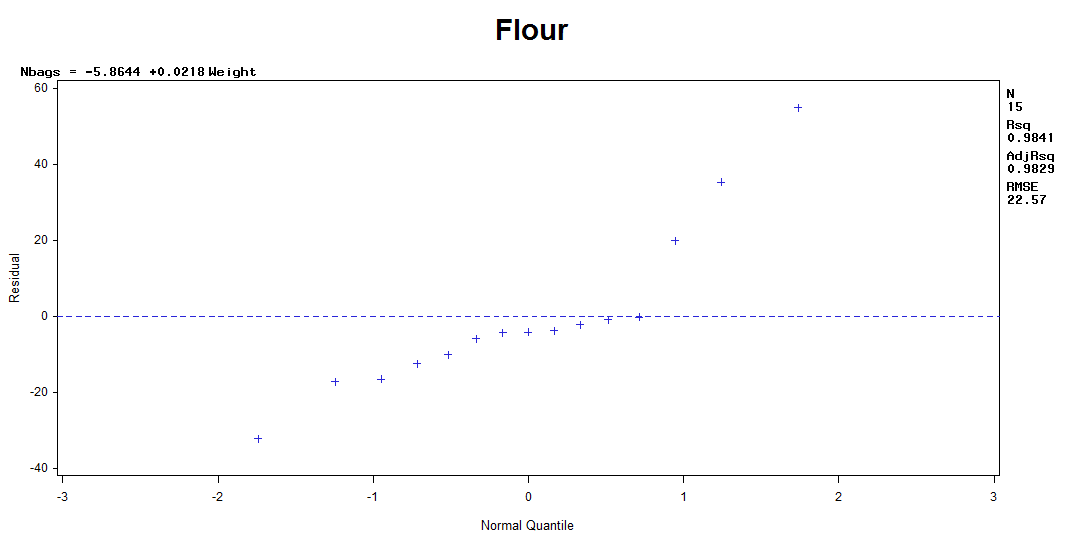
y - y = (rxy sy / sx)(x - x)

y-287=0.992(172.4798/7850.55)(x-13437)

Y=0.02179x-5.854

1. For the simple linear regression model, create the residual and normal plot of the residuals.



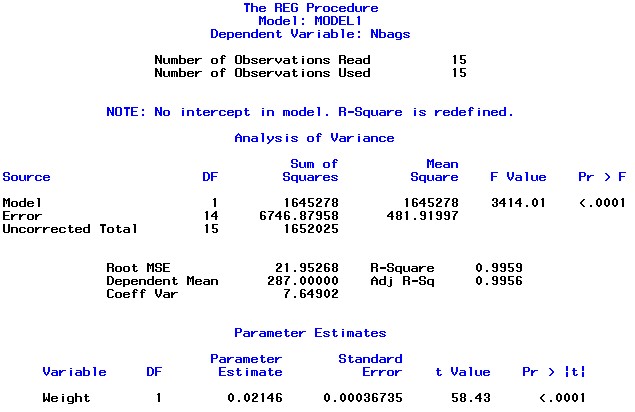


1. Compute the regression through the origin model by hand using the formula

a = (x1 y1 + ... + xn yn) / (x12 + ... + xn2),     y = ax.

a = (76652695.05/3571211205) = 0.0214

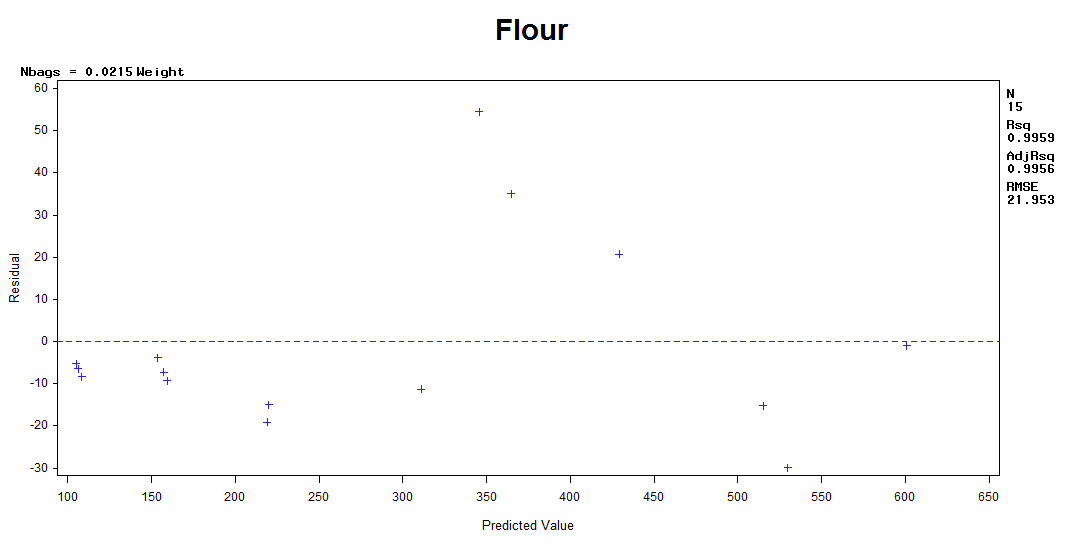
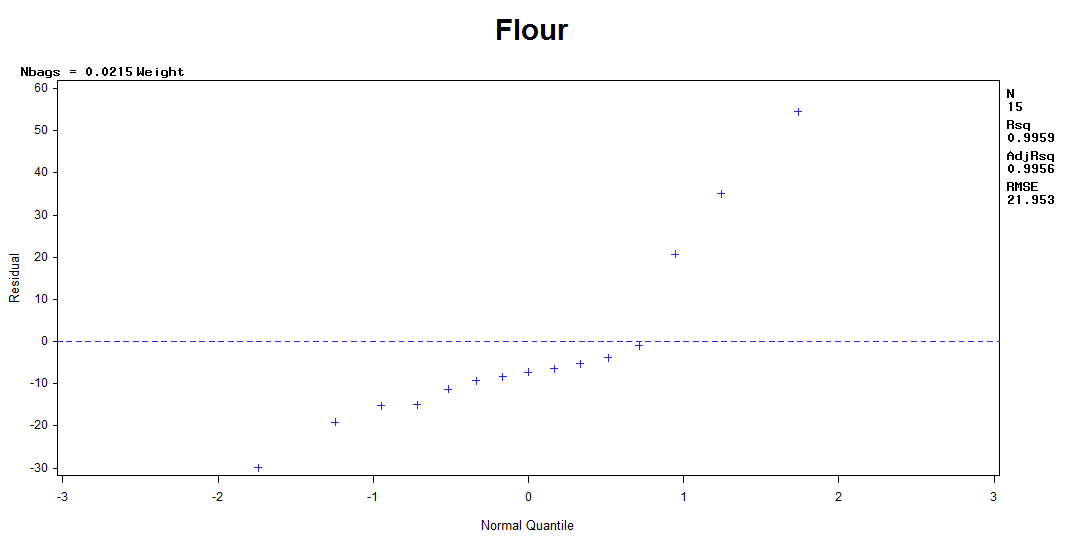
1. Use SAS or R to find the regression through the origin model for predicting Nbags from weight.



Answer:

Nbags = 0.0215Weight

1. For the regression through the origin model, create the residual and normal plot of the residuals.



**Part B: Used Car Dataset (40 pts.)**

Collect the following used car data from the internet or elsewhere for at least 20 cars of the same make and model: Price, Year, Miles.

1. Create and print the SAS dataset or R data frame called UsedCars.

Obs Price Year Mile

1 8990 2003 157401

2 9790 2004 108000

3 10931 2004 106289

4 11876 2005 96078

5 11990 2005 110849

6 12490 2004 97608

7 13876 2005 92776

8 13989 2005 88780

9 13995 2005 98066

10 14998 2005 74726

11 29990 2011 9401

12 29990 2009 16681

13 30748 2010 30748

14 30996 2010 18956

15 12888 2005 117429

16 13981 2006 130150

17 16996 2007 111203

18 16998 2004 52174

19 17928 2007 63374

20 19999 2008 64721

21 20000 2008 61684

22 20995 2007 62472

23 21998 2008 43554

24 21995 2008 21995

25 22966 2008 34806

26 23995 2009 34295

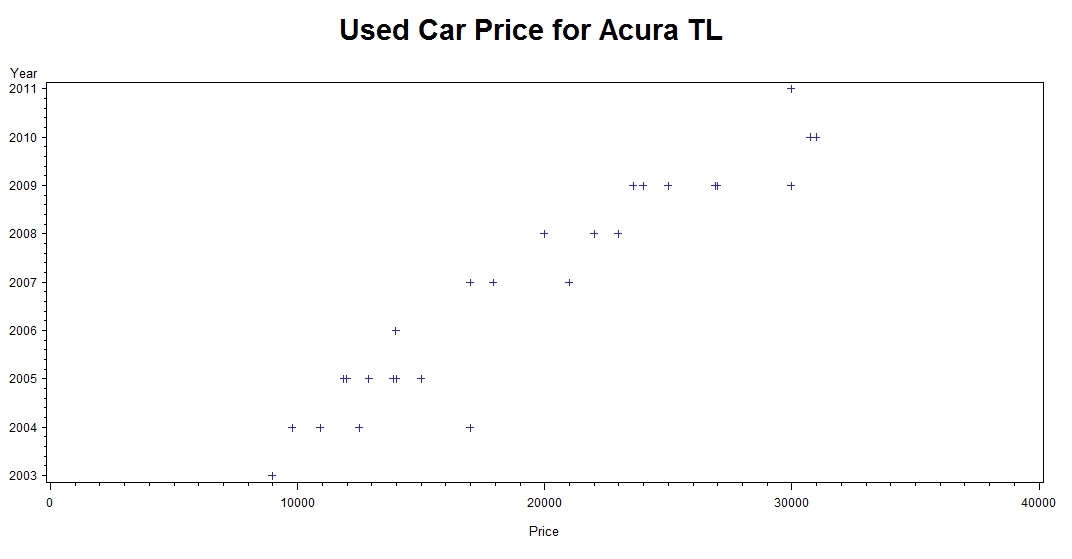
27 23601 2009 34716

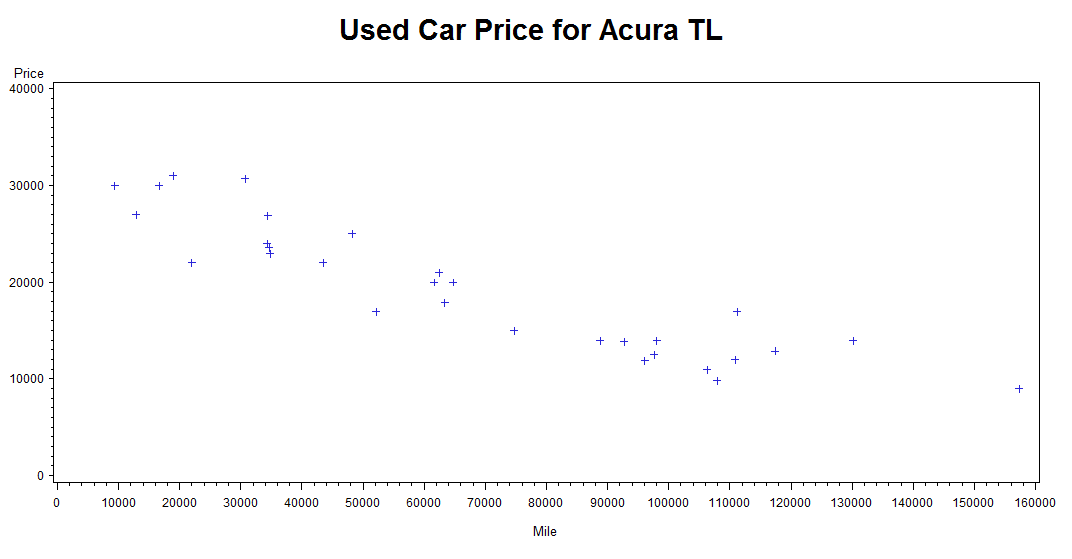
28 24990 2009 48290

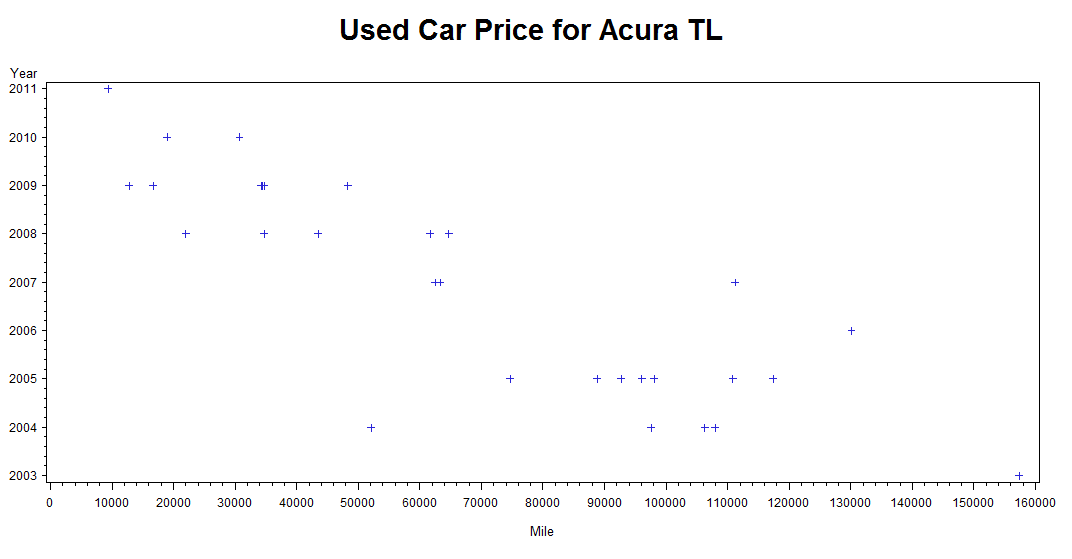
29 26910 2009 34414

30 26995 2009 1287

1. Create the pairwise scatterplots of Year, Miles, and Price.







1. Find the pairwise correlations of Year, Miles, and Price with SAS or R. Interpret them.

Correlation

Variable Mile Price Year

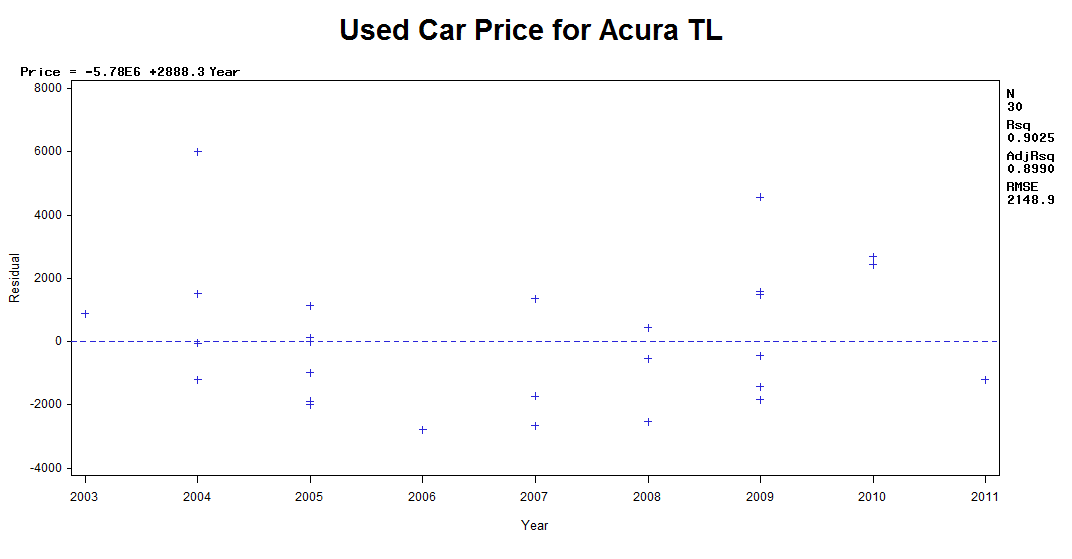
Mile 1.0000 -0.9109 -0.8457

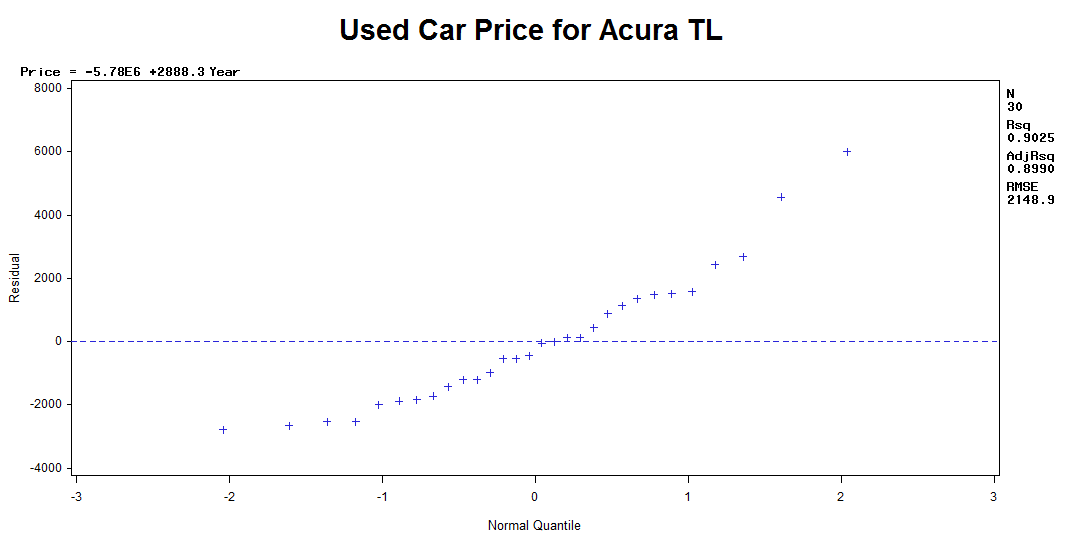
Price -0.9109 1.0000 0.9500

Year -0.8457 0.9500 1.0000

1. Find the simple linear regression model Price=Year with SAS or R.

Answer: Price = -5777120 + 2888.3 Year

1. Create the residual and normal plot of the residuals. Interpret these plots. 

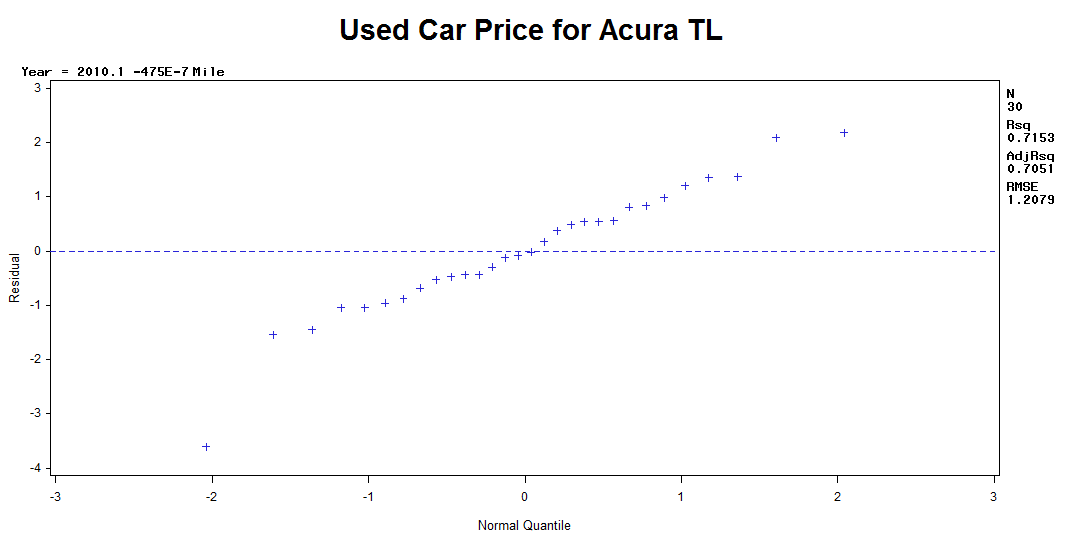


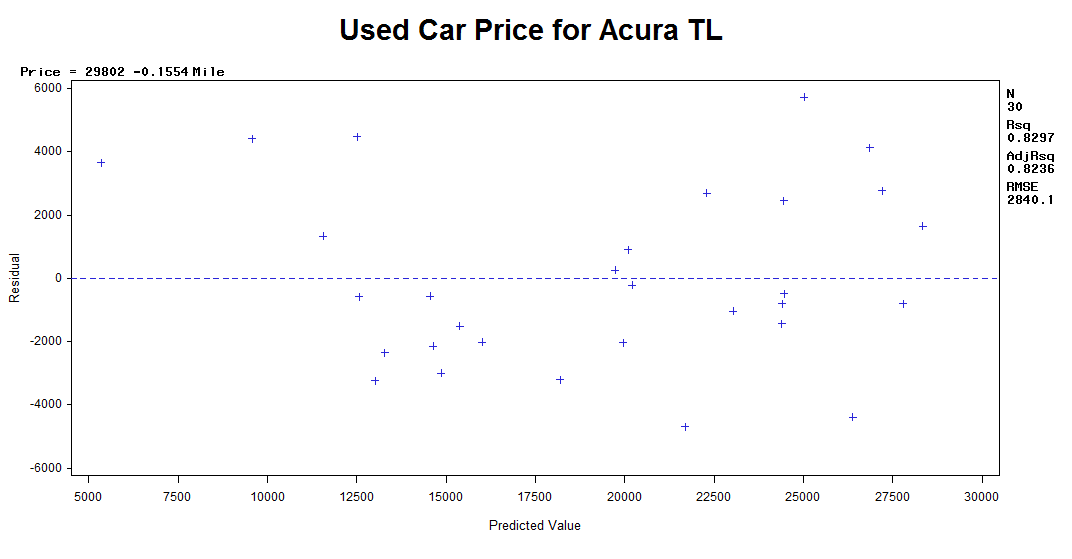
Answer:

The plot graphs tell that the residuals are a little bit biased and sort of heteroscedastic, meaning that for high mileage, the predicted price would vary more for low and high mileage cars.

1. Find the simple linear regression model Price=Miles with SAS or R.

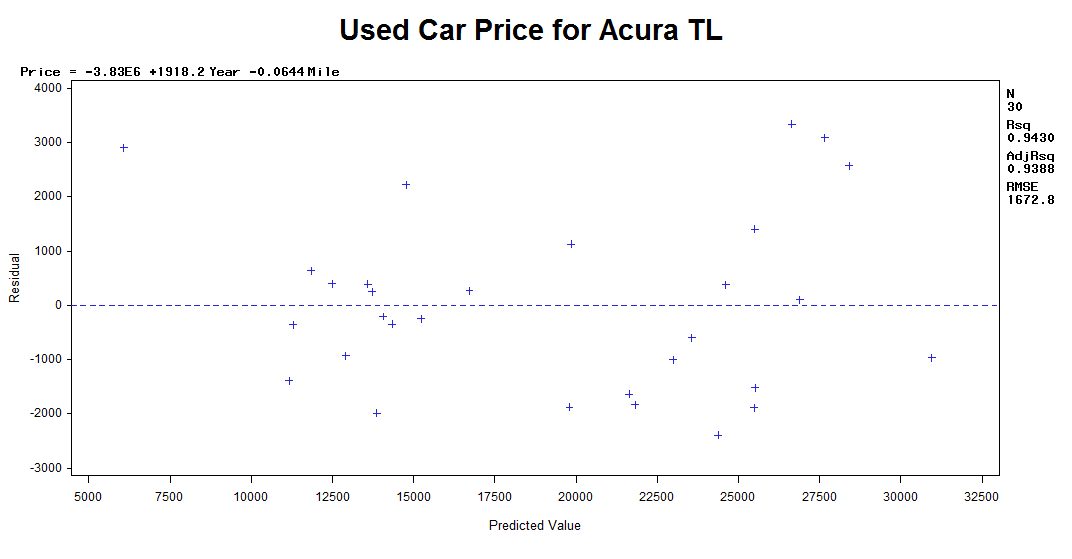
Answer: Price = 29802 - 0.1554Mile

1. Create the residual and normal plot of the residuals. Interpret these plots.   
   



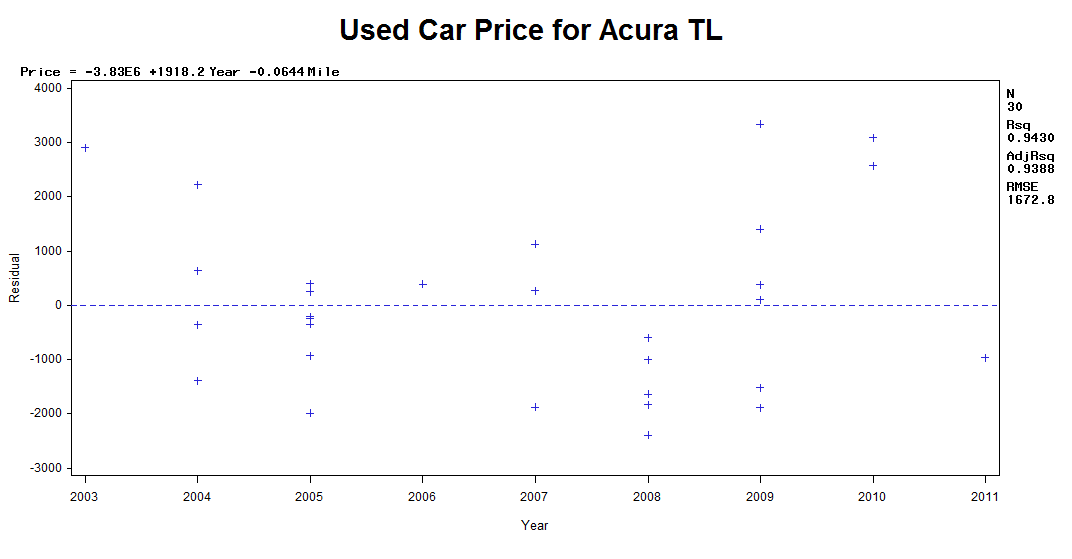
Answer:

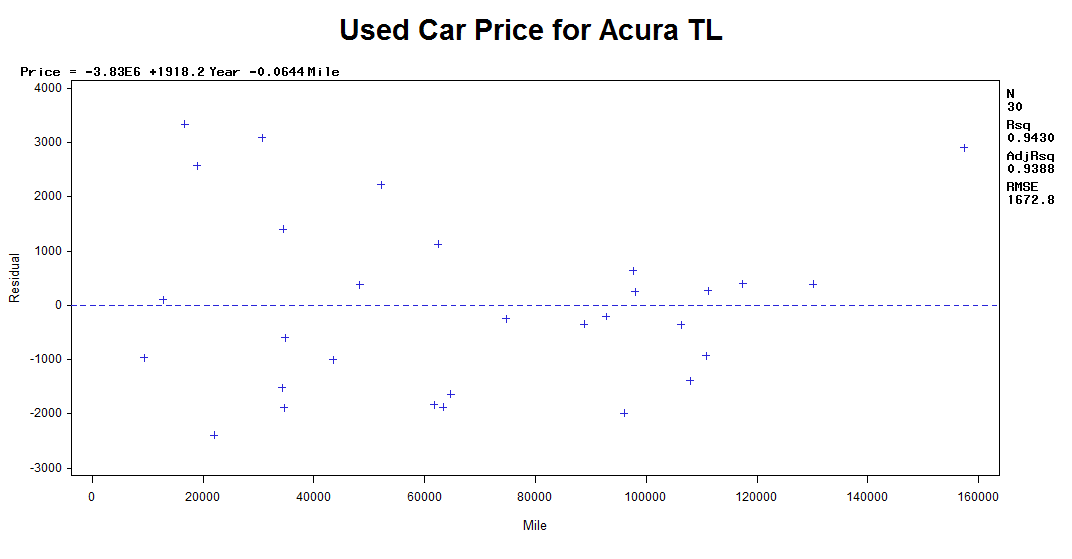
From the residual v. predicted plot graph we can see that the residuals are unbiased but sort of heteroscedastic, meaning that for high mileage, the predicted price would vary more for low and high mileage cars.

1. Find the multiple linear regression model Price=Year Miles with SAS or R.   
   Answer:   
     
   Price = 1918.17Year - 0.06437Mile
2. Create the residual and normal plot of the residuals. Interpret these plots.   
   

Answer: The model is unbiased but a bit heteroscedastic. Maybe it is because different people have different views towards how much their car should be appraised

1. For the model in Problem 8, create the plots of Residual=Year and Residual=Miles.





1. In your opinion, which is the best regression model out of Price=Year, Price=Miles, Price=Year Miles for predicting Price.

Answer: I prefer the multi regression model since R2 is best, although all the models look a little bit biased and heteroscedastic.

**Part C: Short Essay Questions (30 pts.)**

For each of these questions, your audience are persons that are not experts in statistics. Write with complete sentences and paragraphs. Cite any references that you use.

1. (10 pts.) Write a review (maximum one page) of the following article:

 Young and Karr, [*Deming, data and observational studies. A process out of control and needing fixing*](http://facweb.cs.depaul.edu/sjost/csc423/documents/data-observational-studies.pdf)   
Explain why the conclusions of this article are important for researchers using regression analysis.

Answer: The article claims that multiple variables that were used in previous analysis were based on biased data, saying that both data selection and data cleaning are not well selected and that observations in models were of limit size, which will possibly lead to a wrong conclusion. Plus most of these misleading conclusions were publicized without proper hypothesis test. Consequently, they reflected no natural features of the observed subject.

This article emphasizes the importance of the concept of "unbiased". Most of the models that were conducted before contained improper data, which were even not paid attention to. Thus, the article influenced statistics by making our conclusion more precise

2. (10 pts.) You use SAS or R to fit a regression model to a dataset and find that R2 = 0.79. Is this a good regression model or not? If you can't tell, what additional information do you need?

Answer:

As the definition of R2 tells, we could simply add more variables to the R2 to decorate it to a satisfied number like 0.79. So I do not think of 0.79 as sufficient for our consideration. We may use weighted R2 or observe residuals or homoscedasticity to make conclusion.

3. (10 pts.) Look up and explain the regression fallacy. Give at least one example.

Answer:

Regression fallacy means Things like stock market prices, golf scores, the earth's temperature, and chronic back pain fluctuate naturally and usually regress towards the mean. The logical flaw is to make predictions that expect exceptional results to continue as if they were average.

As a example, if I had enjoyed party everyday last quarter and I got a poor score in CSC 423 then I felt shamed and worked diligently this quarter, which is even cannot be represented in figures, I would possibly get a decent score this quarter, then regression fallacy happens.