

Stata v 12**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.

Stata Data Set:
ers.dta
In this illustration, the data set *ers.dta* is accessed from the PubHlth 640 website directly. It is then saved to your current working directory..

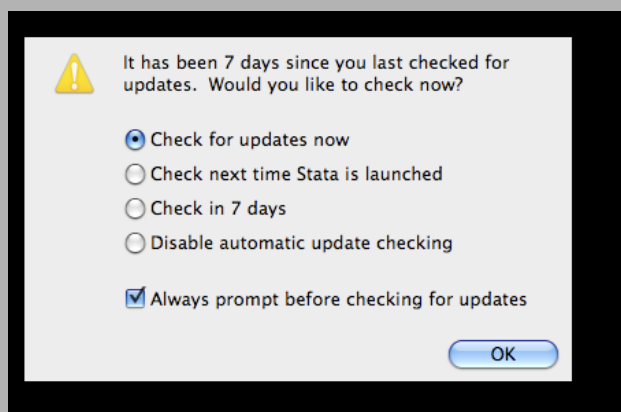
Simple Linear Regression Variables:
Outcome Y = calls
Predictor X = low.

Before you Begin
This illustration assumes that you have installed Stata successfully.

Launch Stata

| PC Users | Mac Users |
|--|--|
| <ul style="list-style-type: none">___ START > ALL PROGRAMS > Stata; <i>or</i>___ Double click on the Stata icon on your desktop | <ul style="list-style-type: none">___ APPLICATIONS > STATA folder > Stata; <i>or</i>___ Double click on the Stata icon on your dock |

You might see something like the following:



Stata will ask you if you would like to check for updates. **Click OK.**

Tip! Checking for updates is a good idea. And it doesn't take long.

If this check is not done for you automatically, you can do it manually as follows from the main toolbar at top:

HELP > OFFICIAL UPDATES

How to use this illustration

In the pages that follow, you will see **green**, **black**, and **blue**. These colorations are things that I did using MS Word.

Green: These are comments that I typed into the STATA command line. Notice that they all begin with an asterisk. I encourage you to use comments liberally. It will save you a lot of grief later when trying to recall what you did and why.

Tip! Each command is introduced to you twice, once as a “generic” and presented in green, and once in illustration of its actual use in black.

Black: Bold black are actual Stata commands that I typed and executed.
Note – You do *not* type the leading period.

Blue: I’ve colored in BLUE the output that Stata produces so that you can compare it with the output you get

I have also inserted some remarks in this weird font (arial, I think).

```
.
. ***** Set working directory, access data from website, and save copy to working directory.
. ***** cd/YOURDIRECTORY
. cd/Users/carolbigelow/Desktop
. use "http://people.umass.edu/biep640w/datasets/ers"
. save "ers.dta", replace
(note: file ers.dta not found)
file ers.dta saved

.
. * Describe data set
. codebook, compact
```

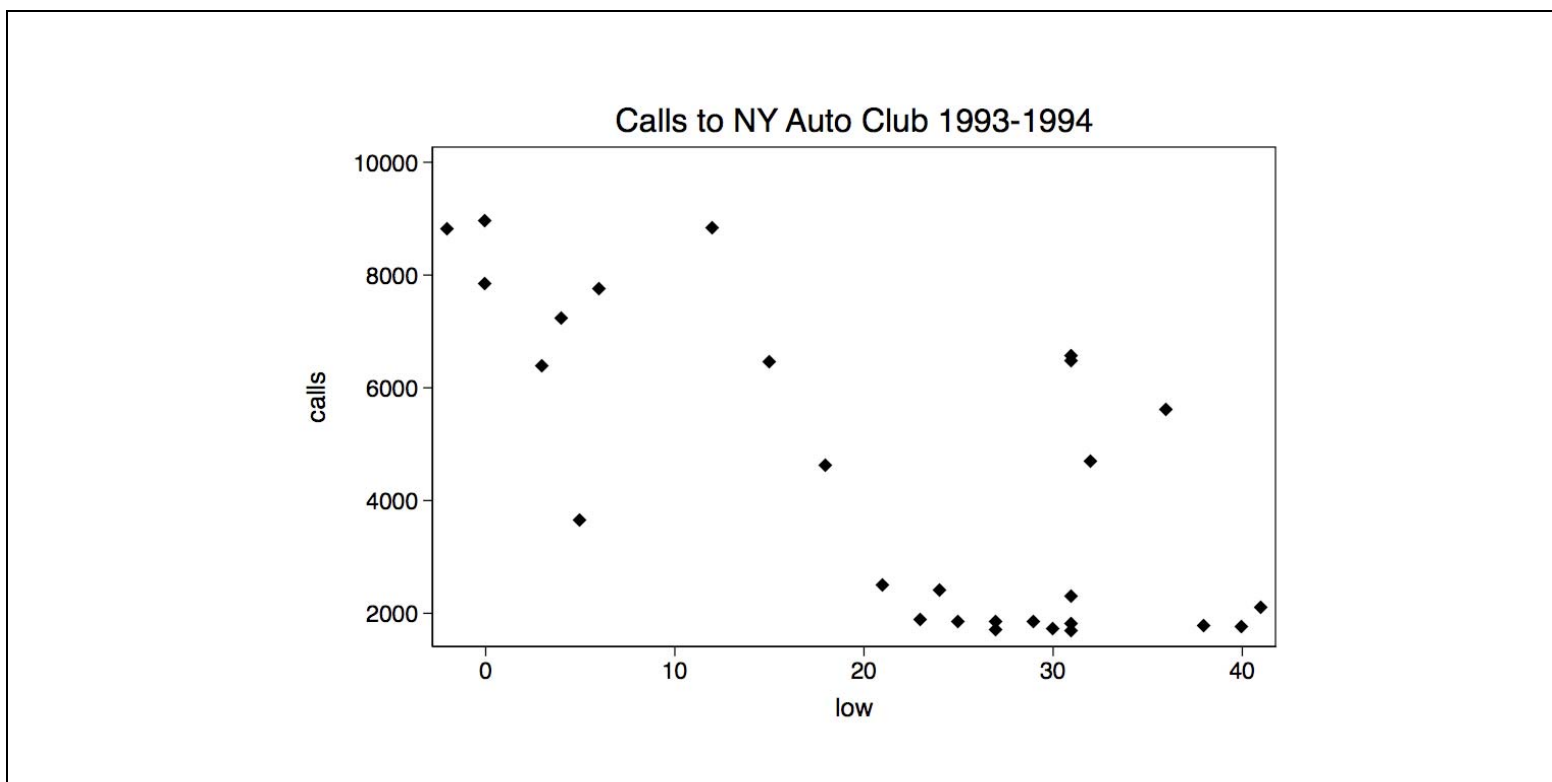
| Variable | Obs | Unique | Mean | Min | Max | Label |
|----------|-----|--------|----------|-------|-------|-------|
| day | 28 | 28 | 12258 | 12069 | 12447 | |
| calls | 28 | 27 | 4318.75 | 1674 | 8947 | |
| fhigh | 28 | 21 | 34.96429 | 10 | 53 | |
| flow | 28 | 19 | 24.46429 | 4 | 40 | |
| high | 28 | 19 | 37.46429 | 10 | 55 | |
| low | 28 | 22 | 21.75 | -2 | 41 | |
| rain | 28 | 2 | .3214286 | 0 | 1 | |
| snow | 28 | 2 | .2142857 | 0 | 1 | |
| weekday | 28 | 2 | .6428571 | 0 | 1 | |
| year | 28 | 2 | .5 | 0 | 1 | |
| sunday | 28 | 2 | .1428571 | 0 | 1 | |
| subzero | 28 | 2 | .1785714 | 0 | 1 | |

We see that this data set has $n=28$ observations on several variables.
For this illustration of simple linear regression, we will consider just two variables: *calls* and *low*
BEWARE – Stata is case sensitive!

```
. ***** Numerical Summaries
. ***** tabstat XVARIABLE YVARIABLE, stat(n mean sd min max)
. tabstat low calls, stat(n mean sd min max)
```

| stats | low | calls |
|-------|----------|----------|
| N | 28 | 28 |
| mean | 21.75 | 4318.75 |
| sd | 13.27383 | 2692.564 |
| min | -2 | 1674 |
| max | 41 | 8947 |

```
. ***** Scatterplot
. ***** graph twoway (scatter YVARIABLE XVARIABLE, symbol(d)), title("TITLE")
. graph twoway (scatter calls low, symbol(d)), title("Calls to NY Auto Club 1993-1994")
```

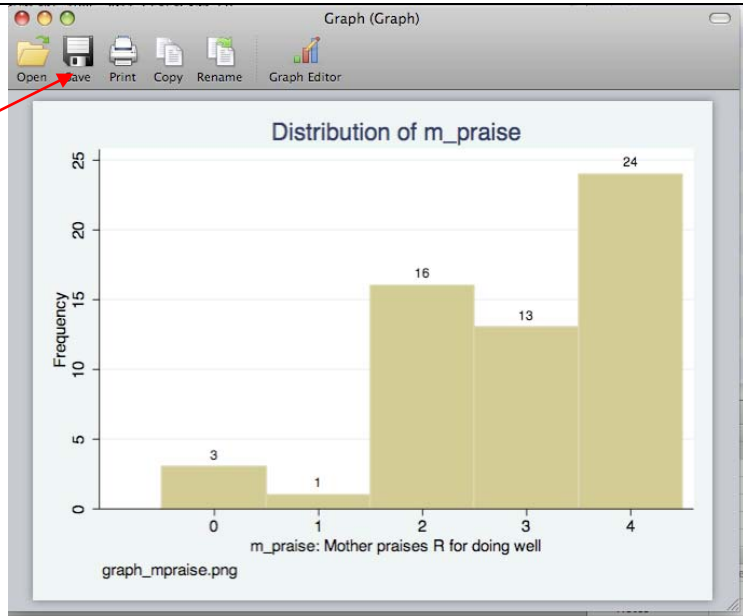


The scatterplot suggests, as we might expect, that lower temperatures are associated with more calls to the NY Auto Club. We also see that the data are a bit messy.

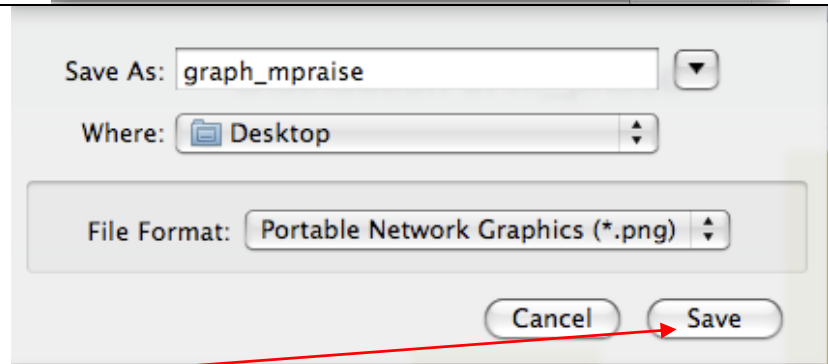
How to Save Your Graph

Tip! Save your graph with the extension “.png”

Step 1 – Click anywhere in the graph to make it active. Click on SAVE icon.

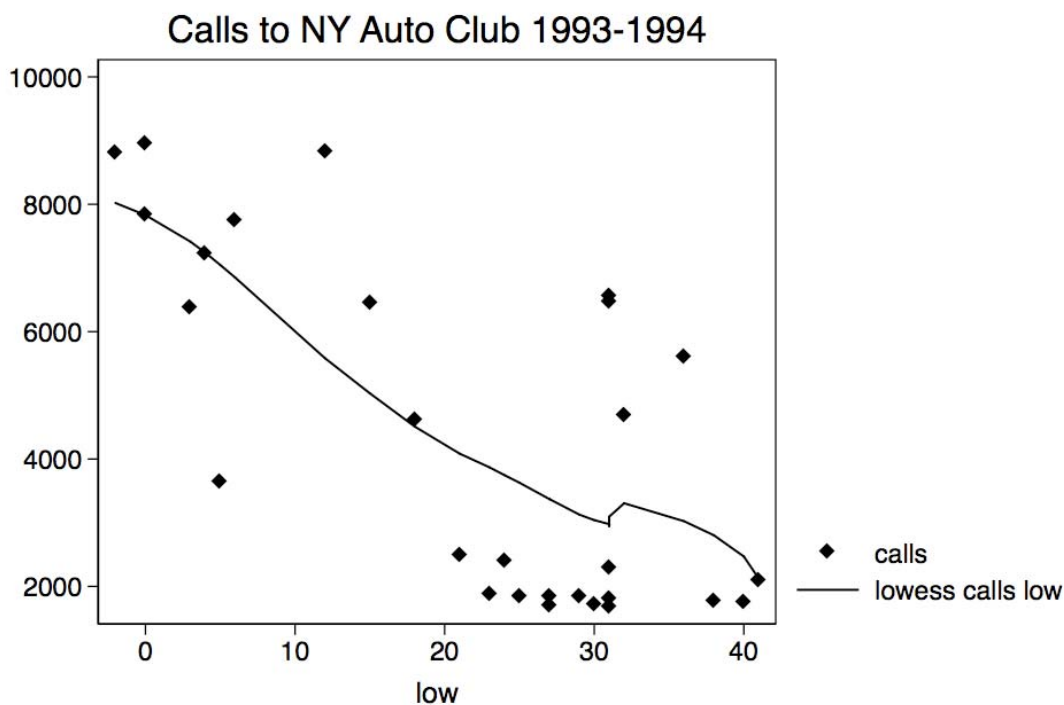


Step 2 –
 (1) At SAVE AS: type graph name without the extension,
 (2) At WHERE: choose directory location,
 (3) At FILE FORMAT drop down menu, choose “portable network graphics (recommended). Click on SAVE icon



Step 3 – SAVE

```
. ***** Scatterplot with Lowess Regression
. ***** graph twoway (scatter YVARIABLE XVARIABLE, symbol(d)) (lowess YVARIABLE XVARIABLE,
bwidth(.99) lpattern(solid)), title("TITLE") subtitle("TITLE")
. graph twoway (scatter calls low, symbol(d)) (lowess calls low, bwidth(.99) lpattern(solid)),
title("Calls to NY Auto Club 1993-1994")
```



Unfamiliar with LOWESS regression? LOWESS regression stands for “locally weighted scatterplot smoother”. It is a technique for drawing a smooth line through the scatter plot to obtain a sense for the nature of the functional form that relates X to Y, not necessarily linear. The method involves the following. At each observation (x,y), the observed data point is fit to a line using some “adjacent” points. It’s handy for seeing where in the data linearity holds and where it no longer holds.

```
. ***** Shapiro Wilk Test of Normality of Y (Reject normality for small p-value)
. ***** swilk YVARIABLE
. swilk calls
```

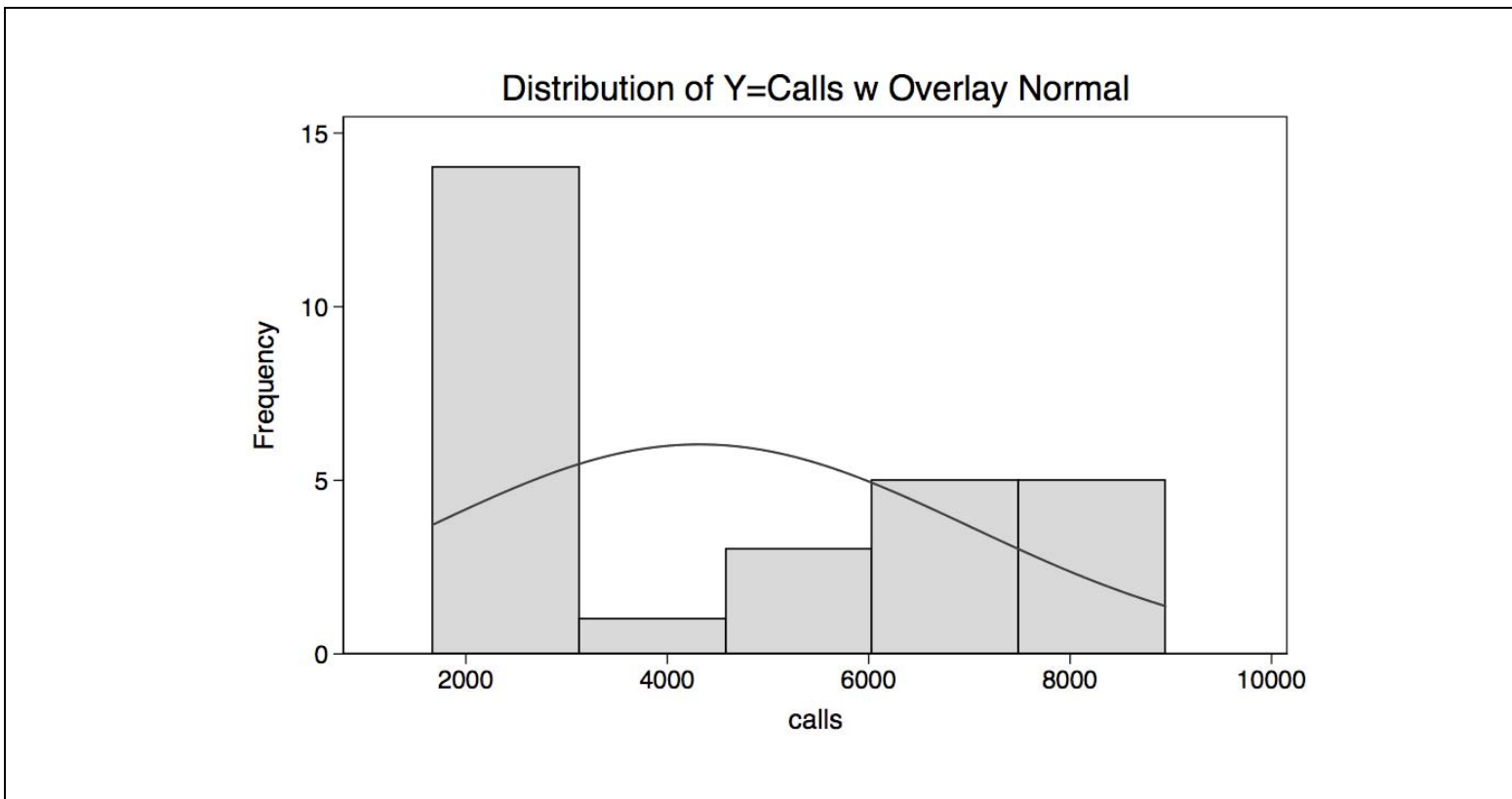
Shapiro-Wilk W test for normal data

| Variable | Obs | W | V | z | Prob>z |
|----------|-----|---------|-------|-------|---------|
| calls | 28 | 0.82916 | 5.159 | 3.378 | 0.00037 |



The null hypothesis of normality of Y=calls is rejected (p-value = .00037). Tip-sometimes the cure is worse than the original violation. For now, we'll charge on.

```
. ***** Histogram with Overlay Normal for Assessment of Normality of Outcome
. ***** histogram YVARIABLE, frequency normal title("TITLE")
. histogram calls, frequency normal title("Distribution of Y=Calls w Overlay Normal")
(bin=5, start=1674, width=1454.6)
```



No surprise here, given that the Shapiro Wilk test rejected normality. This graph confirms non-linearity of the distribution of Y =calls.

```
. ***** Fit and ANOVA Table
. ***** regress YVARIABLE XVARIABLE
. regress calls low
```

| Source | SS | df | MS | Number of obs = | 28 |
|----------|------------|----|------------|-----------------|--------|
| Model | 100233719 | 1 | 100233719 | F(1, 26) = | 27.28 |
| Residual | 95513596.2 | 26 | 3673599.85 | Prob > F = | 0.0000 |
| Total | 195747315 | 27 | 7249900.56 | R-squared = | 0.5121 |
| | | | | Adj R-squared = | 0.4933 |
| | | | | Root MSE = | 1916.7 |

| calls | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] |
|-------|----------|-----------|-------|-------|----------------------|
| low | -145.154 | 27.78868 | -5.22 | 0.000 | -202.2744 -88.03352 |
| _cons | 7475.849 | 704.6304 | 10.61 | 0.000 | 6027.46 8924.237 |

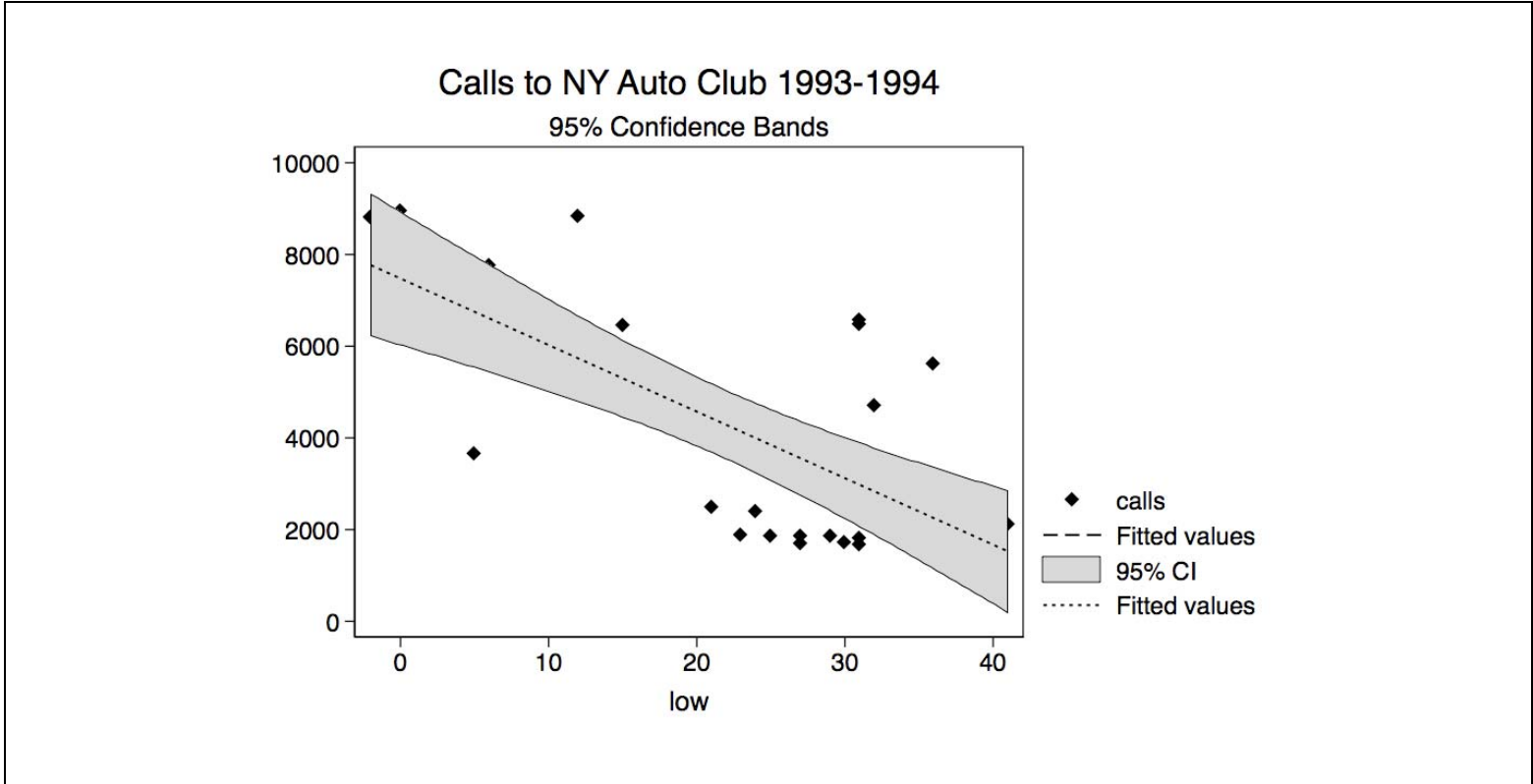
Remarks

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

| Source | Df | Sum of Squares | Mean Square |
|-----------------------|------------|--|----------------------------|
| Model “Regression” | 1 | $MSS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = 100,233,719$ | $MSS/1 = 100,233,719$ |
| Residual “Error” | (n-2) = 26 | $RSS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = 95,513,596.2$ | $RSS/(n-2) = 3,673,599.85$ |
| Total, corrected | (n-1) = 27 | $TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2 = 195,747,315$ | |

```

. * Scatterplot with overlay fit and overlay 95% confidence band
. ***** graph twoway (scatter YVARIABLE XVARIABLE, symbol(d)) (lfit YVARIABLE XVARIABLE)
(lfitci YVARIABLE XVARIABLE), title("TITLE") subtitle("TITLE")
. graph twoway (scatter calls low, symbol(d)) (lfit calls low) (lfitci calls low), title("Calls
to NY Auto Club 1993-1994") subtitle("95% Confidence Bands")
    
```

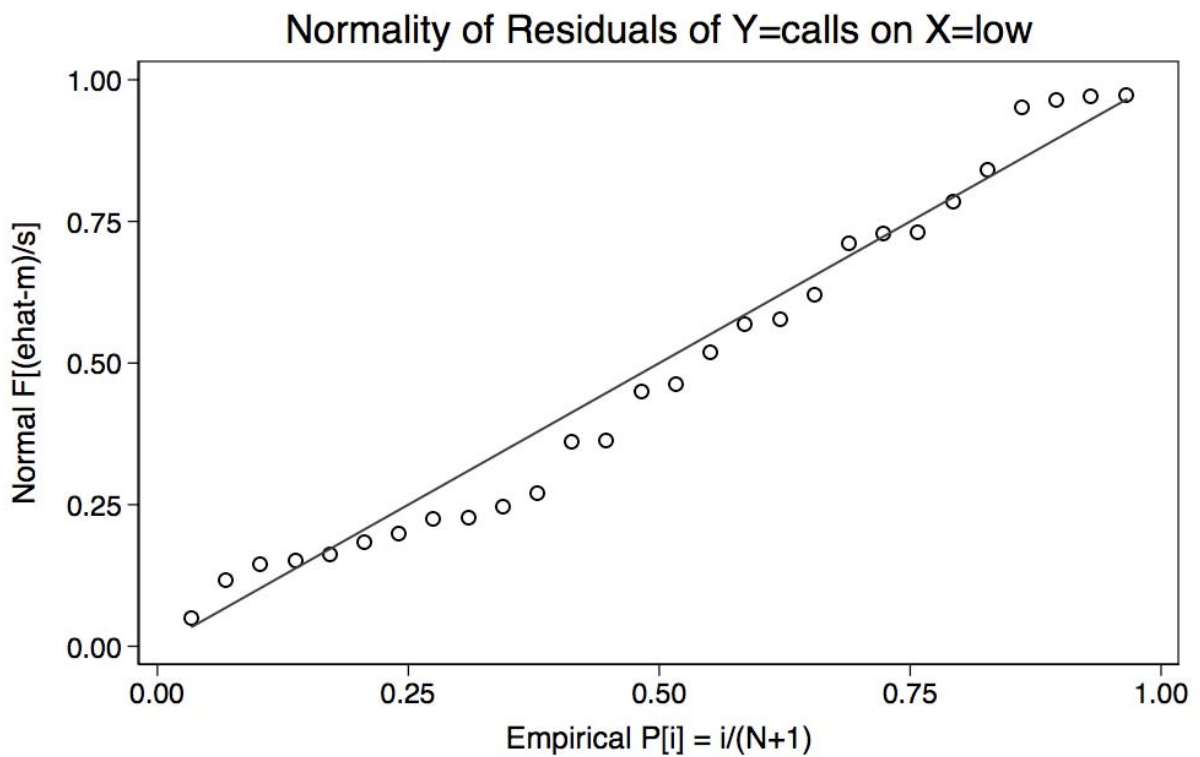


Remarks

- 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.

```
. ***** Residuals Analysis - Normalilty of residuals
. ***** Look for points falling on the line
. ***** predict NAME, residuals
. predict ehat, residuals

. ***** pnorm NAME, title("TITLE")
. pnorm ehat, title("Normality of Residuals of Y=calls on X=low")
```



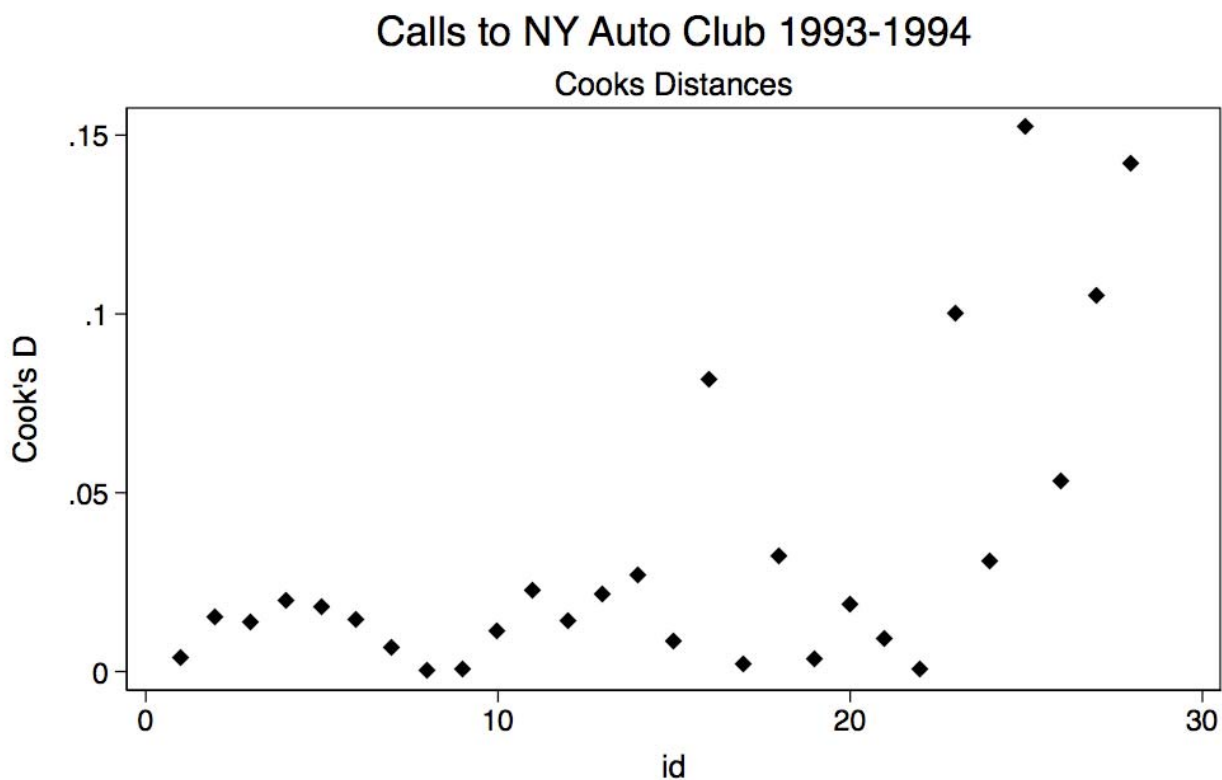
Not bad actually!

```

. ***** Residuals Analysis - Cook Distances
. ***** Look for even band of Cook Distance values with no extremes
. ***** predict NAMECOOK, cooks_d
. predict cookhat, cooks_d
. generate id=_n

. ***** graph twoway (scatter NAMECOOK id, symbol(d)), title("TITLE IN QUOTES")
. subtitle("TITLE IN QUOTES")
. graph twoway (scatter cookhat id, symbol(d)), title("Calls to NY Auto Club 1993-1994")
. subtitle("Cooks Distances")

```



Remarks

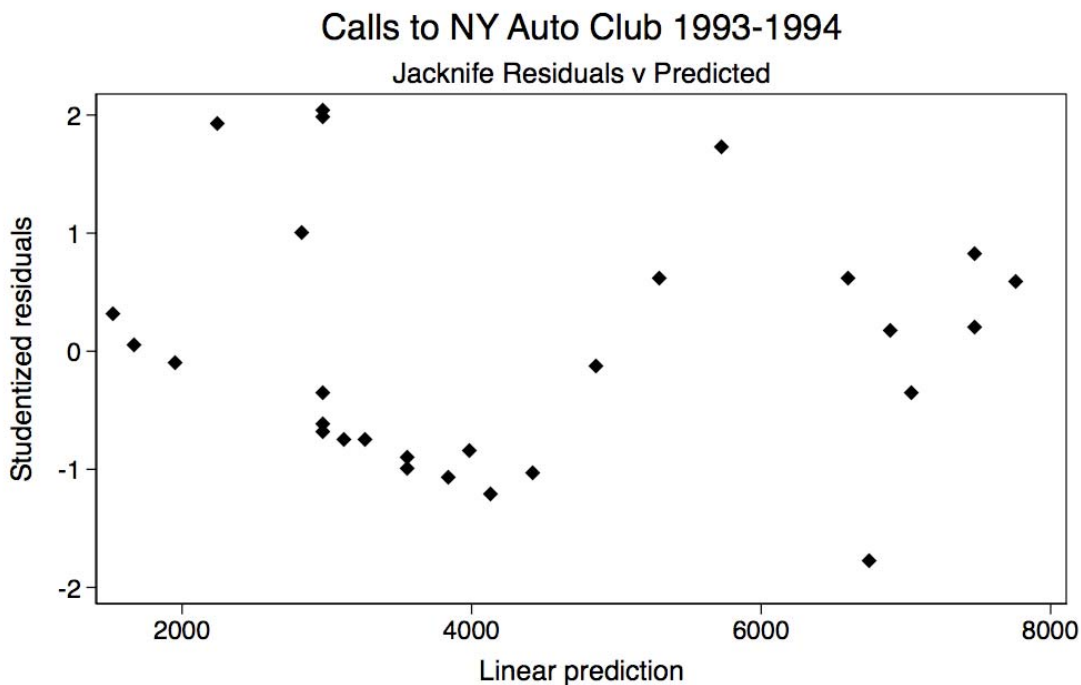
- For straight line regression, the suggestion is to regard Cook's Distance values > 1 as significant.
- Here, there are no unusually large Cook Distance values.
- Not shown but useful, too, are examinations of leverage and jackknife residuals.

```

. ***** Check Linearity, Heteroscedascity & Independence Using Jackknife Residuals
. ***** note - Stata calls these studentized
. ***** predict NAMEPREDICTED, xb
. ***** predict NAMEJACKKNIFE, rstudent
. predict yhat, xb
. predict jack, rstudent

. ***** graph twoway (scatter NAMEJACKKNIFE NAMEPREDICTED, symbol(d)), title("TITLE")
. graph twoway (scatter jack yhat, symbol(d)), title("Calls to NY Auto Club 1993-1994")
. subtitle("Jackknife Residuals v Predicted")

```



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.*
- *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.*

```
. log close  
. exit
```