

Unit 9

Stata for Normal Theory Regression

version 16

SUPPLEMENT

Illustrations: Simple and Multiple Linear Regression

<b>I- Simple Linear Regression .....</b>	<b><u>2</u></b>
1. Introduction to Example .....	2
2. Preliminaries: Descriptives .....	3
3. Model Fitting (Estimation) .....	7
4. Model Examination .....	8
5. Checking Model Assumptions and Fit .....	9
<b>II – Multiple Linear Regression .....</b>	<b><u>12</u></b>
1. Introduction to Example .....	12
2. Preliminaries: Descriptives .....	13
3. Handling of Categorical Predictors: Indicator Variables .....	17
4. Model Fitting (Estimation) .....	18
5. Checking Model Assumptions and Fit .....	23

Data sets used (Download from course website page, 9. Stata for Linear Regression, here:  
<https://people.umass.edu/biostat690c/webpages/statalinear.html>

1. ers.dta

2. p53paper\_small.dta

## I – Simple Linear Regression

### 1. Introduction to Example

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 BIOSTATS 690C course website directly. It is then saved to your current working directory.

Simple Linear Regression Variables:

Outcome Y = calls

Predictor X = low.

Launch Stata and input Stata data set ers.dta

```
. ***** Set working directory to directory of choice
. ***** Command is cd/YOURLIBRARY
. cd/Users/cbigelow/Desktop

. ***** Download data ers.dta from BIOSTATS 640 course website.
. ***** Launch Stata. Input data using FILE > OPEN

. ***** Save the inputted data to the directory you have chosen above
. ***** Command is save "NAME", replace
. save "ers.dta", replace
(note: file ers.dta not found)
file ers.dta saved
```

## 2. Preliminaries: Descriptives

```
.
. * 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	

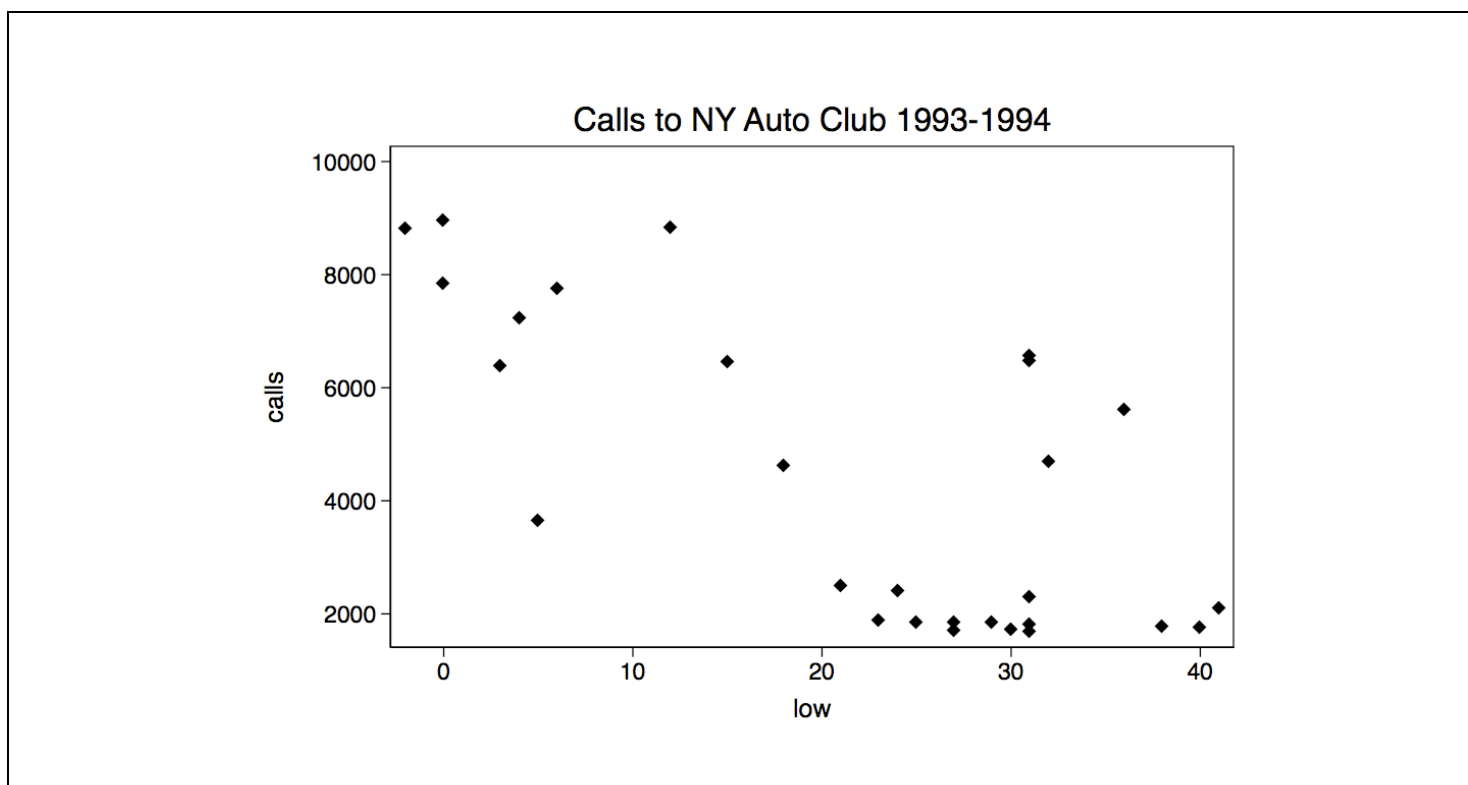
We see that this data set has n=28 observations on several variables. For this illustration of simple linear regression, we will consider just two variables: calls and low  
BEWARE – Stata is case sensitive!

```
. ***** 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

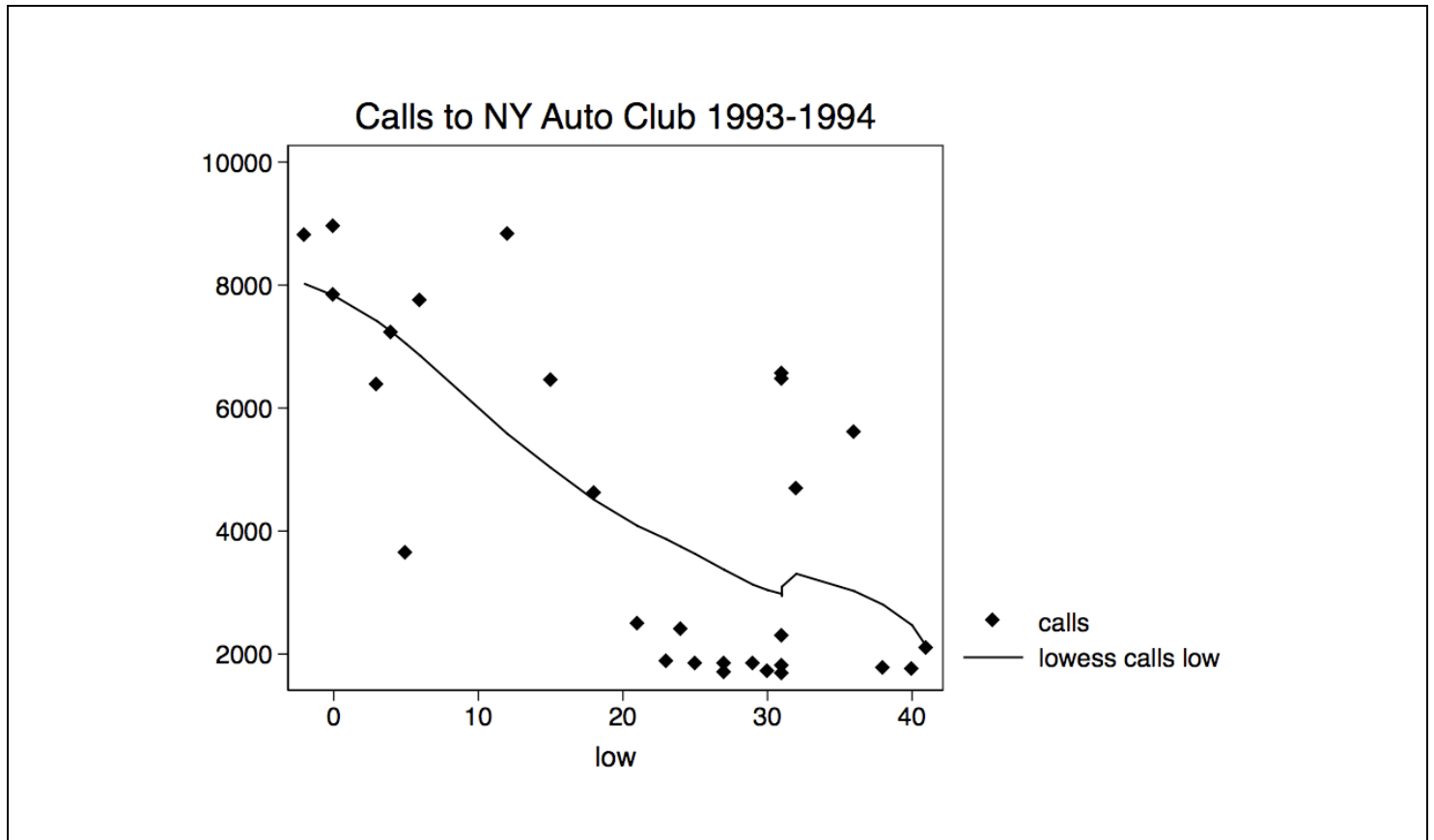
• **Design** ..... **Data Collection** ..... **Data Management** ..... **Data Summarization** ..... **Statistical Analysis** ..... **Reporting**

```
. ***** 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.*

```
. ***** 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
```

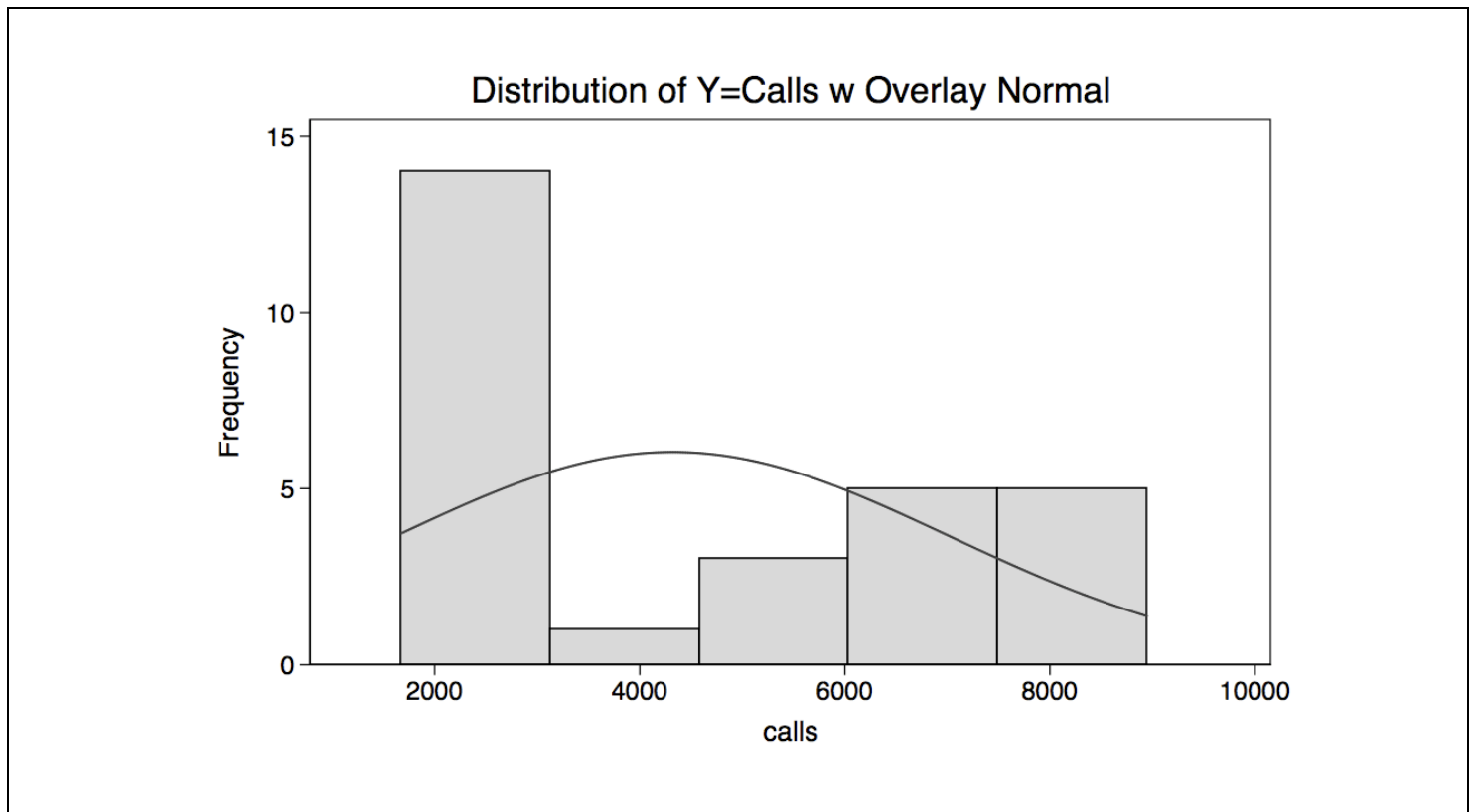
Shapiro-Wilk W test for normal data

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.*

### 3. Model Fitting (Estimation)

```
. ***** 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
				R-squared =	0.5121
				Adj R-squared =	0.4933
Total	195747315	27	7249900.56	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

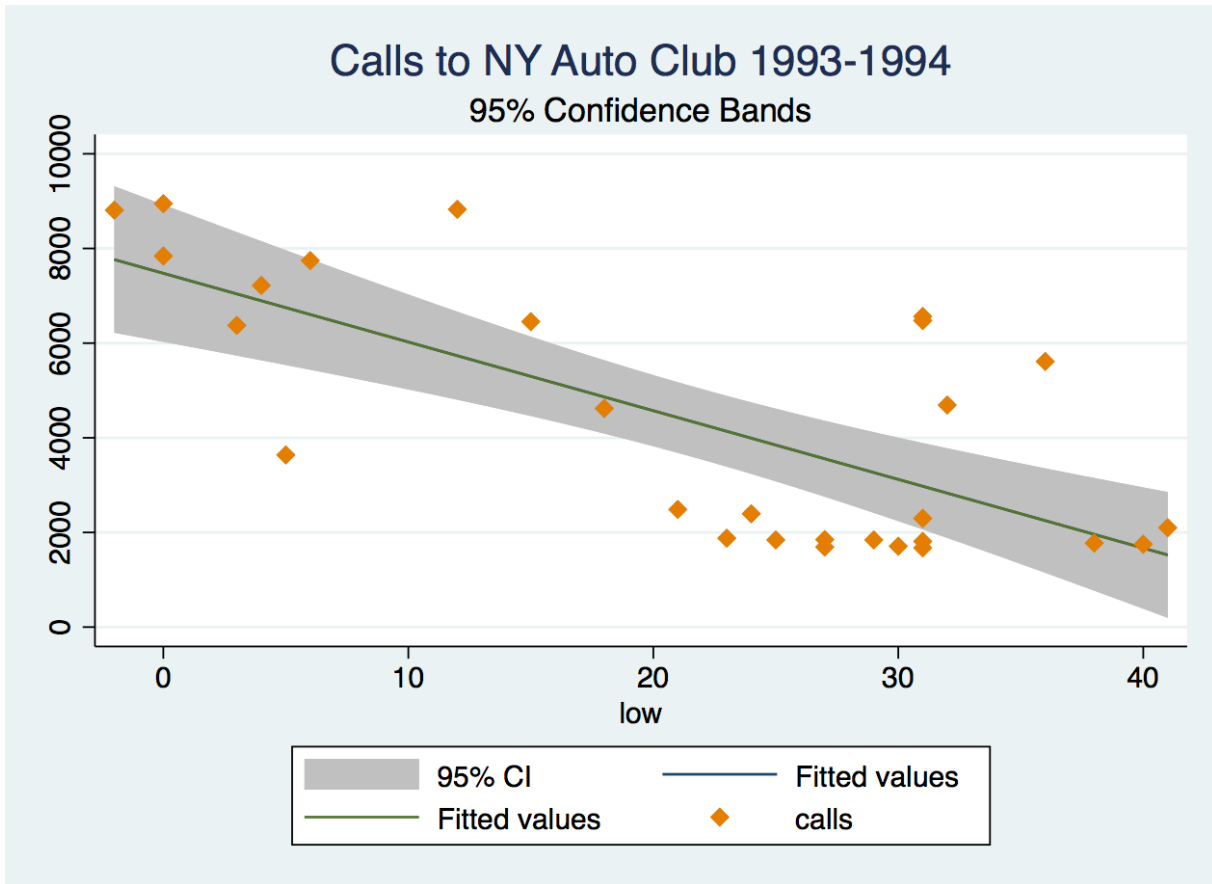
#### Remarks

- The fitted line is  $\text{calls} = 7,475.85 - 145.15[\text{low}]$
- $R^2 = .51$  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	$TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2 = 195,747,315$	

#### 4. Model Examination

```
.
. * Scatterplot with overlay fit and overlay 95% confidence band
. * Tip! - Because of layering: confidence interval first, then fit, then data points
. ***** graph twoway (scatter YVARIABLE XVARIABLE, symbol(d)) (lfit YVARIABLE XVARIABLE)
(lfitci YVARIABLE XVARIABLE), title("TITLE") subtitle("TITLE")
. graph twoway (lfitci calls low) (lfit calls low) (scatter calls low, symbol(d)),
title("Calls to NY Auto Club 1993-1994") subtitle("95% Confidence Bands")
```



#### Remarks

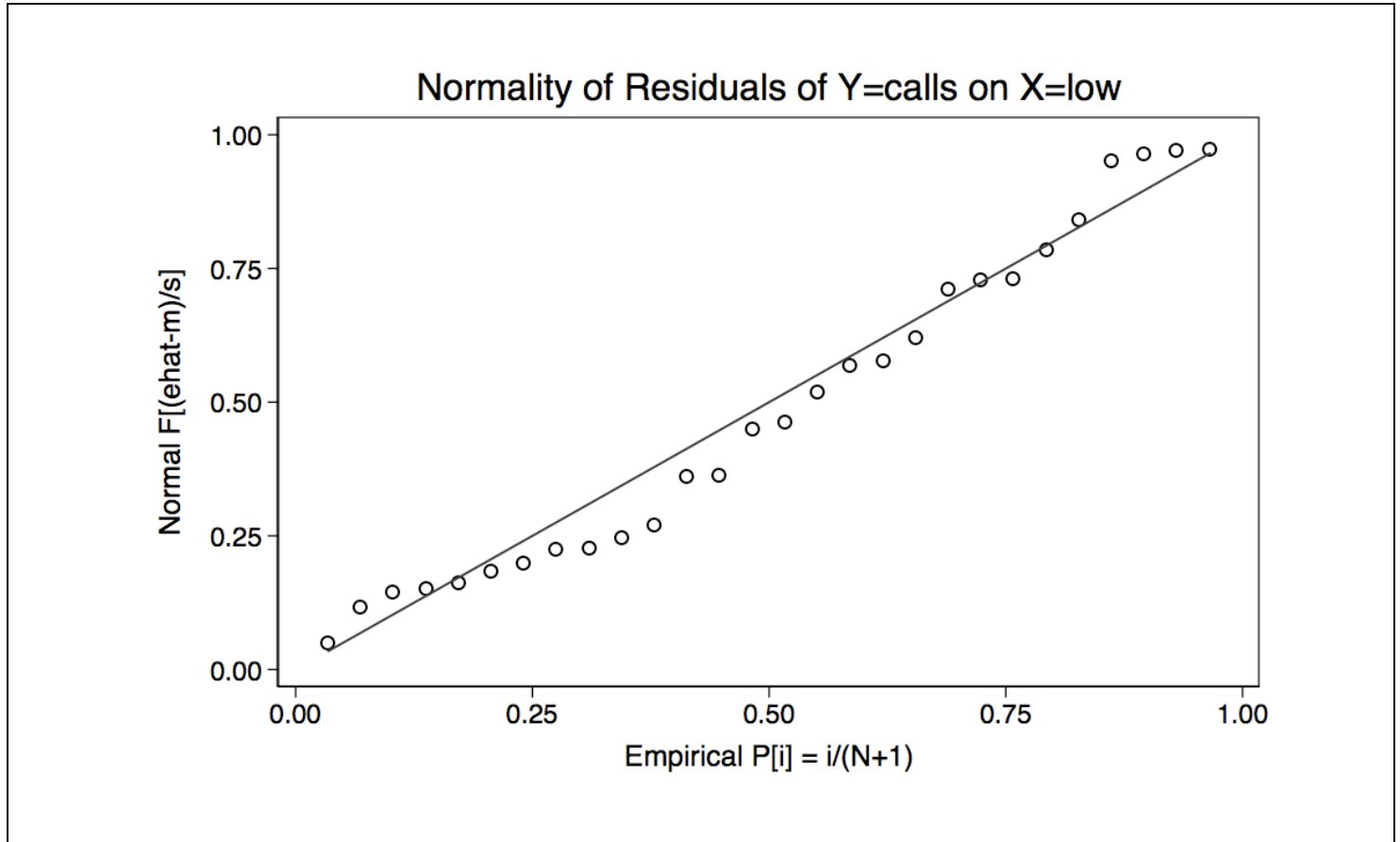
- *The overlay of the straight line fit is reasonable but substantial variability is seen, too.*
- *There is a lot we still don't know, including but not limited to the following ---*
- *Case influence, omitted variables, variance heterogeneity, incorrect functional form, etc.*



## 5. Checking Model Assumptions and Fit

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

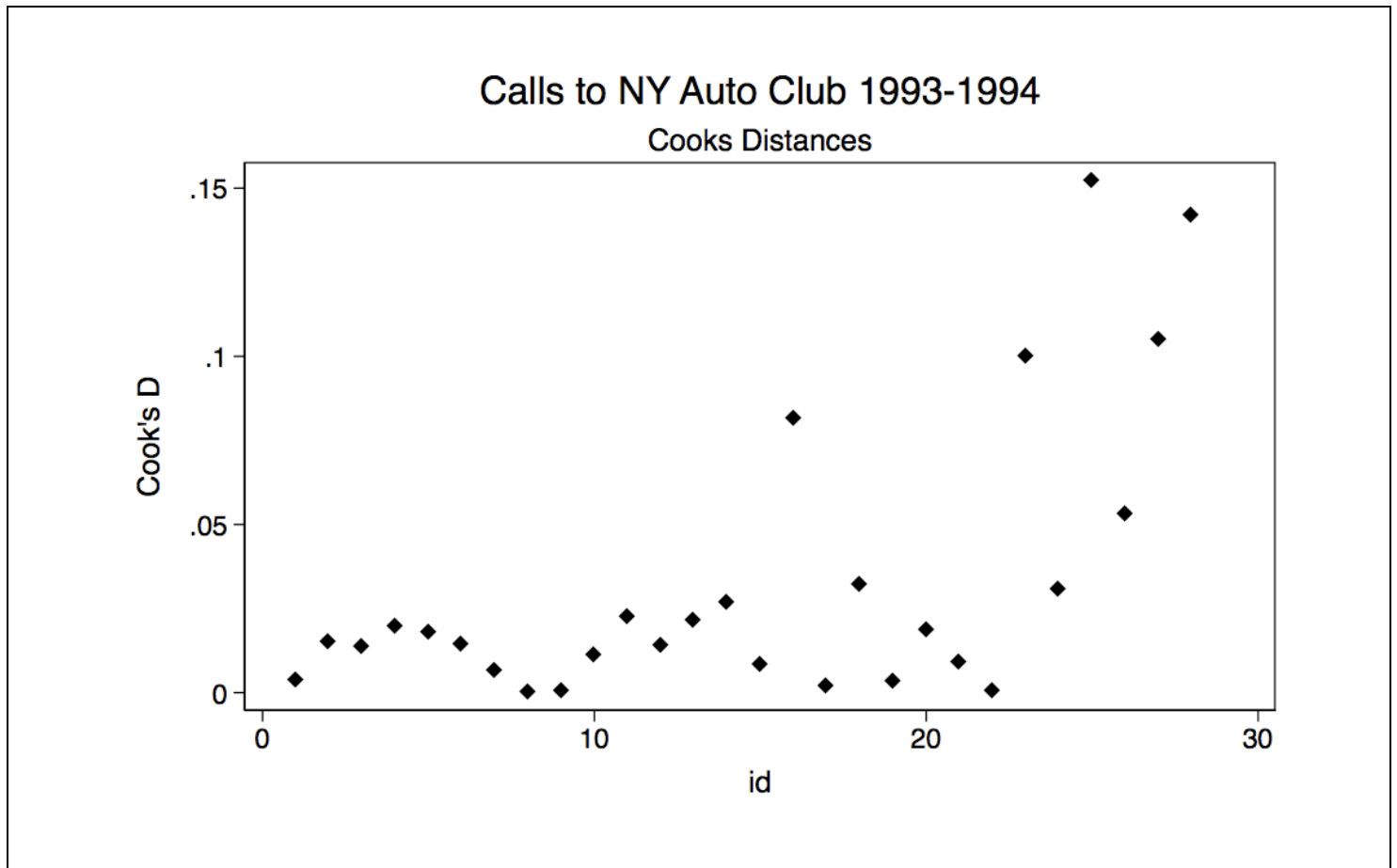
. ***** pnorm NAME, title("TITLE")
. pnorm ehat, title("Normality of Residuals of Y=calls on X=low")
```



*Not bad actually!*

```
. ***** Residuals Analysis - Cook Distances
. ***** Look for even band of Cook Distance values with no extremes
. ***** predict NAMECOOK, cooksd
. predict cookhat, cooksd
. 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")
```

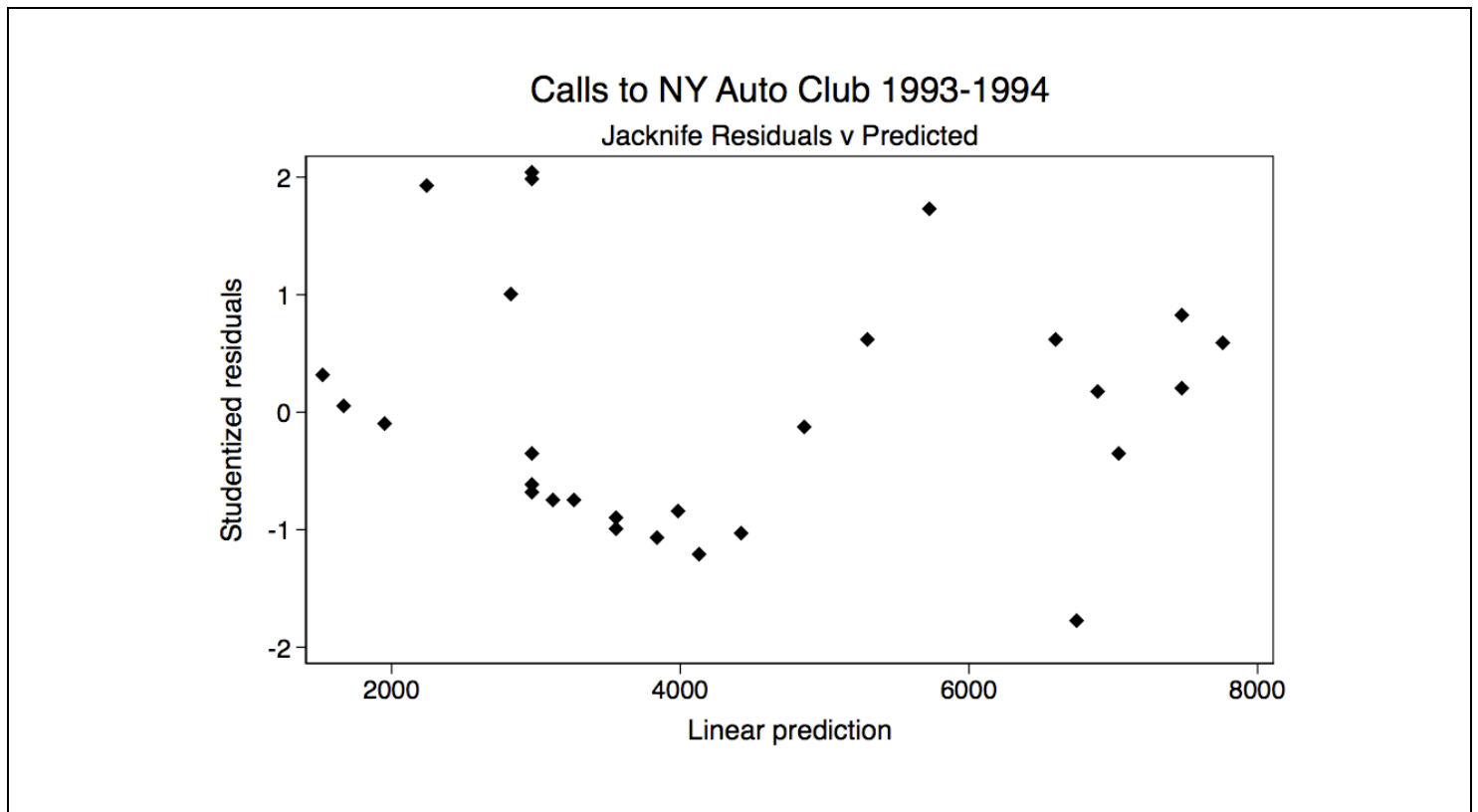


*Remarks*

- 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 NAMEJACKNIFE, rstudent
. predict yhat, xb
. predict jack, rstudent

. ***** graph twoway (scatter NAMEJACKNIFE NAMEPREDICTED, symbol(d)), title("TITLE")
. subtitle("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.

## II – Multiple Linear Regression

### 1. Introduction to Example

#### Source:

Matthews et al. Parity Induced Protection Against Breast Cancer 2007.

#### Research Question:

What is the relationship of **Y=p53 expression** to **parity** and **age at first pregnancy**, after adjustment for the potentially confounding effects of **current age** and **menopausal status**. Age at first pregnancy has been grouped and is either  $\leq 24$  years or  $> 24$  years.

Input Stata data set p53paper\_small.dta

```
. ***** Just to be safe! save the ers.dta data again
. save "ers.dta", replace
file ers.dta saved

. ***** Clear the workspace
. ***** Command is clear
. clear

. ***** Download data 53paper_small.dta from BIOSTATS 690C course website.
. ***** Input data using FILE > OPEN

. ***** Save the inputted data to the directory you have chosen above
. ***** Command is save "NAME", replace
. save "p53paper_small.dta", replace
(note: file p53paper_small.dta not found)
file p53paper_small.dta saved
```

## 2. Preliminaries: Descriptives

```
. *
. ***** Explore the data for shape, range, outliers and completeness.
. summarize p53 pregnum agefirst agecurr menop
```

Variable	Obs	Mean	Std. Dev.	Min	Max
p53	67	3.251493	1.054454	1	6
pregnum	67	1.656716	1.122122	0	3
agefirst	67	1.044776	.7268203	0	2
agecurr	67	39.62687	13.69786	15	75
menop	67	.2835821	.4541382	0	1

Data are complete; n=67 for every variable. Y=p53 has a limited range, so that the assumption of normality is a bit dicey, but we'll proceed anyway. Current age (agecurr) ranges 15 to 75.

```
. *
. ***** Pairwise correlations for all the variables
. pwcorr p53 pregnum agefirst agecurr menop, star(0.05) sig
```

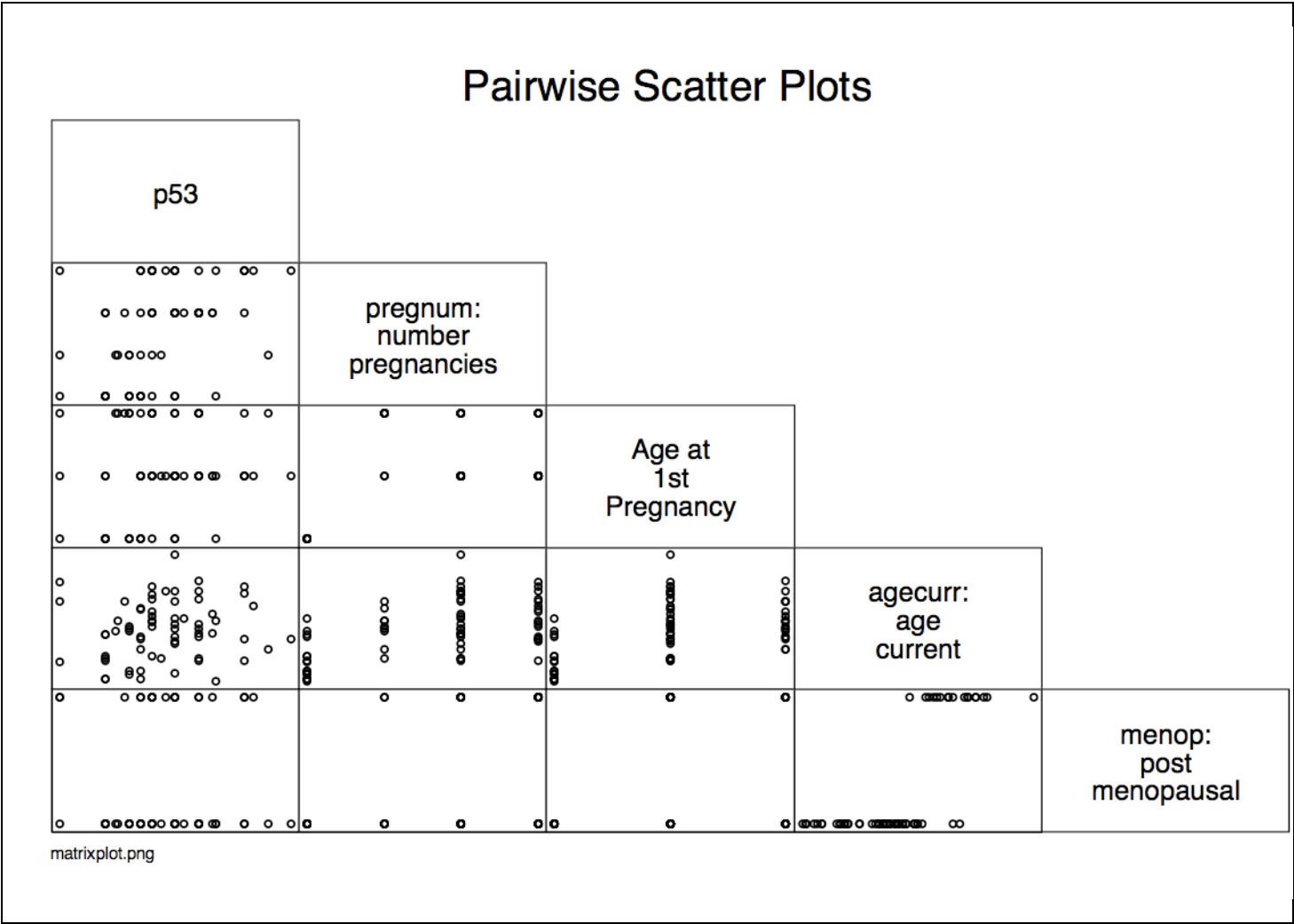
	p53	pregnum	agefirst	agecurr	menop
p53	1.0000	correlation(pregnum, p53) = 0.4419			
pregnum	0.4419*	1.0000			
	0.0002				
		p-value for null (zero correlation) = .0002 → Reject null.			
agefirst	0.2021	0.5765*	1.0000		
	0.1011	0.0000			
agecurr	0.1340	0.5416*	0.4765*	1.0000	
	0.2798	0.0000	0.0000		
menop	0.0450	0.4021*	0.2823*	0.7285*	1.0000
	0.7178	0.0007	0.0207	0.0000	

Only one correlation with Y=p53 is statistically significant  $r(p53, pregnum) = .44$  with  $p\text{-value} = .0002$ . Note that some of the predictors are statistically significantly correlated with each other:  $r(agefirst, pregnum) = .58$  with  $p\text{-value} < .0001$ .

```

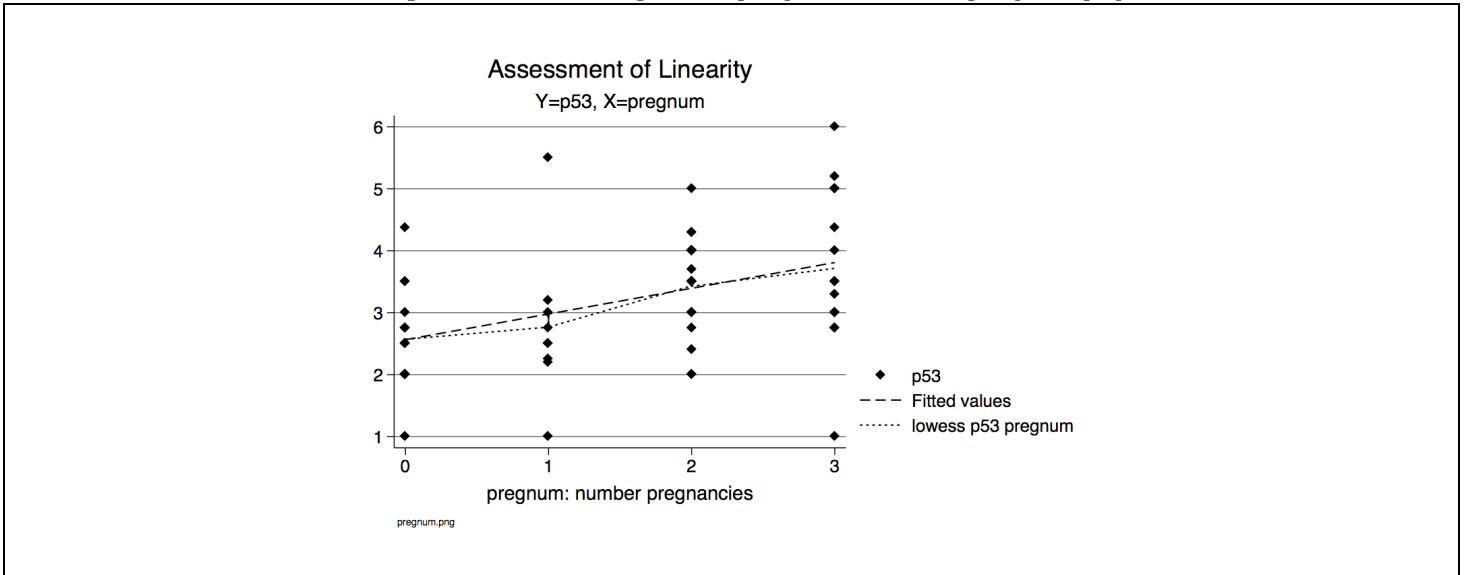
. *
. ***** Pairwise scatterplots for all the variables
. set scheme lean2
. graph matrix p53 pregnum agefirst agecurr menop, half maxis(ylabel(none) xlabel(none))
title("Pairwise Scatter Plots") note("matrixplot.png", size(vsmall))

```



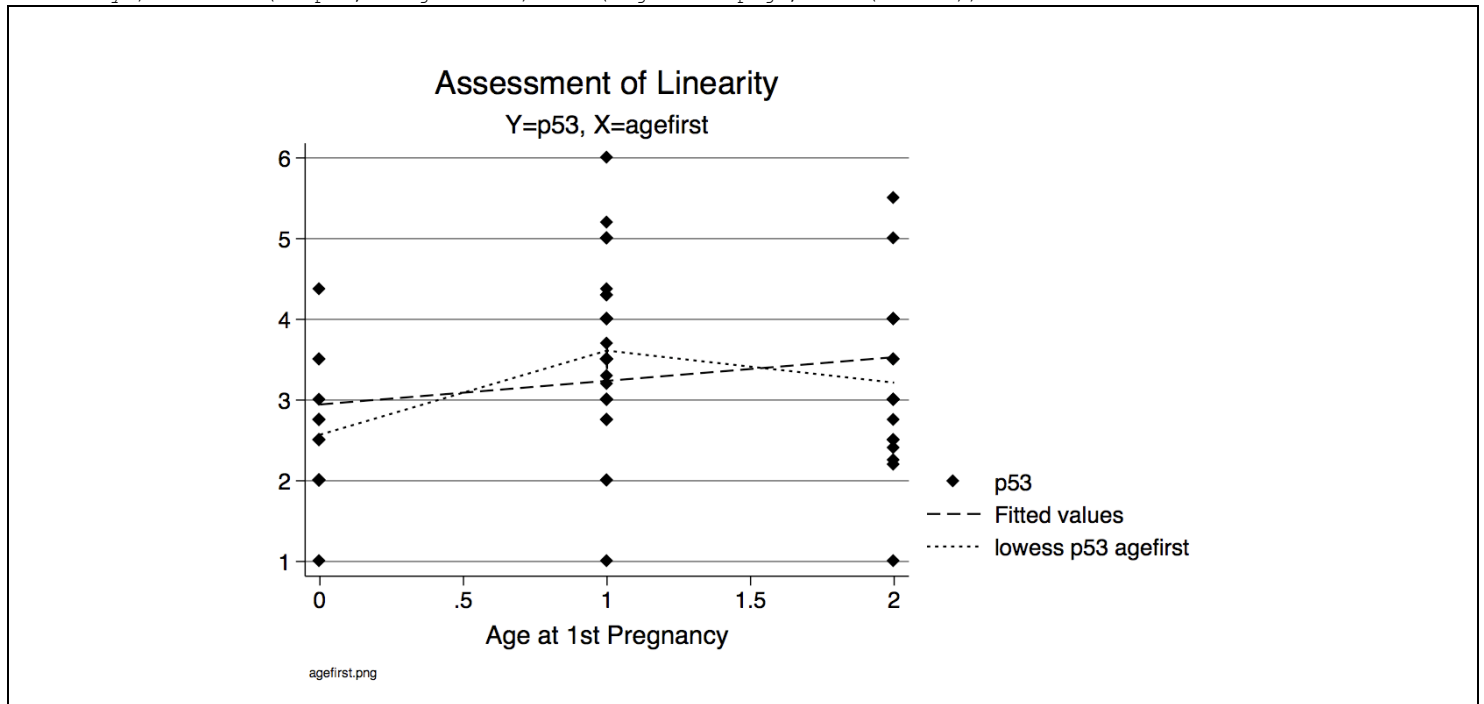
Admittedly, it’s a little hard to see a lot going on here.

```
. *
. ***** Graphical assessment of linearity of Y = p53 in predictors, with line and lowess fits
. *
. ***** pregnum
. graph twoway (scatter p53 pregnum, symbol(d)) (lfit p53 pregnum) (lowess p53 pregnum),
title("Assessment of Linearity") subtitle("Y=p53, X=pregnum") note("pregnum.png", size(vsmall))
```

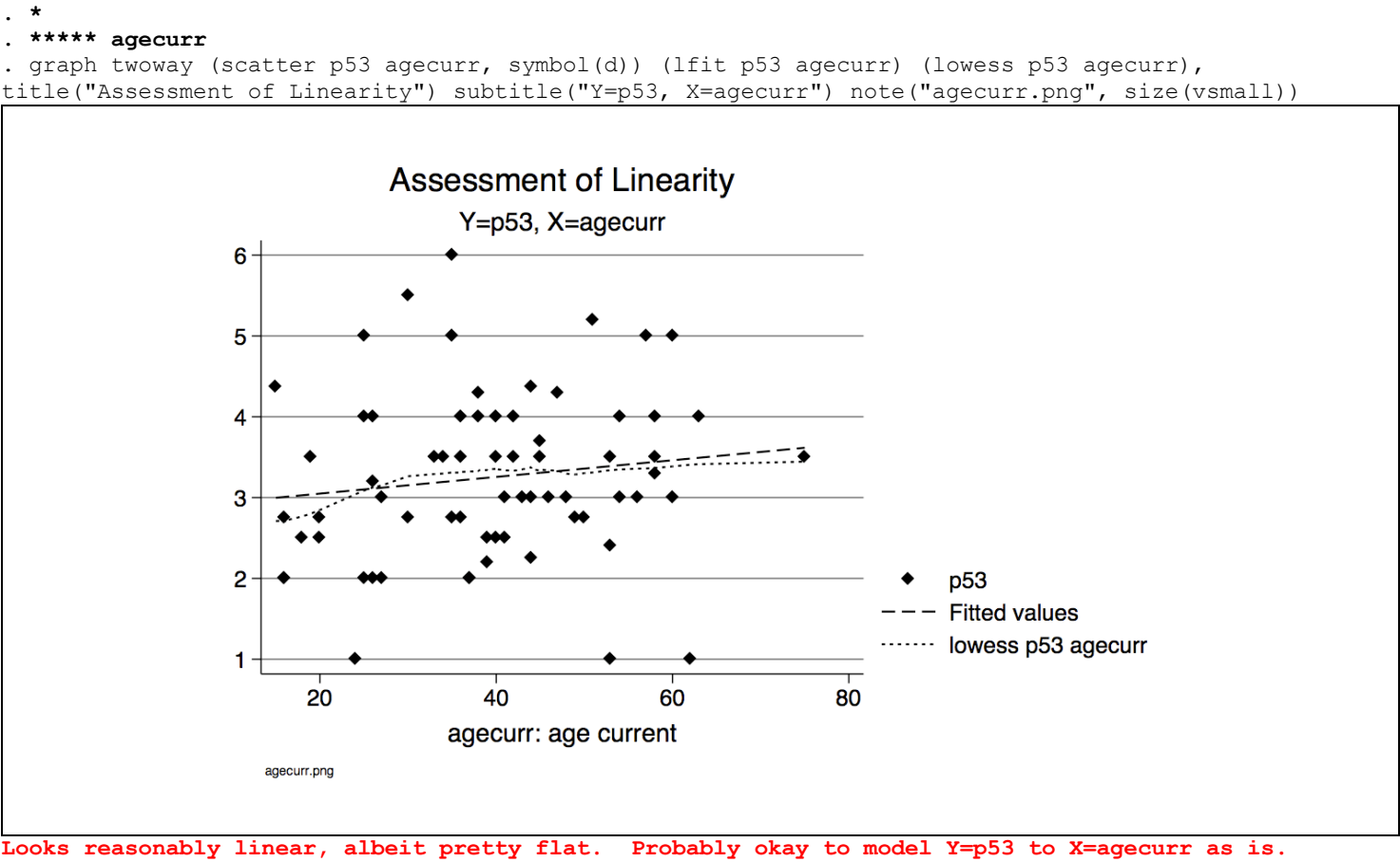


Looks reasonably linear. Probably okay to model  $Y=p53$  to  $X=pregnum$  as is, instead of with dummies.

```
. *
. ***** agefirst
. graph twoway (scatter p53 agefirst, symbol(d)) (lfit p53 agefirst) (lowess p53 agefirst), title("Assessment of Linearity") subtitle("Y=p53, X=agefirst") note("agefirst.png", size(vsmall))
```



This does not look linear. So we will create dummies for age at 1<sup>st</sup> pregnancy.



Looks reasonably linear, albeit pretty flat. Probably okay to model Y=p53 to X=agecurr as is.



### 3. Handling of Categorical Predictors: Indicator Variables

```
. *
. ***** Create Dummy variables for age at first pregnancy: early, late. Check.
. generate early=0
. replace early=1 if agefirst==1
(32 real changes made)

. generate late=0
. replace late=1 if agefirst==2
(19 real changes made)

. tab2 agefirst early
-> tabulation of agefirst by early
```

Age at 1st Pregnancy	early		Total
	0	1	
never pregnant	16	0	16
age ≤ 24	0	32	32
age > 24	19	0	19
Total	35	32	67

Check using tab2 confirms that the new variable, early, is well defined.

```
. tab2 agefirst late
-> tabulation of agefirst by late
```

Age at 1st Pregnancy	late		Total
	0	1	
never pregnant	16	0	16
age ≤ 24	32	0	32
age > 24	0	19	19
Total	48	19	67

Ditto. The new variable, late, is well defined.

```
. label variable early "Age ≤ 24"
. label variable late "Age > 24"
```

## 4. Model Fitting (Estimation)

```
. * -----
. * Model Estimation Set I: Determination of best model in the predictors of interest.
. * Goal is to obtain best parameterization before considering covariates.
. * -----
```

```
. *
. ***** Maximal model: Regression of Y=p53 on all:  pregnum + [early, late]
. regress p53 pregnum early late
```

Source	SS	df	MS	Number of obs =	67
Model	14.8967116	3	4.96557054	F( 3, 63) =	5.35
Residual	58.486889	63	.928363317	Prob > F =	0.0024
Total	73.3836006	66	1.11187274	R-squared =	0.2030
				Adj R-squared =	0.1650
				Root MSE =	.96352

p53	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pregnum	.3764082	.2008711	1.87	0.066	-.0250006 .7778171
early	.160762	.5555887	0.29	0.773	-.9494935 1.271017
late	-.0677176	.5017357	-0.13	0.893	-1.070356 .9349211
_cons	2.570313	.240879	10.67	0.000	2.088954 3.051671

The fitted line is  $p53 = 2.57 + (0.38)*pregnum + (0.16)*early - (0.07)*late$ .

20% of the variability in Y=p53 is explained by this model (R-squared = .20)

This model is statistically significantly better than the null model (p-value of F test = .0024)

**NOTE!!** We see a consequence of the multi-collinearity of our predictors [early, late], pregnum [early, late] have NON-significant t-statistic p-values: early and late pregnum has a t-statistic p-value that is only marginally significant.

```
. *
. ***** 2 df Partial F-test ( Null: [early, late] are not significant, controlling for pregnum).
. testparm early late
```

```
( 1)  early = 0
( 2)  late = 0
```

```
F( 2, 63) = 0.31
Prob > F = 0.7381
```

Not significant (p-value = .74). Conclude that, in the adjusted model containing pregnum, [early, late] are not statistically significantly associated with Y=p53.

```
. *
.***** 1 df Partial F-test (Null: pregnum is not significant, controlling for [early, late] )
. testparm pregnum

( 1)  pregnum = 0

      F( 1,    63) =    3.51
      Prob > F =    0.0656
```

Marginally statistically significant (p=value = .0656). The null hypothesis is rejected. Conclude that, in the model that contains [early, late], pregnum is marginally statistically significantly associated with Y=p53.

```
. *
.***** Save results from model above to "model1" for tabulation later.
. eststo model1
```

```
. *
.***** Regression of Y=p53 on pregnum only. [early, late] dropped.
. regress p53 pregnum
```

Source	SS	df	MS	Number of obs =	67
Model	14.330079	1	14.330079	F( 1, 65) =	15.77
Residual	59.0535216	65	.908515716	Prob > F =	0.0002
				R-squared =	0.1953
				Adj R-squared =	0.1829
Total	73.3836006	66	1.11187274	Root MSE =	.95316

p53	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pregnum	.4152523	.1045572	3.97	0.000	.2064372 .6240675
_cons	2.563537	.2087239	12.28	0.000	2.146687 2.980388

The fitted line is  $p53 = 2.56 + (0.41) \cdot \text{pregnum}$ .

19.5% of the variability in Y=p53 is explained by this model (R-squared = .1953)

This model is statistically significantly more explanatory than the null model (p-value = .0002)

```
. *
.***** Save results from model above to "model2" for tabulation later.
. eststo model2
```

```
. *
. ***** Regression of Y=p53 on design variables [early, late] only.  pregnum dropped.
. regress p53 early late
```

Source	SS	df	MS	Number of obs =	67
Model	11.6368338	2	5.81841692	F( 2, 64) =	6.03
Residual	61.7467667	64	.96479323	Prob > F =	0.0040
				R-squared =	0.1586
				Adj R-squared =	0.1323
Total	73.3836006	66	1.11187274	Root MSE =	.98224

p53	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
early	1.042969	.300748	3.47	0.001	.4421555 1.643782
late	.645477	.3332839	1.94	0.057	-.0203342 1.311288
_cons	2.570313	.2455597	10.47	0.000	2.079751 3.060874

```
. *
. ***** Save results from model above to "model2" for tabulation later.
. eststo model3

. *
. ***** SUMMARY of Model Estimation Set I.
. esttab, r2 se scalar(rmse)
```

	(1) p53	(2) p53	(3) p53
pregnum	0.376 (0.201)	0.415*** (0.105)	
early	0.161 (0.556)		1.043*** (0.301)
late	-0.0677 (0.502)		0.645 (0.333)
_cons	2.570*** (0.241)	2.564*** (0.209)	2.570*** (0.246)
N	67	67	67
R-sq	0.203	0.195	0.159
rmse	0.964	0.953	0.982

Standard errors in parentheses  
\* p<0.05, \*\* p<0.01, \*\*\* p<0.001

Choose model "(2)" as a good "minimally adequate" model: Y=p53 and X=pregnum. This is why.

(1) Model "(1)" is the maximal model. R-squared = .20  
(2) Model "(2)" drops [early,late]. R-squared is minimally lower: R-squared = .195  
(3) Model "(3)" drops pregnum. R-square drop is more substantial: R-squared = .159

```
. * -----
. * Model Estimation Set II: Regression of Y=p53 on parity with adjustment for
. * covariates
. * -----

. *
. ***** Preliminary: Clear the saved models.
. eststo clear

. *
. ***** Maximal model: Regression of Y=p53 on pregnum + covariates
. regress p53 pregnum agecurr menop
```

Source	SS	df	MS	Number of obs =	67
Model	15.9827039	3	5.32756796	F( 3, 63) =	5.85
Residual	57.4008967	63	.911125345	Prob > F =	0.0014
				R-squared =	0.2178
				Adj R-squared =	0.1805
Total	73.3836006	66	1.11187274	Root MSE =	.95453

```
-----
```

p53	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pregnum	.4923299	.1245663	3.95	0.000	.2434041 .7412557
agecurr	-.0047726	.0136385	-0.35	0.728	-.032027 .0224819
menop	-.2797843	.3776867	-0.74	0.462	-1.034531 .4749624
_cons	2.704305	.440352	6.14	0.000	1.824332 3.584279

```
-----

. eststo model1

. *
. ***** Regression of Y=p53 on pregnum + menop only. Agecurr dropped.
. regress p53 pregnum menop
```

Source	SS	df	MS	Number of obs =	67
Model	15.8711336	2	7.93556682	F( 2, 64) =	8.83
Residual	57.5124669	64	.898632296	Prob > F =	0.0004
				R-squared =	0.2163
				Adj R-squared =	0.1918
Total	73.3836006	66	1.11187274	Root MSE =	.94796

```
-----
```

p53	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pregnum	.4750472	.1135703	4.18	0.000	.2481644 .70193
menop	-.3674811	.280619	-1.31	0.195	-.9280819 .1931197
_cons	2.568685	.2076227	12.37	0.000	2.153911 2.983459

```
-----

. eststo model2
```

```
. *
.***** SUMMARY of Model Estimation Set II.
. esttab, r2 se scalar(rmse)
```

	(1)	(2)
	p53	p53
pregnum	0.492*** (0.125)	0.475*** (0.114)
agecurr	-0.00477 (0.0136)	
menop	-0.280 (0.378)	-0.367 (0.281)
_cons	2.704*** (0.440)	2.569*** (0.208)
N	67	67
R-sq	0.218	0.216
rmse	0.955	0.948

```
Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

Choose as "candidate" final model, model (2): Y=p53 and X=pregnum.

## 5. Checking Model Assumptions and Fit

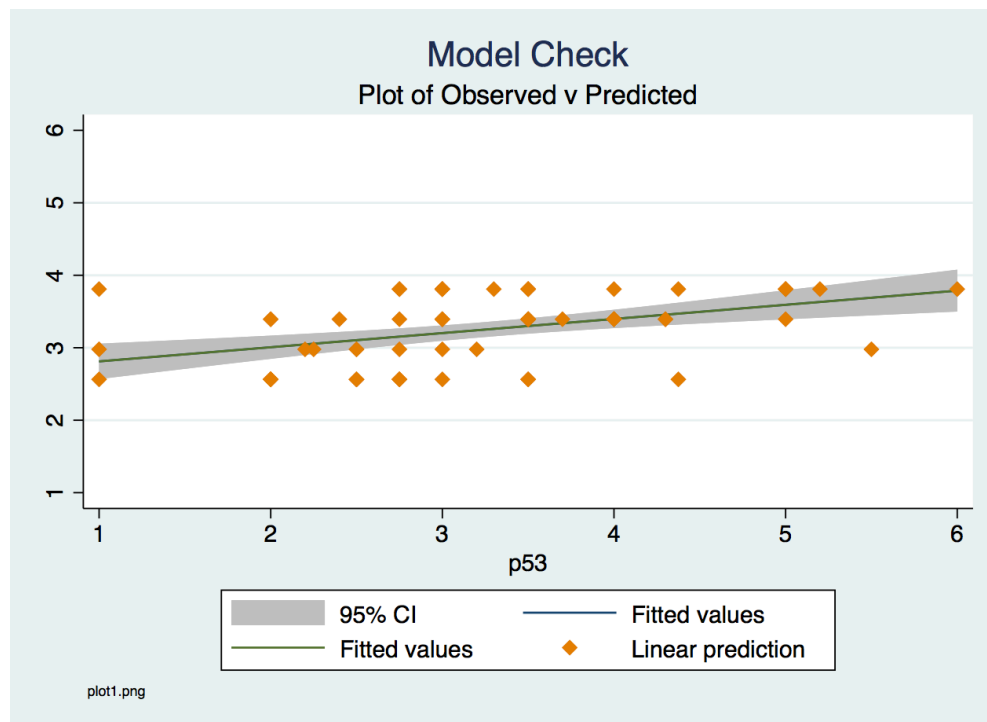
```
. *
. ***** PRELIMINARY to checks: Model checks require that you have just fit the model you are checking.
. regress p53 pregnum
```

Source	SS	df	MS	Number of obs =	67
Model	14.330079	1	14.330079	F( 1, 65) =	15.77
Residual	59.0535216	65	.908515716	Prob > F =	0.0002
				R-squared =	0.1953
				Adj R-squared =	0.1829
Total	73.3836006	66	1.11187274	Root MSE =	.95316

p53	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pregnum	.4152523	.1045572	3.97	0.000	.2064372 .6240675
_cons	2.563537	.2087239	12.28	0.000	2.146687 2.980388

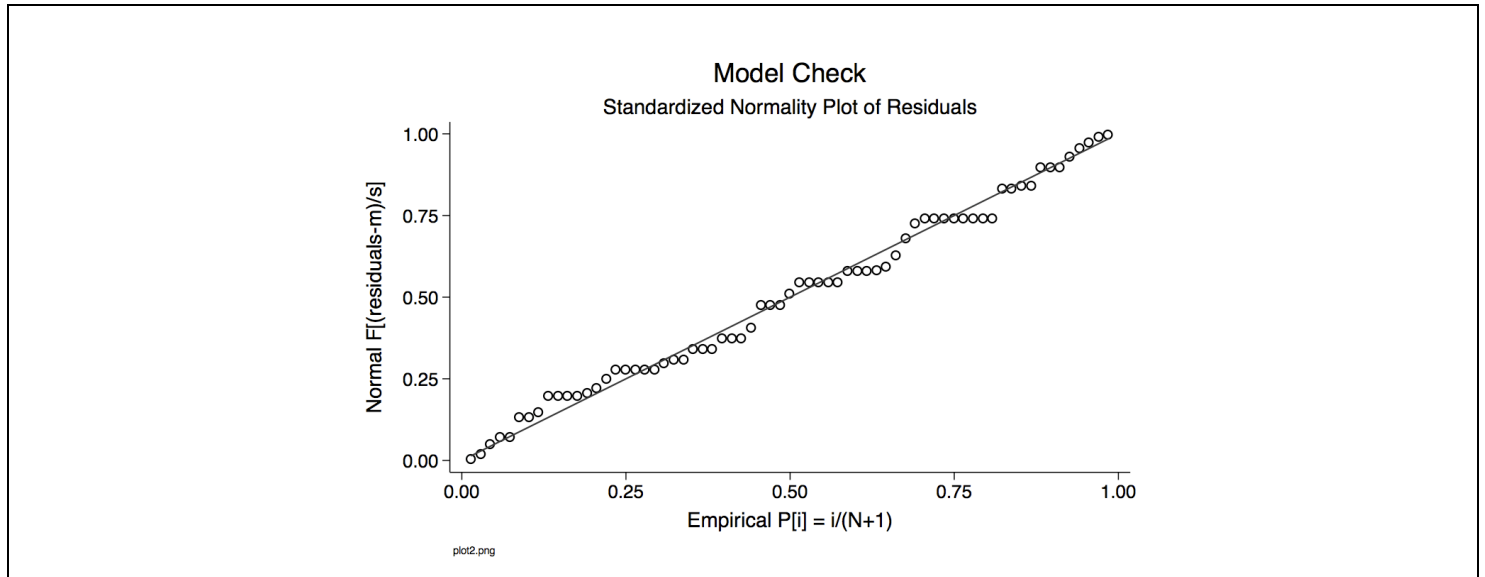
```
. *
. ***** Save the predicted values of Y in a new variable called yhat.
. predict yhat
(option xb assumed; fitted values)
```

```
. *
. ***** Plot Observed versus Predicted - Ideally, points will fall on the X=Y line.
. * Again ...Because of layering: confidence interval first, then fit, then data points
. graph twoway (lfitci yhat p53) (lfit yhat p53) (scatter yhat p53, symbol(d)), title("Model Check")
subtitled("Plot of Observed v Predicted") xlabel(1(1)6) ylabel(1(1)6) note("plot1.png", size(vsmall))
```



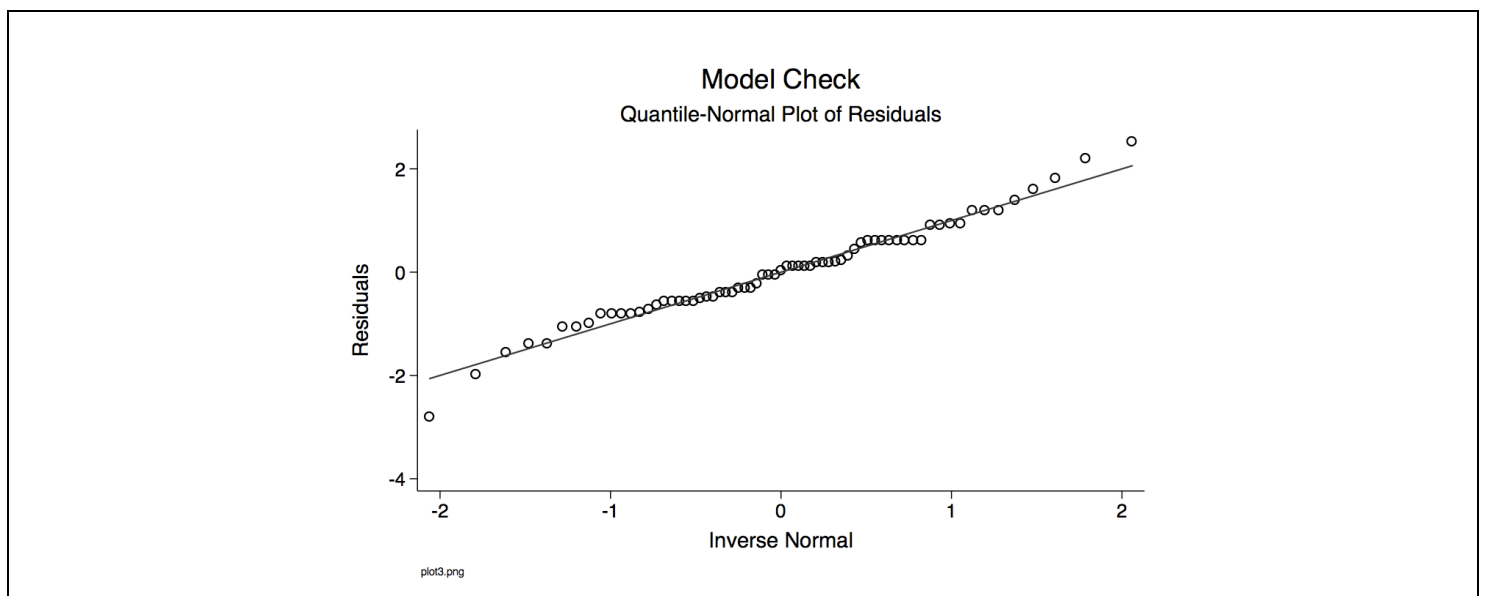
```
. ***** Normality of Residuals - Look for normality
.***** Preliminary: Save the residuals in a new variable called residuals.
. predict residuals, resid

.*
.***** pnorm plot check of normality of residuals in middle range - Ideally points fall on X=Y line
. pnorm residuals, title("Model Check") subtitle("Standardized Normality Plot of Residuals")
note("plot2.png", size(vsmall))
```



Very reasonable. No worries here.

```
. *
.***** qnorm plot check of normality of residuals in the tails - Ideally, points fall on X=Y line
. qnorm residuals, title("Model Check") subtitle("Quantile-Normal Plot of Residuals")
note("plot3.png", size(vsmall))
```



A little off the line in the tails, but okay.



```
. *
. ***** Shapiro Wilk test of normality of residuals (Null: residuals are normal)
. swilk residuals
```

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
residuals	67	0.98800	0.713	-0.735	0.76875

Good. The null hypothesis of normality of the residuals is NOT rejected (p-value = .77)

```
. *
. ***** Test of model misspecification (Null: No misspecification _hatsq is nonsignif.)
. linktest
```

Source	SS	df	MS	Number of obs =	67
Model	14.4804705	2	7.24023523	F( 2, 64) =	7.87
Residual	58.9031301	64	.920361408	Prob > F =	0.0009
Total	73.3836006	66	1.11187274	R-squared =	0.1973
				Adj R-squared =	0.1722
				Root MSE =	.95935

p53	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
_hat	2.748201	4.332136	0.63	0.528	-5.906235 11.40264
_hatsq	-.2753261	.6811046	-0.40	0.687	-1.635989 1.085337
_cons	-2.71457	6.766714	-0.40	0.690	-16.23263 10.8035

Also good. The null hypothesis that \_hatsq is not significant is NOT rejected.

```
. *
. ***** Cook-Weisberg Test for Homogeneity of variance of residuals (Null: homogeneity)
. estat hettest
```

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity

Ho: Constant variance

Variables: fitted values of p53

chi2(1) = 1.56

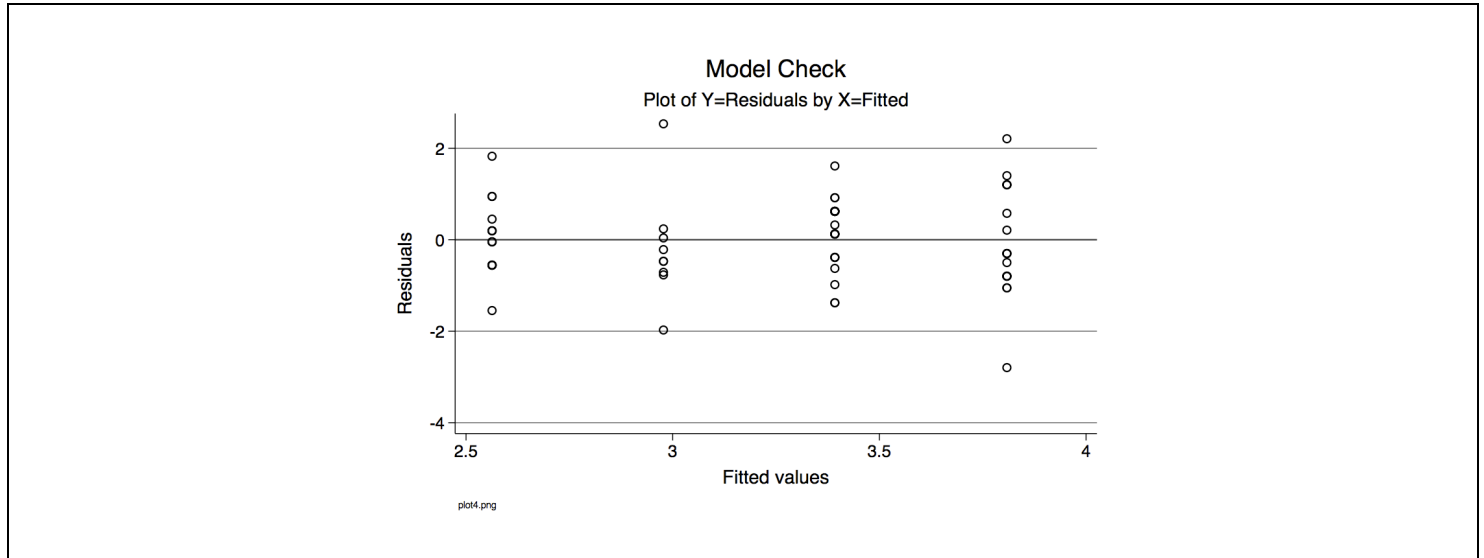
Prob > chi2 = 0.2115

Good. The null hypothesis of homogeneity of variance of residuals is NOT rejected (p-value = .21)

```
. *
```

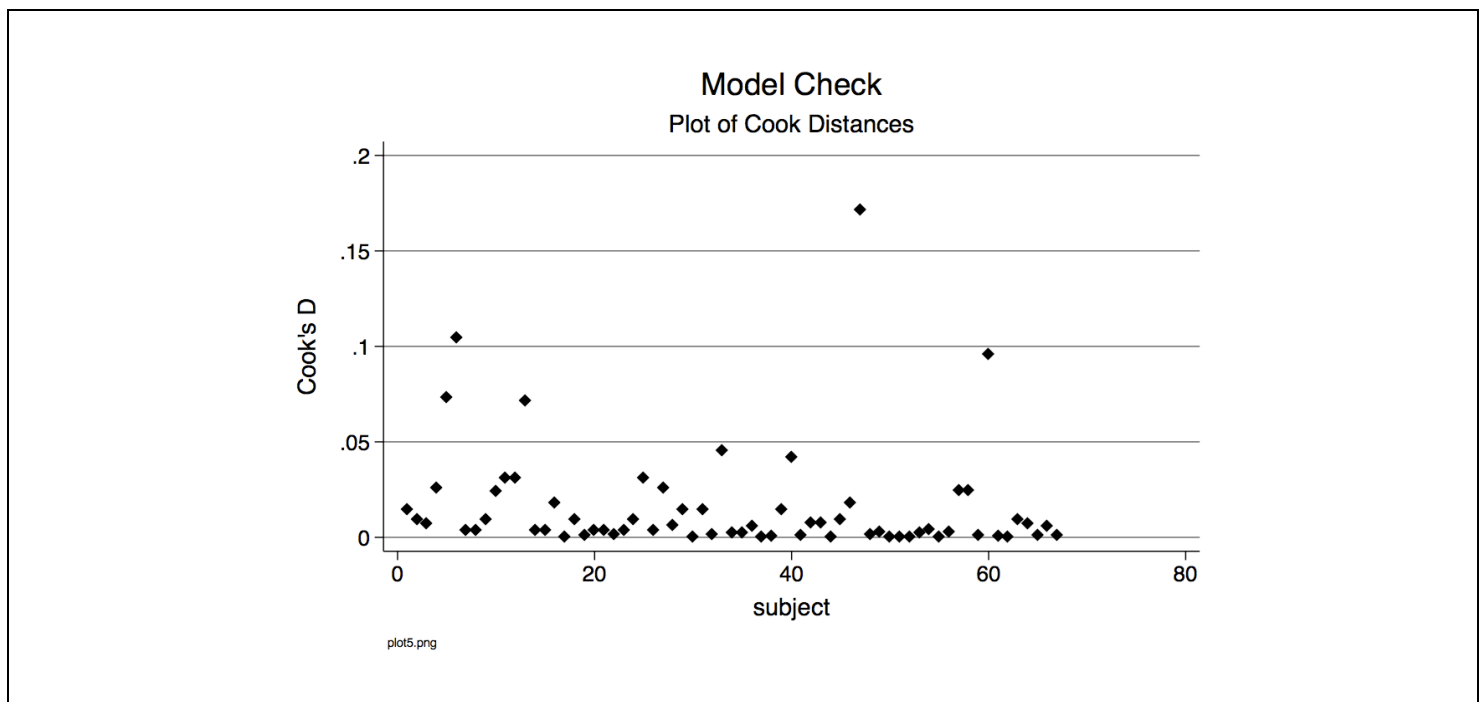
Design ..... Data Collection ..... Data Management ..... Data Summarization ..... Statistical Analysis ..... Reporting

```
. ***** Graphical Assessment of constant variance of the residuals
. rvfplot, yline(0) title("Model Check") subtitle("Plot of Y=Residuals by X=Fitted") note("plot4.png",
size(vsmall))
```



Variability of the residuals looks reasonably homogeneous, confirming the Cook-Weisberg test result

```
. *
. ***** Check for Outlying, Leverage and Influential points - Look for values > 4
. predict cook, cooks_d
. generate subject=_n
. graph twoway (scatter cook subject, symbol(d)), title("Model Check") subtitle("Plot of Cook
Distances") note("plot5.png", size(vsmall))
```



Very nice. Not only are there no Cook distances greater than 4, they are all less than 1!