

Stata v 12
Illustration
Multiple Linear Regression
Human p53 and Breast Cancer Risk

Source:

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

Background:

- Substantial epidemiologic evidence suggests that early first pregnancy confers a reduced life time risk of breast cancer.
- In laboratory studies of mice, similar observations have been made.
- Laboratory studies of mice have also explored the relationship between parity, expression of the tumor suppressor gene p53 and subsequent breast cancer tumor development.
- Lesley et al hypothesized that mammary tissue cultured from women who had an early full term pregnancy would have increased levels of p53 as compared to nulliparous women and as compared to women whose first full term pregnancy was later in life.

Research Question:

What is the relationship of $Y=p53$ expression to parity and age at first pregnancy, after adjustment for current age and established breast cancer risk, specifically the following: age at first mensis, family history of breast cancer, menopausal status, and history of oral contraceptive use?

Note – Age at first pregnancy is considered in each of two ways: (1) continuous, in years; and (2) age at first pregnancy ≤ 24 years versus age at first pregnancy > 24 years.

Design:

Observational cohort.

Stata Data Set:

p53paper.dta

Beware! Stata is case sensitive. All variable names are lower case.

Variable	Label	Definition/Codings
p53	P53	continuous
parous	Parity status	1 = ever parous 0 = not
pregnum	Number of pregnancies	0 = 0 pregnancies 1 = 1 pregnancy 2 = 2 pregnancies 3 = 3+ pregnancies
one	0/1 indicator of 1 pregnancy	= 1 if (pregnum=1) 0 otherwise
two	0/1 indicator of 2 pregnancies	= 1 if (pregnum=2) 0 otherwise
threep	0/1 indicator of 2 or more pregnancies	= 1 if (pregnum=3) 0 otherwise
agepreg1	Age at first pregnancy	Continuous, years = “missing” for never parous
early	0/1 indicator first pregnancy at age ≤ 24	1 = yes 0 = no = “missing” for never parous
late	0/1 indicator first pregnancy at age >24	1 = yes 0 = no = “missing” for never parous
agecurr	Current age	continuous, years
agemen	Age at first mensis	Continuous, years
famhx01	0/1 indicator of family history of breast cancer	= 1 if any family hx of breast ca 0 otherwise
menop	0/1 indicator of post-menopause	= 1 if yes 0 otherwise
oc	0/1 indicator of ever used oral contraceptives	= 1 if yes 0 otherwise
hrt	0/1 indicator of ever used hormone replacement therapy	= 1 if yes 0 otherwise

Key –

Green: comments in stata begin with an asterisk

Black: stata command syntax. Note – You do NOT need to type the leading period

Blue: Output

I have also inserted comments

```
. *****
. ***** Turn off screen by screen pausing of output
. set more off

. **
. ***** Set working directory, access data from website, save copy to working directory
. cd/users/carolbigelow/Desktop
/Users/carolbigelow/Desktop

. use "http://people.umass.edu/biep640w/datasets/p53paper.dta"
. save p53paper
file p53paper.dta saved

. **
. ***** Look at data set structure
. codebook, compact
```

Variable	Obs	Unique	Mean	Min	Max	Label
id	68	68	.	.	.	
agecurr	68	38	39.27941	15	75	agecurr: age current
agepreg1	51	21	23.44118	15	40.5	agepreg1: age at 1st preg
pregnum	68	4	1.632353	0	3	pregnum: number pregnancies
agemen	67	8	12	9	16	agemen: age at 1st mensis
menop	68	2	.2794118	0	1	menop: post menopausal
oc	68	2	.8235294	0	1	oc: oral contraceptives
hrt	68	2	.0441176	0	1	hrt: hormone replacement
cycle	42	26	20.40476	0	300	cycle: cycle days
famhx	68	7	1.014706	0	6	famhx: family hx breast ca
p53	67	19	3.251493	1	6	
parity	68	3	1.029412	0	2	parity: parity, grouped
early	68	2	.4705882	0	1	early: early parity
late	68	2	.2794118	0	1	late: late parity
parous	68	2	.75	0	1	parous 0/1
one	68	2	.1323529	0	1	one: 1 pregnancy
two	68	2	.3529412	0	1	two: 2 pregnancies
threep	68	2	.2647059	0	1	threep: 3+ pregnancies
famhx01	68	2	.2941176	0	1	famhx01: any family hx
twop	68	2	.6176471	0	1	twop: 2+ pregnancies

Be sure to assess completeness of study data. The sample size is $n=68$. We won't use **cycle** in this analysis and we may not do much with **agepreg1** as the number of missing values is $(68-51) = 17$. Notice also that the study id variable "id" is not numeric. That's okay; we just need to use something else for Cook distance calculations.

```
. **
. ***** Descriptives of study variables
. tabstat p53 agecurr agemen agepreg1, stat(n mean sd semean min q max) col(stat) format(%8.2f)
```

variable	N	mean	sd	se(mean)	min	p25	p50	p75	max
p53	67.00	3.25	1.05	0.13	1.00	2.50	3.00	4.00	6.00
agecurr	68.00	39.28	13.89	1.68	15.00	27.00	39.50	49.50	75.00
agemen	67.00	12.00	1.37	0.17	9.00	11.00	12.00	13.00	16.00
agepreg1	51.00	23.44	6.16	0.86	15.00	19.00	23.00	27.00	40.50

```
. tab1 famhx01 hrt menop oc parous early late pregnum, missing
-> tabulation of famhx01
```

famhx01:			
any family			
hx	Freq.	Percent	Cum.
0	48	70.59	70.59
1	20	29.41	100.00
Total	68	100.00	

-> tabulation of hrt

hrt: hormone replacement	Freq.	Percent	Cum.
0	65	95.59	95.59
1	3	4.41	100.00
Total	68	100.00	

Only 3 have history of hormone replacement therapy → not going to be useful in this analysis

-> tabulation of menop

menop: post menopausal	Freq.	Percent	Cum.
0	49	72.06	72.06
1	19	27.94	100.00
Total	68	100.00	

-> tabulation of oc

oc: oral contracepti ves	Freq.	Percent	Cum.
0	12	17.65	17.65
1	56	82.35	100.00
Total	68	100.00	

-> tabulation of parous

parous 0/1	Freq.	Percent	Cum.
0	17	25.00	25.00
1	51	75.00	100.00
Total	68	100.00	

-> tabulation of early

early: early parity	Freq.	Percent	Cum.
0	36	52.94	52.94
1	32	47.06	100.00
Total	68	100.00	

-> tabulation of late

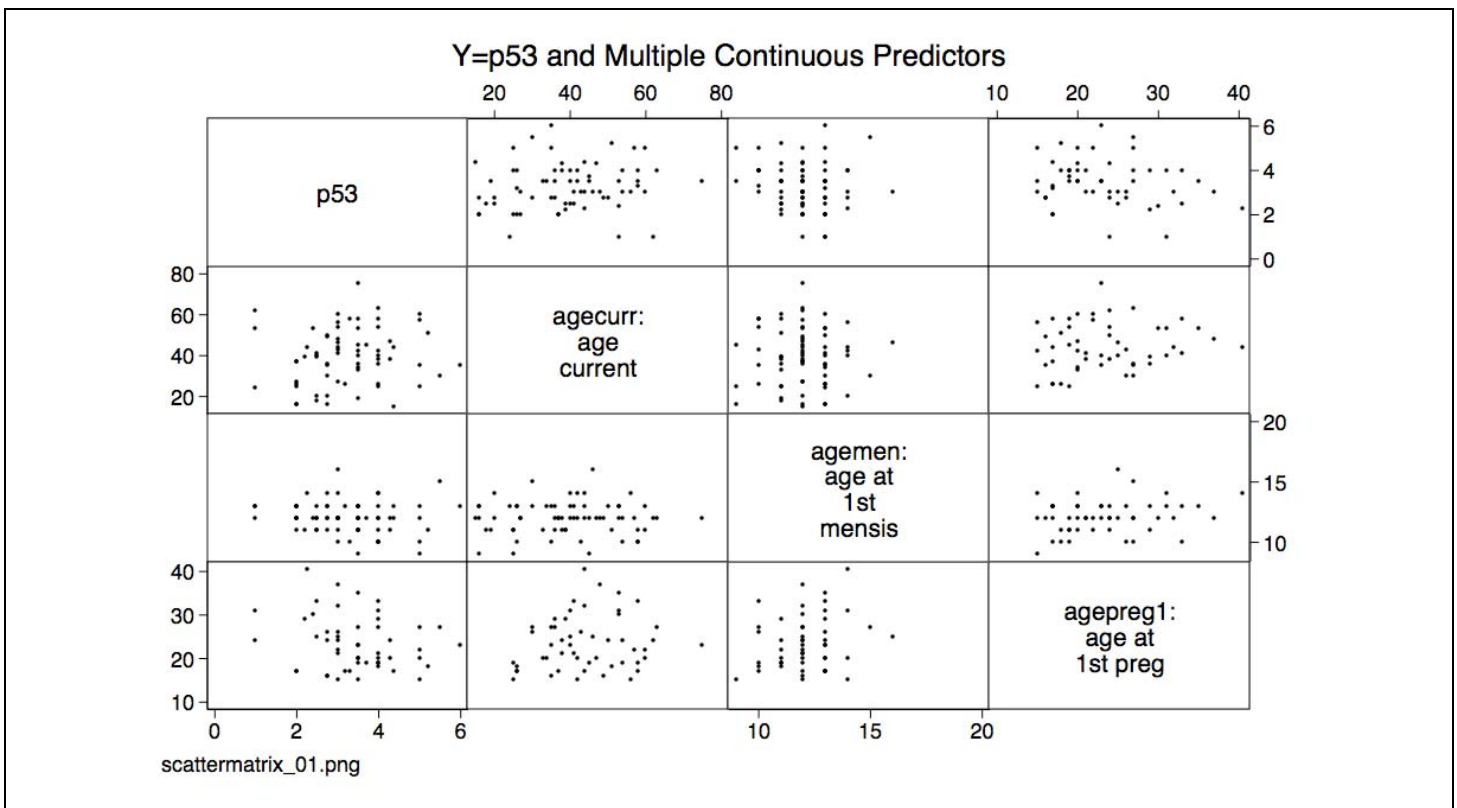
late: late parity	Freq.	Percent	Cum.
0	49	72.06	72.06
1	19	27.94	100.00
Total	68	100.00	

-> tabulation of pregnum

pregnum: number pregnancies	Freq.	Percent	Cum.
0	17	25.00	25.00
1	9	13.24	38.24
2	24	35.29	73.53
3	18	26.47	100.00
Total	68	100.00	

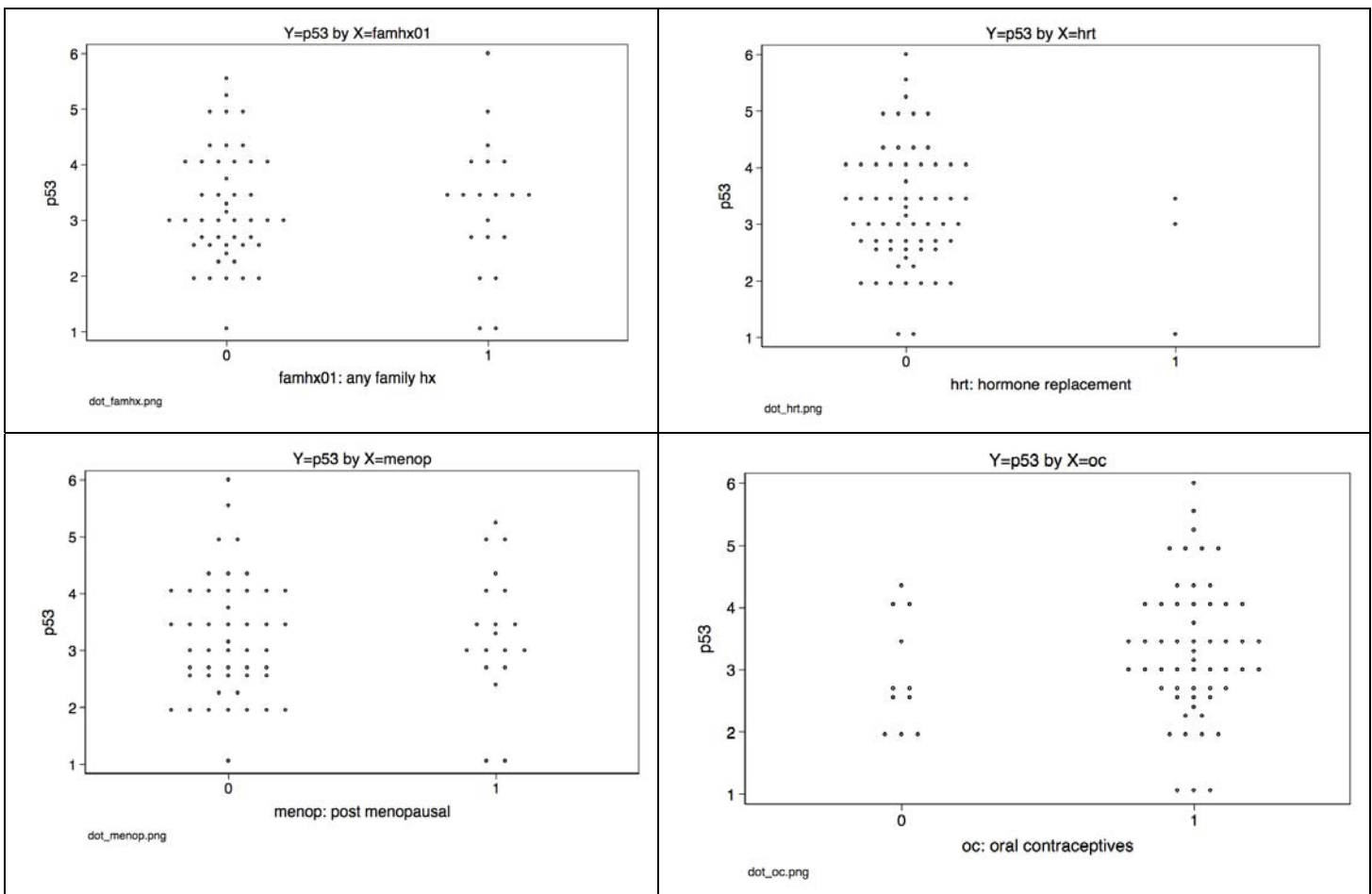
In creating design variables the referent group will be pregnum=0

```
. **
. ***** Scatterplot Matrix Graphs
. graph matrix p53 agecurr agegen agepreg1, msize(vsmall) subtitle("Y=p53 and Multiple Continuous Predictors")
. note("scattermatrix_01.png")
```

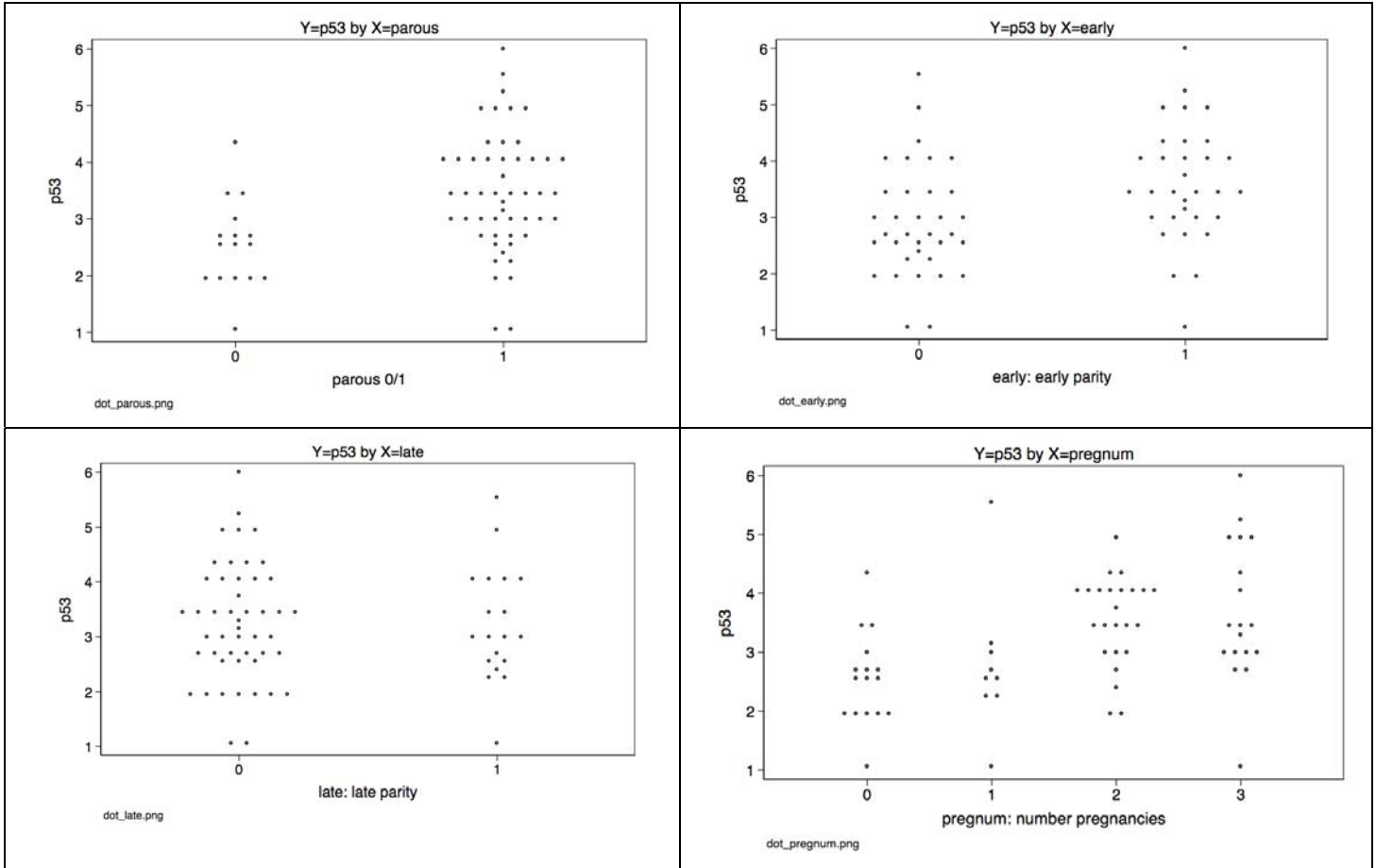


```

. **
. ***** Side-by-side dot plots of Y=p53 over levels of discrete predictors
. sort famhx01
. dotplot p53, over(famhx01) center msize(vsmall) subtitle("Y=p53 by X=famhx01") note("dot_famhx.png")
. sort hrt
. dotplot p53, over(hrt) center msize(vsmall) subtitle("Y=p53 by X=hrt") note("dot_hrt.png")
. sort menop
. dotplot p53, over(menop) center msize(vsmall) subtitle("Y=p53 by X=menop") note("dot_menop.png")
. sort oc
. dotplot p53, over(oc) center msize(vsmall) subtitle("Y=p53 by X=oc") note("dot_oc.png")
. sort parous
. dotplot p53, over(parous) center msize(vsmall) subtitle("Y=p53 by X=parous") note("dot_parous.png")
. sort early
. dotplot p53, over(early) center msize(vsmall) subtitle("Y=p53 by X=early") note("dot_early.png")
. sort late
. dotplot p53, over(late) center msize(vsmall) subtitle("Y=p53 by X=late") note("dot_late.png")
. sort pregnum
. dotplot p53, over(pregnum) center msize(vsmall) subtitle("Y=p53 by X=pregnum") note("dot_pregnum.png")
    
```



*Y=p53 does not appear (from the dot plots) to be associated with **famhx01, hrt, menop, or oc***



Now it's looking more interesting. These dot plots suggest that $Y=p53$ might be positively associated with number of pregnancies. There's also a hint of an association with younger (age ≤ 24) at first pregnancy. But possibly, too, this is spurious if the two predictors **pregnum** and **early** are themselves correlated.

```

. **
. ***** Assessment of Normality of Y=p53
. tabstat p53, stat(n mean sd min p50 max skewness kurtosis)

```

variable	N	mean	sd	min	p50	max	skewness	kurtosis
p53	67	3.251493	1.054454	1	3	6	.2106504	3.011225

```

. swilk p53

```

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
p53	67	0.98958	0.619	-1.040	0.85093

```

. sfrancia p53

```

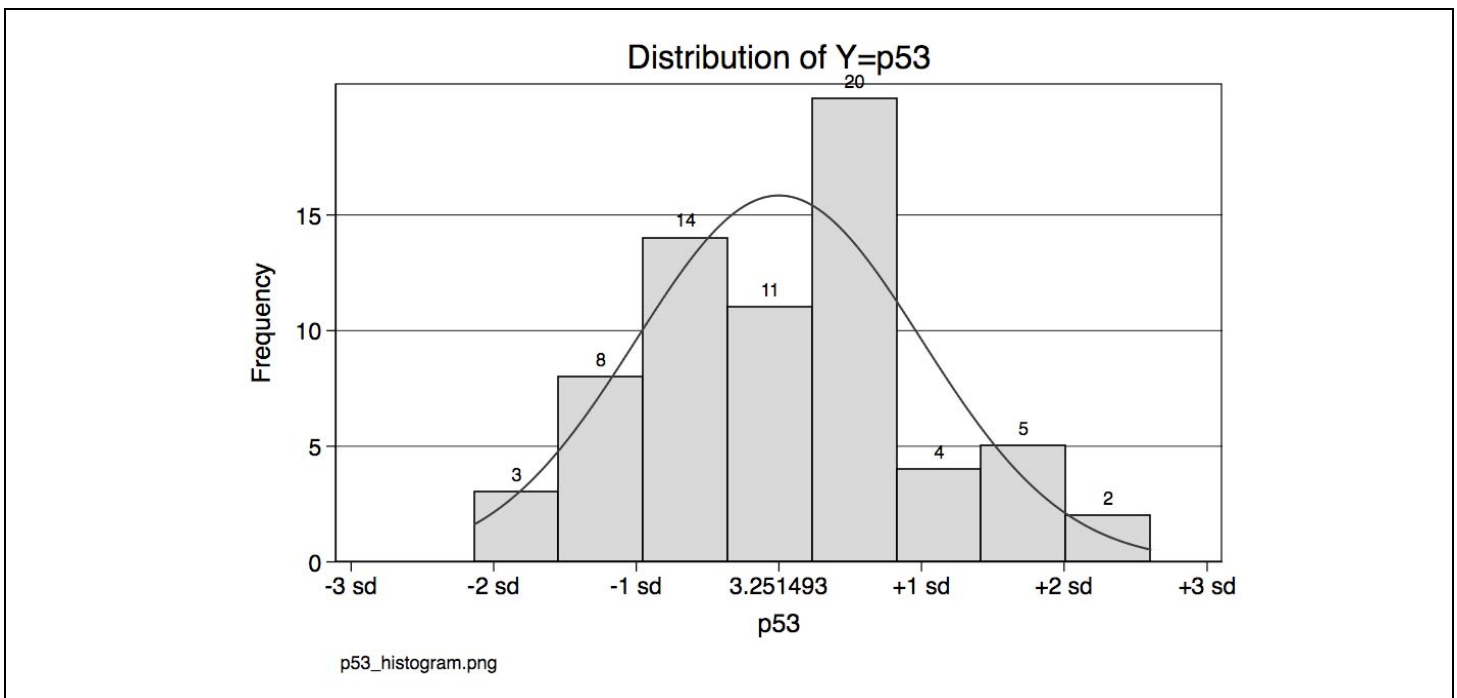
Shapiro-Francia W' test for normal data

Variable	Obs	W'	V'	z	Prob>z
p53	67	0.99165	0.549	-1.156	0.87609

Good news. Both Shapiro-Wilk and Francia tests of normality fail to reject the null hypothesis of normality; p -values are .85 and .88, respectively. We can reasonably assume that the assumption of normality of the dependent variable $Y=p53$ is satisfied.

```
. **
. ***** Fancy histogram with overlay normal and tick marks at each sd increment
. display 3.251493 - (1*1.054454)
2.197039
. display 3.251493 - (2*1.054454)
1.142585
. display 3.251493 - (3*1.054454)
.088131
. display 3.251493 + (1*1.054454)
4.305947
. display 3.251493 + (2*1.054454)
5.360401
. display 3.251493 + (3*1.054454)
6.41855

. histogram p53, start(1) bin(8) frequency addlabels normal ylabel(0(5)15, grid) xlabel(3.251493 "3.251493"
2.197039 "-1 sd" 1.142585 "-2 sd" 0.088131 "-3 sd" 4.305947 "+1 sd" 5.360401 "+2 sd" 6.41844 "+3 sd")
title("Distribution of Y=p53") note("p53_histogram.png")
(bin=8, start=1, width=.625)
```



Nice picture. The overlay normal on the histogram of $Y=p53$ is a reasonable fit. Not surprising since the Shapiro-Wilk and Francia tests failed to reject the null hypothesis of normality.

```
. **
. ***** ONE PREDICTOR MODELS
. regress p53 parous
```

Source	SS	df	MS			
Model	9.75322943	1	9.75322943	Number of obs =	67	
Residual	63.6303711	65	.978928787	F(1, 65) =	9.96	
Total	73.3836006	66	1.11187274	Prob > F =	0.0024	
				R-squared =	0.1329	
				Adj R-squared =	0.1196	
				Root MSE =	.98941	

p53	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
parous	.8948836	.2835097	3.16	0.002	.3286757	1.461091
_cons	2.570313	.2473521	10.39	0.000	2.076316	3.064309

```
. ***** Number of pregnancies is modeled as a nominal predictor. Because it has four values in this data set,
(4-1)=3 design variables are required. Referent group is pregnum=0 for "zero pregnancies"
regress p53 one two threep
```

Source	SS	df	MS			
Model	15.4032579	3	5.13441928	Number of obs =	67	
Residual	57.9803427	63	.9203229	F(3, 63) =	5.58	
Total	73.3836006	66	1.11187274	Prob > F =	0.0019	
				R-squared =	0.2099	
				Adj R-squared =	0.1723	
				Root MSE =	.95933	

p53	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
one	.1963542	.3997228	0.49	0.625	-.602428	.9951364
two	.9692709	.3096239	3.13	0.003	.3505368	1.588005
threep	1.144965	.3296198	3.47	0.001	.4862726	1.803658
_cons	2.570312	.2398337	10.72	0.000	2.091043	3.049582

```
. regress p53 agepreg1
```

Source	SS	df	MS			
Model	2.34529956	1	2.34529956	Number of obs =	51	
Residual	51.9110481	49	1.05940915	F(1, 49) =	2.21	
Total	54.2563477	50	1.08512695	Prob > F =	0.1432	
				R-squared =	0.0432	
				Adj R-squared =	0.0237	
				Root MSE =	1.0293	

p53	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
agepreg1	-.0351398	.0236174	-1.49	0.143	-.0826008	.0123211
_cons	4.288915	.5720733	7.50	0.000	3.139291	5.438539

```
. ***** Age at first pregnancy is also modeled as a nominal predictor. The referent group here is "missing"
for "never parous"
. regress p53 early late
```

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

. regress p53 agecurr

Source	SS	df	MS			
Model	1.31709199	1	1.31709199	Number of obs =	67	
Residual	72.0665086	65	1.10871552	F(1, 65) =	1.19	
Total	73.3836006	66	1.11187274	Prob > F =	0.2798	
				R-squared =	0.0179	
				Adj R-squared =	0.0028	
				Root MSE =	1.053	

p53	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
agecurr	.010313	.0094621	1.09	0.280	-.0085841	.02921
_cons	2.842822	.3964046	7.17	0.000	2.051148	3.634497

. regress p53 agemen

Source	SS	df	MS			
Model	1.10229524	1	1.10229524	Number of obs =	66	
Residual	72.026	64	1.12540625	F(1, 64) =	0.98	
Total	73.1282953	65	1.1250507	Prob > F =	0.3261	
				R-squared =	0.0151	
				Adj R-squared =	-0.0003	
				Root MSE =	1.0609	

p53	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
agemen	-.097962	.0989836	-0.99	0.326	-.2957045	.0997805
_cons	4.439088	1.199432	3.70	0.000	2.042947	6.835229

. regress p53 famhx01

Source	SS	df	MS			
Model	.019284263	1	.019284263	Number of obs =	67	
Residual	73.3643163	65	1.12868179	F(1, 65) =	0.02	
Total	73.3836006	66	1.11187274	Prob > F =	0.8964	
				R-squared =	0.0003	
				Adj R-squared =	-0.0151	
				Root MSE =	1.0624	

p53	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
famhx01	.0370745	.2836346	0.13	0.896	-.529383	.6035319
_cons	3.240426	.1549661	20.91	0.000	2.930937	3.549914

. regress p53 menop

Source	SS	df	MS			
Model	.148477378	1	.148477378	Number of obs =	67	
Residual	73.2351232	65	1.1266942	F(1, 65) =	0.13	
Total	73.3836006	66	1.11187274	Prob > F =	0.7178	
				R-squared =	0.0020	
				Adj R-squared =	-0.0133	
				Root MSE =	1.0615	

p53	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
menop	.1044408	.2877021	0.36	0.718	-.47014	.6790215
_cons	3.221875	.1532083	21.03	0.000	2.915897	3.527853

. regress p53 oc

Source	SS	df	MS	Number of obs =	67
Model	1.25099832	1	1.25099832	F(1, 65) =	1.13
Residual	72.1326023	65	1.10973234	Prob > F =	0.2923
Total	73.3836006	66	1.11187274	R-squared =	0.0170
				Adj R-squared =	0.0019
				Root MSE =	1.0534

p53	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
oc	.3688718	.3474211	1.06	0.292	-.3249761 1.06272
_cons	2.943182	.3176236	9.27	0.000	2.308844 3.57752

Summary of Fit of One Predictor Models

Predictor or Set of Design Variables	Significance of Overall F Test	Remark
parous	.002	Do not consider further since the design variables one , two , and threep will be considered further
one, two, threep	.002	Consider further. Note that t-test for one was not significant
agepreg1	.14	Because the # missing for 17 (25%) of the sample, do not consider further, despite the marginal pvalue
early, late	.004	Consider further
agecurr	.28	Do not consider further
agemen	.33	Do not consider further
famhx01	.90	Do not consider further
menop	.72	Do not consider further
oc	.29	Do not consider further

. **
 . ***** Initial multiple predictor model contains predictors that are crudely significant at p < .25

. regress p53 two threep early late

Source	SS	df	MS	Number of obs =	67
Model	15.6625023	4	3.91562558	F(4, 62) =	4.21
Residual	57.7210983	62	.930985456	Prob > F =	0.0045
Total	73.3836006	66	1.11187274	R-squared =	0.2134
				Adj R-squared =	0.1627
				Root MSE =	.96488

p53	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
two	.6951322	.404922	1.72	0.091	-.1142954 1.50456
threep	.8686042	.42208	2.06	0.044	-.0248783 1.71233
early	.3208093	.4661041	0.69	0.494	-.6109195 1.252538
late	.1607956	.4076397	0.39	0.695	-.6540646 .9756557
_cons	2.570313	.241219	10.66	0.000	2.088123 3.052502

```
. **
. ***** Partial F test for EARLY and LATE, controlling for TWO and THREEP
. testparm early late

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

      F( 2,    62) =    0.26
      Prob > F =    0.7730
```

Use the command TESTPARG to obtain partial F-tests in STATA. Here, the small model has predictors TWO and THREEP. The extra predictors being tested for significance in adjusted analysis are EARLY and LATE. The partial F-test p-value of .77 suggests that, controlling for TWO and THREEP, the extra predictors EARLY and LATE do not have statistical significance for the prediction of Y=p53.

```
. **
. ***** At this point, our candidate final model contains just TWO and THREEP
. regress p53 two threep
```

Source	SS	df	MS			
Model	15.1811813	2	7.59059063	Number of obs =	67	
Residual	58.2024193	64	.909412802	F(2, 64) =	8.35	
Total	73.3836006	66	1.11187274	Prob > F =	0.0006	
				R-squared =	0.2069	
				Adj R-squared =	0.1821	
				Root MSE =	.95363	

p53	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
two	.8985834	.2725229	3.30	0.002	.3541563	1.44301
threep	1.074278	.2947871	3.64	0.001	.4853728	1.663183
_cons	2.641	.1907263	13.85	0.000	2.25998	3.02202

```
. **
. ***** CHECK of candidate smaller model: Is it confounded by EARLY
. ** (a) Fit smaller model WITHOUT confounder
```

```
. regress p53 two threep
```

Source	SS	df	MS			
Model	15.1811813	2	7.59059063	Number of obs =	67	
Residual	58.2024193	64	.909412802	F(2, 64) =	8.35	
Total	73.3836006	66	1.11187274	Prob > F =	0.0006	
				R-squared =	0.2069	
				Adj R-squared =	0.1821	
				Root MSE =	.95363	

p53	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
two	.8985834	.2725229	3.30	0.002	.3541563	1.44301
threep	1.074278	.2947871	3.64	0.001	.4853728	1.663183
_cons	2.641	.1907263	13.85	0.000	2.25998	3.02202

```
. ** (b) fit smaller model WITH confounder
```

```
. regress p53 two threep early
```

Source	SS	df	MS			
Model	15.5176458	3	5.17254858	Number of obs =	67	
Residual	57.8659548	63	.918507219	F(3, 63) =	5.63	
Total	73.3836006	66	1.11187274	Prob > F =	0.0017	
				R-squared =	0.2115	
				Adj R-squared =	0.1739	
				Root MSE =	.95839	

p53	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
two	.8985834	.2725229	3.30	0.002	.3541563	1.44301
threep	1.074278	.2947871	3.64	0.001	.4853728	1.663183
_cons	2.641	.1907263	13.85	0.000	2.25998	3.02202

```

-----+-----
      two | .7856178 .3314332 2.37 0.021 .1233014 1.447934
    threep | .9588152 .3523665 2.72 0.008 .254667 1.662963
     early | .179786 .2970486 0.61 0.547 -.4138183 .7733903
      _cons | 2.626617 .1931451 13.60 0.000 2.240648 3.012587
-----+-----

```

```

. *** partial F test to assess potential confounding
. testparm early

```

```

( 1) early = 0
      F( 1, 63) = 0.37
      Prob > F = 0.5472

```

Quick show and tell to show you that the TESTPARM command does what we think it should.

$$\begin{aligned}
 \text{Partial } F_{1,63} &= \frac{\Delta \text{Model SSQ} / \Delta \text{Model df}}{\text{Residual SSQ}(\text{larger model}) / \text{Residual df}(\text{larger model})} \\
 &= \frac{(15.5176458 - 15.1811813) / (3 - 2)}{57.865948 / 63} \\
 &= \frac{0.3364645}{0.9185071} \\
 &= 0.3663 \text{ match}
 \end{aligned}$$

```

. **
. ***** CHECK of candidate smaller model: Is it confounded by LATE
. ** (a) Fit smaller model WITHOUT confounder

```

```

. regress p53 two threep

```

```

-----+-----
Source |      SS      df      MS                Number of obs =      67
-----+-----+-----+-----                F( 2, 64) =      8.35
Model | 15.1811813    2  7.59059063                Prob > F      = 0.0006
Residual | 58.2024193   64  .909412802                R-squared     = 0.2069
-----+-----+-----+-----                Adj R-squared = 0.1821
Total | 73.3836006   66  1.11187274                Root MSE     = .95363

-----+-----
p53 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----+-----+-----+-----
two | .8985834   .2725229     3.30  0.002   .3541563   1.44301
threep | 1.074278   .2947871     3.64  0.001   .4853728   1.663183
_cons | 2.641      .1907263    13.85  0.000   2.25998    3.02202
-----+-----

```

```

. ** (b) fit smaller model WITH confounder

```

```

. regress p53 two threep late

```

```

-----+-----
Source |      SS      df      MS                Number of obs =      67
-----+-----+-----+-----                F( 3, 63) =      5.50
Model | 15.2214694    3  5.07382313                Prob > F      = 0.0020
Residual | 58.1621312   63  .923208431                R-squared     = 0.2074
-----+-----+-----+-----                Adj R-squared = 0.1697
Total | 73.3836006   66  1.11187274                Root MSE     = .96084

-----+-----
p53 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----+-----+-----+-----
two | .8992181   .274599     3.27  0.002   .3504759   1.44796
threep | 1.074157   .2970152     3.62  0.001   .4806193   1.667694
late | -.0544087   .2604532    -0.21  0.835  -.5748829   .4660655
_cons | 2.656234   .2055399    12.92  0.000   2.245496   3.066973
-----+-----

```

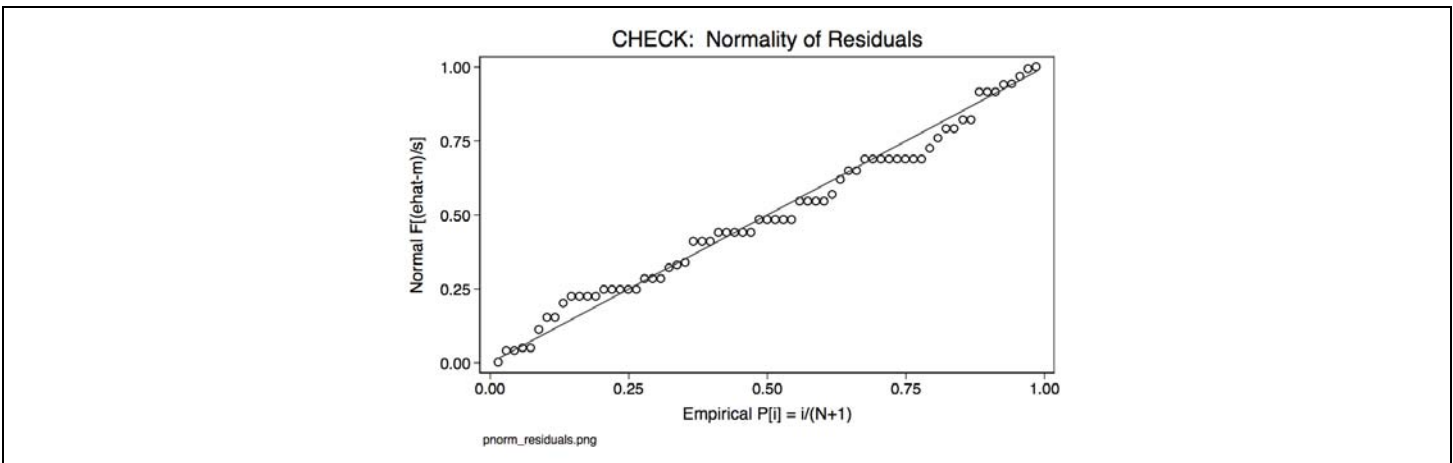
```
. *** partial F test to assess potential confounding
. testparm late
( 1) late = 0
      F( 1, 63) = 0.04
      Prob > F = 0.8352

. **
. *** Almost there. Fit of "candidate" final model - necessary preliminary to diagnostics
. regress p53 two threep
```

Source	SS	df	MS	Number of obs =	67
Model	15.1811813	2	7.59059063	F(2, 64) =	8.35
Residual	58.2024193	64	.909412802	Prob > F =	0.0006
Total	73.3836006	66	1.11187274	R-squared =	0.2069
				Adj R-squared =	0.1821
				Root MSE =	.95363

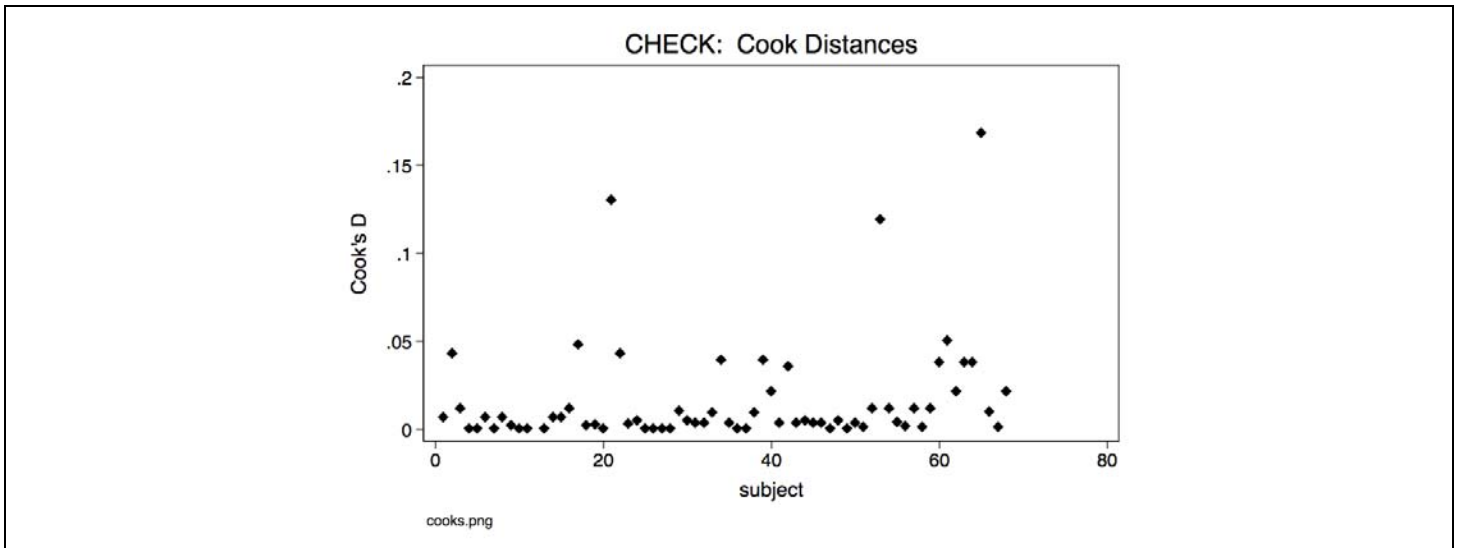
p53	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
two	.8985834	.2725229	3.30	0.002	.3541563 1.44301
threep	1.074278	.2947871	3.64	0.001	.4853728 1.663183
_cons	2.641	.1907263	13.85	0.000	2.25998 3.02202

```
. **
. *** CHECK - Residuals should "look" reasonably distributed normal. Pnorm points should fall on line.
. predict ehat, residuals
(1 missing value generated)
. pnorm ehat, title("CHECK: Normality of Residuals") note("pnorm_residuals.png")
```



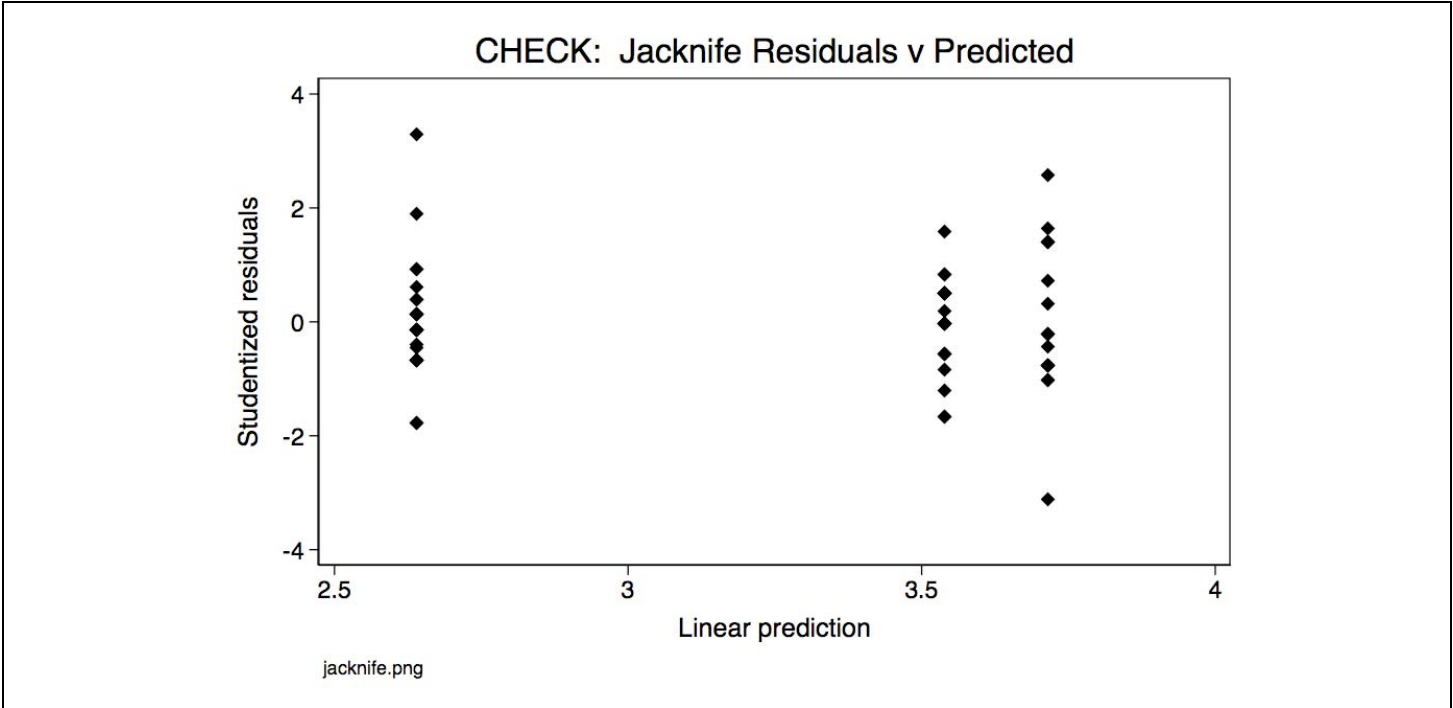
Nice. The points fall pretty much on the line as we'd hoped.

```
. **  
. *** CHECK - Cook Distance values should be nice parallel band with NO huge spikes  
. predict cook, cooksD  
(1 missing value generated)  
. generate subject=_n  
. graph twoway (scatter cook subject, symbol(d)), title("CHECK: Cook Distances") note("cooks.png")
```



Not bad. Even though there appear to be some suggestions of spikes, the magnitude of these cook distances are all small.

```
. **
. *** CHECK - Plot of Jackknife Residuals v Predicted should be nice parallel band centered at 0
. predict yhat, xb
. predict jack, rstudent
(1 missing value generated)
. graph twoway (scatter jack yhat, symbol(d)), title("CHECK: Jackknife Residuals v Predicted")
note("jackknife.png")
```



Not bad.

What to conclude? The final model is consistent with what our pictures suggested, namely that only number of pregnancies is associated with p53 in this data set and that this is a positive association. The final fitted model is:

$$\text{Predicted p53} = 2.641 + 0.90 \cdot \text{two} + 1.07 \cdot \text{threep}$$

% Variance explained = 20.69%
Significance of overall F test is .0006

The remaining 79% of the variability in p53 that is not explained remains to be elucidated. This isn't surprising inasmuch as p53 is such an important molecule with so many functions.