5. Nonlinear Models and Limited Dependent Variables
Part 1: Nonlinear Models

LPO.7870: Research Design and Data Analysis II

Sean P. Corcoran

Last time

Statistical power: do we have enough information to detect "real" effects?

- Type I and Type II errors
- Power calculations (examples using a hypothesis test about the mean)
- Things that affect statistical power: effect size, the standard error (especially n), significance level, 1- vs. 2-sided test.
- Finding the minimum required sample size for a given power.
- Finding the minimum detectable effect size.

Practical significance and effect size

Tonight's sample datasets

We will refer to one dataset tonight (on Github):

1. `nyvoucher.dta`: pre and post-test scores from the New York Scholarship Program, a randomized experiment of private school vouchers in NYC (see M&W ch. 4). Same as Lecture 2.

Nonlinear regression models

Linear regression models

The regression models we've considered thus far have been *linear* in x:

Simple regression:

$$y = \beta_0 + \beta_1 x + u$$

Multiple regression:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

In these models, the slopes are *constant*. That is, β_1 is the predicted change in y for a unit change in x_1 (holding constant other x), at any value of x_1. The slope does not depend on the value of x_1 (or any other x).

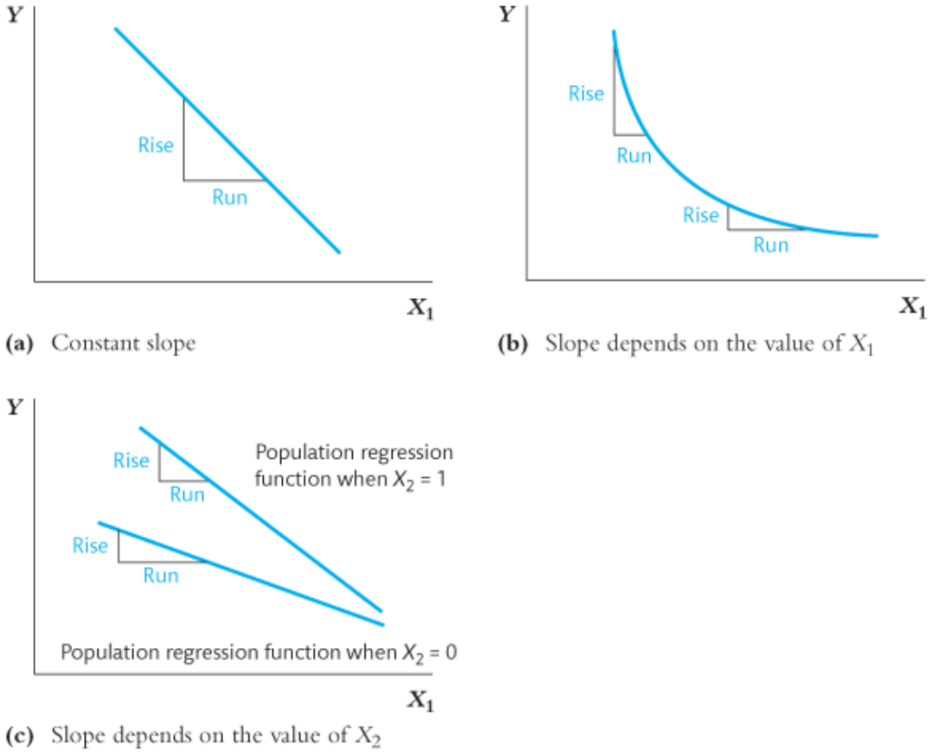
Nonlinear regression models

But there are many applications where we think the slope may depend on the value of one or more regressors. Example:

- Diminishing returns: there may be diminishing returns to the intensity of an input, program, or policy (e.g., parental time working with children).
- Heterogeneous effects: populations may differ in the extent to which they benefit from an input, program, or policy.

In general, a **nonlinear regression model** is one in which one or more slopes are not constant.

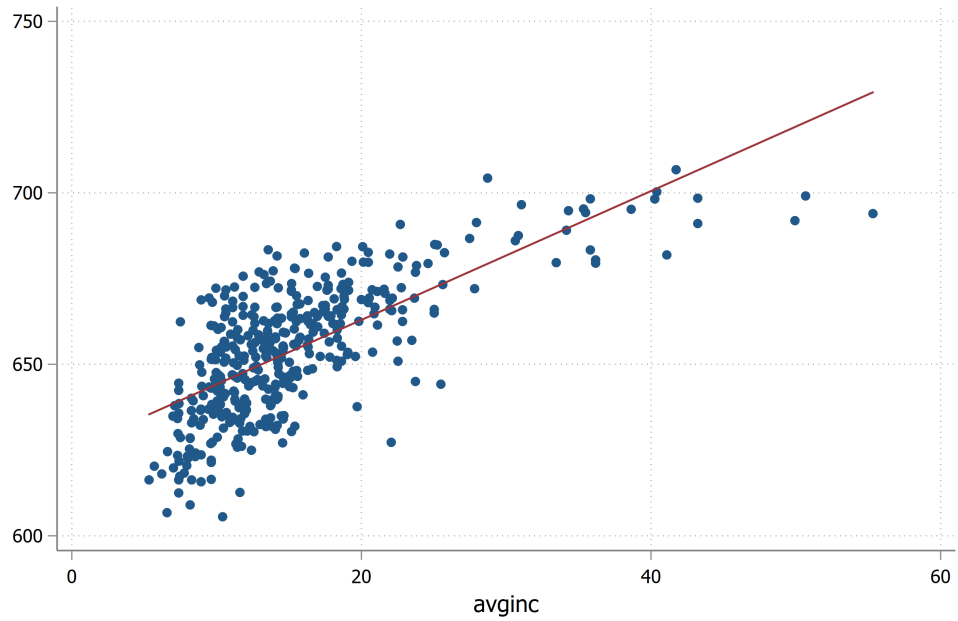
Nonlinear regression models



Source: Stock & Watson chapter 8.

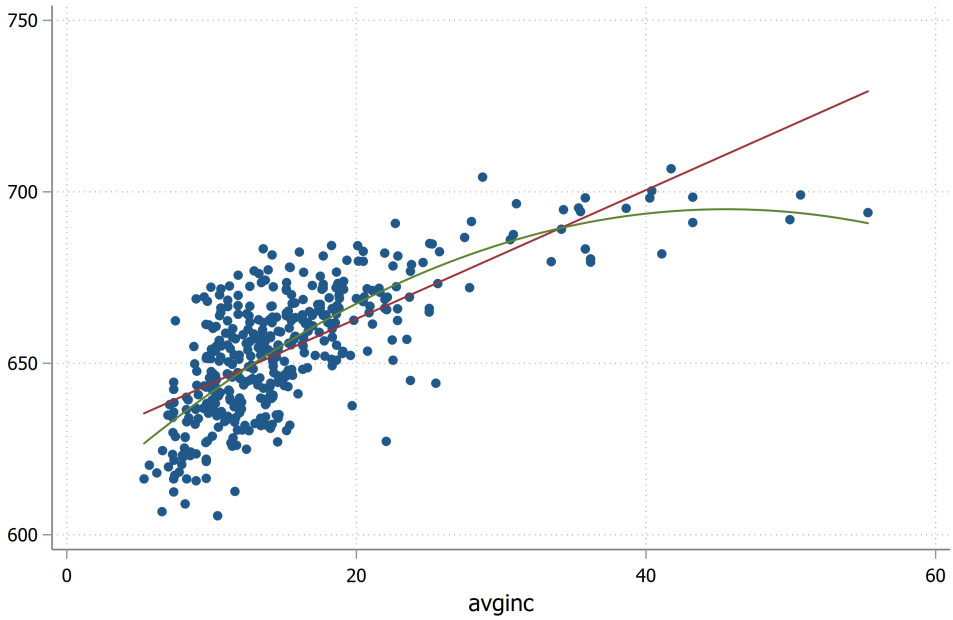
Nonlinear regression models

Consider the relationship between test scores and average annual income in *caschool.dta*:



Nonlinear regression models

A quadratic function does a better job capturing the relationship between these variables:



Regressions on polynomials

Regressions on polynomials

In general, the multiple regression model can accommodate **polynomial functions** of x. A pth order polynomial is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 ... + \beta_p x_1^p + u$$

Quadratic:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + u$$

Cubic:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + u$$

Regressions on polynomials

Some things to know about this regression specification:

- Estimation is exactly the same (OLS)! This is a standard multiple regression where the regressors happen to be powers of x.
- Interpretation is a little different: it doesn't make sense to say β_1 is the predicted change in y for a unit change in x_1 "holding other variables constant" since the other variables are powers of x_1.
- Adding higher-power terms of x gives you more *flexibility*, since a polynomial of order p can have up to $p - 1$ inflection points/bends.
- In practice, analysts don't often go higher than $p = 4$ to avoid **over-fitting**. It's rare to see more than a quadratic or cubed term.

Example: test scores and average income

In Stata: fit a regression in which test scores are a quadratic function of average income.

- Preferred: use Stata's factor variable notation. E.g., `c.avginc##c.avginc` tells Stata to include the "main effect" of *avginc* and the "interaction" with *avginc* (which is just the squared term).
- Alternatively: you can create the squared term yourself, but there are advantages to Stata's factor variable notation.

Stata's `margins` command

Stata's `margins` command is useful for working with nonlinear models. It can give you the predicted y (or the slope) at specified levels of x

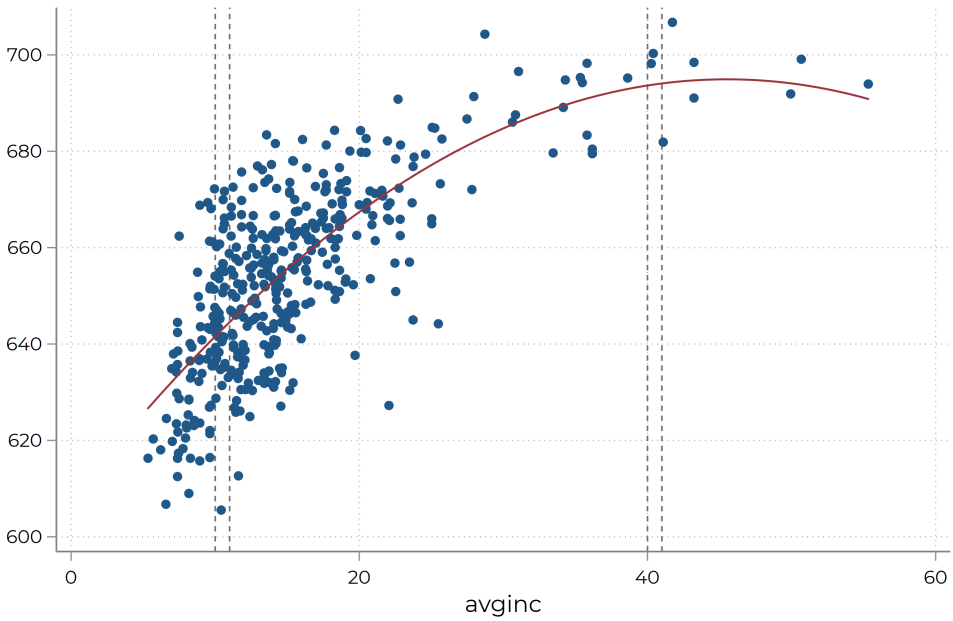
- `margins, at(avginc=(10 11))`
- `margins , at(avginc=(10)) dydx(avginc)`

The first of these gives you $\hat{y}$ at $x = 10$ and $x = 11$. The second gives you the *slope* of the regression line at $x = 10$. Note this will differ a bit from the difference between $(\hat{y}|x = 11) - (\hat{y}|x = 10)$.

In general, the slope of y with respect to x_1 in a quadratic model is: $\beta_1 + 2\beta_2 x_1$. It depends on the level of x_1.

Example: test scores and average income

Compare the slope at $x = 10$ and $x = 40$:



Regressions on polynomials

We can test for the significance of higher-order polynomial terms to decide whether or not they should be included.

```
. reg testscr c.avginc##c.avginc##c.avginc
```

Source	SS	df	MS		
Model	84939.9014	3	28313.3005	Number of obs	= 420
Residual	67169.6923	416	161.465606	F(3, 416)	= 175.35
				Prob > F	= 0.0000
				R-squared	= 0.5584
				Adj R-squared	= 0.5552
Total	152109.594	419	363.030056	Root MSE	= 12.707

testscr	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
avginc	5.018677	.8594538	5.84	0.000	3.329263	6.70809
c.avginc#c.avginc	-.0958052	.0373592	-2.56	0.011	-.1692415	-.0223688
c.avginc#c.avginc#c.avginc	.0006855	.000472	1.45	0.147	-.0002422	.0016132
_cons	600.079	5.829588	102.94	0.000	588.6199	611.5381

Using logarithms

Using logarithms

In some applications it can be useful to transform variables using a **natural logarithm**. Why?

- Logs convert changes into *percentage changes*, which are often the most relevant change to think about (e.g., earnings and work experience)
- Sometimes the conditional distribution of y is *skewed*, or there are a lot of outliers. Log transformations can reduce skewness and make the distribution look more normal. (This is because the log transformation shrinks large values more than small values).
- Sometimes a log transformation can mitigate heteroskedasticity.

A change in logs is an approximate percentage change

Suppose x increases by Δx from x_1 to x_2:

$$x_2 = x_1 + \Delta x$$

This can be written as:

$$x_2 = x_1(1 + r)$$

where $r = \Delta x/x$, the change expressed as a proportion of x. Take logs of both sides:

$$\begin{aligned} \ln(x_2) &= \ln(x_1) + \ln(1 + r) \\ \ln(x_2) - \ln(x_1) &= \ln(1 + r) \\ &\approxeq r \end{aligned}$$

It turns out that—for small values of r, the difference in logs is the approximate proportion change in x.

A change in logs is an approximate percentage change

A few additional notes:

- Logs are only defined for values > 0. Cannot use with negative or zero values.
- The approximation tends to *understate* r when r is positive, and *overstate* r when r is negative.
- The approximation gets worse the larger the change you are considering.
- To convert a logged variable back to the original units, take the exponential function: $exp(\ln x) = x$

Regressions with logs

Three possible uses of logs in regression models:

Linear-log model:

$$y = \beta_0 + \beta_1 \ln(x_1) + u$$

Log-linear model:

$$\ln(y) = \beta_0 + \beta_1 x_1 + u$$

Log-log model:

$$\ln(y) = \beta_0 + \beta_1 \ln(x_1) + u$$

Regressions with logs

Again there is no change in our estimation procedure for these regressions (OLS)! These are standard regression models—it just so happens that some of the variables are measured in log units.

Interpretation

Bearing in mind that a change of 0.01 in the *log* of a variable represents an (approximate) one-percentage point change in that variable:

- **Linear-log**: $0.01 \times \beta_1$ is the predicted change in y for an (approximate) one-percentage point change in x_1.
- **Log-linear**: β_1 is the predicted change in $\ln(y)$ for a one-unit change in x_1. So—for example—if β_1 is 0.05, we predict an (approximate) 5 percentage point change in y for a one-unit change in x_1.
- **Log-log**: $0.01 \times \beta_1$ is the predicted (approximate) percentage-point change in y for an (approximate) one-percentage point change in x_1.

Note: a 1 log-point change in x_1 is very large (more than 100%)! Use caution when interpreting coefficients with logs, and when assessing practical significance.

Interpretation

Note the practical change in our interpretation:

- **Linear-log**: every *percentage point* increase in x_1 has the same effect on y (constant slope). This is different from saying every 1-unit change in x_1 has the same effect on y.
- **Log-linear**: every one-unit change in x_1 has the same effect on y *in percentage point terms*.
- **Log-log**: every *percentage point* increase in x_1 has the same effect on y *in percentage point terms*. (In economics, known as an elasticity).

Example: test scores and average income

In Stata: fit a regression in which test scores are a function of (logged) average income.

- No factor variable notation to use here. Just create your log-transformed variable: `gen logavginc=ln(avginc)`

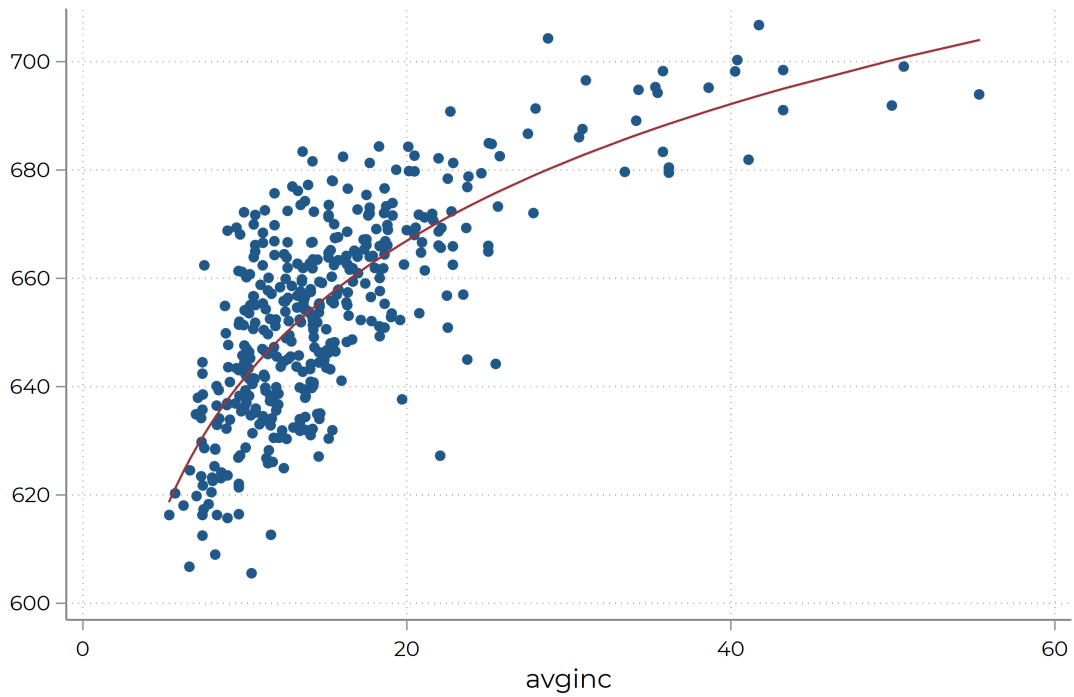
```
. ** regress test scores on logged avginc
. reg testscr logavginc
```

Source	SS	df	MS	Number of obs	=	420
				F(1, 418)	=	537.44
Model	85562.7343	1	85562.7343	Prob > F	=	0.0000
Residual	66546.8593	418	159.203013	R-squared	=	0.5625
				Adj R-squared	=	0.5615
Total	152109.594	419	363.030056	Root MSE	=	12.618

testscr	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
logavginc	36.41968	1.570976	23.18	0.000	33.33169	39.50768
_cons	557.8323	4.200348	132.81	0.000	549.5758	566.0887

Example: test scores and average income

Fitted model: plotting predicted test score against (original) average income variable:



Comparing model fit

The adjusted R^2 is useful for comparing the "predictive power" of two regression models (how much of the variation in y is explained).

You can only compare the R^2, however, for two models with the same dependent variable y. You cannot compare the R^2 for a model in which y is logged to the R^2 for a model in which y is not logged.

Interaction effects

Interaction effects

Use interaction effects if you think the slope coefficient on one variable (e.g., x_1) depends on the level of another variable (e.g., x_2). Examples:

- The effect of an intervention is larger for one group than another (i.e., group is a **moderator**)
- Inputs are complementary: the effect of one input depends on the level of another.
- Time trends differ by group (e.g., region)

Types of (two variable) interactions:

- Two binary variables
- One binary one continuous variable
- Two continuous variables

Interaction effects

General form (two variable interaction):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$$

What is the predicted change in y for a one unit change in x_1?

$$\beta_1 + \beta_3 x_2$$

And the predicted change in y for a one unit change in x_2?

$$\beta_2 + \beta_3 x_1$$

The slope on x_1 depends on the level of x_2, and vice versa.

Example: test scores and class size (1)

Does the impact of larger classes depend on the share of the student population that is learning English? Example using two binary variables interacted.

- Let *hiel* = 1 if the %EL in the district is greater than 10% (zero otherwise)
- Let *lgstr* = 1 if the average class size in the district is "large" (≥ 20) (zero otherwise)

$$testscr = \beta_0 + \beta_1 lgstr + \beta_2 hiel + \beta_3 lgstr \times hiel + u$$

In Stata: can use factor variable notation: `i.lgstr##i.hiel`

Example: test scores and class size (1)

Interpret each coefficient:

```
. reg testscr i.lgstr##i.hiel
```

Source	SS	df	MS	Number of obs	=	420
				F(3, 416)	=	58.18
Model	44956.7879	3	14985.596	Prob > F	=	0.0000
Residual	107152.806	416	257.57886	R-squared	=	0.2956
				Adj R-squared	=	0.2905
Total	152109.594	419	363.030056	Root MSE	=	16.049

testscr	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
1.lgstr	-1.907842	2.233654	-0.85	0.394	-6.298497	2.482813
1.hiel	-18.16295	2.150084	-8.45	0.000	-22.38933	-13.93656
lgstr#hiel						
1 1	-3.494335	3.22244	-1.08	0.279	-9.82863	2.83996
_cons	664.1433	1.314807	505.13	0.000	661.5588	666.7278

Be sure to pay attention to the "omitted group"!

Stata's `margins` command, revisited

Can use Stata's `margins` command again to get predicted ys at specified levels of x:

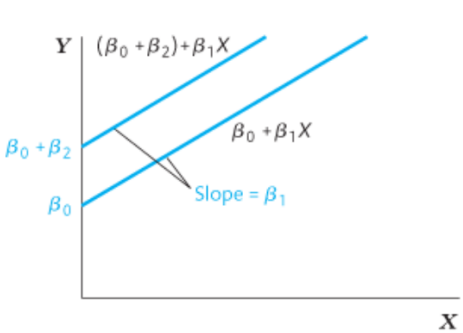
- `margins, at(lgstr=0 hiel=1)`

Notice the margins command reports standard errors and a confidence interval for the prediction!

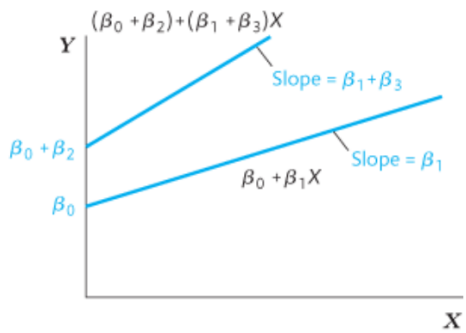
Notice also the lack of statistical significance for the *lgstr* coefficients above—this seems counterintuitive! One issue issue is that there is high collinearity between the *lgstr* variable and its interaction, which increases standard errors on both. We can do a joint F-test for the significance of both variables containing *lgstr*.

Interaction effects: one continuous and one binary variable

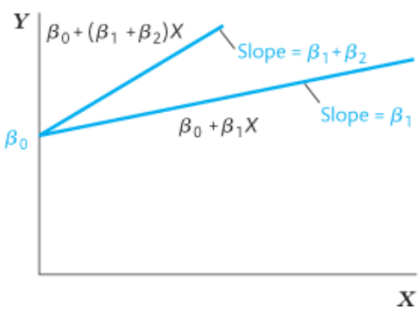
One continuous variable (x_1) and one binary variable (x_2):



(a) Different intercepts, same slope



(b) Different intercepts, different slopes



(c) Same intercept, different slopes

Example: test scores and class size (2)

Does the impact of larger classes depend on the share of the student population that is learning English? Example using one continuous and one binary variables interacted.

- Let *hiel* = 1 if the %EL in the district is greater than 10% (zero otherwise)
- Use the original (continuous) *str*

$$testscr = \beta_0 + \beta_1 str + \beta_2 hiel + \beta_3 str \times hiel + u$$

In Stata: can use factor variable notation: `c.str##i.hiel` (the *c* is for continuous)

Example: test scores and class size (2)

Interpret each coefficient:

```
. reg testscr c.str##i.hiel
```

Source	SS	df	MS		
Model	47205.8516	3	15735.2839	Number of obs	= 420
Residual	104903.742	416	252.172457	F(3, 416)	= 62.40
				Prob > F	= 0.0000
				R-squared	= 0.3103
				Adj R-squared	= 0.3054
Total	152109.594	419	363.030056	Root MSE	= 15.88

testscr	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
str	-.9684601	.539787	-1.79	0.074	-2.02951	.0925899
1.hiel	5.639141	16.71767	0.34	0.736	-27.2225	38.50078
hiel#c.str						
1	-1.276613	.8440608	-1.51	0.131	-2.935769	.3825425
_cons	682.2458	10.51094	64.91	0.000	661.5847	702.907

Be sure to pay attention to the "omitted group"!

Example: test scores and class size (3)

Does the impact of larger classes depend on the share of the student population that is learning English? Example using two continuous variables interacted.

- Use the original (continuous) *el_pct*
- Use the original (continuous) *str*

$$testscr = \beta_0 + \beta_1 str + \beta_2 el\_pct + \beta_3 str \times el\_pct + u$$

In Stata: can use factor variable notation: `c.str##c.el_pct`

Example: test scores and class size (3)

Interpret each coefficient:

```
. reg testscr c.str##c.el_pct
```

Source	SS	df	MS	Number of obs	=	420
				F(3, 416)	=	103.10
Model	64864.891	3	21621.6303	Prob > F	=	0.0000
Residual	87244.7026	416	209.722843	R-squared	=	0.4264
				Adj R-squared	=	0.4223
Total	152109.594	419	363.030056	Root MSE	=	14.482

testscr	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
str	-1.117018	.4825368	-2.31	0.021	-2.065533	-.1685039
el_pct	-.6729114	.4379849	-1.54	0.125	-1.533851	.1880281
c.str#c.el_pct	.0011618	.0219052	0.05	0.958	-.0418969	.0442204
_cons	686.3385	9.402605	72.99	0.000	667.856	704.8211

Be sure to pay attention to the "omitted group"!

Comparing model specifications

See handout (Table 8.3 from Stock & Watson)

- col (1): control for %EL, %FRPL
- col (2): add avg district income
- col (3): drop the above, add binary *hiel* and *hiel***str* interaction
- col (4): same but add back income variables
- col (5): cubic in *str*, *hiel*
- col (6): cubic in *str*, *hiel***str* (cubic)

Next time

- Midterm!
- Spring break!
- Continue Lecture 5: limited dependent variables (chapter 11)