

MULTIVARIATE ANALYSIS continued

THE MULTIPLE REGRESSION MODEL

The first step in specifying the model that I will refer to as Model kp (k parameters) is to decide which K-1 predictor variables, X_1, X_2, \dots, X_{K-1} to include in the model as factors that we believe affect our dependent variable Y.

The next step is to decide, for each variable, whether it should enter the model in its "natural" form or whether a transformation will make its distribution more "normal" and will make the estimated errors more normal.

The model:
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \dots + \beta_{K-1} X_{i,K-1} + \epsilon_i$$

where the variables now may be transformed rather than the originals.

This model, like all the others we have considered, proposes that each observation is the sum of its *expected value* + a term representing *chance error*:

$$Y_i = E[Y_i] + \epsilon_i$$

The *expected value* is a linear function of the X's. In all cases, the *expected value* of Y, for a specific combination of X's, lies on this line and we assume the errors, ϵ_i , are normally distributed about the expected value, with mean 0 and standard deviation (or rms error) of σ .

Repeating the example (which has 25 observations) in h1-2.dta

```
. list y x1 x2 in 1/5
```

	y	x1	x2
1.	190	73	20
2.	254	57	23
3.	181	67	23
4.	209	27	24
5.	302	69	25

```
. regress y x1 x2
```

Source	SS	df	MS	Number of obs = 25		
Model	102570.815	2	51285.4073	F(2, 22) =	26.36	
Residual	42806.2253	22	1945.73752	Prob > F =	0.0000	
Total	145377.04	24	6057.37667	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

ASSUMPTIONS OF ORDINARY LEAST SQUARES (OLS) REGRESSION

There are 6 basic assumptions:

1. The X values are fixed - i.e. we usually don't treat them as a *sample*. Moreover, they are measured without error.
2. There is not *perfect multicollinearity* (i.e., there is no exact linear relationship between two or more of the independent variables).
3. Errors have zero mean - i.e. $E[\epsilon_i] = 0$

Assumptions 1 and 3 are sufficient to ensure that our estimates of the β_k are *unbiased*. They are not sufficient to prove that the OLS estimates are more *efficient* than other possible unbiased estimators.

UNBIASED: $E[b_k] = \beta_k$

EFFICIENT: b_k is more efficient than another unbiased estimator, a_k , if b_k has smaller variance.

4. Errors have the same constant variance (homoscedasticity)
 $\text{Var}[\epsilon_i] = \sigma^2$ for all i .
5. Errors are uncorrelated one with another
 $\text{Cov}[\epsilon_i, \epsilon_j] = 0$ for all $i \neq j$

When we add assumptions 3 and 4 to 1 and 2, it can be proved that

the standard errors of the coefficients are also unbiased, and
OLS is more efficient than any other linear unbiased estimator - the estimates are BLUE

If we also add the assumption that the predictors, the X variables, are not correlated with the error terms, i.e. $\text{Cov}[X_{ik}, \epsilon_i] = 0$ for all i, k , then the estimates of the coefficients and their standard errors are *consistent*, which is defined as follows:

b_k is a *consistent* estimator of β_k if the probability that they are very close approaches 1 as the sample size increases toward an infinite number.

6. Errors are normally distributed $\epsilon_i \sim N(0, \sigma^2)$ for all i

This assumption justifies our use of F tests and t tests, especially with small samples.
It also leads to the proof that OLS is more efficient than any other unbiased estimator, linear or not.

These assumptions are summarized in the statement:

We assume the linear model is correct, with normal, independent, and identically distributed (normal i.i.d.) errors.

Much of what we do is to try to check whether these assumptions are met - and what to do when they are not!
If these assumptions are violated, other types of estimators (e.g. robust estimators, ridge regression estimators) may be better - more efficient.

STANDARDIZED REGRESSION COEFFICIENTS

As in the case of bivariate regression, we can calculate regressions using all variables expressed as *standard scores*. In that case, the coefficient of the standardized variable X_k is

$$b_k^* = b_k s_k / s_y,$$

where b_k is the unstandardized coefficient and s_k is the standard deviation of X_k and s_y is the standard deviation of Y .

They are useful, within a given model, to assess whether one variable's effect on Y is *larger* than another's.

They CANNOT be compared across different samples because they depend on the sample standard deviations, which may well vary in different samples.

```
. regress y x1 x2, beta
```

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

y	Coef.	Std. Err.	t	P> t	Beta
x1	.4173621	.7287761	0.573	0.573	.0682491
x2	5.216591	.7572445	6.889	0.000	.8209727
_cons	77.98254	52.42964	1.487	0.151	.

```
. 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

The interpretation of the coefficient of x_1 now is: for each change of one s_{x_1} in x_1 (where $s_{x_1} = 12.73$), we expect a change in y of $.418 s_y$, if x_2 is held constant.

The interpretation of the β 's, whether they are in natural or standardized units, is that these are *partial effects*. A particular coefficient, say β_k , represents the relationship between X_k and Y when *all other variables are held constant* or when the effects of all the other variables in the model are removed from both X_k and Y .

The coefficients in multiple regression are also referred to as the *partial coefficients*. We can see why by removing the effect of X_1 from BOTH Y and X_2 and then looking at the relationship between the residual Y and the residual X_2 .

PARTIAL COEFFICIENTS

```
. regress y x1
```

Source	SS	df	MS	Number of obs =	25
Model	10231.7262	1	10231.7262	F(1, 23) =	1.74
Residual	135145.314	23	5875.88321	Prob > F =	0.2000
Total	145377.04	24	6057.37667	R-square =	0.0704
				Adj R-square =	0.0300
				Root MSE =	76.654

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x1	1.622343	1.229433	1.320	0.200	-.9209323	4.165618
_cons	199.2975	85.81792	2.322	0.029	21.76962	376.8254

```
. predict yx1e, resid
```

What we have left is y , *purged of the effect of the variable x1*. Do the same for $x2$:

```
. regress x2 x1
```

Source	SS	df	MS	Number of obs = 25		
Model	207.419835	1	207.419835	F(1, 23) =	1.41	
Residual	3393.22017	23	147.531312	Prob > F =	0.2478	
Total	3600.64	24	150.026667	R-square =	0.0576	
				Adj R-square =	0.0166	
				Root MSE =	12.146	

x2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x1	.2309901	.1948096	1.186	0.248	-.1720042	.6339843
_cons	23.2556	13.59827	1.710	0.101	-4.874551	51.38576

```
. predict x2x1e, resid
```

```
. regress yx1e x2x1e
```

In this step, we are looking at the relationship of y and $x2$, having eliminated the relationship of *both* to $x1$

Source	SS	df	MS	Number of obs = 25		
Model	92339.0869	1	92339.0869	F(1, 23) =	49.61	
Residual	42806.2247	23	1861.1402	Prob > F =	0.0000	
Total	135145.312	24	5631.05465	R-square =	0.6833	
				Adj R-square =	0.6695	
				Root MSE =	43.141	

yx1e	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x2x1e	5.216591	.7405997	7.044	0.000	3.684544	6.748638
_cons	-7.63e-08	8.628187	-0.000	1.000	-17.84876	17.84876

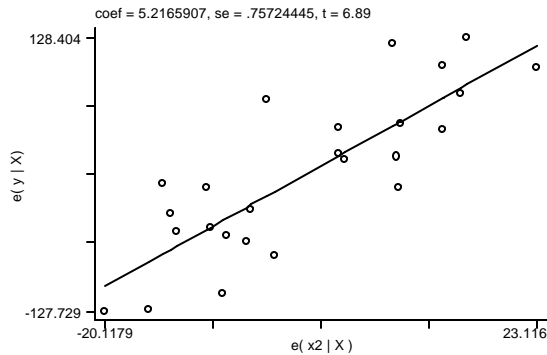
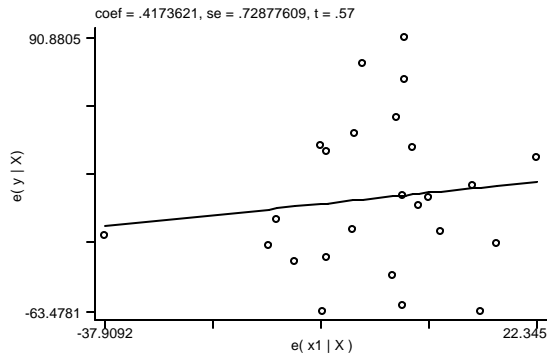
STATA will produce these plots, which are referred to as **leverage plots** if we ask for `avplots` after `regress`.

```
. regress y x1 x2
      output shown earlier
```

```
. avplots
```

ESTIMATION

Again, to find the b 's, the *estimates*, based on our sample, of the population coefficients (the β 's), we turn to the *least squares criterion*. We choose as our set of estimates the set of b 's that make the residual sum of squares (the



RSS, the sum of the squared deviations of the observed Y 's from the predicted) the *smallest*. In other words, we find the b 's that minimize

$$RSS = \sum (Y - \hat{Y})^2$$

where now \hat{Y} comes from the model with K parameters and $K-1$ predictor variables, $X_1 \dots X_{K-1}$

To find these values, we turn to matrix algebra, which permits us to write down the equations in the same form, no matter how many predictor variables ($X_1 \dots X_{K-1}$) and how many observations (n) we have. We will go over this approach later in the semester.

 t -TESTS AND CONFIDENCE INTERVALS

These are similar in multiple regression to their bivariate counterparts. The standard error of b_k is obtained from the matrix methods we will discuss next time and printed in the appropriate column of STATA output. The formula for the standard error is given as eq. 3.19 in *Regression with Graphics* and the formula for the t -statistic is in eq. 3.22. This statistic can, in particular, be used to test the null hypothesis that $\beta_k = 0$.

Standard errors for the predicted mean Y for a given set of X values and for an individual Y can also be obtained using the matrix methods. As in bivariate regression, the difference in the two is that the term in the variance for the individual has an extra σ^2 , representing variability of individuals around the expected value predicted by the linear expression. These can be obtained in STATA using the same predict commands discussed for bivariate regression.

F-TESTS FOR SETS OF COEFFICIENTS

The *t*-test described above tests one coefficient at a time. The test of the hypothesis, $\beta_k = 0$ answers the question whether adding the variable X_k to the model improves the fit. Suppose $k=3$. Then the *t*-test actually compares two models. each with residual sums of squares:

Model Kp: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \dots + \beta_{K-1} X_{i,K-1} + \epsilon_i$ RSS(K)

Model (K-1)p $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{K-1} X_{i,K-1} + \epsilon_i$ RSS(K-1)

It asks whether the smaller model with K-1 parameters (in which $\beta_3=0$) fits as well as the larger K parameter model, in which $\beta_3 \neq 0$. It asks whether the RSS from the larger model is *smaller* than the RSS from the smaller model. If $\beta_3 \neq 0$, then certainly the squared deviations of the predicted values based on the larger model from the observed Y's should be smaller. That is, we are asking if the RSS based on K-1 predictors (and K parameters), RSS(K), is smaller than the one based on K-2 predictors, RSS(K-1). Another way of looking at the question is that we are asking if the amount *explained* by adding the variable X_3 to the model is different from zero. The amount explained by adding the variable is

$$RSS(K-1) - RSS(K).$$

If this difference is small, then we say X_3 does not add to our ability to predict Y. We will return to a test of this difference in a moment.

The Larger model and the Smaller model are *nested* models - all predictor variables in the smaller model are included in the Larger. I have repeated the results for the regression of y on x1 and x2 and on x1 alone. The Residual SS is 135145 in model 2p (x1 only) and drops to 42806 in model 3p, which adds x2. The R^2 increases from a nonsignificant .07 in the smaller model to .71 when both x1 and x2 are included in the regression.

```
. 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

```
. regress y x1
```

Source	SS	df	MS	
Model	10231.7262	1	10231.7262	Number of obs = 25
Residual	135145.314	23	5875.88321	F(1, 23) = 1.74
Total	145377.04	24	6057.37667	Prob > F = 0.2000
				R-square = 0.0704
				Adj R-square = 0.0300
				Root MSE = 76.654

Frequently we want to ask whether there are *several* variables that are not necessary. For example, we could see whether the smaller model could have H fewer predictors and fit as well.

We use the difference $RSS(K-H) - RSS(K)$, which can be thought of as the additional sum of squares explained by adding the H variables to the model. If this difference is large, then we can say the variables help in prediction. If it is close to zero, we would say that the variables do not add predictive power and stick with the smaller model.

We need a "yardstick" to determine what a small difference is. That turns out to be the RSS from the larger model - the residual sum of squares when we use all the variables at our disposal. Further, it turns out that the ratio of these terms, each divided by their degrees of freedom, follows a particular distribution, called the F distribution, if the null hypothesis that the H coefficients are all zero and if the errors are normally distributed.

$$F_{n-K}^H = \frac{(\text{RSS}\{K-H\} - \text{RSS}\{K\})/H}{(\text{RSS}\{K\})/(n-K)}$$

This statistic follows an F distribution with H degrees of freedom in the numerator, n-K in the denominator. The test can be carried out in STATA after the regression for Model Kp.

```
. test x1 x2

( 1)    x1 = 0.0
( 2)    x2 = 0.0

      F( 2,    22) =    26.36
      Prob > F      =    0.0000
```

There is one particular test that is carried out routinely by STATA, namely a comparison of the particular model we fit to the one that predicts all Y values to be equal to the mean. This is equivalent to a test that $\beta_1 = \beta_2 = \dots = \beta_{K-1} = 0$. In that case, H=K-1 and RSS(1) is the total sum of squares in our regression output. In that case,

$$F_{n-K}^{K-1} = \frac{(\text{RSS}\{1\} - \text{RSS}\{K\})/(K-1)}{\text{RSS}\{K\}/(n-K)} = \frac{\text{ESS}/(K-1)}{\text{RSS}/(n-K)}$$

since RSS{K} is simply the usual residual sum of squares and the numerator is the *explained* or *model* sum of squares. The sums of squares divided by their degrees of freedom are found in the MS or mean square column of STATA output.

This, in our case, is equivalent to the test we just did on x1 and x2, since they are the only variables in our larger model.

