Lecture Notes 6
Econ B2000, MA Econometrics
Kevin R Foster, CCNY
Fall 2012

To Recap for univariate OLS:

- A zero slope for the line is saying that there is no relationship.
- A line has a simple equation, that $Y = \beta_0 + \beta_1 X$
- How can we "best" find a value of β?
- We know that the line will not always fit every point, so we need to be a bit more careful and write that our observed Y values, Y_i (i=1, ..., N), are related to the X values, X_i, as: $Y_i = \beta_0 + \beta_1 X_i + \upsilon_i$. The υ_i term is an error – it represents everything that we haven't yet taken into consideration.
- Suppose that we chose values for β_0 and β_1 that minimized the squared values of the errors. This would mean $\min_{\beta_0,\beta_1} \sum_{i=1}^{N} u_i^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{N} \left(Y_i - \beta_0 - \beta_1 X_i\right)^2$. This will generally give us unique values of β (as opposed to the eyeball method, where different people can give different answers).
- The β_0 term is the intercept and the β_1 term is the slope, $\frac{dY}{dX}$.
- These values of β are the Ordinary Least Squares (OLS) estimates. If the Greek letters denote the true (but unknown) parameters that we're trying to estimate, then denote $\hat{\beta}_0$ and $\hat{\beta}_1$ as our estimators that are based on the particular data. We denote $\hat{Y}_i$ as the predicted value of what we would guess Yi would be, given our estimates of β_0 and β_1, so that $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$.
- There are formulas that help people calculate $\hat{\beta}_0$ and $\hat{\beta}_1$ (rather than just guessing numbers); these are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2} \text{ and}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \text{ so that } \frac{1}{N}\sum_{i=1}^{N} \hat{Y}_i = \bar{Y} \text{ and } \frac{1}{N}\sum_{i=1}^{N} \hat{u}_i = 0$$

Why OLS? It has a variety of desirable properties, if the data being analyzed satisfy some very basic assumptions. Largely because of this (and also because it is quite easy to calculate) it is

widely used in many different fields. (The method of least squares was first developed for astronomy.)

- OLS requires some basic assumptions (more below)
- These assumptions are costly; what do they buy us? First, if true then the OLS estimates are distributed normally in large samples. Second, it tells us when to be careful.
- Must distinguish between dependent and independent variables (no simultaneity).
- There are formulas that you can use, for calculating the standard errors of the β estimates, however for now there's no need for you to worry about them. The computer will calculate them. (Also note that the textbook uses a more complicated formula than other texts, which covers more general cases. We'll talk about that later.)

Regression Details

Hypotheses about regression coefficients: t-stats, p-values, and confidence intervals again! Usually two-sided (rarely one-sided).

We will regularly be testing if the coefficients are significant; i.e. is there evidence in the data that the best estimate of the coefficient is different from zero? This goes back to our original "Jump into OLS" where we looked at the difference between the Hong Kong/Singapore stock returns and the US stock returns/interest rate. A zero slope is evidence against any relationship – this shows that the best guess of the value of Y does not depend on current information about the level of X. So coefficient estimates that are statistically indistinguishable from zero are not evidence that the particular X variable is useful in prediction.

A hypothesis test of some statistical estimate uses this estimator (call it $\hat{X}$) and the estimator's standard error (denote it as $se_{\hat{X}}$) to test against some null hypothesis value, X_{null}.

To make the hypothesis test, form $Z = \frac{\hat{X} - X_{null}}{se_{\hat{X}}}$, and – here is the magic! – under certain conditions this Z will have a Standard Normal distribution (or sometimes, if there are few degrees of freedom, a t-distribution; later in more advanced stats courses, some other distribution). The magic happens because if Z has a Standard Normal distribution then this allows me to measure if the estimate of X, $\hat{X}$, is very far away from X_{null}. It's generally tough to specify a common unit that allows me to say sensible things about "how big is big?" without some statistical measure. The p-value of the null hypothesis tells me, "If the null hypothesis were actually true, how likely is it that I would see this $\hat{X}$ value?" A low p-value tells me that it's very unlikely that my hypothesis could be true and yet I'd see the observed values, which is evidence against the null hypothesis.

Often the formula, $Z = \frac{\hat{X} - X_{null}}{se_{\hat{X}}}$, gets simpler when X_{null} is zero, since it is just $Z' = \frac{\hat{X} - 0}{se_{\hat{X}}} = \frac{\hat{X}}{se_{\hat{X}}}$, and this is what SPSS prints out in the regression output labeled as "t". This generally has a t-distribution (with enough degrees of freedom, a Standard Normal) so SPSS calculates the area in the tails beyond this value and labels it "Sig".

This is in Chapter 5 of Stock & Watson.

We know that the standard normal distribution has some important values in it, for example the values that are so extreme, that there is just a 5% chance that we could observe what we saw, yet the true value were actually zero. This 5% critical value is just below 2, at 1.96. So if we find a t-statistic that is bigger than 1.96 (in absolute value) then the slope would be "statistically significant"; if we find a t-statistic that is smaller than 1.96 (in absolute value) then the slope would not be "statistically significant". We can re-write these statements into values of the slope itself instead of the t-statistic.

We know from above that

$$\frac{\hat{\beta}_1 - 0}{se(\beta_1)} = \frac{\hat{\beta}_1}{se(\beta_1)} = t,$$

and we've just stated that the slope is not statistically significant if:

$$|t| < 1.96.$$

This latter statement is equivalent to:

$$-1.96 < t < 1.96$$

Which we can re-write as:

$$-1.96 < \frac{\hat{\beta}_1}{se\left(\hat{\beta}_1\right)} < 1.96$$

Which is equivalent to:

$$-1.96\left(se\left(\hat{\beta}_1\right)\right) < \hat{\beta}_1 < 1.96\left(se\left(\hat{\beta}_1\right)\right)$$

So this gives us a "Confidence Interval" – if we observe a slope within 1.96 standard errors of zero, then the slope is not statistically significant; if we observe a slope farther from zero than 1.96 standard errors, then the slope is statistically significant.

This is called a "95% Confidence Interval" because this shows the range within which the observed values would fall, 95% of the time, if the true value were zero. Different confidence intervals can be calculated with different critical values: a 90% Confidence Interval would need the critical value from the standard normal, so that 90% of the probability is within it (this is 1.64).

OLS is nothing particularly special. The Gauss-Markov Theorem tells us that OLS is **BLUE**: **B**est **L**inear **U**nbiased **E**stimator (and need to assume homoskedasticity). Sounds good, right? Among the linear unbiased estimators, OLS is "best" (defined as minimizing the squared error). But this is like being the best-looking economist – best within a very small and very particular group is not worth much! Nonlinear estimators may be good in various situations, or we might even consider biased estimators.

If X is a binary dummy variable

Sometimes the variable X is a binary variable, a dummy, D_i, equal to either one or zero (for example, female). So the model is $Y_i = \beta_0 + \beta_1 D_i + u_i$ can be expressed as

$Y_i = \begin{cases} \beta_0 + \beta_1 + u_i & if\ D_i = 1 \\ \beta_0 + u_i & if\ D_i = 0 \end{cases}$. So this is just saying that Y has mean $\beta_0 + \beta_1$ in some cases and mean β_0 in other cases. So β_1 is interpreted as the difference in mean between the two groups (those with D=1 and those with D=0). Since it is the difference, it doesn't matter which group is specified as 1 and which is 0 – this just allows measurement of the difference between them.

Other 'tricks' of time trends (& functional form)

- If the X-variable is just a linear change [for example, (1,2,3,…25) or (1985, 1986,1987,…2010)] then regressing a Y variable on this is equivalent to taking out a linear trend: the errors are the deviations from this trend.
- If the Y-variable is a log function then the regression is interpreted as explaining percent deviations (since derivative of lnY = dY/Y, the percent change). (So what would a linear trend on a logarithmic form look like?)
- If both Y and X are logs then can interpret the coefficient as the elasticity.
- examine errors to check functional form – e.g. height as a function of age works well for age < 12 but then breaks down
- plots of X vs. both Y and predicted-Y are useful, as are plots of X vs. error (note how to do these in SPSS – the dialog box for Linear Regression includes a button at the right that says "Save…", then click to save the unstandardized predicted values and unstandardized residuals).

In addition to the standard errors of the slope and intercept estimators, the regression line itself has a standard error.

A commonly overall assessment of the quality of the regression is the R^2 (displayed on the charts at the beginning automatically by SPSS). This is the fraction of the variance in Y that is explained by the model so $0 \leq R^2 \leq 1$. Bigger is usually better, although different models have different expectations (i.e. it's graded on a curve).

Statistical significance for a univariate regression is the same as overall regression significance – if the slope coefficient estimate is statistically significantly different from zero, then this is

equivalent to the statement that the overall regression explains a statistically significant part of the data variation.

- Excel calculates OLS both as regression (from Data Analysis TookPak), as just the slope and intercept coefficients (formula values), and from within a chart

Multiple Regression – more than one X variable
Regressing just one variable on another can be helpful and useful (and provides a great graphical intuition) but it doesn't get us very far.

Consider this example, using data from the March 2010 CPS. We limit ourselves to only examining people with a non-zero annual wage/salary who are working fulltime (`WSAL_VAL > 0 & HRCHECK = 2`). We look at the different wages reported by people who label themselves as white, African-American, Asian, Native American, and Hispanic. There are 62,043 whites, 9,101 African-Americans, 4476 Asians, 2149 Native Americans, and 12,401 Hispanics in the data who fulfill this condition.

The average yearly salary for whites is $50,782; for African-Americans it is $39,131; for Asians $57,541; for Native Americans $38,036; for Hispanics it is $36,678. Conventional statistical tests find that these averages are significantly different. Does this prove discrimination? No; there are many other reasons why groups of people could have different incomes such as educational level or age or a multitude of other factors. (But it is not inconsistent with a hypothesis of racism: remember the difference, when evaluating hypotheses, between 'not rejecting' or 'accepting'). We might reasonably break these numbers down further.

These groups of people are different in a variety of ways. Their average ages are different between Hispanics, averaging 38.72 years, and non-Hispanics, averaging 42.41 years. So how much of the wage difference, for Hispanics, is due to the fact that they're younger? We could do an ANOVA on this but that would omit other factors.

The populations also different in gender ratios. For whites, 57% were male; for African-Americans 46% were male; for Hispanics 59% were male. Since gender also affects income, we might think some of the wage gap could be due, not to racial discrimination, but to gender discrimination.

But then they're also different in educational attainment! Among the Hispanic workers, 30% had not finished high school; for African-Americans 8.8% had not; for whites 9% had not finished with a diploma. And 12% of whites had an advanced degree while 8.3% of African Americans and 4.2% of Hispanics had such credentials. The different fractions in educational attainment add credibility to the hypothesis that not all racial/ethnic variation means discrimination (in the labor market, at least – there could be discrimination in education so certain groups get less or worse education).

Finally they're different in what section of the country they live in, as measured by Census region.

So how can we keep all of these different factors straight?

Multiple Regression

From the standpoint of just using SPSS, there is no difference for the user between a univariate and multivariate linear regression. Again use "`Analyze\ Regression\ Linear ...`" but then add a bunch of variables to the "`Independent(s)`" box.

In formulas, model has k explanatory variables for each of $i = (1, 2, \ldots n)$ observations (must have n > k)

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_k x_{k,i} + \varepsilon_i$$

Each coefficient estimate, notated as $\hat{\beta}_j$, has standardized distribution as t with (n – k) degrees of freedom.

Each coefficient represents the amount by which the y would be expected to change, for a small change in the particular x-variable (i.e. $\beta_j = \frac{\partial y}{\partial x_j}$).

Note that you must be a bit careful specifying the variables. The CPS codes educational attainment with a bunch of numbers from 31 to 46 but these numbers have no inherent meaning. So too race, geography, industry, and occupation. If a person graduates high school then their grade coding changes from 38 to 39 but this must be coded with a dummy variable. If a person moves from New York to North Dakota then this increases their state code from 36 to 38; this is not the same change as would occur for someone moving from North Dakota to Oklahoma (40) nor is it half of the change as would occur for someone moving from New York to North Carolina (37). Each state needs a dummy variable.

A multivariate regression can control for all of the different changes to focus on each item individually. So we might model a person's wage/salary value as a function of their age, their gender, race/ethnicity (African-American, Asian, Native American, Hispanic), if they're an immigrant, six educational variables (high school diploma, some college but no degree, Associate's in vocational field, Associate's in academic field, a 4-year degree, or advanced degree), if they're married or divorced/widowed/separated, if they're a union member, and if they're a veteran. Results (from the sample above, of March 2010 fulltime workers with non-zero wage), are given by SPSS as:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.454[a]	.206	.206	46820.442

a. Predictors: (Constant), Veteran (any), African American, Education: Associate in vocational, Union member, Education: Associate in academic, Native American Indian or Alaskan or Hawaiian, Divorced or Widowed or Separated, Asian, Education: Advanced Degree, Hispanic, Female, Education: Some College but no degree, Demographics, Age, Education: 4-yr degree, Immigrant, Married, Education: High School Diploma

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.416E13	17	2.598E12	1185.074	.000[a]
	Residual	1.704E14	77751	2.192E9		
	Total	2.146E14	77768			

a. Predictors: (Constant), Veteran (any), African American, Education: Associate in vocational, Union member, Education: Associate in academic, Native American Indian or Alaskan or Hawaiian, Divorced or Widowed or Separated, Asian, Education: Advanced Degree, Hispanic, Female, Education: Some College but no degree, Demographics, Age, Education: 4-yr degree, Immigrant, Married, Education: High School Diploma

b. Dependent Variable: Total wage and salary earnings amount - Person

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10081.754	872.477		11.555	.000
	Demographics, Age	441.240	15.422	.104	28.610	.000
	Female	-17224.424	351.880	-.163	-48.950	.000
	African American	-5110.741	539.942	-.031	-9.465	.000
	Asian	309.850	819.738	.001	.378	.705
	Native American Indian or Alaskan or Hawaiian	-4359.733	1029.987	-.014	-4.233	.000

Hispanic	-3786.424	554.159	-.026	-6.833	.000
Immigrant	-3552.544	560.433	-.026	-6.339	.000
Education: High School Diploma	8753.275	676.683	.075	12.936	.000
Education: Some College but no degree	15834.431	726.533	.116	21.795	.000
Education: Associate in vocational	17391.255	976.059	.072	17.818	.000
Education: Associate in academic	21511.527	948.261	.093	22.685	.000
Education: 4-yr degree	37136.959	712.417	.293	52.128	.000
Education: Advanced Degree	64795.030	788.824	.400	82.141	.000
Married	10981.432	453.882	.102	24.194	.000
Divorced or Widowed or Separated	4210.238	606.045	.028	6.947	.000
Union member	-2828.590	1169.228	-.008	-2.419	.016
Veteran (any)	-2863.140	666.884	-.014	-4.293	.000

a. Dependent Variable: Total wage and salary earnings amount - Person

For the "Coefficients" table, the "Unstandardized coefficient B" is the estimate of $\hat{\beta}$, the "Std. Error" of the unstandardized coefficient is the standard error of that estimate, $se\left(\hat{\beta}\right)$. *(In economics we don't generally use the standardized beta, which divides the coefficient estimate by the standard error of X.)* The "t" given in the table is the t-statistic, $t = \frac{\hat{\beta}}{se\left(\hat{\beta}\right)}$ and "Sig." is its p-value – the probability, if the coefficient were actually zero, of seeing an estimate as large as the one that you got. (Go back and review if you don't remember all of the details of this.)

So see Excel sheet to show how to get predicted wages for different groups. Can then interpret the residual from the regression.

- Statistical significance of coefficient estimates is more complicated for multiple regression, we can ask whether a group of variables are jointly significant, which takes a more complicated test.

The difference between the overall regression fit and the significance of any particular estimate is that a hypothesis test of one particular coefficient tests if that parameter is zero; is

$\beta_i = 0$? This uses the t-statistic $t = \frac{\hat{\beta}}{se\left(\hat{\beta}\right)}$ and compares it to a Normal or t distribution (depending on the degrees of freedom). The test of the regression significance tests if ALL of the slope coefficients are simultaneously zero; if $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_K = 0$. The latter is much more restrictive. (See Chapter 7 of Stock & Watson.)

The predicted value of y is notated as $\hat{y}$, where $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$. Its standard error is the standard error of the regression, given by SPSS as "Standard Error of the Estimate."

The residual is $y - \hat{y} = y - \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$. The residual of, for example, a wage regression can be interpreted as the part of the wage that is not explained by the factors within the model.

Residuals are often used in analyses of productivity. Suppose I am analyzing a chain's stores to figure out which are managed best. I know that there are many reasons for variation in revenues and cost so I can get data on those: how many workers are there and their pay, the location of the store relative to traffic, the rent paid, any sales or promotions going on, etc. If I run a regression on all of those factors then I get an estimate, $\hat{y}$, of what profit would have been expected, given external factors. Then the difference represents the unexplained or residual amount of variation: some stores would have been expected to be profitable and are indeed; some are not living up to potential; some would not have been expected to do so well but something is going on so they're doing much better than expected.

Why do we always leave out a dummy variable? Multicollinearity. (See Chapter 6 of Stock & Watson.)

- OLS basic assumptions:
    - The conditional distribution of u_i given X_i has a mean of zero. This is a complicated way of saying something very basic: I have no additional information outside of the model, which would allow me to make better guesses. It can also be expressed as implying a zero correlation between X_i and u_i. We will work up to other methods that incorporate additional information.
    - The X and errors are i.i.d. This is often not precisely true; on the other hand it might be roughly right, and it gives us a place to start.
    - X and errors don't have values that are "too extreme." This is technical (about existence of fourth moments) and broadly true, whenever the X and Y data have a limit on the amount of variation, although there might be particular circumstances where it is questionable (sometimes in finance).
- So if these are true then the OLS are unbiased and consistent. So $E\left[\hat{\beta}_0\right] = \beta_0$ and $E\left[\hat{\beta}_1\right] = \beta_1$. The normal distribution, as the sample gets large, allows us to make

hypothesis tests about the values of the betas. In particular, if you look back to the "eyeball" data at the beginning, you will recall that a zero value for the slope, β_1, is important. It implies no relationship between the variables. So we will commonly test the estimated values of β against a null hypothesis that they are zero.

Heteroskedasticity-consistent errors

You can choose to use heteroskedasticity-consistent errors as in the textbook, using hcreg.sps.