ECON 7310 Elements of Econometrics
Week 6: Nonlinear regression functions

David Du [1]

[1] University of Queensland

Draft

Outline

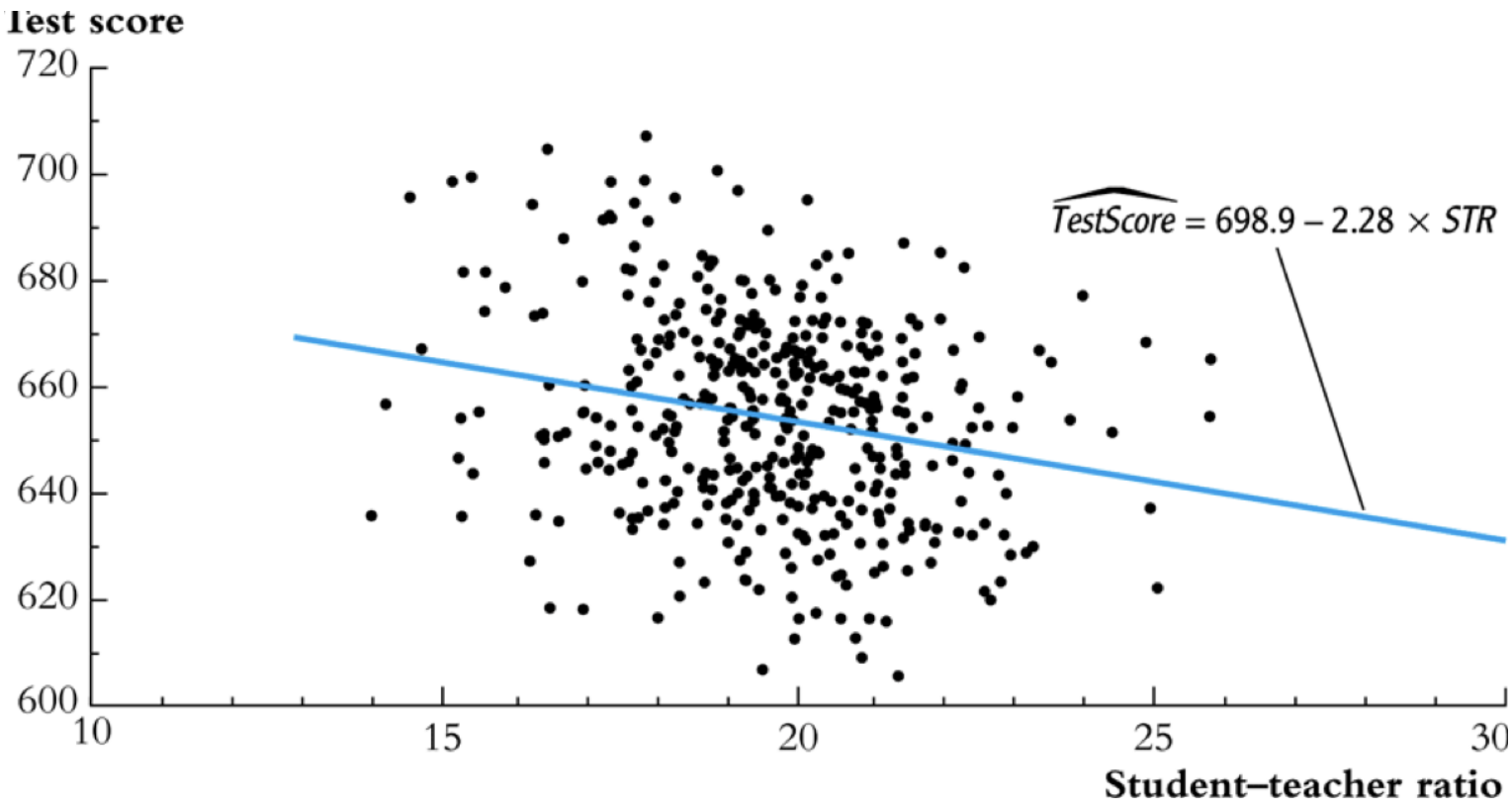
Nonlinear regression functions (SW Chapter 8)

- Nonlinear regression functions – general comments
- Nonlinear functions of one variable
- Nonlinear functions of two variables: interactions
- Application to the California Test Score data set (Read SW Section 8.4)

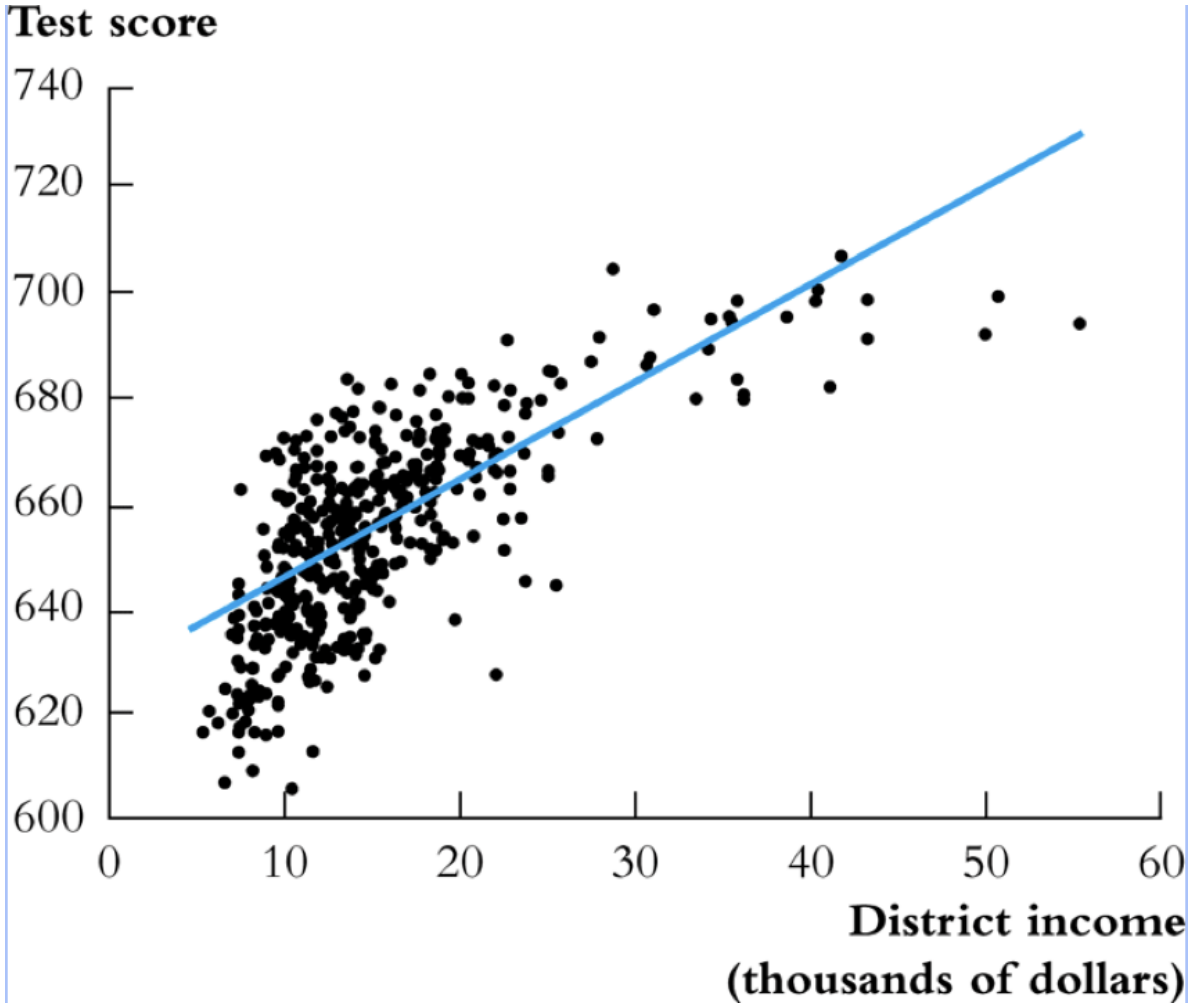
Nonlinear regression functions

- The regression functions so far have been linear in the X's
- But the linear approximation is not always a good one
- The multiple regression model can 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...



Nonlinear Regression Population Regression Functions – General Ideas (SW Section 8.1)

If a relation between Y and X is nonlinear:

- The effect on Y of a change in X often depends on the value of X – that is, the marginal effect of X is not constant
- A linear regression is mis-specified: the functional form is wrong
- The estimator of the effect on Y of X is biased: in general it is not even right on average (omitted variables bias).
- The solution is to estimate a regression function that is nonlinear in X

The general nonlinear population regression function

- Model:
$$Y_i = f(X_{i1}, X_{i2}, \ldots, X_{ik}) + u_i, \quad i = 1, \ldots, n$$
- **Assumptions:**
  - $E(u_i|X_{i1}, X_{i2}, \ldots, X_{ik}) = 0$ (same); implies that f is the conditional expectation of Y given the regressors.
  - $(X_{i1}, X_{i2}, \ldots, X_{ik}, Y_i)$ are i.i.d. (same).
  - Big outliers are rare (same idea; the precise mathematical condition depends on the specific f).
  - No perfect multicollinearity (same idea; the precise statement depends on the specific f).
- The change in Y associated with a change ΔX_1 in X_1, holding $X_2, \ldots, X_k$ constant is:
$$\Delta Y = f(X_1 + \Delta X_1, X_2, \ldots, X_k) - f(X_1, X_2, \ldots, X_k)$$

Nonlinear Functions of a Single Independent Variable (SW Section 8.2)

We will look at two approaches:

- **Polynomials in X:**
  The population regression function is approximated by a quadratic, cubic, or higher-degree polynomial
- **Logarithmic transformations:**
  Y and/or X is transformed by taking its logarithm this gives a "percentages" interpretation that makes sense in many applications

1. Polynomials in X

- Approximate the population regression function by a polynomial:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots \beta_r X_i^r + u_i$$

- This is just the linear multiple regression model – except that the regressors are powers of X!
- Estimation, hypothesis testing, etc. proceeds as in the multiple regression model using OLS
- The coefficients are difficult to interpret, but the regression function itself is interpretable.

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

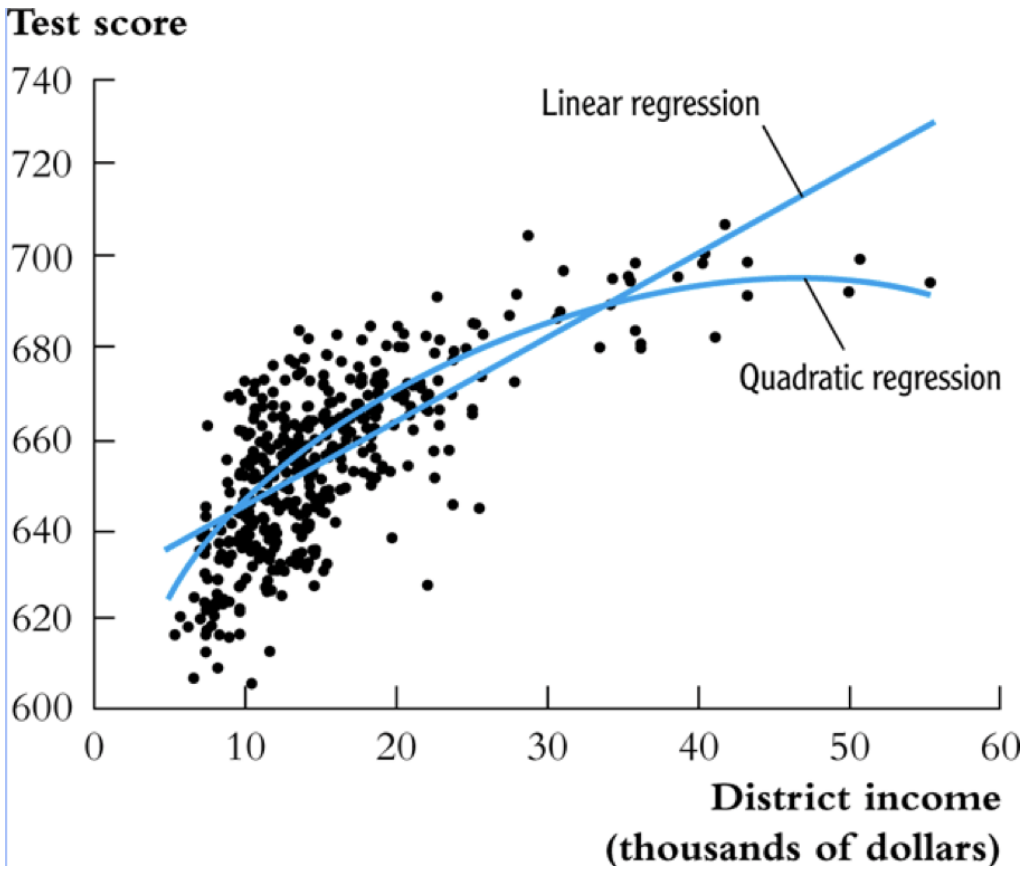
```
generate avginc2 = avginc*avginc;       Create a new regressor
reg testscr avginc avginc2, r;

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:

(a) Plot the predicted values;

$$\underset{}{TestScore_i} = \underset{(2.9)}{607.3} + \underset{(0.27)}{3.85 Income_i} - \underset{(0.0048)}{0.0423 (Income_i)^2}$$



Interpreting the estimated regression function:

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

$$\begin{array}{lll} TestScore_i = & 607.3 + & 3.85 Income_i - \\ & (2.9) & (0.27) \end{array} \quad \begin{array}{l} 0.0423 (Income_i)^2 \\ (0.0048) \end{array}$$

Predicted change in *TestScore* for a change in income from \$5,000 per capita to \$6,000 per capita:

$$\begin{aligned} \Delta TestScore =& 607.3 + 3.85 \times 6 - 0.0423 \times 6^2 \\ & - (607.3 + 3.85 \times 5 - 0.0423 \times 5^2) \\ =& 3.4 \end{aligned}$$

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

Predicted "effects" for different values of X:

Change in *Income* ($1000 per capita)	$\Delta \widehat{TestScore}$
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 income level than high income levels
- Perhaps, a declining marginal benefit of an increase in school budgets?
- Caution! What is the effect of a change from 65 to 66?
- Don't extrapolate outside the range of the data!

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

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

$$H_0 : \text{population coefficients on Income}^2 \text{ and Income}^3 = 0$$
$$H_1 : \text{at least one of these coefficients is nonzero.}$$

```
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.

Summary: polynomial regression functions

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots \beta_r X_i^r + u_i$$

- Estimation: by OLS after defining new regressors
- Coefficients have complicated interpretations
- To interpret the estimated regression function:
  - plot predicted values as a function of x
  - compute predicted $\Delta Y / \Delta X$ at different values of x
- Hypotheses concerning degree r can be tested by t- and F-tests on the appropriate (blocks of) variable(s).
- Choice of degree r
  - plot the data; t- and F-tests, check sensitivity of estimated effects; judgment.
  - Or use model selection criteria (later)

2. Logarithmic functions of Y and/or X

- $\ln(X) =$ the natural logarithm of X
- Logarithmic transforms permit modeling relations in "percentage" terms (like elasticities), rather than linearly

$$\ln(x + \Delta x) - \ln(x) \approx \frac{\Delta x}{x} \qquad \text{when } \Delta x \text{ is small}$$

- For example

$$\ln(1.01) = .00995 \approx .01$$
$$\ln(1.10) = .0953 \approx .10$$

The three log regression specifications:

Case	Population regression function
I. linear-log	$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$
II. log-linear	$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$
III. log-log	$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$

- The interpretation of the slope coefficient differs in each case.
- The interpretation is found by applying the general "before and after" rule: "figure out the change in Y for a given change in X."
- Each case has a natural interpretation (for small changes in X)
- Here, ΔX is always assumed to be small, e.g., one unit change.

Case I: Linear-log population regression function

- Compute Y "before" and "after" changing X:

$$Y = \beta_0 + \beta_1 \ln(X) \qquad (\text{"before"})$$

- New change X:

$$Y + \Delta Y = \beta_0 + \beta_1 \ln(X + \Delta X) \qquad (\text{"after"})$$

- Subtract ("after") - ("before"):

$$\Delta Y = \beta_1[\ln(X + \Delta X) - \ln(X)] \approx \beta_1 \frac{\Delta X}{X} \qquad \text{for small } \Delta x$$

- Suppose X changes by 1%, i.e., $\frac{\Delta X}{X} = 0.01$.
  In this model, a 1% change in X is associated with a change of Y of $0.01\beta_1$

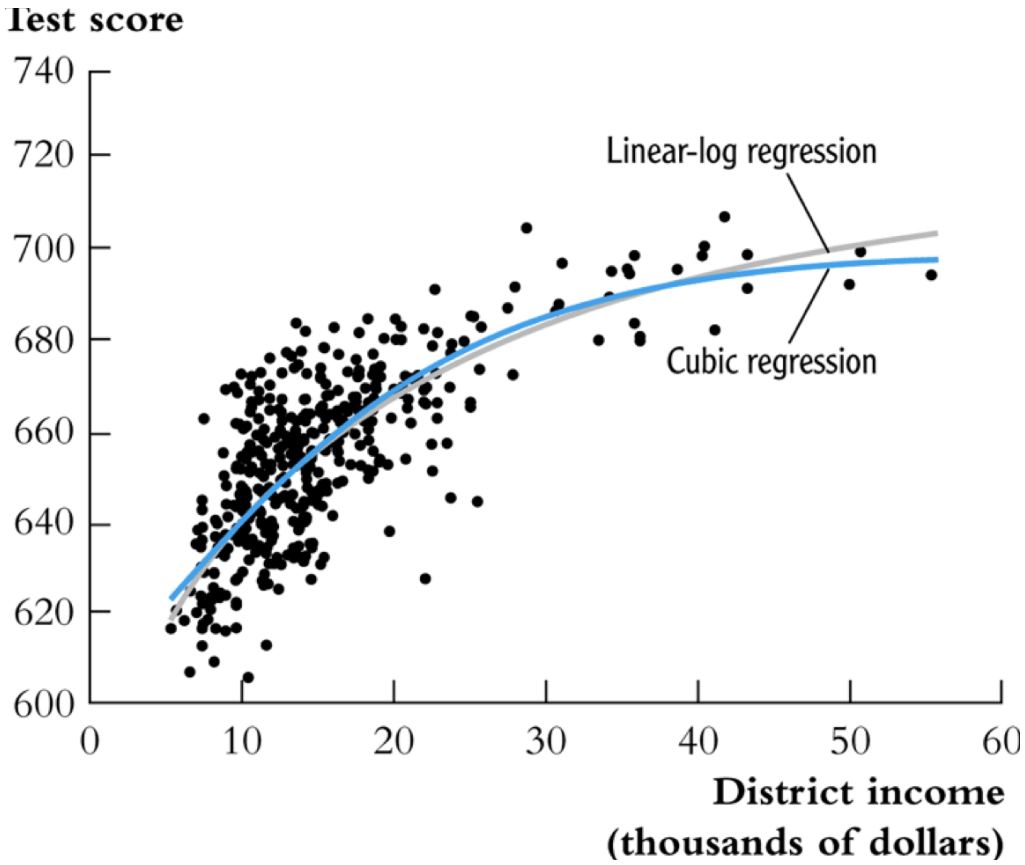
Example: *TestScore* vs. ln(*Income*)

- First defining the new regressor, ln(*Income*)
- The model is now linear in ln(*Income*), so the linear-log model can be estimated by OLS:

$$\begin{array}{lll} \widehat{TestScore}_i = & 557.8 + & 36.42\ln(Income_i) \\ & (3.8) & (1.40) \end{array}$$

- so a 1% increase in Income is associated with an increase in TestScore of 0.36 points on the test
- Standard errors, confidence intervals, R^2 – all the usual tools of regression apply here.
- How does this compare to the cubic model?

The linear-log and cubic regression functions



Case II: Log-linear population regression function

- Compute Y "before" and "after" changing X:

$$\ln(Y) = \beta_0 + \beta_1 X \qquad (\text{"before"})$$

- New change X:

$$\ln(Y + \Delta Y) = \beta_0 + \beta_1 (X + \Delta X) \qquad (\text{"after"})$$

- Subtract ("after") - ("before"):

$$\underbrace{\ln(Y + \Delta Y) - \ln(Y)}_{\approx \Delta Y / Y} = \beta_1 \Delta X \Longrightarrow \frac{\Delta Y}{Y} \approx \beta_1 \Delta X$$

- If X changes by one unit, i.e. $\Delta X = 1$, then $\frac{\Delta Y}{Y}$ changes by β_1.
- A one-unit change in X is associated with a $\beta_1 \times 100\%$ change in Y

Case III: Log-log population regression function

- Compute Y "before" and "after" changing X:

$$\ln(Y) = \beta_0 + \beta_1 \ln(X) \qquad (\text{"before"})$$

- New change X:

$$\ln(Y + \Delta Y) = \beta_0 + \beta_1 \ln(X + \Delta X) \qquad (\text{"after"})$$

- Subtract ("after") - ("before"):

$$\underbrace{\ln(Y + \Delta Y) - \ln(Y)}_{\approx \Delta Y/Y} = \beta_1 \underbrace{[\ln(X + \Delta X) - \ln(X)]}_{\approx \Delta X/X}$$

- So,

$$\beta_1 \approx \frac{\Delta Y/Y}{\Delta X/X} = \frac{\text{percentage change in } Y}{\text{percentage change in } X} = \text{Elasticity of } Y \text{ to } X$$

Example: ln(*TestScore*) vs. ln(*Income*)

- First define ln(*TestScore*) and ln(*Income*)
- The model is now a linear regression of ln(*TestScore*) against ln(*Income*), which can be estimated by OLS:

$$\begin{array}{lll} \ln(TestScore_i) = & 6.336 + & 0.0554 \ln(Income_i) \\ & (0.006) & (0.0021) \end{array}$$

- An 1% increase in Income is associated with an increase of .0554% in TestScore
- Suppose income increases from \$10,000 to \$11,000, or by 10%. Then TestScore increases by $.0554\% = .554\%$. If $TestScore = 650$, this corresponds to an increase of $.005540 = 3.6$ points.

Summary: Logarithmic transformations

- Three cases, differing in whether Y and/or X is transformed by taking logarithms.
- The regression is linear in the new variable(s) $\ln(Y)$ and/or $\ln(X)$, and the coefficients can be estimated by OLS.
- Hypothesis tests and confidence intervals are now implemented and interpreted "as usual."
- The interpretation of β_1 differs from case to case (important!).
- The choice of specification (functional form) should be guided by judgment (which interpretation makes the most sense in your application?), tests, and plotting predicted values.

Interactions Between Independent Variables SW Section 8.3

- Perhaps a class size reduction is more effective in some circumstances than in others...
- Perhaps smaller classes help more if there are many English learners, who need individual attention
- That is, $\frac{\Delta TestScore}{\Delta STR}$ might depend on *PctEL*
- More generally, $\frac{\Delta Y}{\Delta X_1}$ might depend on X_2
- How to model such "interactions" between X_1 and X_2? We first consider binary X's, then continuous X's

(a) Interactions between two binary variables

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + u$$

- D_1 and D_2 are binary (dummies)
- β_1 is the effect of changing $D_1 = 0$ to $D_1 = 1$. In this specification, this effect does not depend on the value of D_2.
- To allow the effect of changing D_1 to depend on D_2, include the interaction term $D_1 D_2$ as a regressor

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 (D_1 \times D_2) + u$$

Interpreting the coefficients

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 (D_1 \times D_2) + u$$

- General rule: compare the various cases

$$E[Y|D_1 = 0, D_2 = d_2] = \beta_0 + \beta_2 d_2 \quad (1)$$
$$E[Y|D_1 = 1, D_2 = d_2] = \beta_0 + \beta_1 + \beta_2 d_2 + \beta_3 d_2 \quad (2)$$

- (2) - (1)

$$E[Y|D_1 = 1, D_2 = d_2] - E[Y|D_1 = 0, D_2 = d_2] = \beta_1 + \beta_3 d_2$$

The effect of D_1 depends on d_2. In particular, β_3 = increment to the effect of D_1, when $D_2 = 1$

Example: *TestScore*, *STR*, English learners

- Let $HiSTR = 1$ if $STR \geq 20$ and $HiSTR = 0$, otherwise.
  Let $HiEL = 1$ if $PctEL \geq 10$ and $HiEL = 0$, otherwise.

$$\begin{array}{lllll} TestScore = & 664.1 - & 18.2 HiEL - & 1.9 HiSTR - & 3.5(HiSTR \times HiEL) \\ & (1.4) & (2.3) & (1.9) & (3.1) \end{array}$$

- Effect of *HiSTR* when $HiEL = 0$ is -1.9
  Effect of *HiSTR* when $HiEL = 1$ is $-1.9 - 3.5 = -5.4$
- Class size reduction is estimated to have a bigger effect when the percentage of English learners is large
- Each of the four different groups has different predicted *TestScore*

	Low STR	***High STR***
Low EL	664.1	662.2
High EL	645.9	640.5

(b) Interactions between continuous and binary variables

$$Y = \beta_0 + \beta_1 D + \beta_2 X + u$$

- D is binary and X is continuous
- As specified above, the effect on Y of X (holding constant D) = β_2, which does not depend on D
- To allow the effect of X to depend on D, include the "interaction term" $D \times X$ as a regressor:

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 (D \times X) + u$$

Binary-continuous interactions: the two regression lines

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 (D \times X) + u$$

- Observations with $D = 0$ (the $D = 0$ group):

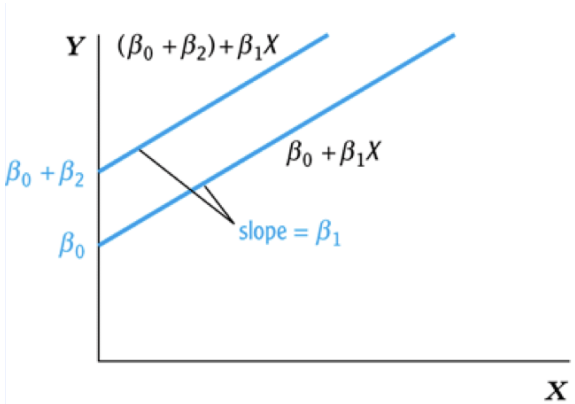
$$Y = \beta_0 + \beta_2 X + u$$

- Observations with $D = 1$ (the $D = 1$ group):

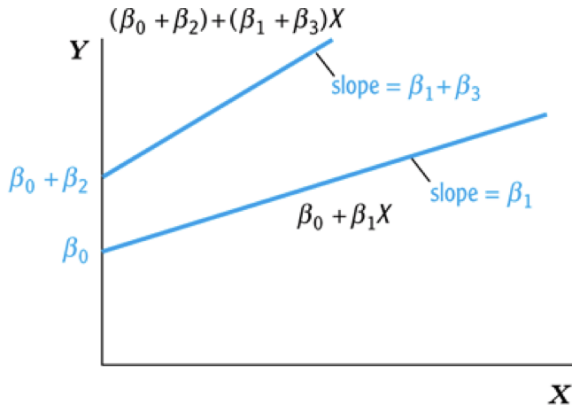
$$Y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X + u$$

- So, each group has a different intercept and a different slope, i.e., we have two regression equations.

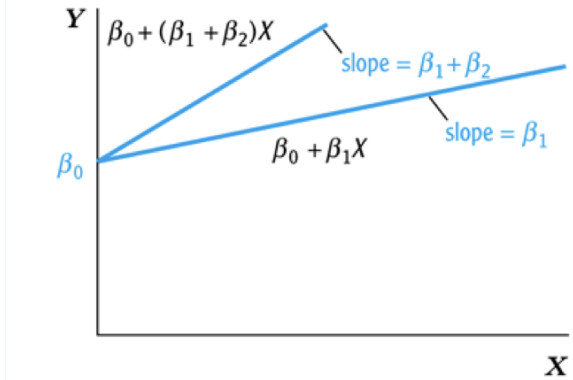
Binary-continuous interactions, ctd.



(a) Different intercepts, same slope



(b) Different intercepts, different slopes



(c) Same intercept, different slopes

Interpreting the coefficients

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 (D \times X) + u$$

- Before change:

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 (D \times X)$$

- After change:

$$Y + \Delta Y = \beta_0 + \beta_1 D + \beta_2 (X + \Delta X) + \beta_3 (D \times (X + \Delta X))$$

- Substract:

$$\Delta Y = \beta_2 \Delta X + \beta_3 D \Delta X \Longleftrightarrow \frac{\Delta Y}{\Delta X} = \beta_2 + \beta_3 D$$

- The effect of X on Y depends on D
  $\beta_3 =$ increment to the effect of X, when $D = 1$

Example: *TestScore*, *STR*, *HiEL*

$$\begin{array}{lllll} TestScore = & 682.2 - & 0.97STR + & 5.6HiEL - & 1.28(STR \times HiEL) \\ & (11.9) & (0.59) & (19.5) & (0.97) \end{array}$$

- For the group with $HiEL = 0$

$$TestScore = 682.2 - 0.97STR$$

- For the group with $HiEL = 1$

$$TestScore = 682.2 - 0.97STR + 5.6HiEL - 1.28(STR \times HiEL)$$
$$TestScore = 687.8 - 2.25STR$$

- Class size reduction is estimated to have a larger effect when the percentage of English learners is large.

Example (continued): Testing hypotheses

$$\begin{array}{lllll} TestScore = & 682.2- & 0.97STR+ & 5.6HiEL- & 1.28(STR \times HiEL) \\ & (11.9) & (0.59) & (19.5) & (0.97) \end{array}$$

The two regression lines

- have the same slope $\Leftrightarrow$ the coefficient on *STRHiEL* is 0:
  $t = -1.28/0.97 = -1.32$ (So, fail to reject)
- have the same intercept $\Leftrightarrow$ the coefficient on *HiEL* is 0:
  $t = -5.6/19.5 = -0.29$ (So, fail to reject)
- are the same $\Leftrightarrow$ the coefficients on *HiEL* and *STRHiEL* are both 0:
  $F = 89.94$, $(p < .001)$ (Reject!!)
- We reject the joint hypothesis but neither individual hypothesis. (Why?)

(c) Interactions between two continuous variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

- X_1 and X_2 are continuous. The effect of X_1 does not depend on X_2
- To allow the effect of X_1 to depend on X_2, include the "interaction term"

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u$$

- Now change X_1:

$$Y + \Delta Y = \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2 + \beta_3 (X_1 + \Delta X_1) X_2 + u$$

- Substract:

$$\Delta Y = \beta_2 \Delta X_1 + \beta_3 X_2 \Delta X_1 \iff \frac{\Delta Y}{\Delta X_1} = \beta_2 + \beta_3 X_2$$

- The effect of X_1 depends on X_2
  $\beta_3 =$ increment to the effect of X_1, from a unit change in X_2

Example: *TestScore*, *STR*, *PctEL*

$$\begin{array}{lllll} \widehat{TestScore} = & 686.3 - & 1.12 STR - & 0.67 PctEL + & 0.0012(STR \times PctEL) \\ & (11.8) & (0.59) & (0.37) & (0.019) \end{array}$$

So,

$$\frac{\Delta \widehat{TestScore}}{\Delta STR} = -1.12 + 0.0012 \times PctEL$$

For example, when $PctEL = 20\%$,

$$\frac{\Delta \widehat{TestScore}}{\Delta STR}\Big|_{PctEL=20\%} = -1.12 + 0.0012 \times 20 = -1.10$$

Example, ctd: hypothesis tests

$$\begin{array}{lllll} \widehat{TestScore} = & 686.3 - & 1.12 STR - & 0.67 PctEL + & 0.0012 (STR \times PctEL) \\ & (11.8) & (0.59) & (0.37) & (0.019) \end{array}$$

- Does population coefficient on $STRPctEL = 0$?
  $t = .0012/.019 = .06$. So, can't reject the null at 5% level
- Does population coefficient on $STR = 0$?
  $t = -1.12/0.59 = -1.90$. So, can't reject the null at 5% level
- Do the coefficients on both STR and $STRPctEL = 0$?
  $F = 3.89$ (p-value = .021). So, reject the null at 5% level(!!)
  (Why? think about imperfect multicollinearity)

Summary: Nonlinear Regression Functions

- Using functions of the independent variables such as $\ln(X)$ or $X_1 X_2$, allows recasting a large family of nonlinear regression functions as multiple regression.
- Estimation and inference proceed in the same way as in the linear multiple regression model.
- Interpretation of the coefficients is model-specific, but the general rule is to compute effects by comparing different cases, i.e.,

$$E[Y|X + \Delta X] - E[Y|X]$$

- Many nonlinear specifications are possible, so you must use judgment:
  - What nonlinear effect you want to analyze?
  - What makes sense in your application?

Assessing Studies Based on Multiple Regression: Internal and External Validity SW Section 9.1

- We will study the most common reasons that multiple regression estimates can result in biased estimates of the causal effect of interest
- In the test score application, we address these threats as best we can.
- **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.
  - Here, "setting" refers to the institutional, legal, social, and economic environment, e.g., tomatoes in the lab $\rightarrow$ tomatoes in the field?

Threats to External Validity

Potential threats arise from differences between the population and setting studied and the population and setting of interest.

- **Differences in populations:**
  - True causal effect can be different in the population studied and in the population of interest.
- **Differences in settings:**
  - a study of effect on college binge drinking of an anti-drinking campaign might not generalize to another identical group of college students if legal penalties for drinking at the two colleges are different.
- We find that *TestScore* $\uparrow$ as *STR* $\downarrow$ from CA (elementary) school districts. Can we generalize this result to
  - elementary schools in MA? (Probably Yes..)
  - high schools in CA? (Well... not sure)
  - universities in CA? (Probably, no)

Threats to Internal Validity of Regression Analysis SW section 9.2

- Studies based on regression analysis are internally valid
  1. if the estimated coefficients are unbiased and consistent and
  2. if statistical inference is valid.
- Five threats:
  - Omitted variable bias
  - Wrong functional form
  - Errors-in-variables bias
  - Sample selection bias
  - Simultaneous causality bias
- All of these imply $E[u_i|X_{i1}, \ldots, X_{ik}] \neq 0$.
  Then, OLS is biased and inconsistent.

Threats to Internal Validity: 1. Omitted variable bias

- Recall that if the omitted variable Z satisfies two conditions,
  1. Z is a determinant of Y (i.e. Z is part of u); and
  2. Z is correlated with the regressor X (i.e. $corr(Z, X) \neq 0$),

  OLS estimators are biased and inconsistent.
- **If there is a set of control variables,** include adequate control variables to address the problem of omitted variable bias.
- In practice, however, adding a variable has both costs and benefits;
  - adding an adequate variable reduces omitted variable bias.
  - adding an irrelevant variable reduces precision of the estimator

  So, there is a trade-off b/w bias and variance of the coefficient of interest

Threats to Internal Validity: 1. Omitted variable bias

Five steps to make a decision whether to add a (set of) variable(s);

1. Be specific about the coefficient(s) of interest.
2. Identify the most likely sources of important omitted variable bias, using economic theory and or other *a priori* knowledge, and set up a base specification and a list of additional candidates for control variables
   - "additional resources for kids" directly affects *TestScore* and also indirectly does so via *STR*. List "income" and "lunch support" as a control variable.
3. Augment the base specification with the additional control variables and examine
   - whether the additional coefficients are statistically significant, and/or
   - whether the estimates of the coefficient of interest substantially change

   If so, include those variables in regression.
4. Present an accurate summary of your results in tabular form.
   This "full disclosure" allows for readers to draw their own conclusions.

Threats to Internal Validity: 1. Omitted variable bias

- **If there is no control variable**, you can use either
  1. panel data approach (Chapter 10), or
  2. instrumental variables regressions (Chapter 12), or
  3. data generated from a randomized controlled experiment (Chapter 13)

Threats to Internal Validity: 2. Misspecification of Functional Form

- A misspecification bias arises if the functional form is incorrect.
  For example, an interaction term is incorrectly omitted; then inferences on causal effects will be biased.
- **Solution:**
  - Include higher order terms and/or interaction terms (Chapter 8)

Threats to Internal Validity:
3. Measurement error and errors-in-variables bias

- There are many possible sources of measurement error.
  - typos, coding errors, imprecise/wrong answers to survey questions (intentionally?), etc.
- Suppose we observe $\widetilde{X}_i = X_i + w_i$ (instead of X_i) and estimate

$$Y_i = \beta_0 + \beta_1 \widetilde{X}_i + v_i,$$

  instead of $Y_i = \beta_0 + \beta_1 X_i + u_i$ (correct one). Then, $v_i = -\beta_1 w_i + u_i$.
- As long as w_i is correlated with $\widetilde{X}_i$, we have $E[v_i|\widetilde{X}_i] \neq 0$.
- Suppose w_i is a pure random error with $E[w_i] = 0$, $V(w_i) = \sigma_w^2$, $Cov(w_i, X_i) = 0$, and $Cov(w_i, u_i) = 0$.
- **Classical measurement error model:** regress Y on $\widetilde{X}$. Then

$$\widehat{\beta}_1 \xrightarrow{p} \left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} \right) \beta_1 \text{. That is, } \widehat{\beta}_1 \text{ will be biased toward 0.}$$

- **Measurement error in Y:** Let $\widetilde{Y}_i = Y_i + w_i$. Suppose we regress $\widetilde{Y}_i$ on X_i. Then, we still have $E[\widehat{\beta}_1] = \beta_1$ and $\widehat{\beta}_1 \xrightarrow{p} \beta_1$, but $V(\widehat{\beta}_1)$ will be larger.
- **Solution:** instrumental variables regression (Chapter 12)

Threats to Internal Validity: 4. Missing Data and Sample Selection

- Data are often missing. We consider three cases:
  1. Data are missing at random
  2. Data are missing based on the value of one or more Xs
  3. Data are missing based in part on the value of Y (or u)
- Cases 1 and 2 do not introduce bias but make standard errors larger
  1. For some reason, you lost half of your sample randomly
     $\rightarrow$ Now, your sample is only smaller: so $\widehat{\beta}$ unbiased but $SE(\widehat{\beta})$ larger.
  2. Suppose we only observe districts with $STR_i > 20$. We can still unbiasedly estimate the effect of class size for districts with $STR > 20$.
- Case 3 introduces "sample selection" bias.
  3. If we estimate the wage equation only using individuals with annual wage larger than \$300k, the estimates will be clearly biased.
- **Solution:** the methods to correct the sample selection bias is beyond the scope of the course, but they are based on techniques in Section 11.3.

Threats to Internal Validity: 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*?

Threats to Internal Validity: 5. Simultaneous causality bias

- Causal effect on Y of X: $Y_i = \beta_0 + \beta_1 X_i + u_i$
  Causal effect on X of Y: $X_i = \gamma_0 + \gamma_1 Y_i + v_i$
  1. Large u_i means large Y_i, which implies large X_i (if $\gamma_1 > 0$)
  2. Thus, u_i and X_i are correlated, $E[u_i|X_i] \neq 0$.
  3. So, $\widehat{\beta}_1$ is biased and inconsistent.
- **Solutions:**
  1. Estimate instrumental variables regression; Chapter 12 (focus on $X \rightarrow Y$)
  2. Run a randomized controlled experiment; Chapter 13 (block $Y \rightarrow X$)
  3. Develop and estimate a complete model of both directions of causality, i.e., $X \rightarrow Y$ and $Y \rightarrow X$. Ex: large macro models.

Additional threats to Internal Validity: Inconsistency of $SE(\widehat{\beta})$

- Even when OLS estimator is consistent, inconsistent SEs will lead to invalid inference (hypothesis testing & confidence intervals).
- Source of inconsistency of SE:
  - **Heteroskedasticity:** SEs would be inconsistent if we do not use (heteroskedasticity) robust standard errors. In Stata, use the option `robust`.
  - **Correlation of u_i across i:** do not happen if sampling is random. But, in practice, sampling can be only partially random:
    - time series data (serial correlation)
    - panel data (individual i is observed over different time points, serial correlation)
    - some individuals are clustered geographically
    - In many cases, this problem can be fixed by using an alternative formula for SE

Internal and External Validity When the Regression is Used for Forecasting SW Section 9.3

- Forecasting and estimation of causal effects are quite different objectives.
- For forecasting,
  - $\overline{R}^2$ matters (a lot!)
  - Omitted variable bias is not a problem! Don't need consistent $\widehat{\beta}$
  - Interpreting coefficients in forecasting models is not important – the important thing is a good fit and a model you can "trust" to work in your application
  - External validity is paramount: the model estimated using historical data must hold into the (near) future
  - More on forecasting when we take up time series data

Example: Test Scores and Class Size SW Section 9.4

External Validity

- Compare results for California and Massachusetts
- Both CA and MA tests are broad measures of student knowledge and analytic skills
- Elementary schools in US are similar in organization of classroom instructions, although different states have different funding and curriculums
- So, if we find similar results from MA data, that would be evidence of external validity of the findings from CA data.

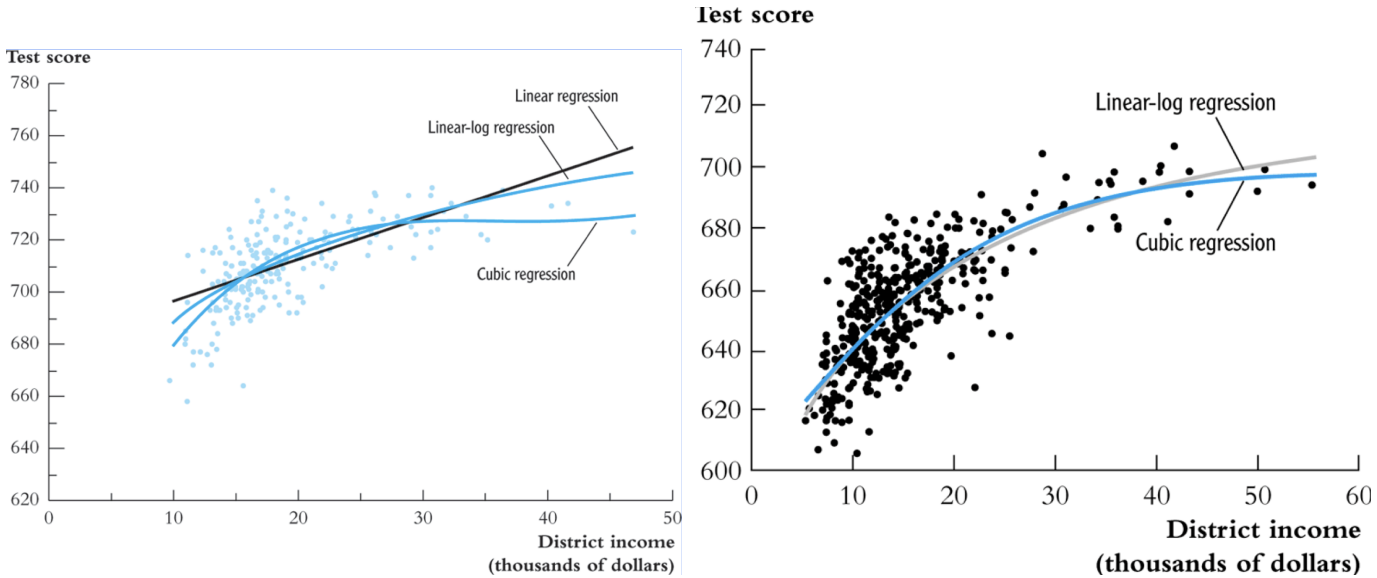
Example: Test Scores and Class Size SW Section 9.4

TABLE 9.1 Summary Statistics for California and Massachusetts Test Score Data Sets

	California		Massachusetts	
	Average	Standard Deviation	Average	Standard Deviation
Test scores	654.1	19.1	709.8	15.1
Student–teacher ratio	19.6	1.9	17.3	2.3
% English learners	15.8%	18.3%	1.1%	2.9%
% Receiving lunch subsidy	44.7%	27.1%	15.3%	15.1%
Average district income ($)	$15,317	$7226	$18,747	$5808
Number of observations	420		220	
Year	1999		1998	

- Average scores cannot be compared directly, as they are different tests
- Average *STR* higher in CA
- Average income lower but more widely spread in CA
- More English Learners and more kids with lunch support in CA

Example: Test Scores and Class Size SW Section 9.4



- (Left, Right) = (MA, CA) for *TestScore* vs *Income*
- They are similar: the relationship is steeper for low value of income
- Best functional forms differ (Cubic for MA, Linear-log for CA)

To assess the external validity of the regression analysis using CA data, run similar multiple regressions using MA data (results presented in SW Section 9.4) and compare two sets of estimation results. We see...

Example: Test Scores and Class Size SW Section 9.4

- Class size effect falls in both CA, MA data when student and district control variables are added.
- Class size effect is statistically significant in both CA, MA data.
- Estimated effect of a 2-student reduction in STR is quantitatively similar for CA, MA.
- Neither data set shows evidence of *STR* – *PctEL* interaction.
- Some evidence of STR nonlinearities in CA data, but not in MA data.
- Overall, this analysis of MA data suggests hat the CA results are externally valid

Example: Test Scores and Class Size SW Section 9.4

Internal Validity

1. Omitted variable bias:

- What causal factors are missing?
  - Student characteristics such as native ability
  - Access to outside learning opportunities
  - Other district quality measures such as teacher quality
- The regressions attempt to control for these omitted factors using control variables that are not necessarily causal but are correlated with the omitted causal variables:
  - district demographics (income, % free lunch eligible)
  - Fraction of English learners

Example: Test Scores and Class Size SW Section 9.4

Are the control variables effective? That is, after including the control variables, is the error term uncorrelated with STR?

$$E[u|X, W] = E[u|W]$$

- Answering this requires using judgment.
- There is some evidence that the control variables might be effective:
  - The STR coefficient doesn't change much when the control variables specifications change
  - The results for California and Massachusetts are similar – so if there is OV bias remaining, that OV bias would be similar in the two data sets
- What additional control variables might you want to use – and what would they be controlling for?

Example: Test Scores and Class Size SW Section 9.4

2. Wrong functional form:

- We have tried quite a few different functional forms, in both the California and Mass. data
- Nonlinear effects are modest
- Plausibly, this is not a major threat at this point.

3. Errors-in-variables bias:

- The data are administrative so it is unlikely that there are substantial reporting/typo type errors.
- STR is a district-wide measure, so students who take the test might not have experienced the measured STR for the district – a complicated type of measurement error
- Ideally we would like data on individual students, by grade level.

Example: Test Scores and Class Size SW Section 9.4

4. Sample selection bias:

- Sample is all elementary public school districts (in CA and in MA.) – there are no missing data
- No reason to think that selection is a problem.

5. Simultaneous causality bias:

- School funding equalization based on test scores could cause simultaneous causality.
- This was not in place in CA or MA. during these samples, so simultaneous causality bias is arguably not important.