

University of North Carolina
Chapel Hill

Soci708-001 Statistics for Sociologists

Fall 2009

Professor François Nielsen

Stata Commands for Module 11 – Multiple Regression

For further information on any command in this handout, simply type `help` followed by the name of the command in Stata.

See also the Stata and SAS Guide pdf (click on Documents in side bar; guide is linked under Software Documentation).

1 Statistical Functions in Stata

The following statistical functions in Stata are useful for regression work. The regression printout itself usually comprises all necessary statistics.

1.1 Normal Distribution Functions

The function `normal(z)` returns $P(Z \leq z)$, the area under the standard normal curve to the left of z . (Compare with Table A.)

```
. display normal(1.207)
.88628393
```

The function `invnormal(p)` returns z such that $P(Z \leq z) = p$, i.e. z such that the area under the standard normal curve to the left of z is p . (Compare with Table A and Table D, bottom row.)

```
. display invnormal(0.975)
1.959964
```

1.2 Student t Distribution Functions

The function `ttail(df, t)` returns $P(T > t)$, the area under the Student's t distribution with df degrees of freedom to the right of t . (Compare with Table D.)

```
. display ttail(7, 1.960)
.04540985
```

The function `invttail(df, p)` returns t such that $P(T > t) = p$, i.e. t such that the area under Student's t distribution with df degrees of freedom to the right of t is p . (Compare with Table D.)

```
. display invttail(7, 0.025)
2.3646243
```

1.3 F Distribution Functions

The function `Ftail(n1, n2, f)` returns $P(F > f)$, the area under the F distribution with `n1` and `n2` degrees of freedom to the right of `f`. (Compare with Table E.)

```
. display Ftail(1, 14, 21.55)
.00038068
```

The function `invFtail(n1, n2, p)` returns `f` such that $P(F > f) = p$, i.e. `f` such that the area under the F distribution with `n1` and `n2` degrees of freedom to the right of `f` is `p`. (Compare with Table E.)

```
. display invFtail(1, 14, .00038068)
21.549979
```

2 Descriptive Statistics, Correlations, and Scatterplot Matrix

I am using as an example the CSDATA from IPS6e (see Appendix D-2 for description). The units are 224 Computer Science majors at a large university. To enter the data in Stata I retrieved the `csdata.xls` file in the CD-ROM, selected the data and copied them to the clipboard (Ctrl-C). Then in Stata I opened the Data Editor (Data -> Data Editor) and pasted the data (Ctrl-V). Then I closed the Data Editor by clicking on \times . (You can save the data as a `*.dta` file if desired with File -> Save ...) Then I listed the first 5 cases.

```
. list in 1/5
```

```
-----+-----
| obs   gpa   hsm   hss   hse   satm   satv   sex |
|-----+-----|
1. |   1   3.32   10   10   10   670   600   1 |
2. |   2   2.26    6    8    5   700   640   1 |
3. |   3   2.35    8    6    8   640   530   1 |
4. |   4   2.08    9   10    7   670   600   1 |
5. |   5   3.38    8    9    8   540   580   1 |
-----+-----
```

The response variable of interest is grade point average after three semesters (`gpa`). The explanatory variables are high school grades in mathematics (`hsm`), science (`hss`) and English or language arts (`hse`); SAT score in math (`satm`) and verbal (`satv`); and `sex` (1=male, 2=female).

First I produced descriptive statistics for all the variables I intend to put in the regression with the command `su` (for summarize).

```
. su hsm hss hse satm satv gpa
```

```
-----+-----
Variable |      Obs      Mean    Std. Dev.    Min     Max
-----+-----
    hsm |      224    8.321429    1.638737         2     10
    hss |      224    8.089286    1.699663         3     10
    hse |      224    8.09375     1.507874         3     10
    satm |      224   595.2857    86.40144        300    800
    satv |      224   504.5491    92.61046        285    760
-----+-----
    gpa |      224    2.635223    .7793949         .12     4
```

Then I produced a matrix of correlation coefficients. Note that I put the response variable `gpa` last, to have the corresponding correlations all on the same row, same as in the scatterplot matrix to be produced next.

```
. corr hsm hss hse satm satv gpa
(obs=224)
```

	hsm	hss	hse	satm	satv	gpa
hsm	1.0000					
hss	0.5757	1.0000				
hse	0.4469	0.5794	1.0000			
satm	0.4535	0.2405	0.1083	1.0000		
satv	0.2211	0.2617	0.2437	0.4639	1.0000	
gpa	0.4365	0.3294	0.2890	0.2517	0.1145	1.0000

Then I produced a scatterplot matrix with the following command. Note that I place the response variable `gpa` last in the list so all the scatterplots will be on the last row, with `gpa` on the vertical axis and the explanatory variable on the horizontal axis of each panel. I used the option `half` to show only the lower half of the matrix, reducing the complexity of the graph and making it more easily comparable to the correlation matrix, and option `xsize(4) ysize(4)` to make the graph 4 in square. The graph is shown in Figure 1.

```
. graph matrix hsm hss hse satm satv gpa, half xsize(4) ysize(4)
```

3 Multiple Regression

To do the multiple regression use the `reg` command with the list of variables, starting with the response variable `gpa`.

```
. reg gpa hsm hss hse satm satv
```

Source	SS	df	MS	Number of obs = 224		
Model	28.6436439	5	5.72872878	F(5, 218)	=	11.69
Residual	106.819145	218	.489996078	Prob > F	=	0.0000
Total	135.462789	223	.607456452	R-squared	=	0.2115
				Adj R-squared	=	0.1934
				Root MSE	=	.7

gpa	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
hsm	.1459611	.039261	3.72	0.000	.0685814	.2233407
hss	.0359053	.0377984	0.95	0.343	-.0385918	.1104024
hse	.0552926	.0395687	1.40	0.164	-.0226936	.1332787
satm	.0009436	.0006857	1.38	0.170	-.0004078	.002295
satv	-.0004078	.0005919	-0.69	0.492	-.0015744	.0007587
_cons	.3267187	.3999964	0.82	0.415	-.4616365	1.115074

By default the `reg` command calculates the 95% CI for each coefficient. You can change the confidence level with the `level` option, e.g. `level(99)` for a 99% CI. The full command would then be (output not shown):

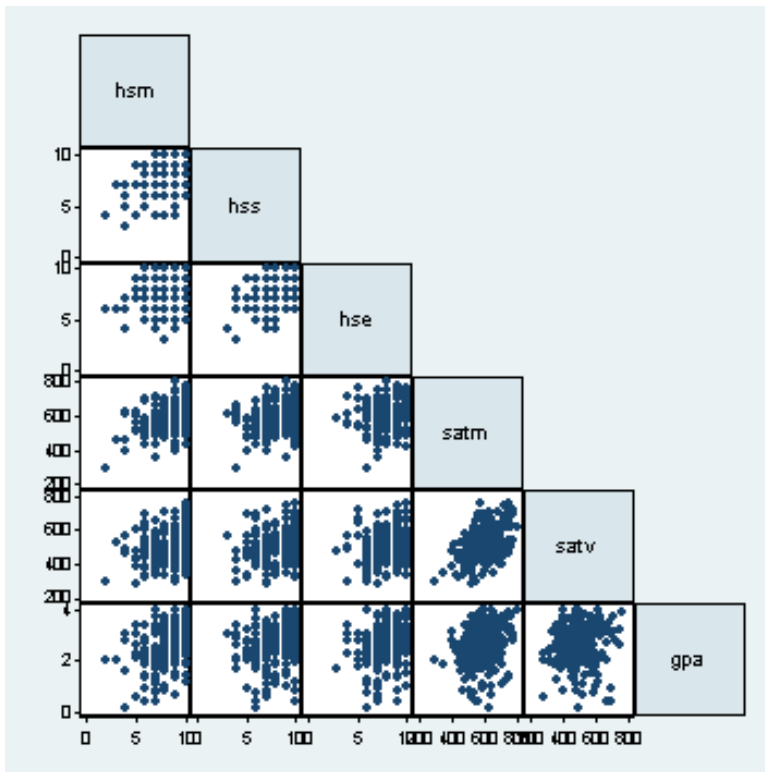


Figure 1: Scatterplot matrix of variables for the gpa regression (CSDATA).
 Produced with Stata command `graph matrix hsm hss hse satm satv gpa, half xsize(4) ysize(4)`

```
. reg gpa hsm hss hse satm satv, level(99)
```

To obtain standardized coefficients (in place of the confidence intervals shown by default) use the option `beta`.

```
. reg gpa hsm hss hse satm satv, beta
```

Source	SS	df	MS		
Model	28.6436439	5	5.72872878	Number of obs =	224
Residual	106.819145	218	.489996078	F(5, 218) =	11.69
Total	135.462789	223	.607456452	Prob > F =	0.0000
				R-squared =	0.2115
				Adj R-squared =	0.1934
				Root MSE =	.7

gpa	Coef.	Std. Err.	t	P> t	Beta
hsm	.1459611	.039261	3.72	0.000	.3068942
hss	.0359053	.0377984	0.95	0.343	.0783004
hse	.0552926	.0395687	1.40	0.164	.106973
satm	.0009436	.0006857	1.38	0.170	.1046039
satv	-.0004078	.0005919	-0.69	0.492	-.0484621
_cons	.3267187	.3999964	0.82	0.415	.

Using the command `vif` after running a regression will calculate the *variance inflation factors* (VIF). These are measures of *collinearity*, the degree to which each explanatory variable is associated with all the other explanatory variables. A VIF above 10 is considered bothersome, but there is no VIF above 10 in this particular example. (We are not going to use `vif` in this class.)

```
. vif
```

Variable	VIF	1/VIF
hsm	1.88	0.530820
hss	1.88	0.532372
hse	1.62	0.617241
satm	1.60	0.626083
satv	1.37	0.731277
Mean VIF	1.67	

4 Analysis of Residuals

To calculate the predicted values of `gpa` and the `gpa` residuals I can use the command `predict`, with the option `xb` and `residuals`, respectively, assigning variable names of my choice. Then to check the distribution of residuals I draw a histogram of the residuals (shown in Figure 2) and a normal quantile plot of the residuals (shown in Figure 3). The only fancy options I use is `xsize(3.5)` `ysize(3.5)` with the normal quantile plot, to make the plot square. Together with the straight line that Stata draws automatically the square format shows deviations of the plot from linearity better than a rectangular plot (compare with IPS6e Figure 11.5 p.620). We can see the left-skew in the distribution.

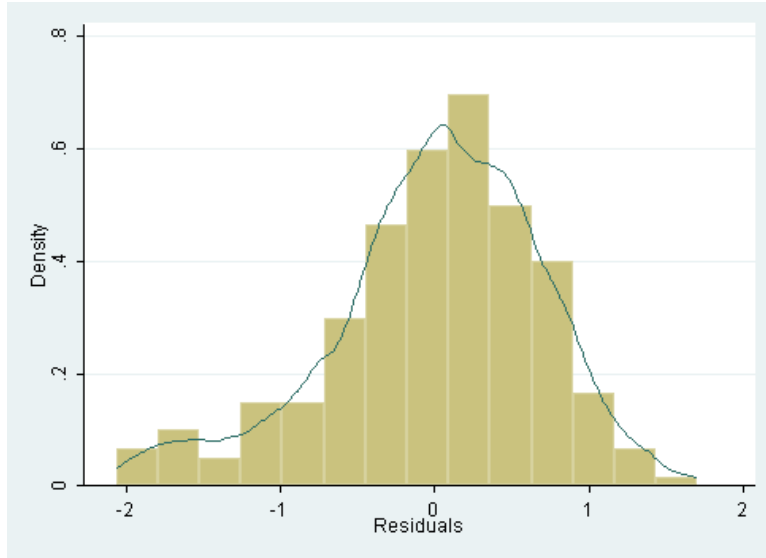


Figure 2: Distribution of residuals for the gpa regression (CSDATA). Produced by Stata command `histogram gparesid, kdensity`

```
. predict gpapredict, xb
. predict gparesid, residuals
. histogram gparesid, kdensity
(bin=14, start=-2.0649281, width=.26931122)
. qnorm gparesid, xsize(3.5) ysize(3.5)
```

Now I want to create a residual plot (plot of residuals against predicted values). I use the following command. The plot is shown in Figure 4.

```
. twoway (scatter gparesid gpapredict), yline(0)
```

I am interested in the six cases I see on the graph that have low predicted gpa. To identify them I use the following command (make sure you understand *why*). What is the story of these students with low predicted gpa? You may want to refer to the descriptive statistics calculated earlier.

```
. list if gpapredict<1.9
```

```
-----+-----
| obs   gpa   hsm   hss   hse   satm   satv   sex   gpapre~t   gparesid |
|-----+-----|
  8. |   8     2     3     7     6    460    530    1    1.565587    .434413 |
 84. |  84     3     4     3     4    620    560    1    1.596081    1.403919 |
107. | 107   2.82     4     5     7    400    470    1    1.662885    1.157115 |
127. | 127    .12     4     6     6    630    490    1    1.852368   -1.732368 |
182. | 182    1.6     4     7     7    460    460    2    1.79539    -.1953901 |
|-----+-----|
183. | 183     2     2     4     6    300    290    2    1.258819    .7411809 |
-----+-----
```

Now I want to calculate the standard error for the estimated mean response ($SE_{\hat{\mu}}$ in IPS6e; standard error of prediction for Stata), and the standard error of prediction for a future observation ($SE_{\hat{y}}$ in IPS6e; standard error of forecast

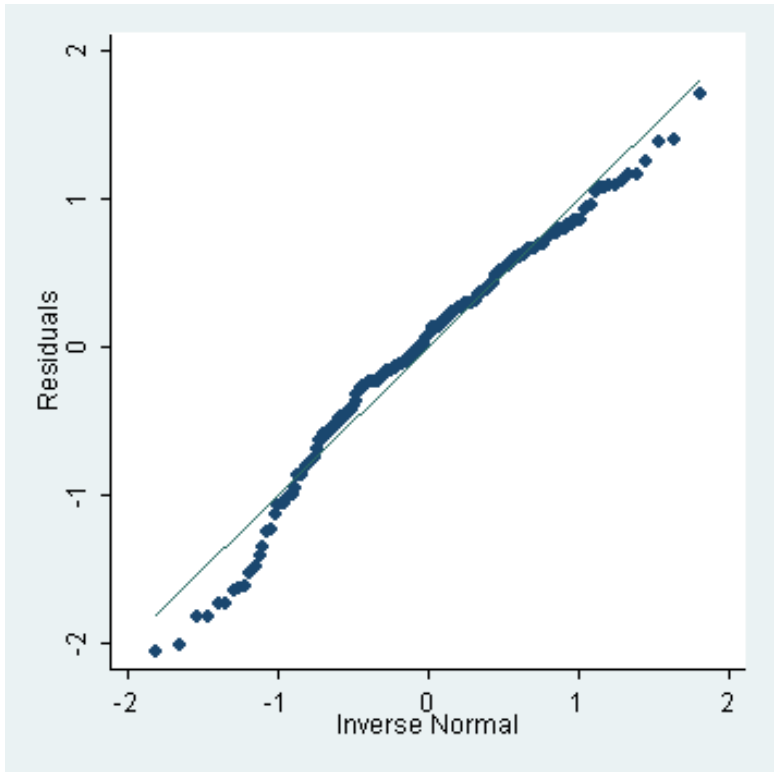


Figure 3: Normal quantile plot of residuals for the gpa regression (CSDATA). Produced by Stata command `qnorm gparesid, xsize(3.5) ysize(3.5)`

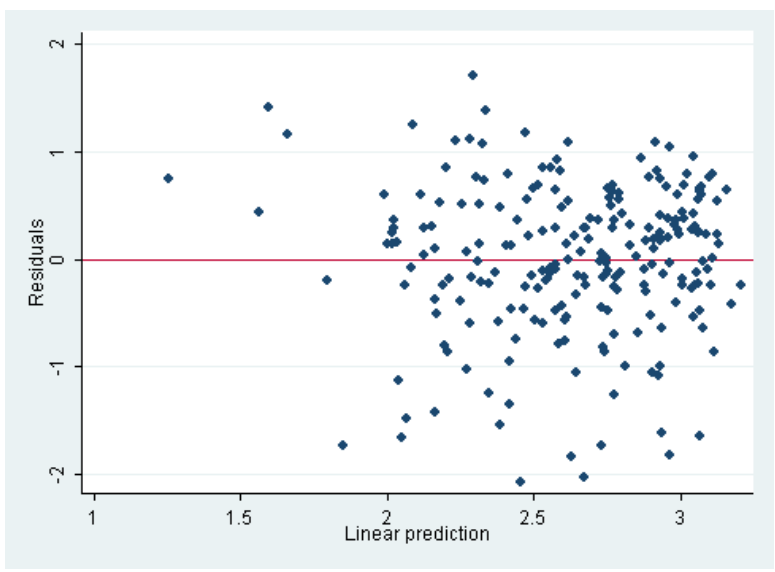


Figure 4: Residual plot for the gpa regression (CSDATA). Produced by Stata command `twoway (scatter gparesid gpapredict), yline(0)`

for Stata; see IPS6e p.594 for formulas). I check the values for the first 5 observations. Note that the SE of forecast is always larger than the SE for the mean response, as the SE of forecast contains individual variation in the response variable in addition to uncertainty about the mean response.

```
. predict gpasepred, stdp
. predict gpaseforecast, stdf
. list gpapredict gpasepred gpaseforecast in 1/5
```

```
+-----+
| gpapre~t  gpasep~d  gpasef~t |
+-----+
1. | 3.085806   .0871256   .7053984 |
2. | 2.165682   .1740038   .7212998 |
3. | 2.539919   .0932556   .7061818 |
4. | 2.773967   .1156441   .7094855 |
5. | 2.532883   .0898973   .7057461 |
+-----+
```