Lecture 12

Annoucements

- PS3 due on 03/09/2020
- Quiz

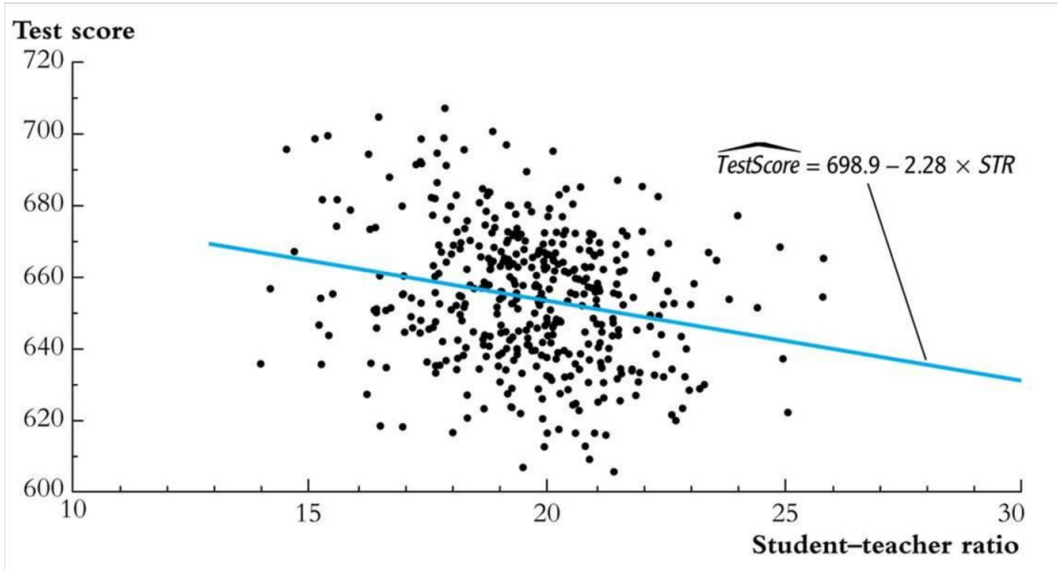
Outline

- Non-linear Models
- Linear Probability model (Dependent Variable is Binary)
- Critique of Multiple Regression
  - How to check threats to internal validity and external validity of multiple regression models

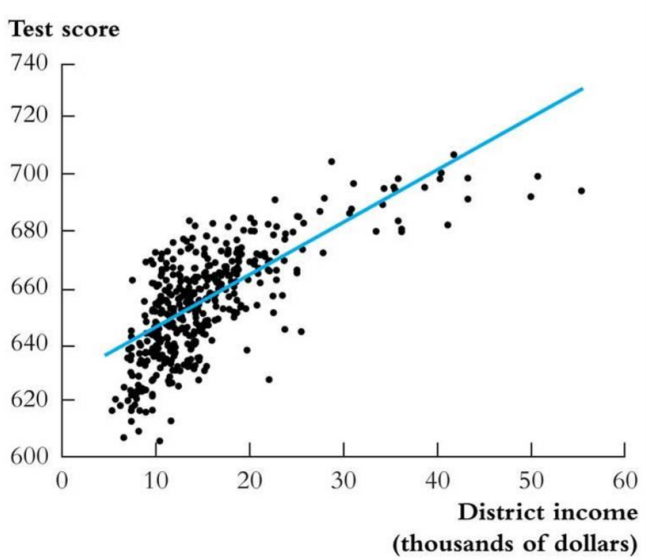
Nonlinear regression functions

- Everything so far has been linear in the *X*'s
- But the linear approximation is not always a good one
- The multiple regression framework can be extended to handle regression functions that are nonlinear in one or more *X*.

The *TestScore* – *STR* relation looks linear (maybe)…



But the *TestScore* – *Income* relation looks nonlinear...



Example: the *TestScore* – *Income* relation

- $Income_i$ = average district income in the i^{th} district (thousands of dollars per capita)

- Quadratic specification:
- $TestScore_i = \beta_0 + \beta_1 Income_i + \beta_2 (Income_i)^2 + u_i$

- Cubic specification:
- $TestScore_i = \beta_0 + \beta_1 Income_i + \beta_2 (Income_i)^2 + \beta_3 (Income_i)^3 + u_i$

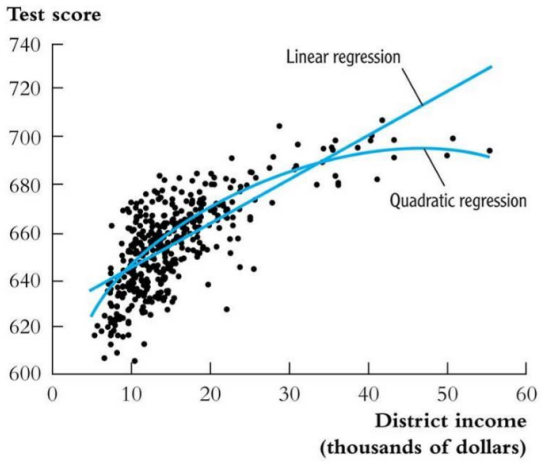
Estimation of the quadratic specification in STATA

```
generate avginc2 = avginc*avginc;       Create a new regressor
reg testscr avginc avginc2, r;

Regression with robust standard errors                Number of obs =     420
                                                      F(  2,   417) =  428.52
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.5562
                                                      Root MSE      =  12.724

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avginc |   3.850995   .2680941    14.36   0.000      3.32401    4.377979
     avginc2 |  -.0423085   .0047803    -8.85   0.000     -.051705   -.0329119
       _cons |   607.3017   2.901754   209.29   0.000     601.5978    613.0056
------------------------------------------------------------------------------
```

Test the null hypothesis of linearity against the alternative that the regression function is a quadratic….

Interpreting the estimated regression function:

- $\widehat{TestScore}_i$ = 607.3+ 3.85$Income_i$ + -0.0423$(Income_i)^2$



Interpreting the estimated regression function, ctd:

- (b) Compute "effects" for different values of *X*

- Predicted change in *TestScore* for a change in income from $5,000 per capita to $6,000 per capita:
- $\Delta Testscore$ =607.3 + 3.85*6 – 0.0423*6^2

  – (607.3 + 3.85*5 – 0.0423*5^2) = 3.4

- Predicted "effects" for different values of *X*:

Change in Income ($1000 per capita)	Δ testscore
from 5 to 6	3.4
from 25 to 26	1.7
from 45 to 46	0.0

- The "effect" of a change in income is greater at low than high income levels (perhaps, a declining marginal benefit of an increase in school budgets?)

Estimation of a cubic specification in STATA

```
gen avginc3 = avginc*avginc2;              Create the cubic regressor
reg testscr avginc avginc2 avginc3, r;

Regression with robust standard errors                Number of obs =     420
                                                      F(  3,   416) =  270.18
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.5584
                                                      Root MSE      =  12.707

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avginc |   5.018677   .7073505     7.10   0.000     3.628251    6.409104
     avginc2 |  -.0958052   .0289537    -3.31   0.001    -.1527191   -.0388913
     avginc3 |   .0006855   .0003471     1.98   0.049     3.27e-06    .0013677
       _cons |    600.079   5.102062   117.61   0.000     590.0499     610.108
------------------------------------------------------------------------------
```

- Testing the null hypothesis of linearity, against the alternative that the population regression is quadratic and/or cubic, that is, it is a polynomial of degree up to 3:

- H_0: pop'n coefficients on *Income*^{2} and *Income*^{3} = 0
- H_1: at least one of these coefficients is nonzero.

```
test avginc2 avginc3;  Execute the test command after running the regression

 ( 1)  avginc2 = 0.0
 ( 2)  avginc3 = 0.0

       F(  2,   416) =   37.69
       Prob > F =    0.0000
```

- The hypothesis that the population regression is linear is rejected at the 1% significance level against the alternative that it is a polynomial of degree up to 3.

TABLE 8.3 Nonlinear Regression Models of Test Scores

Dependent variable: average test score in district; 420 observations.

Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Student–teacher ratio (STR)	−1.00** (0.27)	−0.73** (0.26)	−0.97 (0.59)	−0.53 (0.34)	64.33** (24.86)	83.70** (28.50)	65.29** (25.26)
STR^2					−3.42** (1.25)	−4.38** (1.44)	−3.47** (1.27)
STR^3					0.059** (0.021)	0.075** (0.024)	0.060** (0.021)
% English learners	−0.122** (0.033)	−0.176** (0.034)					−0.166** (0.034)
% English learners ≥ 10%? (Binary, $HiEL$)			5.64 (19.51)	5.50 (9.80)	−5.47** (1.03)	816.1* (327.7)	
$HiEL \times STR$			−1.28 (0.97)	−0.58 (0.50)		−123.3* (50.2)	
$HiEL \times STR^2$						6.12* (2.54)	
$HiEL \times STR^3$						−0.101* (0.043)	
% Eligible for subsidized lunch	−0.547** (0.024)	−0.398** (0.033)		−0.411** (0.029)	−0.420** (0.029)	−0.418** (0.029)	−0.402** (0.033)
Average district income (logarithm)		11.57** (1.81)		12.12** (1.80)	11.75** (1.78)	11.80** (1.78)	11.51** (1.81)
Intercept	700.2** (5.6)	658.6** (8.6)	682.2** (11.9)	653.6** (9.9)	252.0 (163.6)	122.3 (185.5)	244.8 (165.7)

F-Statistics and *p*-Values on Joint Hypotheses							
(a) All *STR* variables and interactions = 0			5.64 (0.004)	5.92 (0.003)	6.31 (< 0.001)	4.96 (< 0.001)	5.91 (0.001)
(b) $STR^2, STR^3 = 0$					6.17 (< 0.001)	5.81 (0.003)	5.96 (0.003)
(c) $HiEL \times STR$, $HiEL \times STR^2$, $HiEL \times STR^3 = 0$						2.69 (0.046)	
SER	9.08	8.64	15.88	8.63	8.56	8.55	8.57
$\overline{R}^2$	0.773	0.794	0.305	0.795	0.798	0.799	0.798

These regressions were estimated using the data on K-8 school districts in California, described in Appendix 4.1. Standard errors are given in parentheses under coefficients, and *p*-values are given in parentheses under *F*-statistics. Individual coefficients are statistically significant at the *5% or **1% significance level.

- Do we have enough evidence that there's a non-linear effect of STR on average test score?

F-Statistics and *p*-Values on Joint Hypotheses							
(a) All *STR* variables and interactions = 0			5.64 (0.004)	5.92 (0.003)	6.31 (< 0.001)	4.96 (< 0.001)	5.91 (0.001)
(b) $STR^2, STR^3 = 0$					6.17 (< 0.001)	5.81 (0.003)	5.96 (0.003)
(c) $HiEL \times STR$, $HiEL \times STR^2$, $HiEL \times STR^3 = 0$						2.69 (0.046)	
SER	9.08	8.64	15.88	8.63	8.56	8.55	8.57
$\overline{R}^2$	0.773	0.794	0.305	0.795	0.798	0.799	0.798

These regressions were estimated using the data on K-8 school districts in California, described in Appendix 4.1. Standard errors are given in parentheses under coefficients, and *p*-values are given in parentheses under *F*-statistics. Individual coefficients are statistically significant at the *5% or **1% significance level.

- Do we have enough evidence that there's a non-linear effect of STR on average test score? Yes because in (5), (6), (7), we all reject the joint hypothesis that the coefficients of squared term and cubic terms are zero.

Linear Probability Model

Introduction

- So far the dependent variable (Y) has been continuous:
  - Earnings
  - Test score
  - Marriage rate

- What if Y is binary?
  - Y=get into college, or not X=parental income.
  - Y=person smokes, or not; X=cigarette tax rate, income.
  - Y=mortgage application is accepted or not; X=race, income, household characteristics, martial status…

Mortgage applications

- Example:
  - Most individuals who want to buy a house apply for a mortgage at a bank
  - Not all mortgage applications are approved.
  - What determines whether or not a mortgage application is approved or denied?

- Suppose we have the below dataset:

Variable	Description	Mean	SD
deny	= 1if mortgage application is denied	0.120	0.325
pi_ratio	anticipated monthly loan payments / monthly income	0.331	0.107
black	= 1if applicant is black, = 0 if applicant is white	0.142	0.350

Mortgage applications

- Does the payment to income ratio affect whether or not a mortgage application is denied?

Linear regression

Number of obs	2380
F(1, 2378)	37.56
Prob > F	0.0000
R-squared	0.0397
Root MSE	.31828

deny	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
pi_ratio	.6035349	.0984826	6.13	0.000	.4104144	.7966555
_cons	-.0799096	.0319666	-2.50	0.012	-.1425949	-.0172243

- How should we interpret the coefficient of pi_ratio?

Mortgage applications

- Does the payment to income ratio affect whether or not a mortgage application is denied?

Linear regression

Number of obs =	2380
F(1, 2378) =	37.56
Prob > F =	0.0000
R-squared =	0.0397
Root MSE =	.31828

deny	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
pi_ratio	.6035349	.0984826	6.13	0.000	.4104144	.7966555
_cons	-.0799096	.0319666	-2.50	0.012	-.1425949	-.0172243

- How should we interpret the coefficient of pi_ratio?
  - A change in the payment to income ratio by 1 is estimated to increase the *probability* that the mortgage application is denied by 0.60.

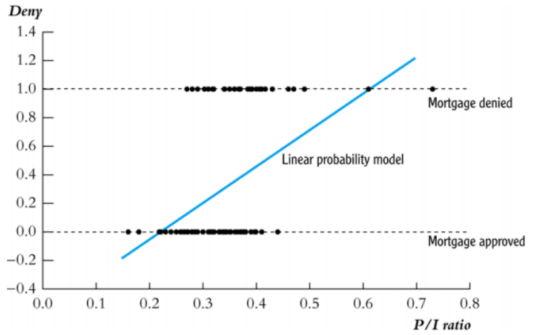
- Advantages of the linear probability model:
  - Easy to estimate
  - Coefficient estimates are easy to interpret

- Disadvantages of the linear probability model:
  - Predicted probability can be above 1 or below 0!

The linear probability model: shortcomings

- Predicted probability can be above 1 or below 0!

Example: linear probability model, HMDA data
Mortgage denial v. ratio of debt payments to income (P/I ratio) in a subset of the HMDA data set (*n* = 127)



- What is the difference in the probability of smoking between workers affected by a workplace smoking ban and workers not affected by a workplace smoking ban? Use a linear probability model to determine whether this difference is statistically significant.

```
. reg smoker smkban
```

Source	SS	df	MS		
Model	14.313036	1	14.313036	Prob > F	0.0000
Residual	1821.59406	9,998	.182195846	R-squared	0.0078
Total	1835.9071	9,999	.183609071	Root MSE	.42684

Number of obs	=	10,000				
F(1, 9998)	=	78.56				
Prob > F	=	0.0000				
R-squared	=	0.0078				
Adj R-squared	=	0.0077				
Root MSE	=	.42684				

smoker	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
smkban	-.0775583	.0087505	-8.86	0.000	-.094711	-.0604056
_cons	.2895951	.0068332	42.38	0.000	.2762006	.3029896

```
. reg smoker smkban
```

Source	SS	df	MS	Number of obs	=	10,000
				F(1, 9998)	=	78.56
Model	14.313036	1	14.313036	Prob > F	=	0.0000
Residual	1821.59406	9,998	.182195846	R-squared	=	0.0078
				Adj R-squared	=	0.0077
Total	1835.9071	9,999	.183609071	Root MSE	=	.42684

smoker	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
smkban	-.0775583	.0087505	-8.86	0.000	-.094711	-.0604056
_cons	.2895951	.0068332	42.38	0.000	.2762006	.3029896

The probability of smoking is 7.8 percentage points less if there is a smoking ban than if there is not. The t-statistic is -8.86 so the hypothesis that this difference is zero in population is rejected at the 1% significance level.

Assessing Studies Based on Multiple Regression

- Let's step back and take a broader look at regression:
- Is there a systematic way to assess (critique) regression studies? We know the strengths – but what are the pitfalls of multiple regression?
- When we put all this together, what have we learned about the effect on test scores of class size reduction?

Is there a systematic way to assess regression studies?

- Multiple regression has some key virtues:
  - It provides an estimate of the effect on *Y* of arbitrary changes ΔX.
  - It resolves the problem of omitted variable bias, if an omitted variable can be measured and included.
  - It can handle nonlinear relations (effects that vary with the *X*'s)

- Still, OLS might yield a *biased* estimator of the true *causal* effect – it might not yield "valid" inferences…

A Framework for Assessing Statistical Studies: Internal and External Validity

- ***Internal validity***: the statistical inferences about causal effects are valid for the population being studied.

- ***External validity***: the statistical inferences can be generalized from the population and setting studied to other populations and settings, where the "setting" refers to the legal, policy, and physical environment and related salient features.

Threats to External Validity of Multiple Regression Studies

- How far can we generalize class size results from California school districts?
- Differences in populations
  - California in 2005?
  - Massachusetts in 2005?
  - Mexico in 2005?
- Differences in settings
  - different legal requirements concerning special education
  - different treatment of bilingual education
  - differences in teacher characteristics

Threats to Internal Validity of Multiple Regression Analysis

- ***Internal validity***: the statistical inferences about causal effects are valid for the population being studied.
-
- *Five threats to the internal validity of regression studies:*
  1. Omitted variable bias
  2. Wrong functional form
  3. Errors-in-variables bias
  4. Sample selection bias
  5. Simultaneous causality bias

1. Omitted variable bias

- Omitted variable bias arises if an omitted variable is ***both***:

  (i) a determinant of *Y* and

  (ii) correlated with at least one included regressor.

- We first discussed omitted variable bias in regression with a single *X*, but OV bias will arise when there are multiple *X*'s as well, if the omitted variable satisfies conditions (i) and (ii) above.

Potential solutions to omitted variable bias

- If the variable can be measured, include it as an additional regressor in multiple regression

2. Wrong functional form

- Arises if the functional form is incorrect – for example, an interaction term is incorrectly omitted; then inferences on causal effects will be biased.

- **Potential solutions to functional form misspecification**
- Continuous dependent variable: use the "appropriate" nonlinear specifications in *X* (logarithms, interactions, etc.)

3. Errors-in-variables bias

- So far we have assumed that X is measured without error.
- In reality, economic data often have measurement error
  - Data entry errors in administrative data
  - Recollection errors in surveys (when did you start your current job?)
  - Ambiguous questions problems (what was your income last year?)
  - Intentionally false response problems with surveys (What is the current value of your financial assets? How often do you drink and drive?)
- In general, measurement error in a regressor results in "errors-in-variables" bias.

Potential solutions to errors-in-variables bias

- Obtain better data.
- Develop a specific model of the measurement error process.
- This is only possible if a lot is known about the nature of the measurement error – for example a subsample of the data are cross-checked using administrative records and the discrepancies are analyzed and modeled.

4. Sample selection bias

- So far we have assumed simple random sampling of the population. In some cases, simple random sampling is thwarted because the sample, in effect, "selects itself."

- *Sample selection bias* arises when a selection process:

  (i) influences the availability of data and

  (ii) that process is related to the dependent variable.

Example #1: Mutual funds

- Do some mutual funds consistently beat other funds and the market?
- Empirical strategy:
  - Sampling scheme: simple random sampling of mutual funds available to the public on a given date.
  - Is there sample selection bias?

Example #1: Mutual funds

- Do some mutual funds consistently beat other funds and the market?
- Empirical strategy:
  - Sampling scheme: simple random sampling of mutual funds available to the public on a given date.
  - Is there sample selection bias?
- Yes, the most poorly performing funds are omitted from the data set because they went out of business or were merged into other funds. This will overstate the mean return of all funds.

Example #2: returns to education

- What is the return to an additional year of education?
- Empirical strategy:
  - Sampling scheme: simple random sample of employed college grads (employed, so we have wage data)
  - Data: earnings and years of education
  - Estimator: regress ln(*earnings*) on *years_education*
  - Ignore issues of omitted variable bias and measurement error – is there sample selection bias?

Example #2: returns to education

- What is the return to an additional year of education?
- Empirical strategy:
  - Sampling scheme: simple random sample of employed college grads (employed, so we have wage data)
  - Data: earnings and years of education
  - Estimator: regress ln(*earnings*) on *years_education*
  - Ignore issues of omitted variable bias and measurement error – is there sample selection bias?
- Yes, "employed" are already selected samples.

Potential solutions to sample selection bias

- Collect the sample in a way that avoids sample selection.
  - *Mutual funds example*: change the sample population from those available at the *end* of the ten-year period, to those available at the *beginning* of the period (include failed funds)
  - *Returns to education example*: sample college graduates, not workers (include the unemployed)

5. Simultaneous causality bias

- So far we have assumed that *X* causes *Y*.
- What if *Y* causes *X*, too?

- *Example*: Class size effect
- Low *STR* results in better test scores
- But suppose districts with low test scores are given extra resources: as a result of a political process they also have low *STR*
- What does this mean for a regression of *TestScore* on *STR*?
- The estimate is likely to be biased.

Potential solutions to simultaneous causality bias

- Randomized controlled experiment. Because X_i is chosen at random by the experimenter, there is no feedback from the outcome variable to Y_i.
- Develop and estimate a complete model of both directions of causality. This is the idea behind many large macro models (e.g. Federal Reserve Bank-US). *This is extremely difficult in practice*.

Review for quizzes

- Know how to check for non-linear effects
- Know how to interpret model with binary dependent variable.
- Know how to check for threats to internal validity and external validity of multiple regression models