

MULTIVARIATE REGRESSION I

Steps in Statistical Analysis, especially model specification and goodness of fit

We will be going over various aspects of regression problems during the next few weeks. These will include:

1. Omitted variables
2. Nonlinear relationships
3. Non-constant error variance
4. Correlation among errors
5. Nonnormal errors
6. Influential cases

First, however, we want to focus on

Steps in statistical analysis

1. Specify the model
 - Usually a linear model with normally distributed homoscedastic error terms
 - Transform data (dependent and/or independent variables) to meet normality assumptions
2. Estimate the parameters
 - Frequently use the least squares criterion - choose as estimates of the parameters those that yield the smallest RSS (residual sum of squares).
 - Later in the course, we will discuss another criterion that is frequently used, the *maximum likelihood criterion* for finding estimates
 - We have also briefly referred to *robust* estimates, those that are little affected by one or two unusual values
3. Consider goodness of fit
 - One criterion is the R^2 value - the proportion of the original variation about the mean (sum of the squared residuals about the mean) that is "explained" by the model
 - Another consideration is the pattern of residuals about the line -
 - are there unusual values (outliers)?
 - is there heteroscedasticity? or are they distributed about the predicted value in the same way, no matter what that predicted value?
 - A third question to ask is whether the model fits better, or as well, as another model
4. Interpret the results
 - The b 's are *partial* regression coefficients. If the coefficient of X_i is estimated as b_i , we say that, if X_i increases by 1, the expected of Y increases by b_i .
 - If the b 's are *standardized* coefficients, we talk about everything in *standardized* units. The interpretation then is, if X_i increases by one standard deviation of X_i , then the expected value of Y increases by (standardized) b_i standard deviations of Y .
 - When we do cross-sectional studies, we can ONLY be sure that the coefficients represent *association* between the X variable and the dependent variable, Y . They do NOT represent causation. We will discuss this point further in later lectures.

We will return to this schema repeatedly.

Today I wanted to focus on points 1 and 3: specifying the model, and goodness of fit.

MODELS

In statistics, we usually talk about two models for our data:

- the model where Y is the sum of a constant and an error term, which I will refer to as Model C (for constant):

Install Equation Editor and double-click here to view equation.

- the model where Y is the sum of its linear relationship to one predictor, X , and an error term, which I will refer

Install Equation Editor and double-click here to view equation.

to as Model 2p (2 parameters, $\hat{\alpha}_0$ and $\hat{\alpha}_1$):

Model C could have been called Model 1p, since it only has one parameter ($\hat{\mu}$).

Each of these models says that Y is the sum of its *expected value* and an error term. The expected value in Model C is a constant, the mean; the expected value in Model 2p is the point on the line corresponding to the particular value of X_i .

In each case, we choose as the estimates of the parameters the values of $\hat{\mu}$ in Model C and $\hat{\alpha}_0$ and $\hat{\alpha}_1$ that *minimize the sum of the squared errors*. For Model C, the estimate of $\hat{\mu}$ is the sample mean. For model 2p, the equations for b_0 and b_1 , the estimates of $\hat{\alpha}_0$ and $\hat{\alpha}_1$, were given last semester. I'd like to go over what we mean by these statements in some detail.

To do so, I will use a small dataset on the relationship of number of breeding pairs to area in 22 colonies.

```
. use h1-1.dta
. list colony area pairs
```

	colony	area	pairs
1.	Out Skerries	1616	284
2.	N.E. Unst	1588	495
3.	Fetlar	931	372
4.	W. Yell	126	134
5.	Vaila	302	278
6.	Fitful Head	596	500
7.	Uyea	898	731
8.	W. Unst	208	311
9.	Buravoe	353	485
10.	St. Ninians	106	250
11.	Papa Stour	809	1036
12.	Gruney	565	1364
13.	Eshaness	1069	2430
14.	S. Havra	242	925
15.	Foula	2927	5570
16.	Moussa	614	1975
17.	Sumburge	1273	3243
18.	Hermaness	1570	3872
19.	Reawick	111	970
20.	Dalsetter	60	970
21.	Fair Isle	3957	17000
22.	Noss	1317	10767

Rename

```
. rename pairs Y
. rename area X
```

Regress Y on X, then find the predicted values

```
. regress Y X
```

Source	SS	df	MS	Number of obs =	22
-----+-----				F(1, 20) =	32.06

Model		213268475	1	213268475	Prob > F	=	0.0000
Residual		133050026	20	6652501.31	R-square	=	0.6158

Total		346318501	21	16491357.2	Adj R-square	=	0.5966
					Root MSE	=	2579.2

Y		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	

X		3.302147	.5832108	5.662	0.000	2.085591	4.518704
_cons		-734.9547	786.9989	-0.934	0.362	-2376.606	906.6963

Here I want to note that the Root MSE in the top panel of the table produced by `regress` is the estimate of the sd of the error about the line, i.e. **the Root MSE estimates the sd of \hat{a}** . It is calculated for you by STATA, but you could calculate it from other information in the table, specifically, the Root MSE

Install Equation Editor and double-click here to view equation.

Predicted value of Y, Yhat

We can find the predicted value of y, `yhat2p`, in 2 ways: brute force do it yourself or by using the command `predict` after a regression calculation; the results are exactly the same (but for rounding error):

```
. gen yhat2pm = -734.9547 + 3.302147*X           Brute force - my calculation
. predict yhat2p                                predict used after regress
```

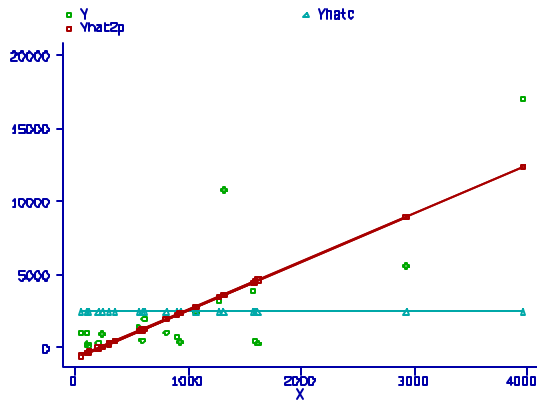
To convince ourselves both procedures give us the same results:

```
. list yhat2p yhat2pm in 1/5   I could also use the command list yhat2p* in 1/5
   yhat2p   yhat2pm
1.   4601.315   4601.315
2.   4508.855   4508.854
3.   2339.344   2339.344
4.   -318.8842  -318.8842
5.    262.2938   262.2937
```

Generate the predicted value under Model C - all areas are predicted to have the same number of pairs and that number is the mean of pairs. Again we can do it two ways - brute force, or by using the command `egen`:

```
. gen yhatcm = 2452.818           Brute force - my calculation
. egen yhatc = mean(Y)           egen with the function mean
. list yhatc yhatcm in 1/5
```

Plot the observed Y's, and those predicted under the two models, Model C and Model 2p, against X



```
. graph Y yhatc yhat2p X, connect(.11) xlabel ylabel
```

Where do the SS given as a result of regress come from?

Install Equation Editor and double-click here to view equation.

Install Equation Editor and double-click here to view equation.

The graph above shows two predicted values for each Y -- one from Model C and the other from Model 2p.

Install Equation Editor and double-click here to view equation.

Install Equation Editor and double-click here to view equation.

There are, therefore, two errors.

Install Equation Editor and double-click here to view equation.

and a part of the difference between Y and the mean that we consider *explained* by Model 2p. This is the difference I have called this the *model* difference below.

```
. gen ec = Y-yhatc
. predict e2p, residual

. gen model = yhat2p-yhatc
```

Again, I can use `predict` only after a `regress` command; here, by using the option `residual`, I am asking STATA to produce the residual, or *error*, terms rather than the predicted values of Y

```
. list ec e2p model
      ec      e2p      model
1. -2168.818 -4317.315  2148.497
2. -1957.818 -4013.855  2056.037
3. -2080.818 -1967.344 -113.4736
4. -2318.818  452.8842 -2771.702
5. -2174.818  15.70623 -2190.524
6. -1952.818 -733.1251 -1219.693
7. -1721.818 -1499.374 -222.4446
8. -2141.818  359.1081 -2500.926
9. -1967.818  54.29672 -2022.115
10. -2202.818  634.9271 -2837.745
11. -1416.818 -900.4824 -516.3357
12. -1088.818  233.2415 -1322.06
13. -22.81812 -365.0407  342.2227
14. -1527.818  860.8351 -2388.653
15.  3117.182 -3360.43  6477.612
16. -477.8181  682.4363 -1160.254
17.  790.1819 -225.6788  1015.861
18.  1419.182 -577.4165  1996.598
19. -1482.818  1338.416 -2821.234
20. -1482.818  1506.826 -2989.644
21.  14547.18  4668.358  9878.823
22.  8314.182  7153.027  1161.155
```

In every case, $ec = e2p + model$.

Install Equation Editor and double-click here to view equation.

Install Equation Editor and double-click here to view equation.

Install Equation Editor and double-click here to view equation.

The sum of squares of each of these is one of the SS given in the regression output.

We can calculate each of these using egen:

```
. egen ssec = sum(ec*ec)
. egen sse2p = sum(e2p^2)
. egen ssmodel = sum(model^2)
```

We can also calculate R^2 as the Model SS divided by the Total SS

```
. gen R2 = ssmodel/ssec
. list ssec sse2p ssmodel R2 in 1/5
      ssec      sse2p      ssmodel      R2
1.  3.46e+08  1.33e+08  2.13e+08  .615816
2.  3.46e+08  1.33e+08  2.13e+08  .615816
3.  3.46e+08  1.33e+08  2.13e+08  .615816
4.  3.46e+08  1.33e+08  2.13e+08  .615816
5.  3.46e+08  1.33e+08  2.13e+08  .615816
```

Please check to make sure you see that these sums of squares correspond to those in the regression output. Also, here I do the calculation that shows that the error terms can be found either by using the `predict` command in STATA or by generating the values ourselves:

```
. predict e2p, resid
. gen e2pm = Y- yhat2pm
. list e2p e2pm in 1/5
```

	e2p	e2pm
1.	-4317.315	-4317.315
2.	-4013.855	-4013.855
3.	-1967.344	-1967.344
4.	452.8842	452.8842
5.	15.70623	15.70624

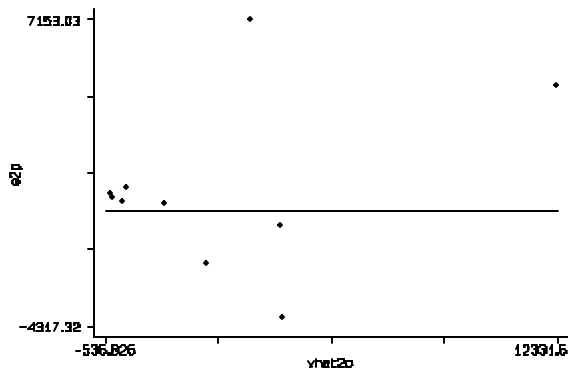
This was the command used earlier to calculate e2p
We calculate the error here

Goodness of fit: The important point here: R^2 is a measure of how much better Model 2p fits our data than does Model C. The measure is based on the reduction in the sum of the squared errors in prediction of Y when we move from a model with only a constant to one that has 2 parameters and assumes a straight line relationship between the two variables Y and X.

Is R^2 significantly different from 0? One test is that the coefficient b_1 is significantly different from 0, as we saw from the regression output. This is equivalent to a test that the reduction in the SS is significant --- IF there is only one predictor variable. Otherwise we use the F test, which will be discussed in handout 2.

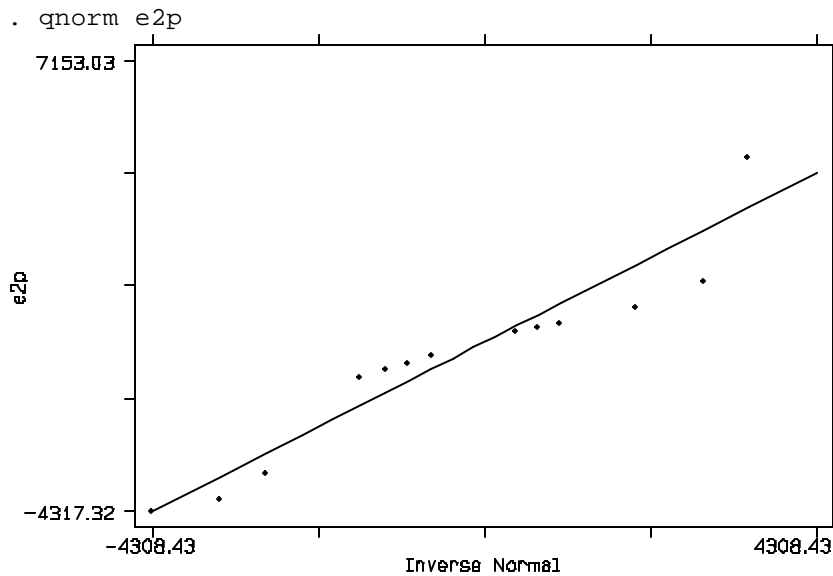
R^2 , in this case, is .62, so that the variation of Y about the line is 62% less than the variation of Y about its sample mean.

But there are other aspects of goodness of fit. We can look at the residuals, the e_{2p} , themselves. One assumption of regression analysis is that the errors are *homoscedastic* -- the distribution should be the same no matter what the values of X - or of the predicted Y. That is, if the model fits well, then the estimated errors, the e_{2p} 's, should not show a pattern. We, therefore, frequently look at the so-called residual plot, that plots e's against the predicted Y, i.e. the points (e_i, \hat{Y}_i) . Usually the line $e=0$ is added to this type of plot to help see whether there is a pattern - which, ideally, should not be present.



```
. graph e2p yhat2p, yline(0)
```

Another plot that is helpful is the qnorm plot of the residuals, which helps us judge whether or not they are normally distributed.



In this case, there seem to be both larger than expected residuals when \hat{Y} is large, and some evidence of non-normality.

"FIXING" NON-NORMALITY OF THE RESIDUALS

We use the scatterplots of Y against X and residual plots to help us determine, in the same way as for a single variable, whether a transformation will make our errors more normal.

Find transformations that normalize each variable -- here I go back to the original variable names and show part of the output of ladder.

```
. ladder area
Transformation      formula      Chi-sq(2)      P(Chi-sq)
-----
log                 log(area)          1.23          0.540

. ladder pairs
Transformation      formula      Chi-sq(2)      P(Chi-sq)
-----
log                 log(pairs)         1.85          0.397
```

```
. gen logpair = log(pairs)
. gen logarea = log(area)
. regress logpair logarea
```

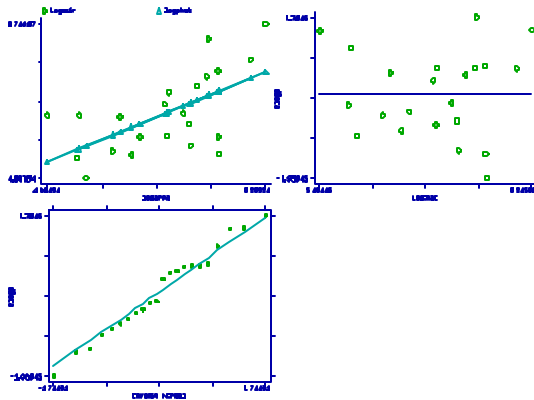
Source	SS	df	MS	Number of obs =	22
Model	12.4906569	1	12.4906569	F(1, 20) =	11.45
Residual	21.8082911	20	1.09041455	Prob > F =	0.0029
				R-square =	0.3642
				Adj R-square =	0.3324
Total	34.2989479	21	1.63328323	Root MSE =	1.0442

logpair	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
logarea	.6783232	.2004194	3.385	0.003	.2602555 1.096391
_cons	2.627175	1.293557	2.031	0.056	-.0711384 5.325488

```
. * find predicted logpairs, error, standard error for predicting an area's
logpair and standard error for predicting the mean for a particular area
. predict logphat
. predict elogp, resid
```

Look at the same diagnostic graphs as before:

```
. graph logpair logphat logarea, connect(.1)
. graph elogp logphat
. qnorm elogp
```



The transformations, although they reduce R^2 , seem to have reduced the deviations from normality with which we were concerned.

CONFIDENCE INTERVALS FOR THE PREDICTED VALUE

We can consider prediction of

1. the mean value of Y for all cases with a specific value of X and
2. the value of Y for an individual with a specific value of X .

The model says the mean lies on the regression line, while an individual has an expected value on the line *and* an error that represents that case's deviation from the expected value.

1 The mean Y for all cases with a specific value of X

Install Equation Editor and double-click here to view equation.

Install Equation Editor and double-click here to view equation.

Recall that

if y and b are variables that are not correlated. Statisticians can prove that the estimates of the mean of Y and of b

Install Equation Editor and double-click here to view equation.

are not correlated. Therefore

Install Equation Editor and double-click here to view equation.

We know what those variances are, so that

The confidence interval for the mean Y corresponding to a specific value of X , X_i , uses, as the SE, the square root of the variance given above.

2. Confidence interval for the *individual* - person i : For a specific person, recall that

$$Y_i = E[Y_i] + \hat{a}$$

For the individual, we have to add the error term to the variance above. {Unfortunately, nearly all statistics texts

Install Equation Editor and double-click here to view equation.

use the same notation for both of these variances.}

Thus, our predictions, whether for the mean or for individuals, are "better" the closer the X value is to the mean of the X 's. We can also plot the confidence intervals.

STATA will calculate for you the standard errors (sqrt of the variance) for both the predicted mean Y and the predicted individual Y . To get confidence intervals, you have to multiply by the value needed to obtain the level of confidence you choose (e.g. 1.96 for a 95% confidence interval when the sample size is large).

```
predict newvar, stdf
predict newvar, stdp
```

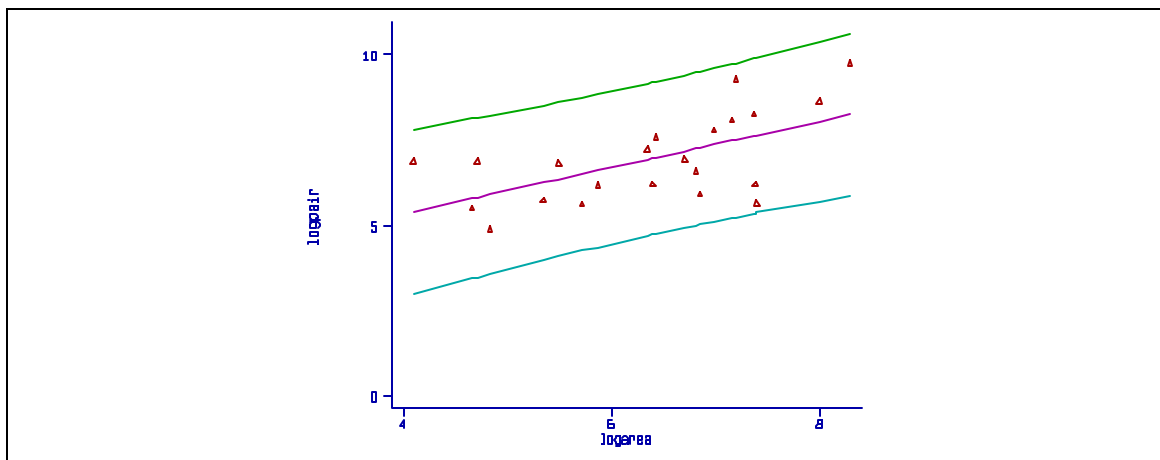
standard error of predicted individual Y
standard error of predicted mean Y

I have created these variables for the relationship of logpair to logarea and graphed the predicted line with the confidence intervals for the predicted values for an individual

```
. predict selogi, stdf                standard error of predicted individual  $Y_i$ 
. predict selogm, stdp                standard error of predicted mean  $Y_i$ 
. list selog* in 1/5
      selogi      selogm

1.  1.087462    .3035774                note that the se for the individual is always much
larger than
2.  1.086803    .3012068                for the mean
3.  1.116194    .3943033
4.  1.071993    .2423927
5.  1.07137     .2396242

. * t value for 95% confidence interval when df=20 is 2.086 (from book)
. gen loghi = logphat + 2.086* selogi
. gen loglo = logphat - 2.086* selogi
. graph loghi loglo logpair logphat logarea, connect(l1.1) symbol(iiTi) xlabel
```



ylabel

MULTIVARIATE REGRESSION

We will start by looking at a simple example, and then return to notation and theory.

The expected value of Y is now a linear function of more than one predictor variable, labeled as $X_1 \dots X_{K-1}$.
Y itself is equal to this linear function plus an error term

The model: $Y_i = \hat{\alpha}_0 + \hat{\alpha}_1 X_{i1} + \hat{\alpha}_2 X_{i2} + \hat{\alpha}_3 X_{i3} + \hat{\alpha}_4 X_{i4} + \dots + \hat{\alpha}_{K-1} X_{i,K-1} + \hat{\alpha}$

As with bivariate regression, we estimate the coefficients using the least squares criterion - finding the SET of b's that produce the smallest RSS.

```
. use h1-2
. list y x1 x2
```

	y	x1	x2
1.	190	73	20
2.	254	57	23
3.	181	67	23
4.	209	27	24
5.	302	69	25
6.	288	63	28
7.	263	70	30
8.	244	63	30
9.	290	89	31
10.	220	82	34
11.	220	60	34
12.	385	72	36
13.	274	85	37
14.	303	55	40
15.	365	75	44
16.	311	59	46
17.	354	84	46
18.	434	69	48
19.	308	75	50
20.	374	79	51
21.	405	65	52
22.	346	65	52
23.	402	79	57
24.	451	76	57
25.	395	59	60

```
. summ y x1 x2
```

Variable	Obs	Mean	Std. Dev.	Min	Max
y	25	310.72	77.82915	181	451
x1	25	68.68	12.72701	27	89
x2	25	39.12	12.24854	20	60

```
. regress y x1 x2
```

Source	SS	df	MS			
Model	102570.815	2	51285.4073	Number of obs =	25	
Residual	42806.2253	22	1945.73752	F(2, 22) =	26.36	
Total	145377.04	24	6057.37667	Prob > F =	0.0000	
				R-square =	0.7056	
				Adj R-square =	0.6788	
				Root MSE =	44.111	

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x1	.4173621	.7287761	0.573	0.573	-1.094027	1.928751
x2	5.216591	.7572445	6.889	0.000	3.646162	6.78702
_cons	77.98254	52.42964	1.487	0.151	-30.74988	186.715

The interpretation of the coefficients above is that y increases by .42 for each increase of 1 in x1 and that y increases by 5.2 for each increase of 1 in x2. The proportion of the variation in y that is *explained by* the association to both x1 and x2 is given by R-square - and is, in this case, .71.

The coefficients in multiple regression are also referred to as the *partial coefficients*. What is meant by this is that the interpretation I gave above is incomplete. It really is that, *holding x2 constant*, y increases by .42 for each increase in x1. Further, it says that this gives the relationship between y and x1 when the effect of x2 is *eliminated* from both y and x1. We will go on to this next time.

APPENDIX: LEAST SQUARES ESTIMATES

Whatever model we chose to use, the least squares approach leads to errors that average 0.

```
. sum e2p
```

Variable	Obs	Mean	Std. Dev.	Min	Max
e2p	22	.0001162	4060.955	-2318.818	14547.18

In addition, the residual sum of squares is smaller than with any other line. You can test this by generating a different line with any intercept and slope you choose, finding the predicted y, eg.

```
. gen myyhat = -600 + 4.302147*area
```

and finding

Install Equation Editor and double-click here to view equation.

This sum *will be larger* than the Residual SS from the least squares line. Similarly, were you to calculate

Install Equation Editor and double-click here to view equation.

with a constant that does not equal the sample mean, you'd find it larger than the TSS.