

## SAS

**Illustration**  
**Simple Linear Regression**  
**Emergency Calls to the New York Auto Club**

Source:

Chatterjee, S; Handcock MS and Simonoff JS *A Casebook for a First Course in Statistics and Data Analysis*. New York, John Wiley, 1995, pp 145-152.

Setting:

Calls to the New York Auto Club are possibly related to the weather, with more calls occurring during bad weather. This example illustrates descriptive analyses and simple linear regression to explore this hypothesis in a data set containing information on calendar day, weather, and numbers of calls.

SAS Data Set:

Ers.sas7bdat

**Tip!** This illustration assumes that you have downloaded the data from the PubHlth 640 website directly. It also assumes that you have saved it to a directory having a name that you chose

Variable Name	Label	Coding/Remarks
DAY	Date using informat <b>MMDDYY6.</b>	Example: 016193 is January 16, 1993
CALLS	Calls answered	
FHIGH	Forecasted high temperature	
FLOW	Forecasted low temperature	
HIGH	High temperature	
LOW	Low temperature	
RAIN	Rain Forecast	0 = NO 1 = RAIN
SNOW	Snow Forecast	0 = NO 1 = SNOW
WEEKDAY	Type of Day	0 = NO 1 = Weekday
YEAR		0 = 1993 1 = 1994
SUNDAY		0 = NO 1 = SUNDAY
SUBZERO		0 = NO 1 = SUBZERO

## 1. Read in the data, create a dictionary of discrete variable values

Tip! You will have to edit the code that is shaded in yellow.

<pre> Libname class "z:\bigelow\teaching\web640\data sets"; data temp;   set class.ers; run; quit;  * * * Create dictionary of variable values for readability * proc format;   value rainf 0='0=no'               1='1=rain';   value snowf 0='0=no'               1='1=snow';   value weekdayf 0='0=no'                  1='1=weekday';   value yearf 0='0=1993'               1='1=1994';   value sundayf 0='0=no'                 1='1=Sunday';   value subzerof 0='0=no'                  1='1=subzero';  run; quit; </pre>	<p>Tip -- MMDDYY6. tells SAS that the variable DAY is a date variable of the form mmdyy</p>
--	---

## 2. For small data sets, produce a listing for review

<pre> proc print data=temp;   format day MMDDYY6.          rain RAINF.          snow SNOWF.          weekday WEEKDAYF.          year YEARF.          sunday SUNDAYF.          subzero SUBZEROF.;   title "Temporary Data set";   title2 " Emergency Calls to New York Auto Club"; run; quit; </pre>	<p>Note - Now we make use of the dictionary we created above by using a FORMAT statement.</p>
---	---

Partial listing of output.

Emergency Calls to New York Auto Club													
Obs	day	calls	fhigh	flow	high	low	rain	snow	weekday	year	sunday	subzero	
1	011693	2298	38	31	39	31	0=no	0=no	0=no	0=1993	0=no	0=no	
2	011793	1709	41	27	41	30	0=no	0=no	0=no	0=1993	1=Sunday	0=no	
3	011893	2395	33	26	38	24	0=no	0=no	0=no	0=1993	0=no	0=no	
4	011993	2486	29	19	36	21	0=no	0=no	1=weekday	0=1993	0=no	0=no	
5	012093	1849	40	19	43	27	0=no	0=no	1=weekday	0=1993	0=no	0=no	
6	012193	1842	44	30	43	29	0=no	0=no	1=weekday	0=1993	0=no	0=no	
7	012293	2100	46	40	53	41	1=rain	0=no	1=weekday	0=1993	0=no	0=no	
8	012393	1752	47	35	46	40	0=no	0=no	0=no	0=1993	0=no	0=no	
9	012493	1776	53	34	55	38	1=rain	0=no	0=no	0=1993	1=Sunday	0=no	
10	012593	1812	38	32	43	31	0=no	0=no	1=weekday	0=1993	0=no	0=no	
11	012693	1842	35	21	35	25	0=no	0=no	1=weekday	0=1993	0=no	0=no	
12	012793	1674	39	27	44	31	1=rain	1=snow	1=weekday	0=1993	0=no	0=no	
13	012893	1692	34	28	40	27	0=no	0=no	1=weekday	0=1993	0=no	0=no	

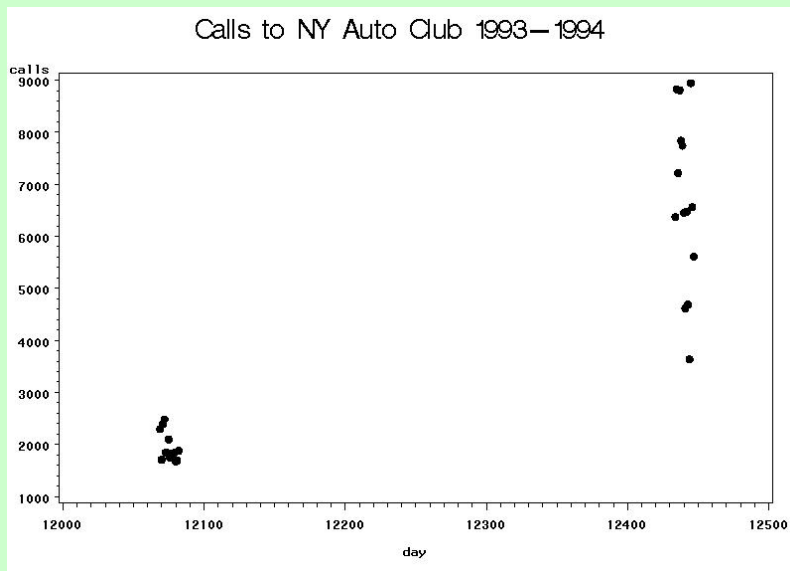
### 3. First look at the data. Plot of calls over time

```

goptions reset=all;
symbol1 value=dot;
proc gplot data=temp;
  title "Calls to NY Auto Club 1993-1994";
  plot calls*day;
run;
quit;
    
```

Note - How to save a graph as ".jpeg"

1. Activate the GRAPH output window at bottom right
2. FILE --> EXPORT AS IMAGE
3. At the "SAVE AS TYPE" dialog box, choose jpeg



Nyauto1.jpg

#### Remarks

- 1993 values are lower
- 1994 values are higher and more variable

### 4. Descriptives of Y=Calls. Tests of Assumption of Normality

<pre>proc capability data=temp NORMAL normaltest;   var calls;   title "Distribution of Y=CALLS"; run; quit;</pre>																																										
<p>The CAPABILITY Procedure Variable: calls</p> <p style="text-align: center;">Moments</p> <table> <tr> <td>N</td> <td>28</td> <td>Sum Weights</td> <td>28</td> </tr> <tr> <td>Mean</td> <td>4318.75</td> <td>Sum Observations</td> <td>120925</td> </tr> <tr> <td>Std Deviation</td> <td>2692.56394</td> <td>Variance</td> <td>7249900.56</td> </tr> <tr> <td>Skewness</td> <td>0.48107831</td> <td>Kurtosis</td> <td>-1.4180211</td> </tr> <tr> <td>Uncorrected SS</td> <td>717992159</td> <td>Corrected SS</td> <td>195747315</td> </tr> <tr> <td>Coeff Variation</td> <td>62.3459089</td> <td>Std Error Mean</td> <td>508.846755</td> </tr> </table> <p>...</p> <p style="text-align: center;">Tests for Normality</p> <table> <thead> <tr> <th>Test</th> <th>--Statistic--</th> <th>-----p Value-----</th> <th></th> </tr> </thead> <tbody> <tr> <td>Shapiro-Wilk</td> <td>W 0.829019</td> <td>Pr &lt; W 0.000</td> <td rowspan="5" style="color: red; vertical-align: top;">Null Hypothesis of normality is rejected.</td> </tr> <tr> <td>Kolmogorov-Smirnov</td> <td>D 0.251960</td> <td>Pr &gt; D &lt;0.010</td> </tr> <tr> <td>Cramer-von Mises</td> <td>W-Sq 0.311246</td> <td>Pr &gt; W-Sq &lt;0.005</td> </tr> <tr> <td>Anderson-Darling</td> <td>A-Sq 1.867261</td> <td>Pr &gt; A-Sq &lt;0.005</td> </tr> </tbody> </table>		N	28	Sum Weights	28	Mean	4318.75	Sum Observations	120925	Std Deviation	2692.56394	Variance	7249900.56	Skewness	0.48107831	Kurtosis	-1.4180211	Uncorrected SS	717992159	Corrected SS	195747315	Coeff Variation	62.3459089	Std Error Mean	508.846755	Test	--Statistic--	-----p Value-----		Shapiro-Wilk	W 0.829019	Pr < W 0.000	Null Hypothesis of normality is rejected.	Kolmogorov-Smirnov	D 0.251960	Pr > D <0.010	Cramer-von Mises	W-Sq 0.311246	Pr > W-Sq <0.005	Anderson-Darling	A-Sq 1.867261	Pr > A-Sq <0.005
N	28	Sum Weights	28																																							
Mean	4318.75	Sum Observations	120925																																							
Std Deviation	2692.56394	Variance	7249900.56																																							
Skewness	0.48107831	Kurtosis	-1.4180211																																							
Uncorrected SS	717992159	Corrected SS	195747315																																							
Coeff Variation	62.3459089	Std Error Mean	508.846755																																							
Test	--Statistic--	-----p Value-----																																								
Shapiro-Wilk	W 0.829019	Pr < W 0.000	Null Hypothesis of normality is rejected.																																							
Kolmogorov-Smirnov	D 0.251960	Pr > D <0.010																																								
Cramer-von Mises	W-Sq 0.311246	Pr > W-Sq <0.005																																								
Anderson-Darling	A-Sq 1.867261	Pr > A-Sq <0.005																																								

### Remarks

- The null hypothesis of normality of  $Y=CALLS$  is rejected by every test.
- Take care, sometimes the cure is worse than the problem.
- For now, we'll continue along anyway; this will give us a chance to see some interesting diagnostics!

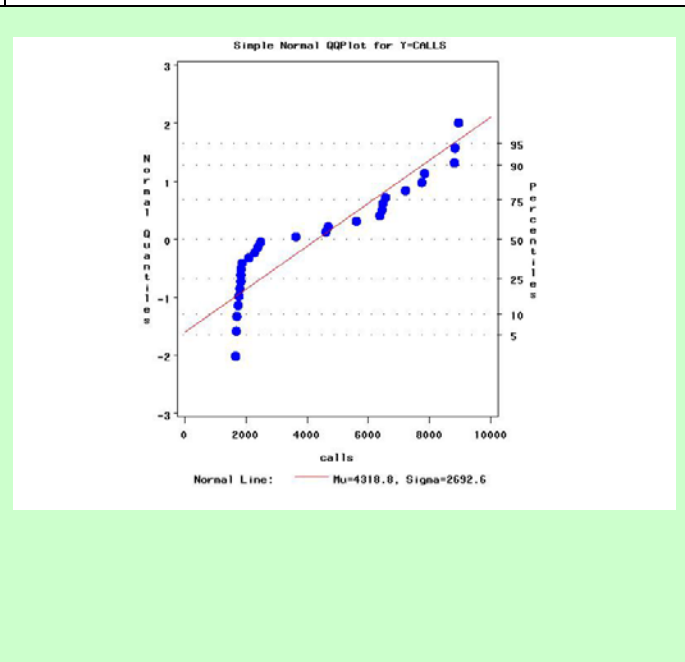
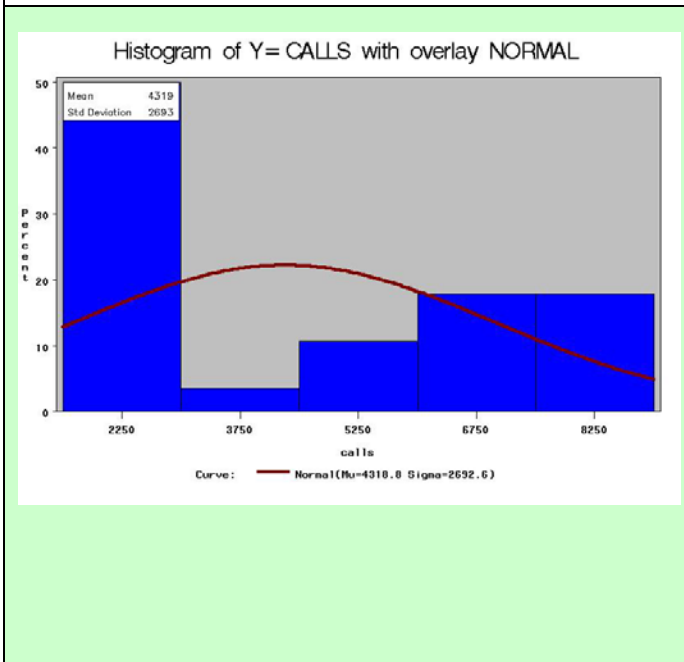
### 5. Graphical Assessments of Normality of Y=Calls.

#### Histogram with overlay normal (note use of NOPRINT)

```
proc capability data=temp NORMAL noprint;
  var calls;
  title "Histogram of Y=CALLS with overlay
        NORMAL";
  histogram calls/normal(color=maroon w=4)
    cfill=blue cframe=LIGR;
  inset mean std/cfill=blank format=5.2;
run;
quit;
```

#### Quantile Quantile Plot w reference = Normal

```
goptions reset=all;
goptions ftext=none htext=1 cell;
symbol1 c=blue v=dot h=1.5;
symbol2 c=red;
proc capability data=temp noprint;
  var calls;
  title "Simple Normal QQPlot for Y=CALLS";
  qqplot calls/
  normal (mu=est sigma=est)
  rotate
  pctlaxis(LABEL='Percentiles' GRID LGRID=35)
  square;
run;
quit;
```



Nyauto2.jpg

nyauto3.jpg

Remarks

- The graphs show what we suspected – nonnormality of Y=CALLS.

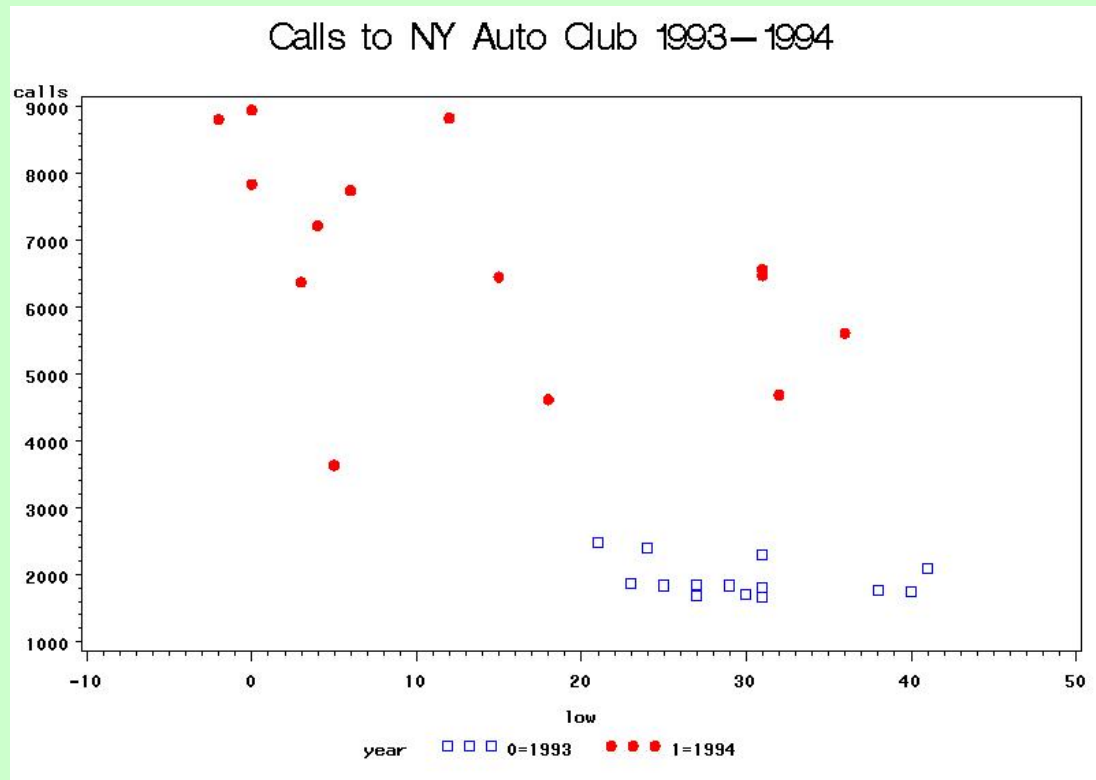
## 5. Scatterplot of Y=CALLS v X=LOW

```

options reset=all;
symbol1 value=square color=blue;
symbol2 value=dot color=red;
proc gplot data=temp;
  title "Calls to NY Auto Club 1993-1994";
  plot calls*low=year;
  format year YEARF.;
run;
quit;

```

Note -  
I used value of YEAR as my plotting symbol.  
I then indicated 1993 and 1994 using different symbols and colors using SYMBOL1 and SYMBOL2 statements.



Nyauto4.jpg

### Remarks

- The scatterplot suggests, as we might expect, that lower temperatures are associated with more calls to the NY Auto club.
- The distinction between 1993 and 1994 points suggests that, perhaps, low temperature is not the only predictor of increased calls.

## 6. Least Squares Estimation and Analysis of Variance Table

<pre>proc reg data=temp SIMPLE;   title "Straight line fit of Y=CALLS to X=LOW";   model calls=low; run; quit;</pre>	<p>Note - The option SIMPLE produces descriptive statistics.</p>																																																																		
<p><i>Partial listing -----.</i></p> <p>The REG Procedure</p> <p>Number of Observations Read            28          Number of Observations Used         28</p> <p style="text-align: center;"><b>Descriptive Statistics</b></p> <table border="1"> <thead> <tr> <th>Variable</th> <th>Sum</th> <th>Mean</th> <th>Uncorrected SS</th> <th>Variance</th> <th>Standard Deviation</th> </tr> </thead> <tbody> <tr> <td>Intercept</td> <td>28.00000</td> <td>1.00000</td> <td>28.00000</td> <td>0</td> <td>0</td> </tr> <tr> <td>low</td> <td>609.00000</td> <td>21.75000</td> <td>18003</td> <td>176.19444</td> <td>13.27383</td> </tr> <tr> <td>calls</td> <td>120925</td> <td>4318.75000</td> <td>717992159</td> <td>7249901</td> <td>2692.56394</td> </tr> </tbody> </table> <p style="text-align: center;"><b>Analysis of Variance</b></p> <table border="1"> <thead> <tr> <th>Source</th> <th>DF</th> <th>Sum of Squares</th> <th>Mean Square</th> <th>F Value</th> <th>Pr &gt; F</th> </tr> </thead> <tbody> <tr> <td>Model</td> <td>1</td> <td>100233719</td> <td>100233719</td> <td>27.28</td> <td>&lt;.0001</td> </tr> <tr> <td>Error</td> <td>26</td> <td>95513596</td> <td>3673600</td> <td></td> <td></td> </tr> <tr> <td>Corrected Total</td> <td>27</td> <td>195747315</td> <td></td> <td></td> <td></td> </tr> </tbody> </table> <p>Root MSE            1916.66373    R-Square            0.5121          Dependent Mean    4318.75000    Adj R-Sq            0.4933          Coeff Var            44.38006</p> <p style="text-align: center;"><b>Parameter Estimates</b></p> <table border="1"> <thead> <tr> <th>Variable</th> <th>DF</th> <th>Parameter Estimate</th> <th>Standard Error</th> <th>t Value</th> <th>Pr &gt;  t </th> </tr> </thead> <tbody> <tr> <td>Intercept</td> <td>1</td> <td>7475.84897</td> <td>704.63038</td> <td>10.61</td> <td>&lt;.0001</td> </tr> <tr> <td>low</td> <td>1</td> <td>-145.15398</td> <td>27.78868</td> <td>-5.22</td> <td>&lt;.0001</td> </tr> </tbody> </table>		Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation	Intercept	28.00000	1.00000	28.00000	0	0	low	609.00000	21.75000	18003	176.19444	13.27383	calls	120925	4318.75000	717992159	7249901	2692.56394	Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	Model	1	100233719	100233719	27.28	<.0001	Error	26	95513596	3673600			Corrected Total	27	195747315				Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Intercept	1	7475.84897	704.63038	10.61	<.0001	low	1	-145.15398	27.78868	-5.22	<.0001
Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation																																																														
Intercept	28.00000	1.00000	28.00000	0	0																																																														
low	609.00000	21.75000	18003	176.19444	13.27383																																																														
calls	120925	4318.75000	717992159	7249901	2692.56394																																																														
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F																																																														
Model	1	100233719	100233719	27.28	<.0001																																																														
Error	26	95513596	3673600																																																																
Corrected Total	27	195747315																																																																	
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t																																																														
Intercept	1	7475.84897	704.63038	10.61	<.0001																																																														
low	1	-145.15398	27.78868	-5.22	<.0001																																																														

Remarks

- The fitted line is  $calls\hat{=} = 7,475.85 - 145.15*[low]$
- $R^2 = .51$  indicates that 51% of the variability in calls is explained.
- The overall F test significance level  $< .0001$  suggests that the straight line fit performs better in explaining variability in calls than does  $\bar{Y}$  = average # calls

**7. Overlay of Straight Line Fit onto Scatterplot of Y=CALLS v X=LOW**

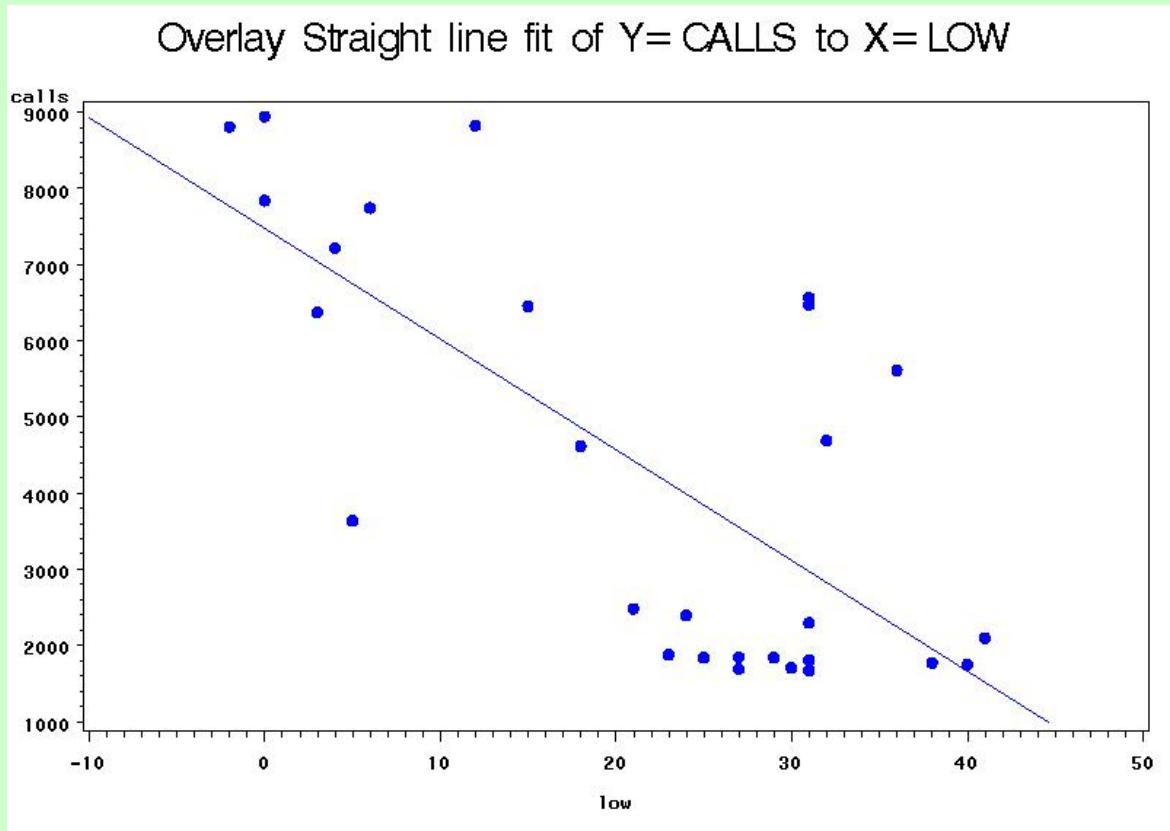
<pre>options reset=all;</pre>	<p>Note -</p>
-------------------------------	---------------

```

symbol1 value=dot color=blue i=r;
proc gplot data=temp;
  title "Overlay Straight line fit of Y=CALLS to
  X=LOW";
  plot calls*low;
run;
quit;

```

The overlay of the line is accomplished in the SYMBOL1 statement as i=r.



Nyauto5.jpg

#### Remarks

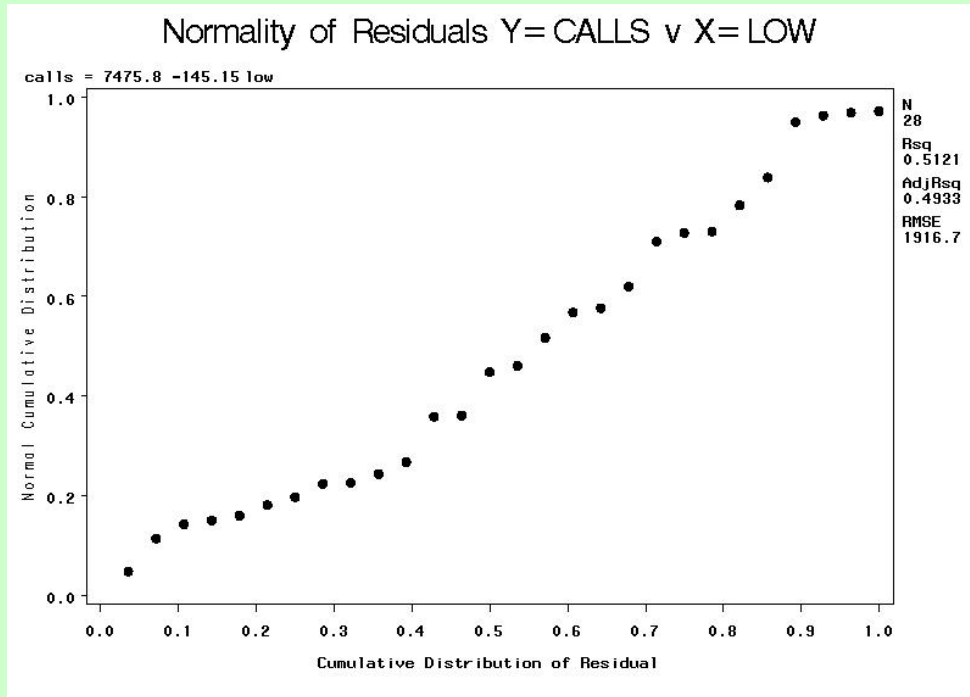
- The overlay of the straight line fit is reasonable but substantial variability is seen, too.
- There is a lot we still don't know, including but not limited to the following ---
- Case influence, omitted variables, variance heterogeneity, incorrect functional form, etc.

## 8. Residuals Analysis – Assessment of Normality of Residuals

```
goptions reset=all;
```

Note -

<pre> symbol1 value=dot; proc reg data=temp noprint;   model calls=low;   plot npp.*residual.;   title "Normality of Residuals Y=CALLS v         X=LOW"; run; quit;                 </pre>	<p>The NOPRINT suppresses numerical output.  npp. (note the period!) refers to normal percentiles  Residual. Refers to residuals</p> <p>Note all nice information that appears on the graph!</p>
--	--



Nyauto6.jpg

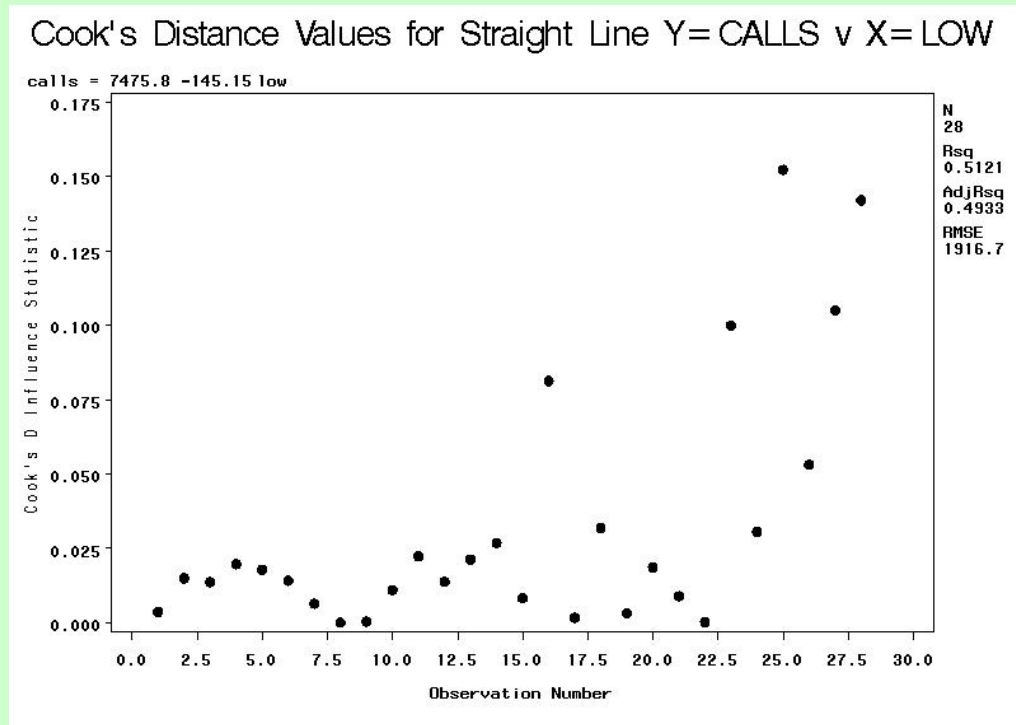
Remarks

- Not bad, actually.

9. Residuals Analysis – Detection of Outliers Using Cook’s Distance

<pre> goptions reset=all; symbol1 value=dot;                 </pre>	<p>Note -  The NOPRINT suppresses numerical output.</p>
---	---

```
proc reg data=temp noprint;
  model calls=low;
  plot cookd.*obs.;
  title "Cook's Distance Values for Straight
        Line Y=CALLS v X=LOW";
run;
quit;
```



Nyauto7.jpg

## Remarks

- For straight line regression, the suggestion is to regard Cook's Distance values  $> 1$  as significant..
- Here, there are no unusually large Cook Distance values.
- Not shown but useful, too, are examinations of leverage and jackknife residuals.

## 10. Assessing Assumptions of Linearity, Heteroscedascity, Independence Using Jackknife Residuals

```
goptions reset=all;
symbol1 value=dot;
proc reg data=temp noprint;
```

Note -  
The NOPRINT suppresses numerical output.  
Jackknife residuals are stored in rstudent.

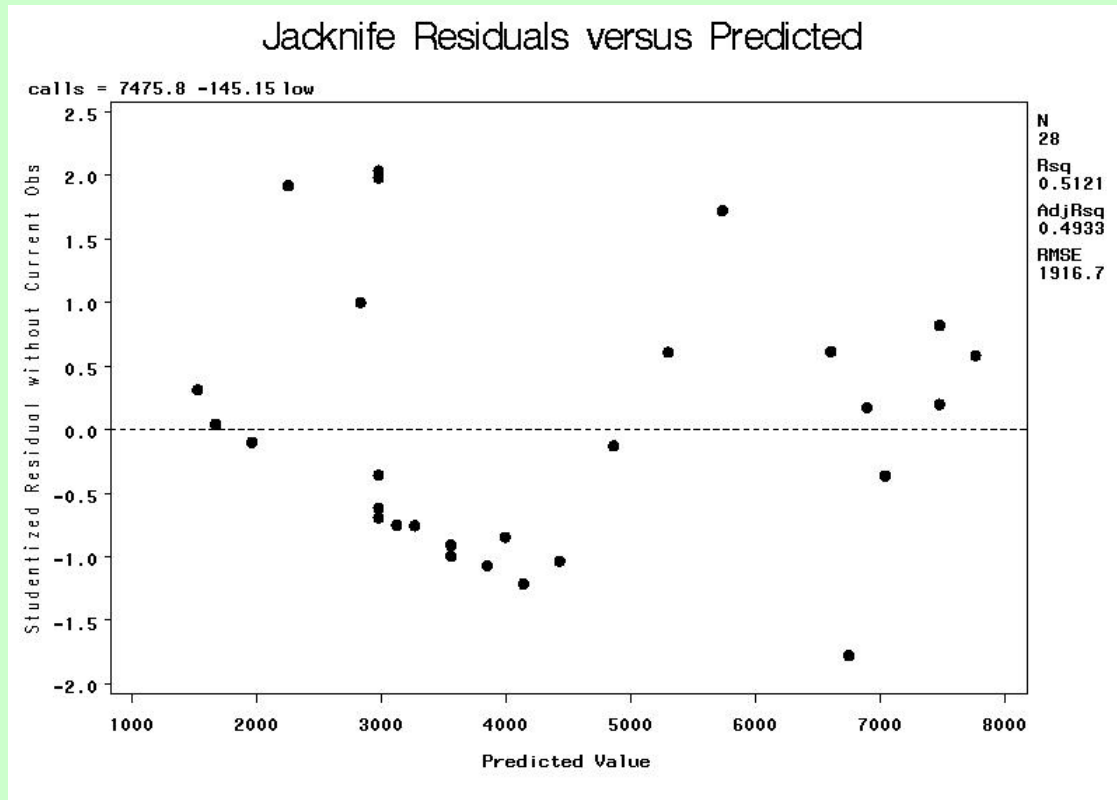
```

model calls=low;
plot rstudent.*predicted.;
title "Jackknife Residuals versus Predicted";
run;
quit;

```

Predicteds are stored in predicted.

Be sure to remember the periods!



Nyauto8.jpg

#### Remarks

- *Recall – A jackknife residual for an individual is a modification of the solution for a studentized residual in which the mean square error is replaced by the mean square error obtained after deleting that individual from the analysis.*
- *This plot in SAS is nice for its inclusion of some useful summaries – the fitted line, the  $R^2$*
- *Departures of this plot from a parallel band about the horizontal line at zero are significant.*
- *The plot here is a bit noisy but not too bad considering the small sample size.*