Solutions for Session 5: Linear Models

08/11/2022

```
. do solution.do

. global basedir http://personalpages.manchester.ac.uk/staff/mark.lunt

. global datadir $basedir/stats/5_LinearModels1/data

. use $datadir/anscombe, clear

. scatter Y1 x1, xlab(0 (5) 20) ylab(0 (5) 15)

. scatter Y2 x1, xlab(0 (5) 20) ylab(0 (5) 15)

. scatter Y3 x1, xlab(0 (5) 20) ylab(0 (5) 15)

. scatter Y4 x2, xlab(0 (5) 20) ylab(0 (5) 15)

. regress Y1 x1
      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,     9) =   17.99
       Model |  27.5100011     1  27.5100011            Prob > F      =  0.0022
    Residual |  13.7626904     9  1.52918783            R-squared     =  0.6665
-------------+------------------------------           Adj R-squared =  0.6295
       Total |  41.2726916    10  4.12726916            Root MSE      =  1.2366

------------------------------------------------------------------------------
          Y1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .5000909   .1179055     4.24   0.002     .2333701    .7668117
       _cons |   3.000091   1.124747     2.67   0.026     .4557369    5.544445
------------------------------------------------------------------------------

. regress Y2 x1
      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,     9) =   17.97
       Model |  27.5000024     1  27.5000024            Prob > F      =  0.0022
    Residual |   13.776294     9  1.53069933            R-squared     =  0.6662
-------------+------------------------------           Adj R-squared =  0.6292
       Total |  41.2762964    10  4.12762964            Root MSE      =  1.2372

------------------------------------------------------------------------------
          Y2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |          .5   .1179638     4.24   0.002     .2331475    .7668526
       _cons |   3.000909   1.125303     2.67   0.026     .4552978     5.54652
------------------------------------------------------------------------------
```

```
. regress Y3 x1
```

Source	SS	df	MS		
Model	27.4700075	1	27.4700075		
Residual	13.7561905	9	1.52846561		
Total	41.2261979	10	4.12261979		

Number of obs	=	11			
F(1, 9)	=	17.97			
Prob > F	=	0.0022			
R-squared	=	0.6663			
Adj R-squared	=	0.6292			
Root MSE	=	1.2363			

Y3	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	.4997273	.1178777	4.24	0.002	.2330695	.7663851
_cons	3.002455	1.124481	2.67	0.026	.4587014	5.546208

```
. regress Y4 x2
```

Source	SS	df	MS		
Model	27.4900007	1	27.4900007		
Residual	13.7424908	9	1.52694342		
Total	41.2324915	10	4.12324915		

Number of obs	=	11			
F(1, 9)	=	18.00			
Prob > F	=	0.0022			
R-squared	=	0.6667			
Adj R-squared	=	0.6297			
Root MSE	=	1.2357			

Y4	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x2	.4999091	.1178189	4.24	0.002	.2333841	.7664341
_cons	3.001727	1.123921	2.67	0.026	.4592411	5.544213

```
. sysuse auto, clear
(1978 Automobile Data)

. regress mpg weight
```

Source	SS	df	MS		
Model	1591.9902	1	1591.9902		
Residual	851.469256	72	11.8259619		
Total	2443.45946	73	33.4720474		

Number of obs	=	74			
F(1, 72)	=	134.62			
Prob > F	=	0.0000			
R-squared	=	0.6515			
Adj R-squared	=	0.6467			
Root MSE	=	3.4389			

mpg	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
weight	-.0060087	.0005179	-11.60	0.000	-.0070411	-.0049763
_cons	39.44028	1.614003	24.44	0.000	36.22283	42.65774

2.1 Yes: the coefficient for weight is very significantly different from 0
2.2. 65.15%: this is given by R-squared
2.3 A reduction of 0.006 mpg

```
. lincom _cons + 3000 * weight
 ( 1)  3000*weight + _cons = 0
```

mpg	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	21.41422	.3998898	53.55	0.000	20.61706	22.21139

2.4 21.4 mpg, with a 95% CI of (20.6, 22.2)
2.5 No, because there are no vehicles this light in the dataset

```
. use "$datadir/constvar"

. regress y x
```

Source	SS	df	MS		
Model	47.9706438	1	47.9706438	Number of obs	= 80
Residual	207.014126	78	2.65402726	F(1, 78)	= 18.07
				Prob > F	= 0.0001
				R-squared	= 0.1881
				Adj R-squared	= 0.1777
Total	254.98477	79	3.22765532	Root MSE	= 1.6291

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	2.676801	.6296237	4.25	0.000	1.423317	3.930286
_cons	1.599564	.1827062	8.75	0.000	1.235824	1.963304

3.1 Yes, p=0.000

```
. predict rstand, rstand

. predict yhat
(option xb assumed; fitted values)

. scatter rstand yhat


. graph export graph1.eps replace
(file graph1.eps written in EPS format)
```

3.2 The variance (the spread of the data) increases as the fitted value increases

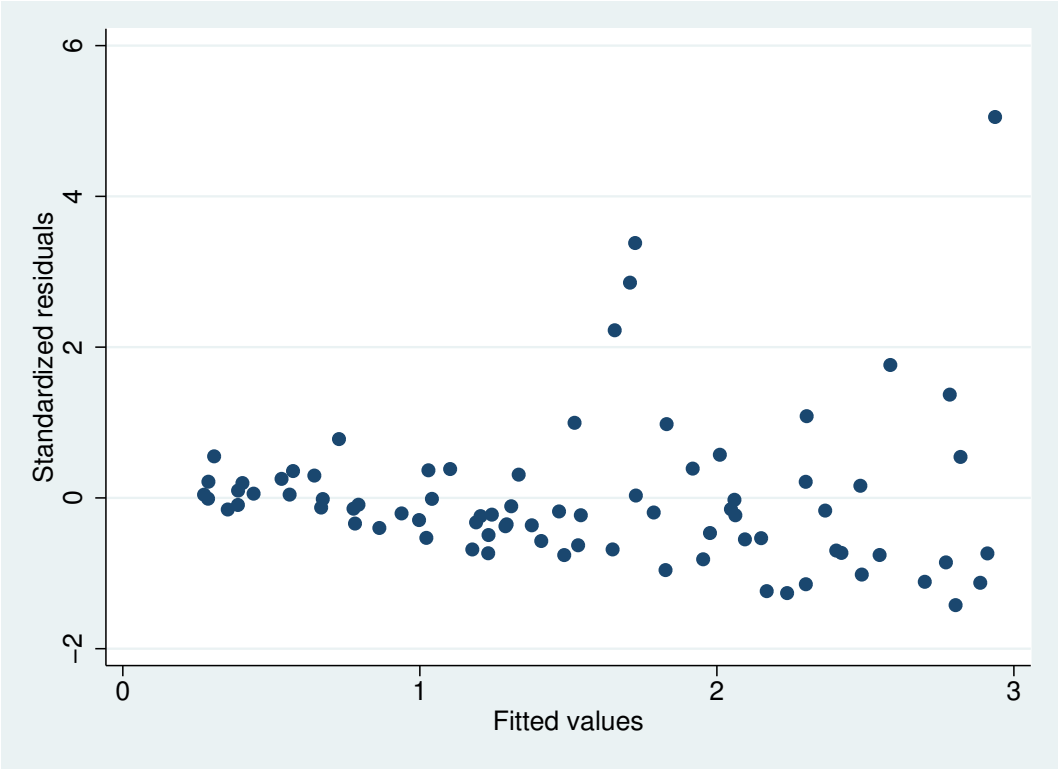

Figure 1: . scatter rstand yhat

```
. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of y

         chi2(1)      =    34.34
         Prob > chi2  =   0.0000
```

3.3 hettest confirms that the variance is not constant

```
. rvfplot
```

3.4 Yes: there is very little difference between these two plots

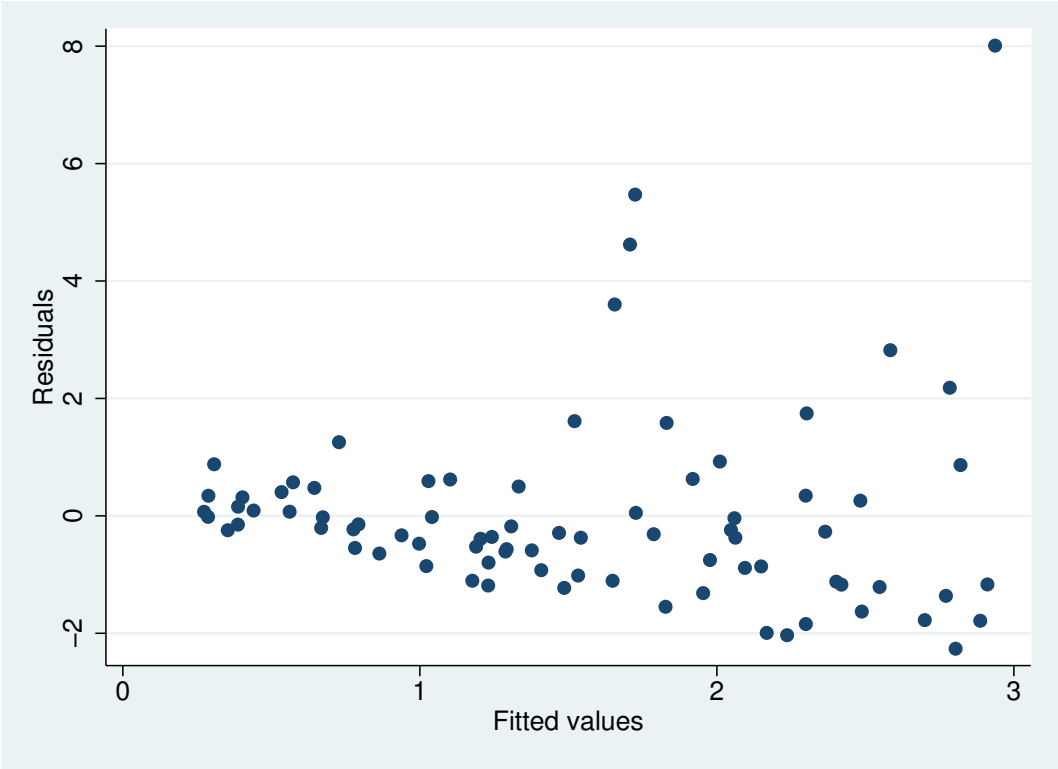

Figure 2: . rvfplot

```
. graph export graph2.eps replace
(file graph2.eps written in EPS format)

. gen ly = ln(y)

. regress ly x
```

Source	SS	df	MS	Number of obs =	80
				F(1, 78) =	21.96
Model	18.8639824	1	18.8639824	Prob > F =	0.0000
Residual	66.9993584	78	.858966134	R-squared =	0.2197
				Adj R-squared =	0.2097
Total	85.8633408	79	1.08687773	Root MSE =	.9268

ly	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	1.678592	.3581924	4.69	0.000	.9654853	2.391698
_cons	-.0323861	.1039414	-0.31	0.756	-.2393176	.1745454

```
. predict rstand2, rstand

. predict yhat2
(option xb assumed; fitted values)
```

```
. scatter rstand2 yhat2
```

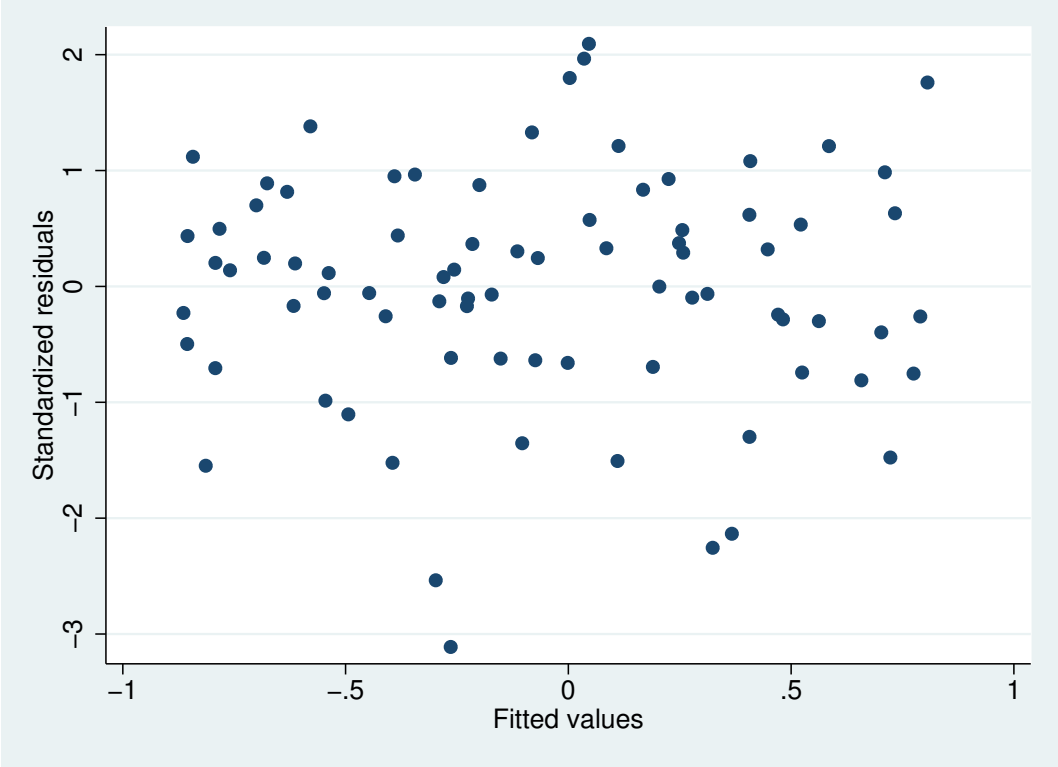

Figure 3: . scatter rstand2 yhat2

```
. graph export graph3.eps replace
(file graph3.eps written in EPS format)
```

3.5 There is no longer evidence of changing variance

```
. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of ly

         chi2(1)      =     0.52
         Prob > chi2  =   0.4696
```

3.6 This is confirmed by hettest

```
. use $datadir/wood73, clear

. scatter Y x1
```

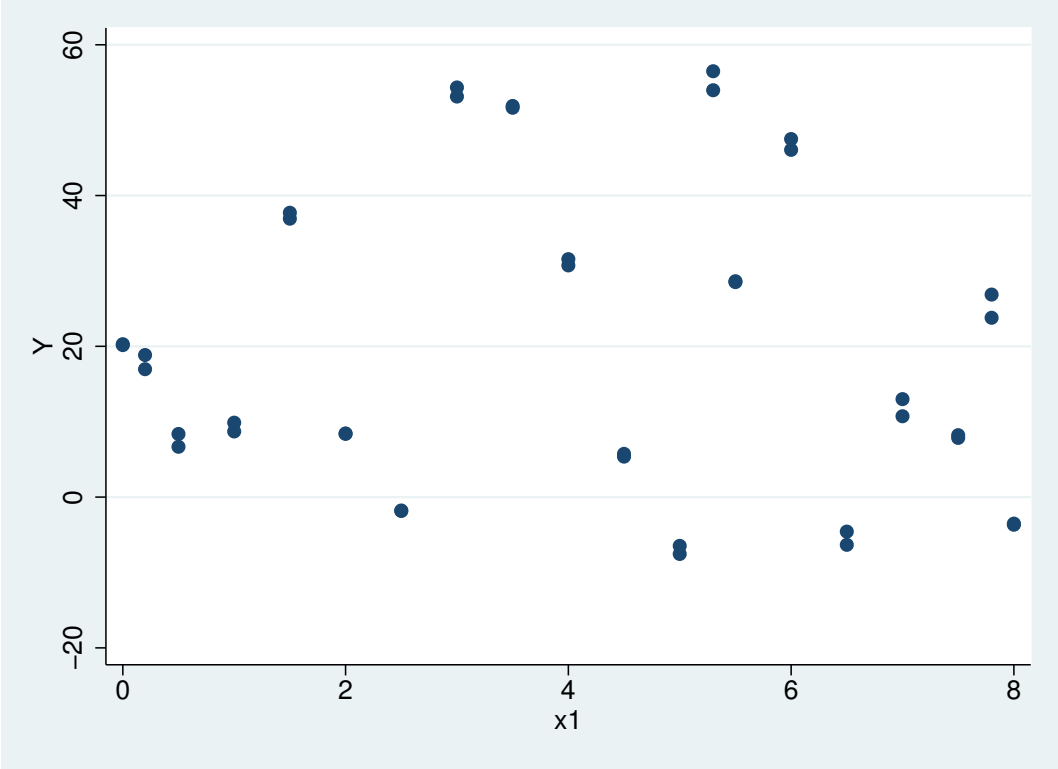

Figure 4: . scatter Y x1

```
. graph export graph4.eps replace
(file graph4.eps written in EPS format)

. scatter Y x2


. graph export graph5.eps replace
(file graph5.eps written in EPS format)
```

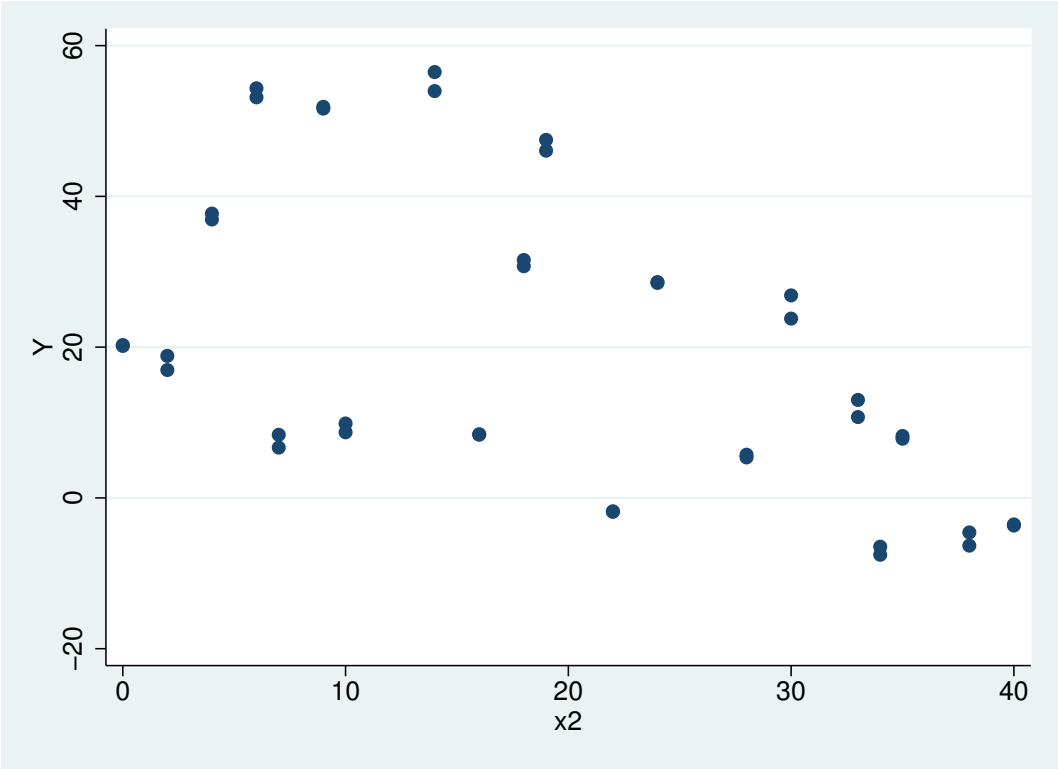

Figure 5: . scatter Y x2

```
. regress Y x1 x2
```

Source	SS	df	MS		
Model	14349.7681	2	7174.88407		
Residual	1405.26007	37	37.9800018		
Total	15755.0282	39	403.975082		

Number of obs	=	40			
F(2, 37)	=	188.91			
Prob > F	=	0.0000			
R-squared	=	0.9108			
Adj R-squared	=	0.9060			
Root MSE	=	6.1628			

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	12.23327	.7632992	16.03	0.000	10.68668	13.77987
x2	-3.049444	.1574177	-19.37	0.000	-3.368402	-2.730485
_cons	29.62759	1.858254	15.94	0.000	25.86241	33.39277

```
. cprplot x1


. graph export graph6.eps replace
(file graph6.eps written in EPS format)
```

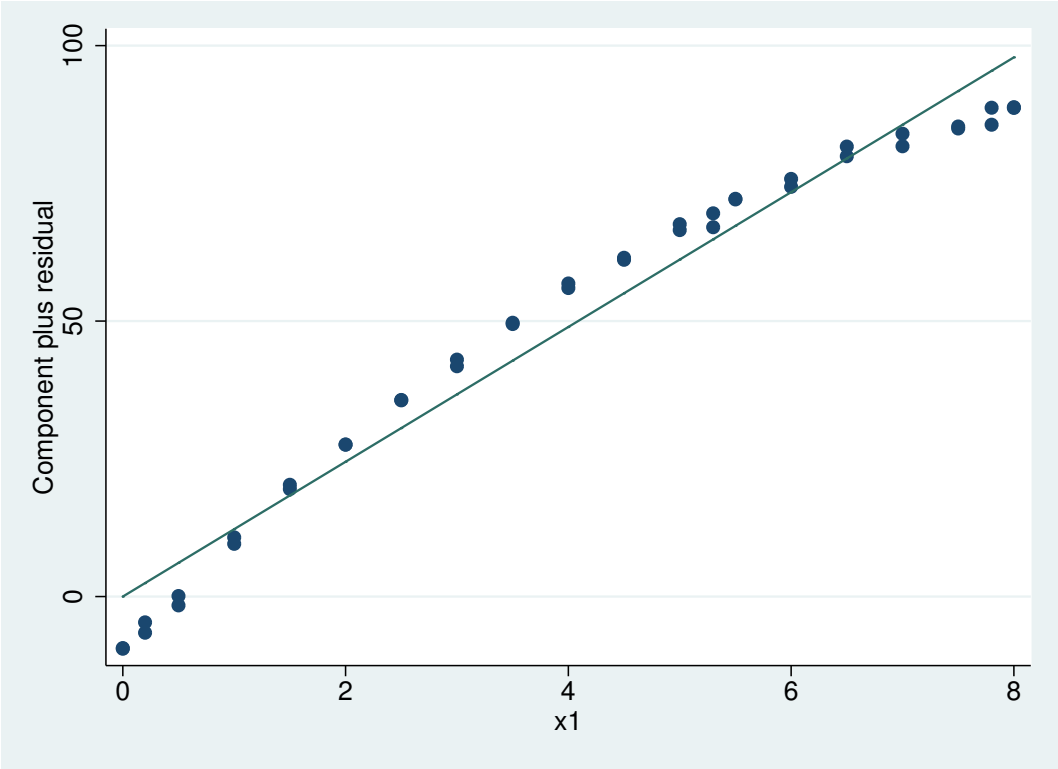

Figure 6: . cprplot x1

3.9 Y against x1 looks non-linear

```
. cprplot x2

. graph export graph7.eps replace
(file graph7.eps written in EPS format)
```

3.9 Y against x2 looks reasonably linear

```
. gen x3 = x1^2
```

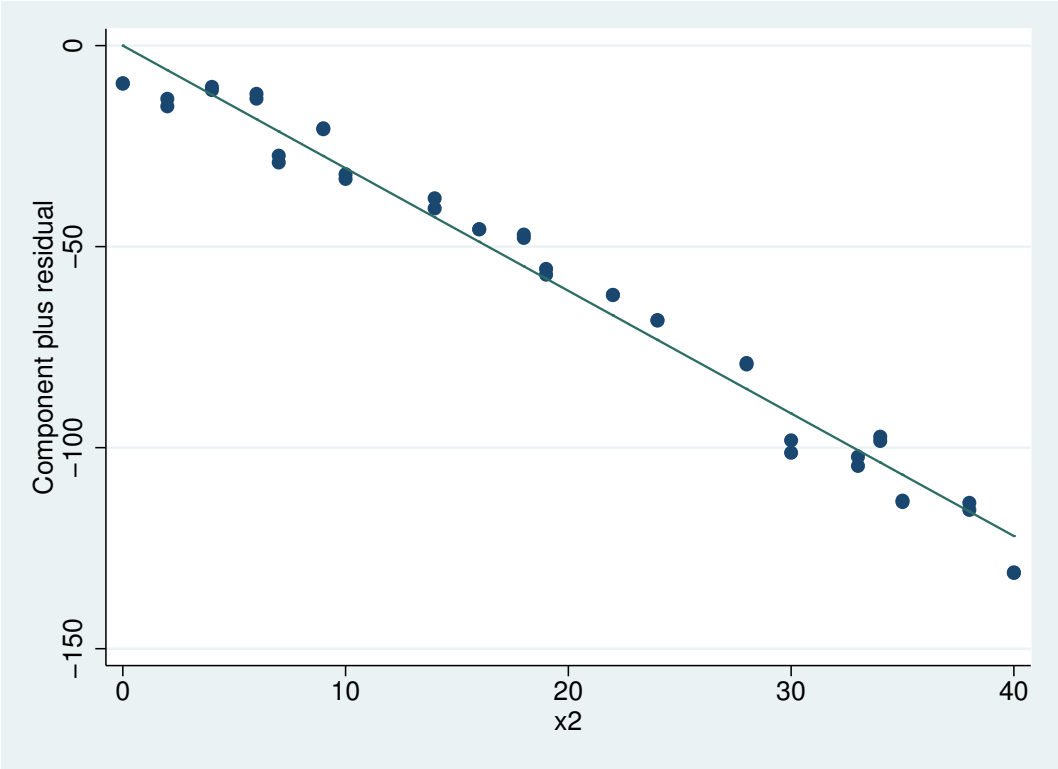

Figure 7: . cprplot x2

```
. regress Y x1 x2 x3
```

Source	SS	df	MS	Number of obs =	40
				F(3, 36) =	5455.28
Model	15720.4479	3	5240.14929	Prob > F =	0.0000
Residual	34.580338	36	.960564943	R-squared =	0.9978
				Adj R-squared =	0.9976
Total	15755.0282	39	403.975082	Root MSE =	.98008

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	20.31001	.2458675	82.61	0.000	19.81137	20.80866
x2	-3.007407	.0250592	-120.01	0.000	-3.05823	-2.956585
x3	-1.038003	.0274786	-37.78	0.000	-1.093733	-.9822743
_cons	20.00627	.3901361	51.28	0.000	19.21504	20.7975

3.10 Yes, the coefficient for x3 is highly significant, so after adjusting for x1 and x3, itis a significant predictor

```
. cprplot x1
```

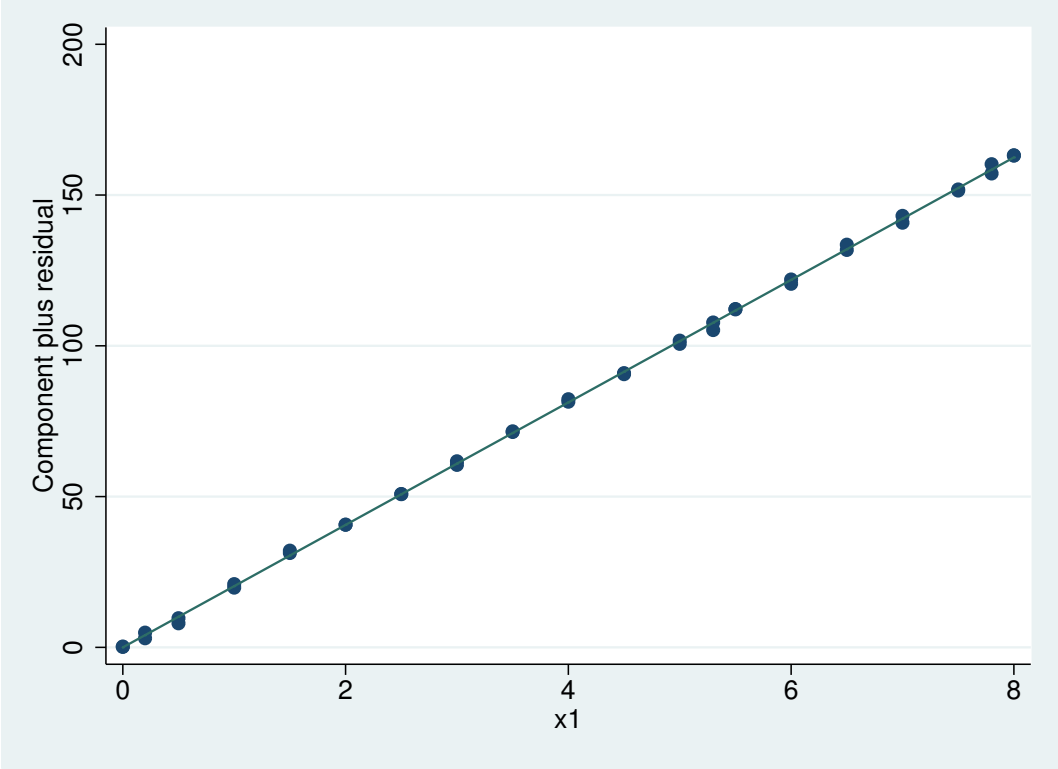

Figure 8: . cprplot x1

```
. graph export graph8.eps replace
(file graph8.eps written in EPS format)

. cprplot x2


. graph export graph9.eps replace
(file graph9.eps written in EPS format)

. cprplot x3


. graph export graph10.eps replace
(file graph10.eps written in EPS format)
```

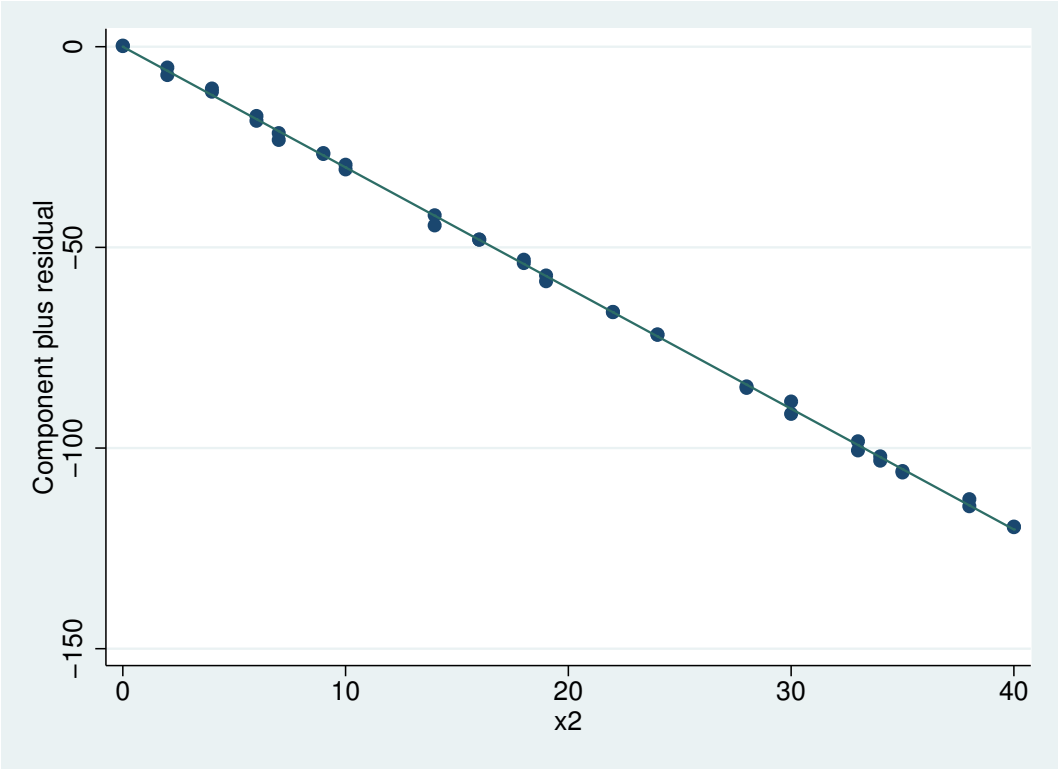

Figure 9: . cprplot x2

3.11 No, the non-linearity has been removed

```
. predict Yhat
(option xb assumed; fitted values)

. scatter Y Yhat


. graph export graph11.eps replace
(file graph11.eps written in EPS format)
```

3.12 The correlation between observed and predicted values is extremely high, so the regression model is producing excellent predictions
This is to be expected, since R-squared was well over 99%

```
. use $datadir/lifeline, clear
```

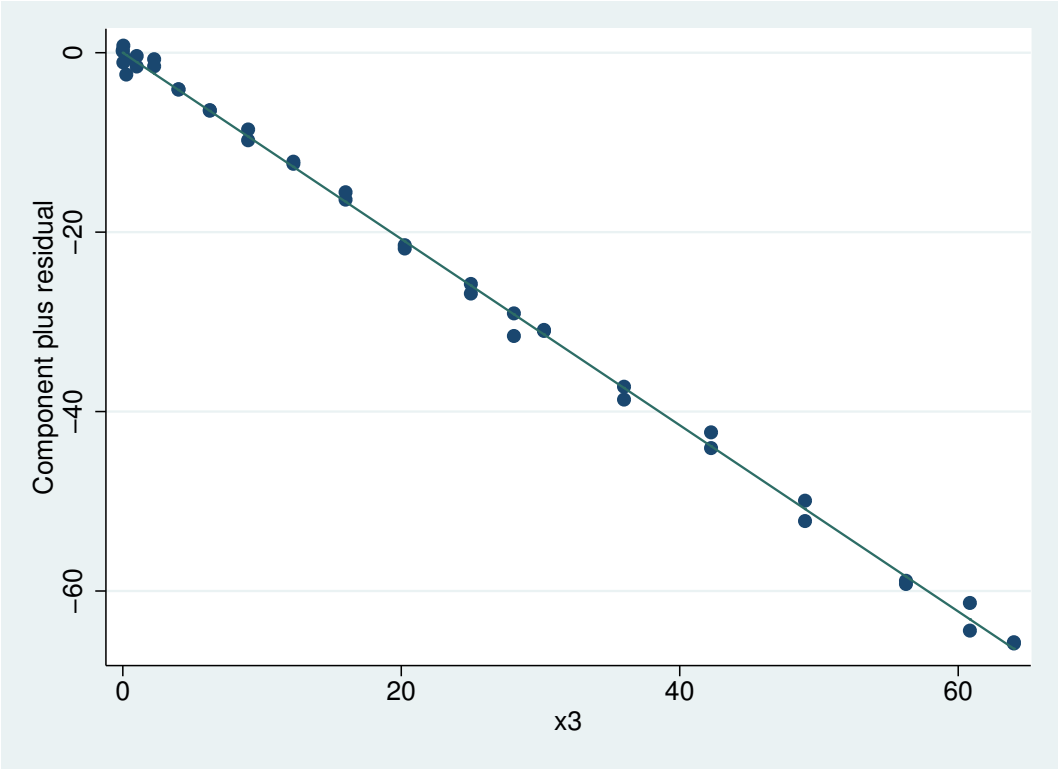

Figure 10: . cprplot x3

```
. regress age lifeline
```

Source	SS	df	MS	Number of obs =	50
				F(1, 48) =	7.39
Model	1301.96859	1	1301.96859	Prob > F =	0.0091
Residual	8453.25141	48	176.109404	R-squared =	0.1335
				Adj R-squared =	0.1154
Total	9755.22	49	199.086122	Root MSE =	13.271

age	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lifeline	-3.272017	1.203391	-2.72	0.009	-5.691596	-.8524384
_cons	97.1552	11.37154	8.54	0.000	74.29119	120.0192

3.13 Yes: p = 0.009

```
. scatter age lifeline
```

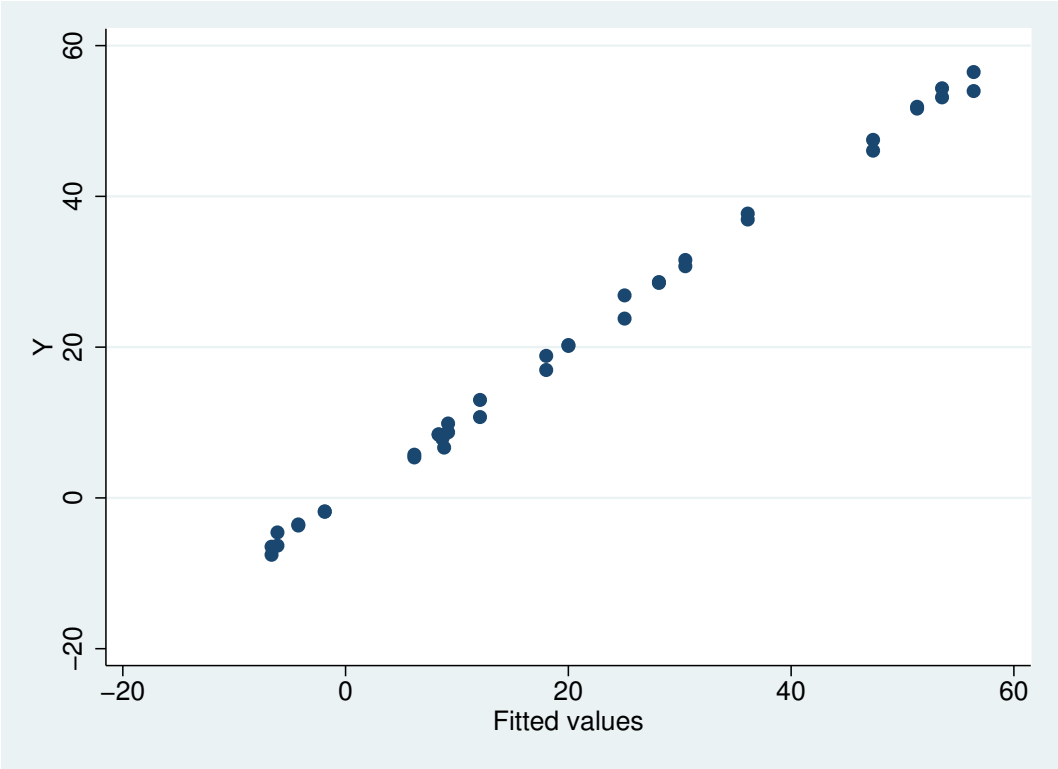

Figure 11: . scatter Y Yhat

```
. graph export graph12.eps replace
(file graph12.eps written in EPS format)
```

3.14 There is a single outlier in the bottm right cormer of the plot
3.15 This point has high leverage, and so should have a large effect on the regression

```
. predict predage
(option xb assumed; fitted values)

. predict cooksd, cooksd

. scatter cooksd predage


. graph export graph13.eps replace
(file graph13.eps written in EPS format)
```

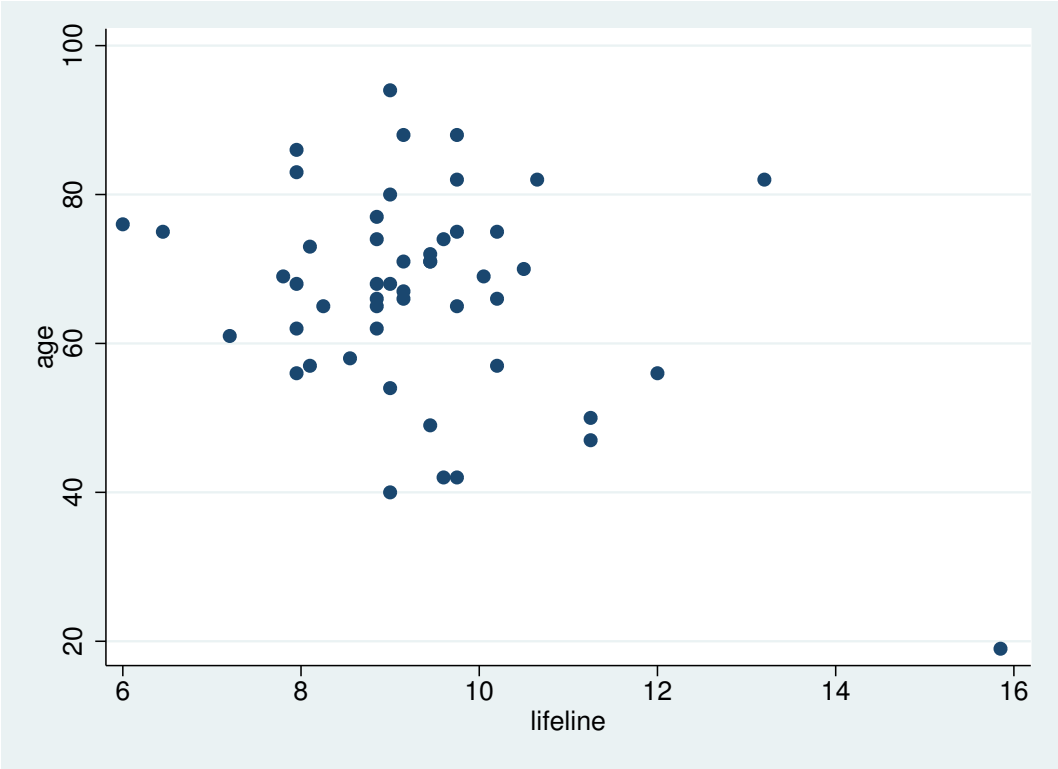

Figure 12: . scatter age lifeline

3.16 Certainly 1, possibly 2

```
. summarize cooksd, det
                          Cook´s D
-------------------------------------------------------------
      Percentiles      Smallest
 1%    2.53e-06       2.53e-06
 5%    4.09e-06       2.80e-06
10%    .0002006       4.09e-06       Obs                  50
25%    .0009213       5.30e-06       Sum of Wgt.          50
50%    .0049755                      Mean           .0563673
                      Largest        Std. Dev.       .264227
75%    .0238684       .0426679
90%    .0376543       .0473808       Variance       .0698159
95%    .0473808       .4377032       Skewness       6.361973
99%    1.836694       1.836694       Kurtosis       43.01234
```

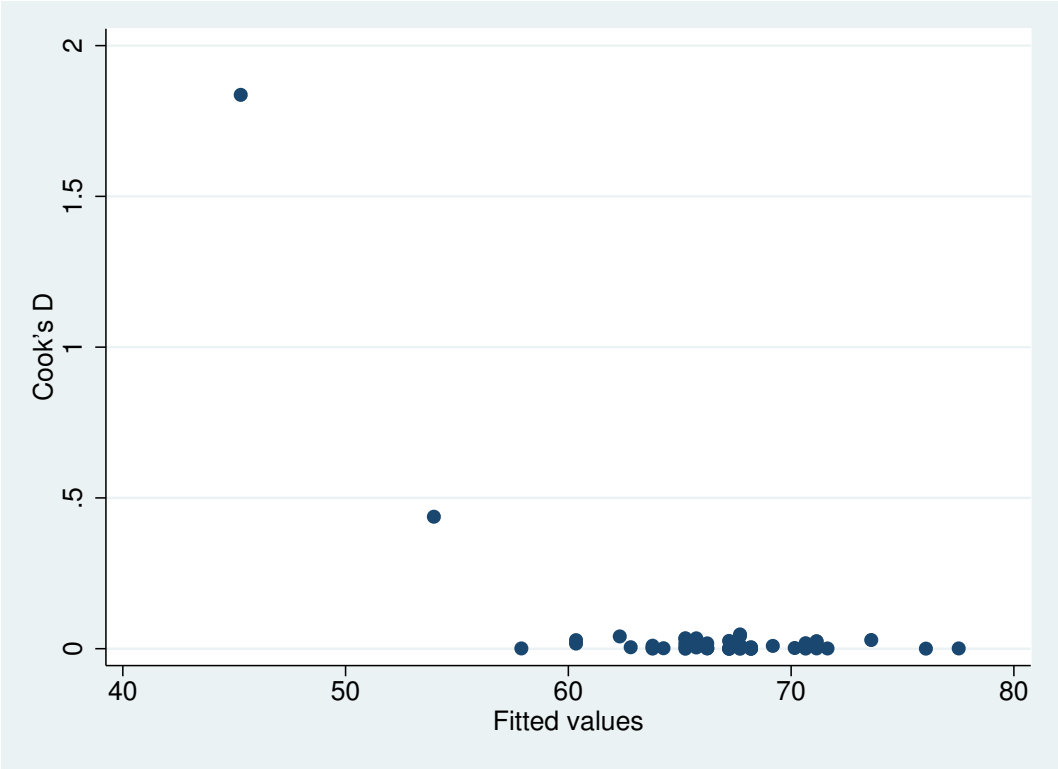

Figure 13: . scatter cooksd predage

```
. regress age lifeline if cooksd < 1
```

Source	SS	df	MS		
Model	82.6429704	1	82.6429704	Number of obs	= 49
				F(1, 47)	= 0.53
Residual	7354.74478	47	156.483932	Prob > F	= 0.4710
				R-squared	= 0.0111
				Adj R-squared	= -0.0099
Total	7437.38776	48	154.945578	Root MSE	= 12.509

age	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lifeline	-1.028681	1.415509	-0.73	0.471	-3.876316	1.818955
_cons	77.08287	13.12612	5.87	0.000	50.67652	103.4892

3.17 Effect of lifeline is no longer significant

```
. regress age lifeline if cooksd < 0.1
```

Source	SS	df	MS	Number of obs =	48
				F(1, 46) =	2.09
Model	314.264999	1	314.264999	Prob > F =	0.1549
Residual	6912.40167	46	150.269601	R-squared =	0.0435
				Adj R-squared =	0.0227
Total	7226.66667	47	153.758865	Root MSE =	12.258

age	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lifeline	-2.25765	1.561149	-1.45	0.155	-5.40008	.8847788
_cons	87.88501	14.32105	6.14	0.000	59.05822	116.7118

3.18 The association between age and lifeline is still not significant

3.19 There is no association between age and lifeline in general, the apparent association was caused by a single unusual observation

```
. regress age lifeline
```

Source	SS	df	MS	Number of obs =	50
				F(1, 48) =	7.39
Model	1301.96859	1	1301.96859	Prob > F =	0.0091
Residual	8453.25141	48	176.109404	R-squared =	0.1335
				Adj R-squared =	0.1154
Total	9755.22	49	199.086122	Root MSE =	13.271

age	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lifeline	-3.272017	1.203391	-2.72	0.009	-5.691596	-.8524384
_cons	97.1552	11.37154	8.54	0.000	74.29119	120.0192

```
. predict rstand, rstand

. qnorm rstand
```

3.20 The plot is reasonably linear: no points stand out as being unusual

```
. swilk rstand
```

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
rstand	50	0.99044	0.449	-1.705	0.95594

3.21 Yes: there is no evidence against the null hypothesis of a normal distribution

```
. use $datadir/hsng, clear
(1980 Census housing data)

. regress rent hsngval hsnggrow hsng faminc
```

Source	SS	df	MS		
Model	55285.8044	4	13821.4511	Number of obs	= 50
Residual	5957.31561	45	132.384791	F(4, 45)	= 104.40
				Prob > F	= 0.0000
				R-squared	= 0.9027
				Adj R-squared	= 0.8941
Total	61243.12	49	1249.85959	Root MSE	= 11.506

rent	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsngval	.0004964	.0001576	3.15	0.003	.000179	.0008139
hsnggrow	.6458343	.0988301	6.53	0.000	.4467803	.8448883
hsng	2.32e-06	9.39e-07	2.47	0.017	4.30e-07	4.21e-06
faminc	.0085855	.0008816	9.74	0.000	.0068098	.0103612
_cons	16.15788	13.70752	1.18	0.245	-11.4505	43.76625

4.1 50
4.2 All 4
4.3 0.65 (0.45, 0.84)
4.4 For each 1% increase in housing growth, the mean rent increases by about 65 cents
The true rent increase is probably between 45 and 84 cents
4.5 R-squared is 0.9, so the model accounts for 90% of the variation in rents

```
. predict rstand, rstand

. predict pred_val
(option xb assumed; fitted values)

. scatter rstand pred_val


. graph export graph14.eps replace
(file graph14.eps written in EPS format)

. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of rent

         chi2(1)      =     3.54
         Prob > chi2  =   0.0598
```

4.6 There is a slight suggestion of less variation for smaller fitted values, but it is onlyslight

Using hettest, it is of borderline significance

```
. rvfplot
```

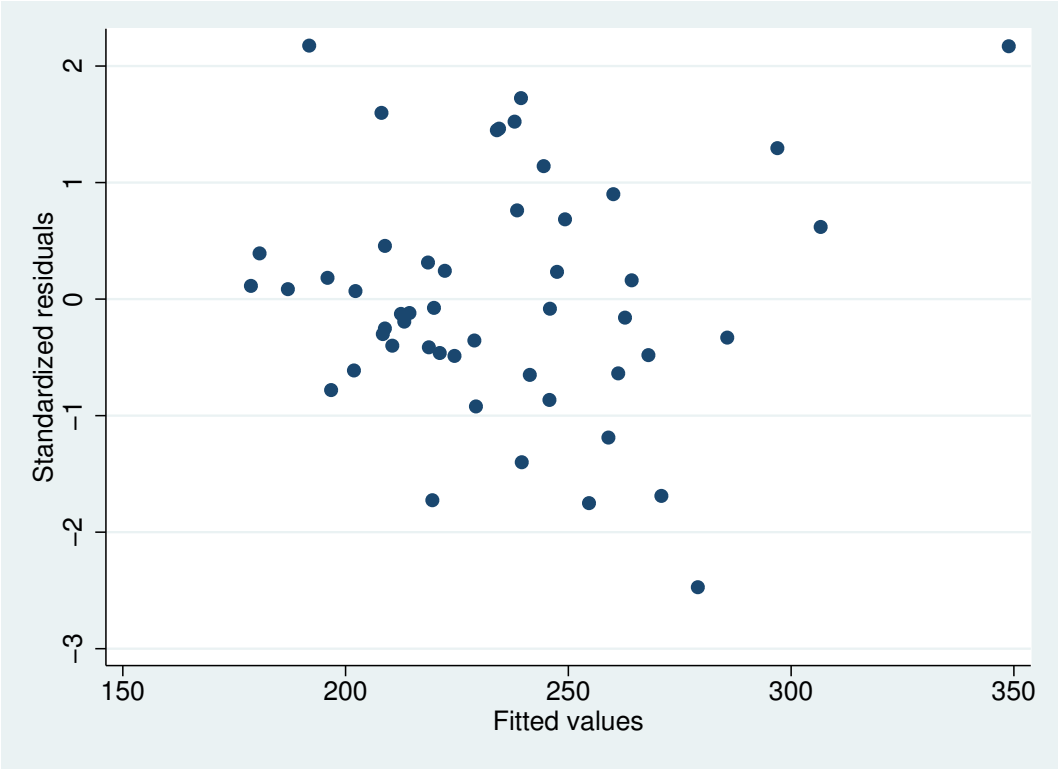

Figure 14: . scatter rstand pred˙val

```
. graph export graph15.eps replace
(file graph15.eps written in EPS format)
```

4.7 This plot is very similar to the previous one

```
. cprplot faminc
```

```
. graph export graph16.eps replace
(file graph16.eps written in EPS format)

. cprplot hsng
```

```
. graph export graph17.eps replace
(file graph17.eps written in EPS format)

. cprplot hsnggrow
```

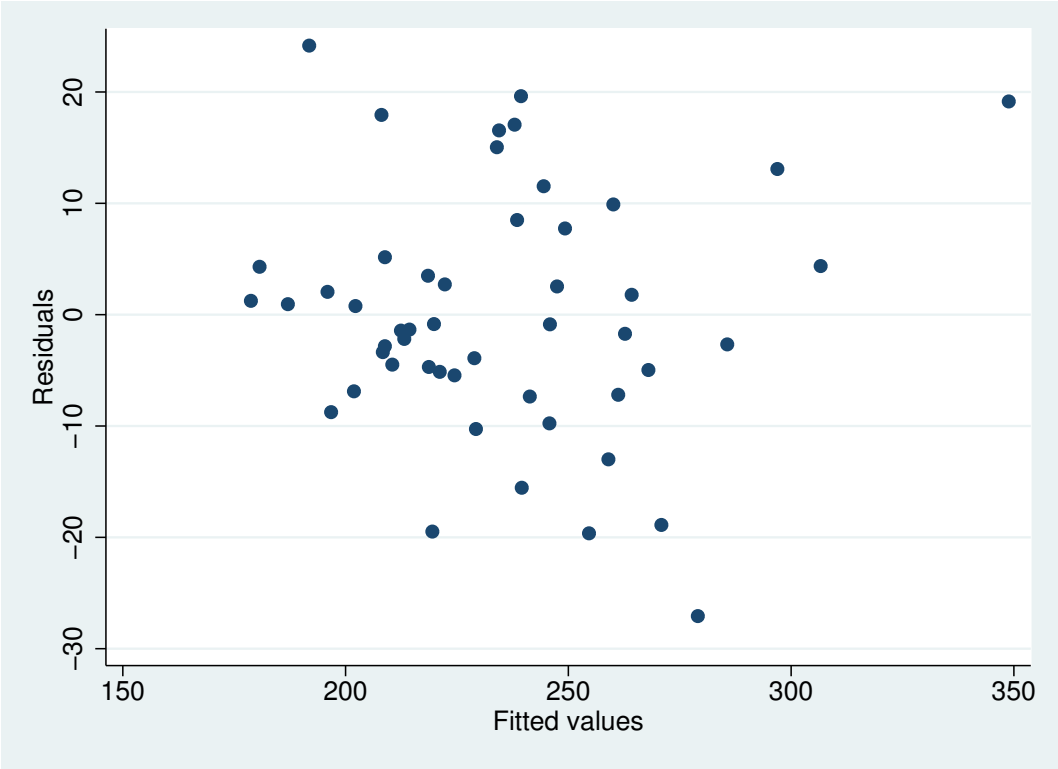

Figure 15: . rvfplot

```
. graph export graph18.eps replace
(file graph18.eps written in EPS format)

. cprplot hsngval


. graph export graph19.eps replace
(file graph19.eps written in EPS format)
```

4.8 There is no sign of non-linearity in any of the plots

```
. predict cooksd, cooksd

. scatter cooksd pred_val
```

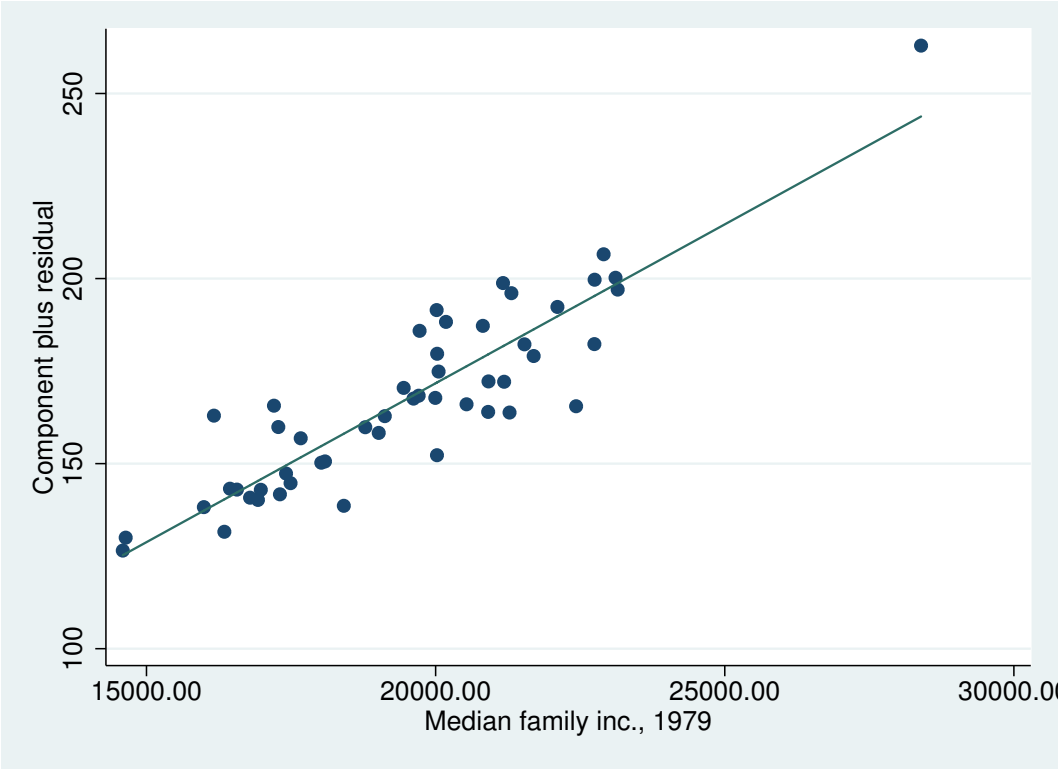

Figure 16: . cprplot faminc

```
. graph export graph20.eps replace
(file graph20.eps written in EPS format)
```

4.9 There is one point with a large Cook´s distance

```
. list if cooksd > 0.4
```

	state	division	region	pop	popgrow	popden	pcturban	faminc	hsng
2.	Alaska	Pacific	West	401851	32.8	7.0	64.3	28395.00	162825

hsnggrow	hsngval	rent	rstand	pred_val	cooksd
79.3	75200.00	368.00	2.169972	348.8493	.6589686

4.10 Alaska

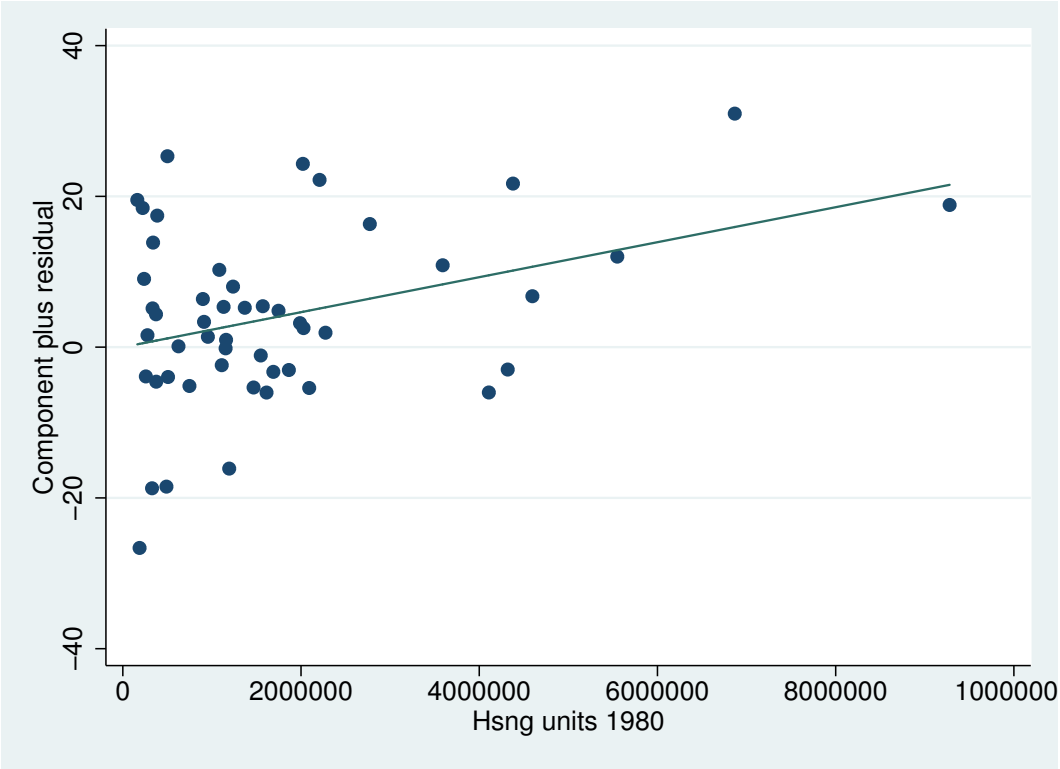

Figure 17: . cprplot hsng

```
. regress rent hsngval hsnggrow hsng faminc
```

Source	SS	df	MS
Model	55285.8044	4	13821.4511
Residual	5957.31561	45	132.384791
Total	61243.12	49	1249.85959

Number of obs = 50
F(4, 45) = 104.40
Prob > F = 0.0000
R-squared = 0.9027
Adj R-squared = 0.8941
Root MSE = 11.506

rent	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsngval	.0004964	.0001576	3.15	0.003	.000179	.0008139
hsnggrow	.6458343	.0988301	6.53	0.000	.4467803	.8448883
hsng	2.32e-06	9.39e-07	2.47	0.017	4.30e-07	4.21e-06
faminc	.0085855	.0008816	9.74	0.000	.0068098	.0103612
_cons	16.15788	13.70752	1.18	0.245	-11.4505	43.76625

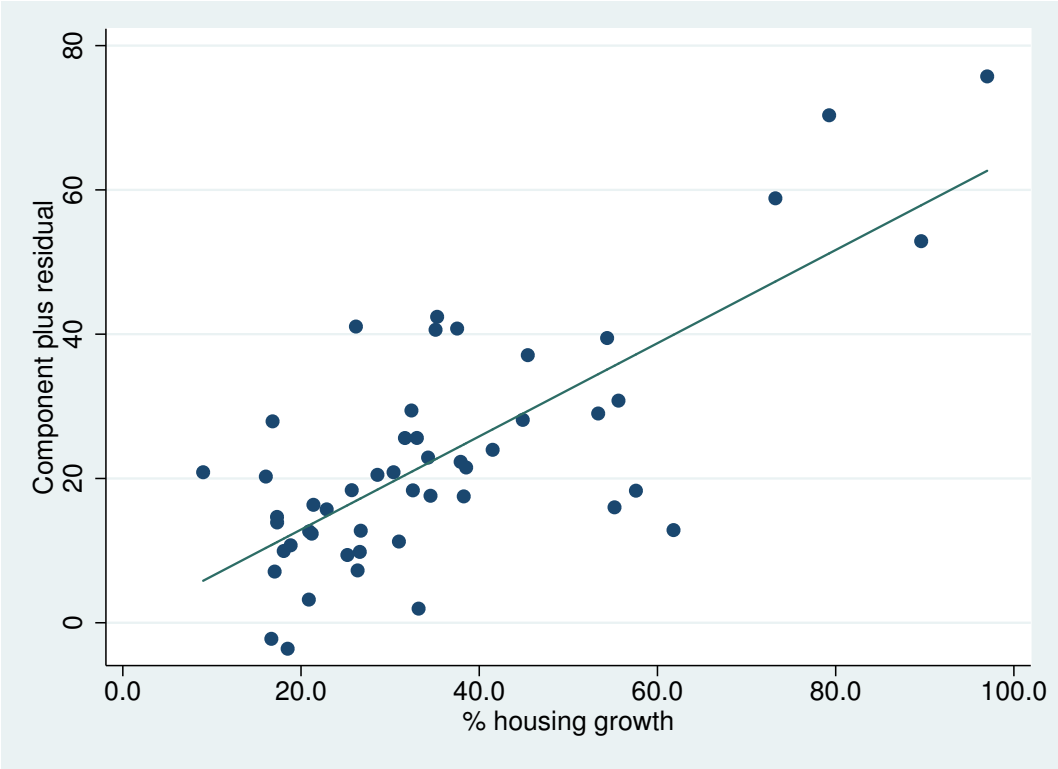

Figure 18: . cprplot hsnggrow

```
. regress rent hsngval hsnggrow hsng faminc if cooksd < 0.5
```

Source	SS	df	MS		
Model	37793.9737	4	9448.49341	Number of obs	= 49
Residual	5333.94471	44	121.226016	F(4, 44)	= 77.94
				Prob > F	= 0.0000
				R-squared	= 0.8763
				Adj R-squared	= 0.8651
Total	43127.9184	48	898.498299	Root MSE	= 11.01

rent	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsngval	.0006095	.0001588	3.84	0.000	.0002894	.0009296
hsnggrow	.5591967	.1019989	5.48	0.000	.3536314	.764762
hsng	2.65e-06	9.10e-07	2.91	0.006	8.13e-07	4.48e-06
faminc	.0072962	.0010174	7.17	0.000	.0052459	.0093466
_cons	37.67935	16.19046	2.33	0.025	5.049616	70.30909

4.11 They all change slightly, but all remain significant, in the same direction, and with nearly the same magnitude

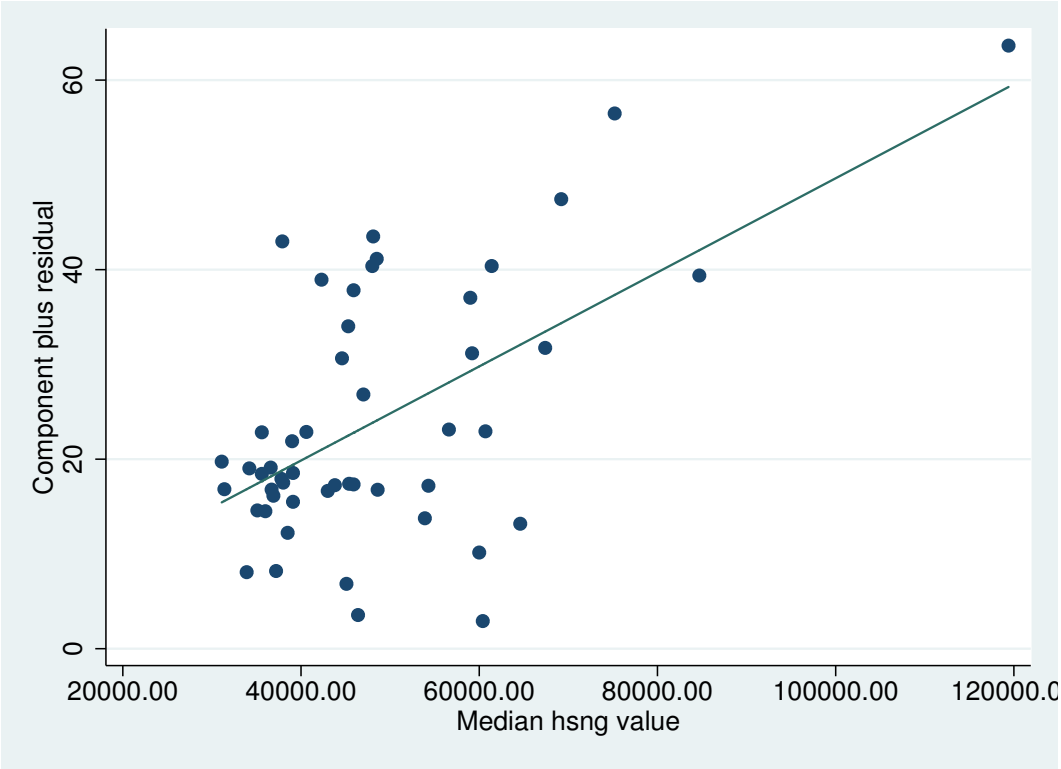

Figure 19: . cprplot hsngval

```
. predict pred2
(option xb assumed; fitted values)

. scatter pred2 pred_val
```

4.12 No: the predicted values including and excluding Alaska are very nearly the same

```
. qnorm rstand

. scatter pred2 pred_val


. graph export graph21.eps replace
(file graph21.eps written in EPS format)

. qnorm rstand
```

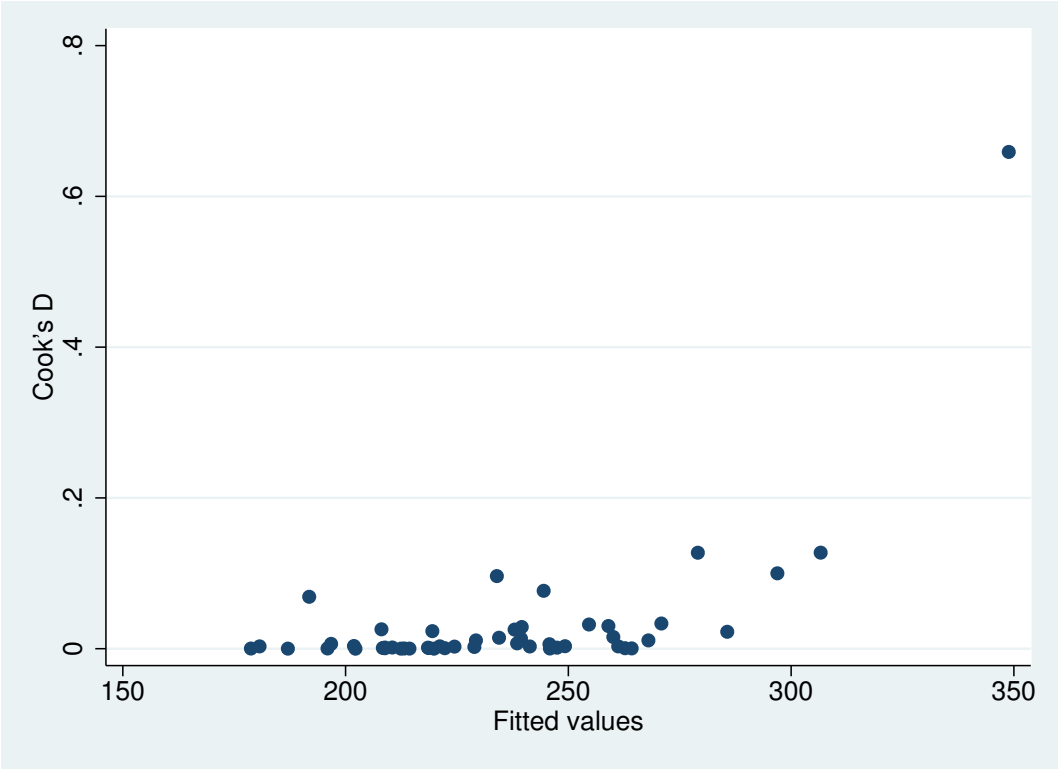

Figure 20: . scatter cooksd pred˙val

```
. graph export graph22.eps replace
(file graph22.eps written in EPS format)
```

4.13 Yes, the residuals appear to be normally distributed

```
. swilk rstand
```

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
rstand	50	0.97838	1.017	0.036	0.48579

4.14 Yes, there is no evidence against the null hypothesis of a normal distribution
end of do-file

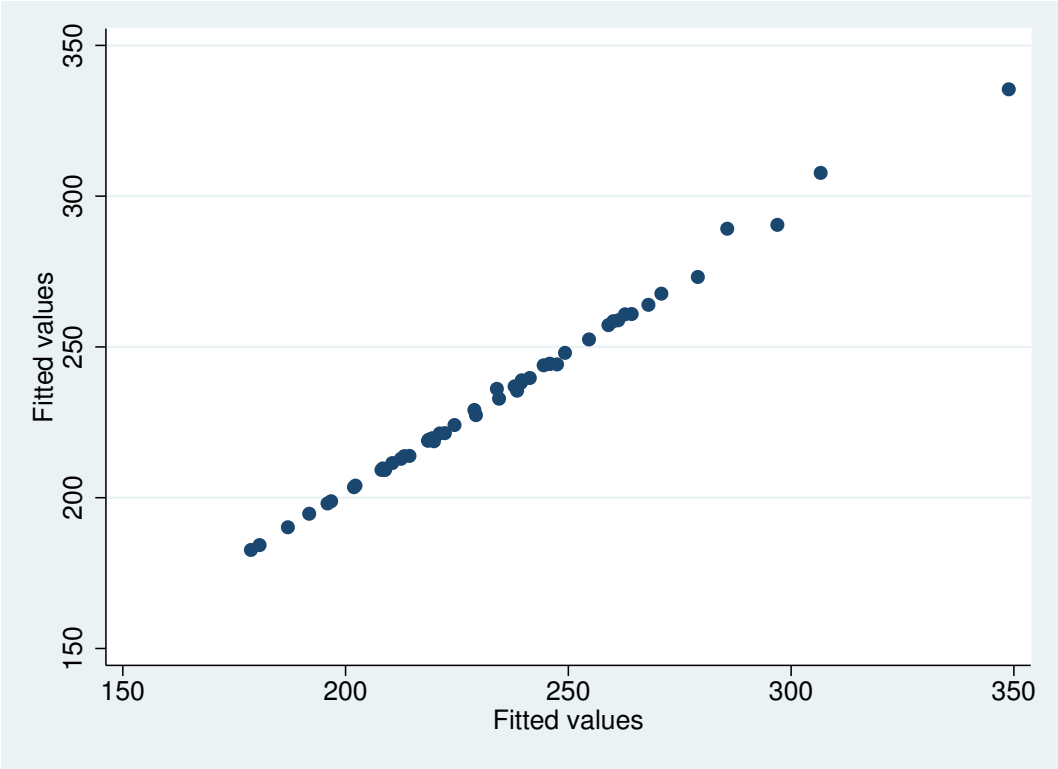

Figure 21: . scatter pred2 pred˙val

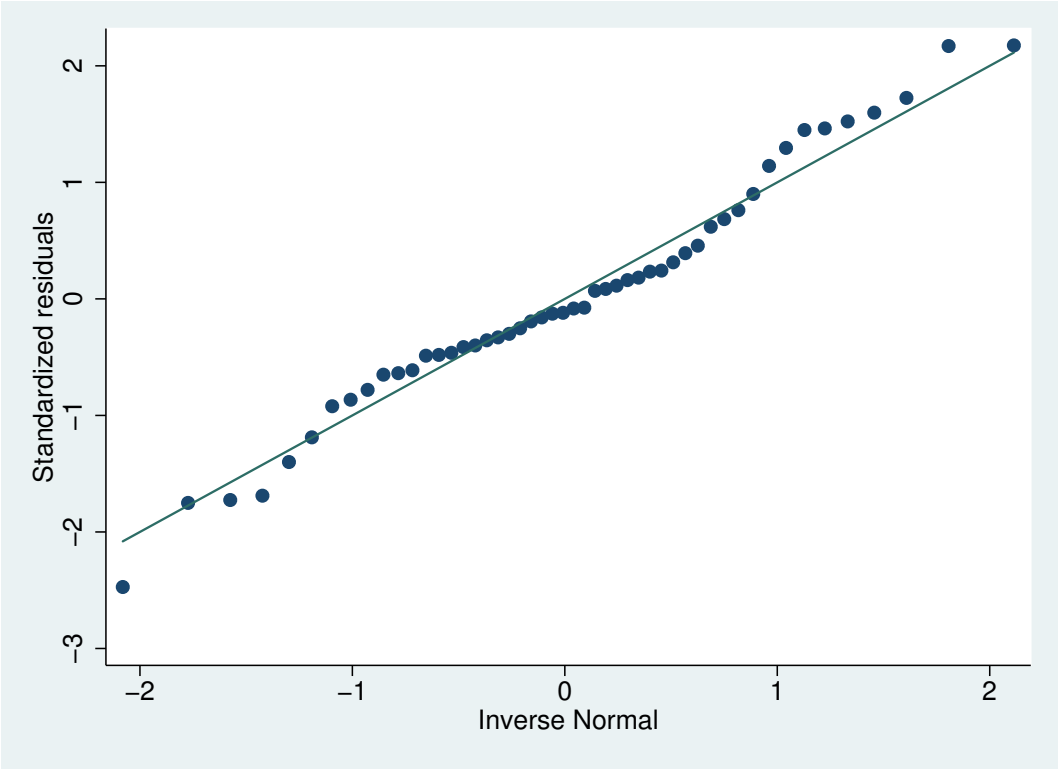

Figure 22: . qnorm rstand