Econ 423 – Lecture Notes

(These notes are slightly modified versions of lecture notes provided by Stock and Watson, 2007. They are for instructional purposes only and are not to be distributed outside of the classroom.)

Review of Linear Regression

The Linear Model with Two Regressors

Consider the case of two regressors:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,\ \ i = 1,\ldots,n$$

- Y is the *dependent variable*
- X_1, X_2 are the two *independent variables* (*regressors*)
- (Y_i, X_{1i}, X_{2i}) denote the i^{th} observation on Y, X_1, and X_2.
- β_0 = unknown population intercept
- β_1 = effect on Y of a change in X_1, holding X_2 constant
- β_2 = effect on Y of a change in X_2, holding X_1 constant
- u_i = the regression error (omitted factors)

Interpretation of coefficients in multiple regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,\ \ i = 1,\ldots,n$$

Consider changing X_1 by ΔX_1 while holding X_2 constant:

Population regression line ***before*** the change:

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

Population regression line, ***after*** the change:

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

Before: $Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2$

After: $Y + \Delta Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2$

Difference: $\Delta Y = \beta_1 \Delta X_1$

So:

$\beta_1 = \frac{\Delta Y}{\Delta X_1}$, **holding X_2 constant**

$\beta_2 = \frac{\Delta Y}{\Delta X_2}$, **holding X_1 constant**

β_0 = predicted value of Y when $X_1 = X_2 = 0$.

The OLS Estimator in Multiple Regression

With two regressors, the OLS estimator solves:

$$\min_{b_0,b_1,b_2} \sum_{i=1}^{n}[Y_i - (b_0 + b_1X_{1i} + b_2X_{2i})]^2$$

- The OLS estimator minimizes the average squared difference between the actual values of Y_i and the prediction (predicted value) based on the estimated line.
- This minimization problem is solved using calculus
- **This yields the OLS estimators of β_0 and β_1.**

Empirical Example: Class size and educational output

- Policy question: What is the effect on test scores (or some other outcome measure) of reducing class size by one student per class? by 8 students/class?
- We must use data to find out (is there any way to answer this *without* data?)

The California Test Score Data Set

All K-6 and K-8 California school districts (*n* = 420)

Variables:

- 5th grade test scores (Stanford-9 achievement test, combined math and reading), district average
- Student-teacher ratio (STR) = no. of students in the district divided by no. full-time equivalent teachers

Initial look at the data:

(*You should already know how to interpret this table*)

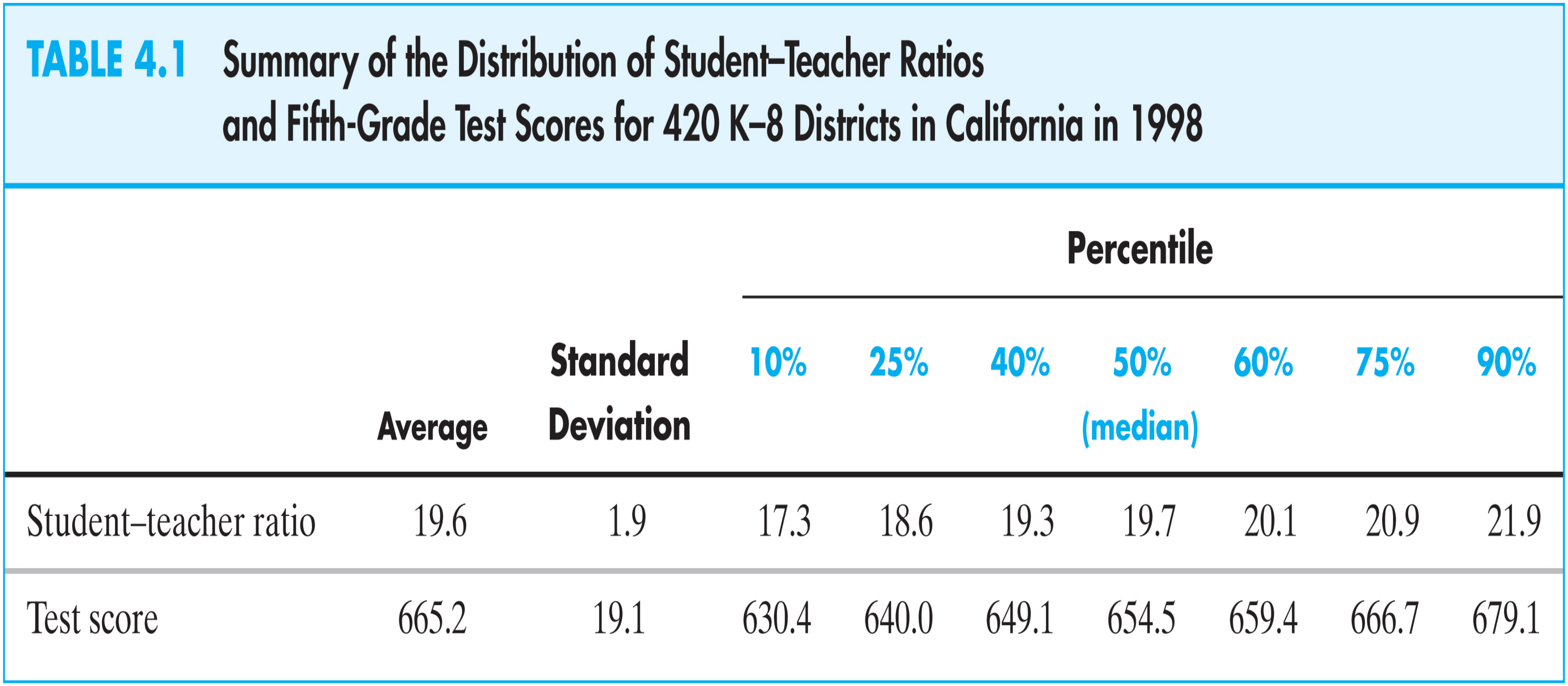


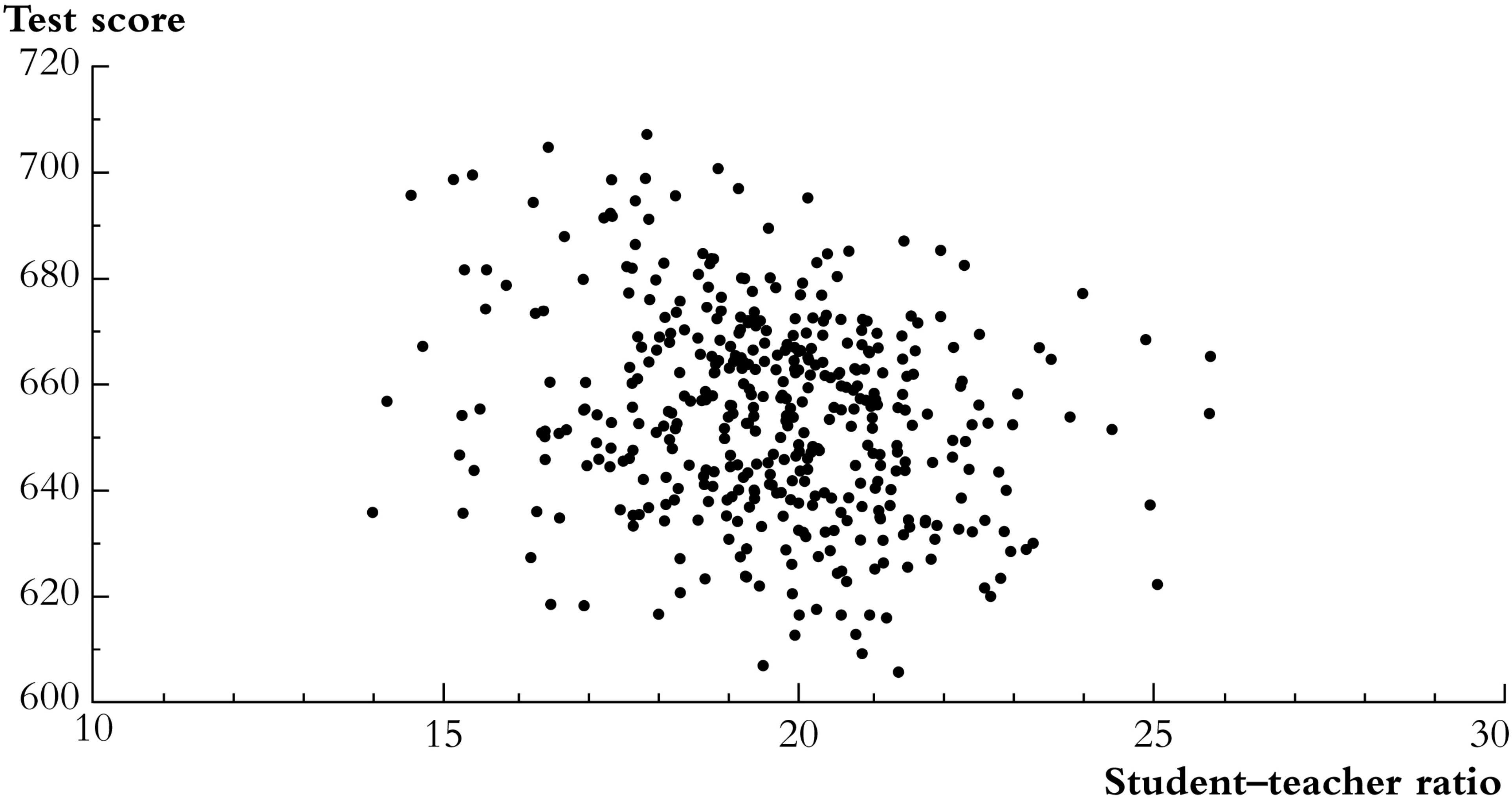
TABLE 4.1 Summary of the Distribution of Student–Teacher Ratios and Fifth-Grade Test Scores for 420 K–8 Districts in California in 1998

			Percentile						
	Average	Standard Deviation	10%	25%	40%	50% (median)	60%	75%	90%
Student–teacher ratio	19.6	1.9	17.3	18.6	19.3	19.7	20.1	20.9	21.9
Test score	665.2	19.1	630.4	640.0	649.1	654.5	659.4	666.7	679.1

This table doesn’t tell us anything about the relationship between test scores and the *STR*.

Do districts with smaller classes have higher test scores?

Scatterplot of test score v. student-teacher ratio



what does this figure show?

Example (con’t): the California test score data

Regression of *TestScore* against *STR*:

$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$

Now include percent English Learners in the district (*PctEL*):

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 PctEL$$

- What happens to the coefficient on *STR*?
- Why? (*Note*: corr(*STR*, *PctEL*) = 0.19)

Multiple regression in STATA

```
reg testscr str pctel, robust;

Regression with robust standard errors                 Number of obs =     420
                                                       F(  2,   417) =  223.82
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4264
                                                       Root MSE      =  14.464

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.       t     P>|t|      [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -1.101296   .4328472    -2.54    0.011     -1.95213   -.2504616
       pctel |  -.6497768   .0310318   -20.94    0.000     -.710775   -.5887786
       _cons |   686.0322   8.728224    78.60    0.000     668.8754     703.189
------------------------------------------------------------------------------
```

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 PctEL$$

More on this printout later…

Measures of Fit for Multiple Regression

Actual = predicted + residual: $Y_i = \hat{Y}_i + \hat{u}_i$

SER = std. deviation of $\hat{u}_i$ (with d.f. correction)

$RMSE$ = std. deviation of $\hat{u}_i$ (without d.f. correction)

R^2 = fraction of variance of Y explained by X

$\bar{R}^2$ = "adjusted R^2" = R^2 with a degrees-of-freedom correction that adjusts for estimation uncertainty; $\bar{R}^2 < R^2$

SER and RMSE

As in regression with a single regressor, the *SER* and the *RMSE* are measures of the spread of the *Y*'s around the regression line:

$$SER = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2}$$

R^2 and $\bar{R}^2$

The R^2 is the fraction of the variance explained – same definition as in regression with a single regressor:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS},$$

where $ESS = \sum_{i=1}^{n}(\hat{Y}_i - \overline{\hat{Y}})^2$, $SSR = \sum_{i=1}^{n}\hat{u}_i^2$, $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

- The R^2 always increases when you add another regressor (*why?)* – a bit of a problem for a measure of "fit"

R^2 and $\bar{R}^2$, ctd.

The $\bar{R}^2$ (the "adjusted R^2") corrects this problem by "penalizing" you for including another regressor – the $\bar{R}^2$ does not necessarily increase when you add another regressor.

$$Adjusted\ R^2\text{:}\quad \bar{R}^2 = 1-\left(\frac{n-1}{n-k-1}\right)\frac{SSR}{TSS}$$

Note that $\bar{R}^2 < R^2$, however if n is large the two will be very close.

Measures of fit, ctd.

Test score example:

(1) $$\widehat{TestScore} = 698.9 - 2.28\times STR,$$

$$R^2 = .05,\ SER = 18.6$$

(2) $$\widehat{TestScore} = 686.0 - 1.10\times STR - 0.65 PctEL,$$

$$R^2 = .426,\ \bar{R}^2 = .424,\ SER = 14.5$$

- *What – precisely – does this tell you about the fit of regression (2) compared with regression (1)?*
- *Why are the R^2 and the $\bar{R}^2$ so close in (2)?*

The Least Squares Assumptions for Multiple Regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \; i = 1,\ldots,n$$

1. The conditional distribution of u given the X's has mean zero, that is, $E(u|X_1 = x_1,\ldots, X_k = x_k) = 0$.
2. $(X_{1i},\ldots,X_{ki},Y_i)$, $i =1,\ldots,n$, are i.i.d.
3. Large outliers are rare: $X_1,\ldots, X_k$, and Y have four moments: $E(X_{1i}^4) < \infty,\ldots, E(X_{ki}^4) < \infty, E(Y_i^4) < \infty$.
4. There is no perfect multicollinearity.

Assumption #1: the conditional mean of *u* given the included *X*'s is zero.

$$E(u|X_1 = x_1,\dots, X_k = x_k) = 0$$

- This has the same interpretation as in regression with a single regressor.
- If an omitted variable (1) belongs in the equation (so is in *u*) and (2) is correlated with an included *X*, then this condition fails
- Failure of this condition leads to omitted variable bias
- The solution – ***if possible*** – is to include the omitted variable in the regression.

Assumption #2: $(X_{1i},\ldots,X_{ki},Y_i)$, i =1,…,n, are i.i.d.

This is satisfied automatically if the data are collected by simple random sampling.

Assumption #3: large outliers are rare (finite fourth moments)

This is the same assumption as we had before for a single regressor. As in the case of a single regressor, OLS can be sensitive to large outliers, so you need to check your data (scatterplots!) to make sure there are no crazy values (typos or coding errors).

Assumption #4: There is no perfect multicollinearity

Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors.

Example: Suppose you accidentally include *STR* twice:

```
regress testscr str str, robust
Regression with robust standard errors                Number of obs =     420
                                                      F(  1,   418) =   19.26
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0512
                                                      Root MSE      =  18.581
-------------------------------------------------------------------------
        |               Robust
testscr |      Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
--------+----------------------------------------------------------------
    str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
    str |  (dropped)
  _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
-------------------------------------------------------------------------
```

Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors.

- In the previous regression, β_1 is the effect on *TestScore* of a unit change in *STR*, holding *STR* constant (???)

Multicollinearity, Perfect and Imperfect

The dummy variable trap

Suppose you have a set of multiple binary (dummy) variables, which are mutually exclusive and exhaustive – that is, there are multiple categories and every observation falls in one and only one category (Freshmen, Sophomores, Juniors, Seniors, Other). If you include all these dummy variables *and* a constant, you will have perfect multicollinearity – this is sometimes called ***the dummy variable trap***.

- *Why is there perfect multicollinearity here*?
- *Solutions to the dummy variable trap*:

  1. Omit one of the groups (e.g. Senior), or
  2. Omit the intercept

Perfect multicollinearity, ctd.

- Perfect multicollinearity usually reflects a mistake in the definitions of the regressors, or an oddity in the data
- If you have perfect multicollinearity, your statistical software will let you know – either by crashing or giving an error message or by “dropping” one of the variables arbitrarily
- The solution to perfect multicollinearity is to modify your list of regressors so that you no longer have perfect multicollinearity.

Imperfect multicollinearity

Imperfect and perfect multicollinearity are quite different despite the similarity of the names.

Imperfect multicollinearity occurs when two or more regressors are very highly correlated.

- Why this term? If two regressors are very highly correlated, then their scatterplot will pretty much look like a straight line – they are collinear – but unless the correlation is exactly ±1, that collinearity is imperfect.

Imperfect multicollinearity, ctd.

Imperfect multicollinearity implies that one or more of the regression coefficients will be imprecisely estimated.

- Intuition: the coefficient on X_1 is the effect of X_1 holding X_2 constant; but if X_1 and X_2 are highly correlated, there is very little variation in X_1 once X_2 is held constant – so the data are pretty much uninformative about what happens when X_1 changes but X_2 doesn't, so the variance of the OLS estimator of the coefficient on X_1 will be large.
- Imperfect multicollinearity (correctly) results in large standard errors for one or more of the OLS coefficients.

The Sampling Distribution of the OLS Estimator

Under the four Least Squares Assumptions,

- The exact (finite sample) distribution of $\hat{\beta}_1$ has mean β_1, var($\hat{\beta}_1$) is inversely proportional to *n*; so too for $\hat{\beta}_2$.
- Other than its mean and variance, the exact (finite-*n*) distribution of $\hat{\beta}_1$ is very complicated; but for large *n*…
- $\hat{\beta}_1$ is consistent: $\hat{\beta}_1 \overset{p}{\to} \beta_1$ (law of large numbers)
- $\dfrac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}}$ is approximately distributed $N(0,1)$ (CLT)
- So too for $\hat{\beta}_2,\ldots,\ \hat{\beta}_k$

Hypothesis Tests and Confidence Intervals in Multiple Regression

Hypothesis Tests and Confidence Intervals for a Single Coefficient in Multiple Regression

- $\dfrac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}}$ is approximately distributed $N(0,1)$ (CLT).
- Thus hypotheses on β_1 can be tested using the usual *t*-statistic, and confidence intervals are constructed as $\{\hat{\beta}_1 \pm 1.96\times SE(\hat{\beta}_1)\}$.
- So too for $\beta_2,\ldots,\beta_k$.
- $\hat{\beta}_1$ and $\hat{\beta}_2$ are generally not independently distributed – so neither are their *t*-statistics (more on this later).

***Example*:** The California class size data

(1) $\widehat{TestScore} = 698.9 - 2.28 \times STR$

$(10.4)\quad(0.52)$

(2) $\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.650 PctEL$

$(8.7)\quad(0.43)\qquad\qquad(0.031)$

- The coefficient on *STR* in (2) is the effect on *TestScores* of a unit change in *STR*, holding constant the percentage of English Learners in the district
- The coefficient on *STR* falls by one-half
- The 95% confidence interval for coefficient on *STR* in (2) is $\{-1.10 \pm 1.96 \times 0.43\} = (-1.95, -0.26)$
- The *t*-statistic testing $\beta_{STR} = 0$ is $t = -1.10/0.43 = -2.54$, so we reject the hypothesis at the 5% significance level

Standard errors in multiple regression in STATA

```
reg testscr str pctel, robust;

Regression with robust standard errors                 Number of obs =     420
                                                       F(  2,   417) =  223.82
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4264
                                                       Root MSE      =  14.464

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -1.101296   .4328472    -2.54   0.011     -1.95213   -.2504616
       pctel |  -.6497768   .0310318   -20.94   0.000     -.710775   -.5887786
       _cons |   686.0322   8.728224    78.60   0.000     668.8754     703.189
------------------------------------------------------------------------------
```

$$\widehat{TestScore} = \underset{(8.7)}{686.0} - \underset{(0.43)}{1.10}\times STR - \underset{(0.031)}{0.650}PctEL$$

We use heteroskedasticity-robust standard errors – for exactly the same reason as in the case of a single regressor.

Tests of Joint Hypotheses

Let *Expn* = expenditures per pupil and consider the population regression model:

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

The null hypothesis that “school resources don’t matter,” and the alternative that they do, corresponds to:

$$H_0\text{: } \beta_1 = 0 \textbf{\textit{ and }} \beta_2 = 0$$

$$\text{vs. } H_1\text{: } \textbf{\textit{either }} \beta_1 \neq 0 \textbf{\textit{ or }} \beta_2 \neq 0 \textbf{\textit{ or both}}$$

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

Tests of joint hypotheses, ctd.

$$H_0\text{: } \beta_1 = 0 \textbf{\textit{ and }} \beta_2 = 0$$

$$\text{vs. } H_1\text{: } \textbf{\textit{either }} \beta_1 \neq 0 \textbf{\textit{ or }} \beta_2 \neq 0 \textbf{\textit{ or both}}$$

- A ***joint hypothesis*** specifies a value for two or more coefficients, that is, it imposes a restriction on two or more coefficients.
- In general, a joint hypothesis will involve q restrictions. In the example above, $q = 2$, and the two restrictions are $\beta_1 = 0$ and $\beta_2 = 0$.
- A “common sense” idea is to reject if either of the individual t-statistics exceeds 1.96 in absolute value.
- But this “one at a time” test isn’t valid: the resulting test rejects too often under the null hypothesis (more than 5%)!

Why can't we just test the coefficients one at a time?

Because the rejection rate under the null isn't 5%. We'll calculate the probability of incorrectly rejecting the null using the "common sense" test based on the two individual t-statistics. To simplify the calculation, suppose that $\hat{\beta}_1$ and $\hat{\beta}_2$ are independently distributed. Let t_1 and t_2 be the t-statistics:

$$t_1 = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \text{ and } t_2 = \frac{\hat{\beta}_2 - 0}{SE(\hat{\beta}_2)}$$

The "one at time" test is:

reject H_0: $\beta_1 = \beta_2 = 0$ if $|t_1| > 1.96$ and/or $|t_2| > 1.96$

What is the probability that this "one at a time" test rejects H_0, when H_0 is actually true? (It *should* be 5%.)

Suppose t_1 and t_2 are independent **(for this calculation).**

The probability of incorrectly rejecting the null hypothesis using the "one at a time" test

$$= \Pr_{H_0}[|t_1| > 1.96 \text{ and/or } |t_2| > 1.96]$$

$$= \Pr_{H_0}[|t_1| > 1.96, |t_2| > 1.96] + \Pr_{H_0}[|t_1| > 1.96, |t_2| \leq 1.96]$$

$$+ \Pr_{H_0}[|t_1| \leq 1.96, |t_2| > 1.96] \qquad \text{(disjoint events)}$$

$$= \Pr_{H_0}[|t_1| > 1.96] \times \Pr_{H_0}[|t_2| > 1.96]$$

$$+ \Pr_{H_0}[|t_1| > 1.96] \times \Pr_{H_0}[|t_2| \leq 1.96]$$

$$+ \Pr_{H_0}[|t_1| \leq 1.96] \times \Pr_{H_0}[|t_2| > 1.96]$$

(t_1, t_2 are independent by assumption)

$$= .05\times.05 + .05\times.95 + .95\times.05$$

= .0975 = 9.75% – which is ***not*** the desired 5%!!

The ***size*** of a test is the actual rejection rate under the null hypothesis.

- The size of the “common sense” test isn’t 5%!
- In fact, its size depends on the correlation between t_1 and t_2 (and thus on the correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$).

Two Solutions:

- Use a different critical value in this procedure – not 1.96 (this is the “Bonferroni method) (this method is rarely used in practice however)
- Use a different test statistic that test both β_1 and β_2 at once: the *F*-statistic (this is common practice)

The *F*-statistic

The *F*-statistic tests all parts of a joint hypothesis at once.

Formula for the special case of the joint hypothesis $\beta_1 = \beta_{1,0}$ and $\beta_2 = \beta_{2,0}$ in a regression with two regressors:

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2}\right)$$

where $\hat{\rho}_{t_1,t_2}$ estimates the correlation between t_1 and t_2.

Reject when *F* is large (how large?)

The *F*-statistic testing β_1 and β_2:

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2}\right)$$

- The *F*-statistic is large when t_1 and/or t_2 is large
- The *F*-statistic corrects (in just the right way) for the correlation between t_1 and t_2.
- The formula for more than two β's is nasty unless you use matrix algebra.
- This gives the *F*-statistic a nice large-sample approximate distribution, which is…

Large-sample distribution of the *F*-statistic

Consider *special case* that t_1 and t_2 are independent, so $\hat{\rho}_{t_1,t_2}$ $\xrightarrow{p}$ 0; in large samples the formula becomes

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2}t_1t_2}{1-\hat{\rho}_{t_1,t_2}^2}\right) \cong \frac{1}{2}(t_1^2 + t_2^2)$$

- Under the null, t_1 and t_2 have standard normal distributions that, in this special case, are independent
- The large-sample distribution of the *F*-statistic is the distribution of the average of two independently distributed squared standard normal random variables.

The ***chi-squared*** distribution with q degrees of freedom (χ_q^2) is defined to be the distribution of the sum of q independent squared standard normal random variables.

In large samples, *F* is distributed as χ_q^2/*q*.

Selected large-sample critical values of χ_q^2/*q*

q	5% critical value	
1	3.84	(*why*?)
2	3.00	(the case q=2 above)
3	2.60	
4	2.37	
5	2.21	

***Computing the p-value using the F-statistic*:**

p-value = tail probability of the χ_q^2/q distribution beyond the F-statistic actually computed.

Implementation in STATA

Use the "test" command after the regression

Example: Test the joint hypothesis that the population coefficients on *STR* and expenditures per pupil (*expn_stu*) are both zero, against the alternative that at least one of the population coefficients is nonzero.

F-test example, California class size data:

```
reg testscr str expn_stu pctel, r;

Regression with robust standard errors                 Number of obs =     420
                                                       F(  3,   416) =  147.20
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4366
                                                       Root MSE      =  14.353

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t     P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -.2863992   .4820728    -0.59   0.553    -1.234001     .661203
    expn_stu |   .0038679   .0015807     2.45   0.015     .0007607    .0069751
       pctel |  -.6560227   .0317844   -20.64   0.000    -.7185008   -.5935446
       _cons |   649.5779   15.45834    42.02   0.000     619.1917    679.9641
------------------------------------------------------------------------------
                                 NOTE

test str expn_stu;               The test command follows the regression

 ( 1)  str = 0.0                 There are q=2 restrictions being tested
 ( 2)  expn_stu = 0.0

       F(  2,   416) =    5.43   The 5% critical value for q=2 is 3.00
       Prob > F =    0.0047      Stata computes the p-value for you
```

More on *F*-statistics: *a simple F-statistic formula that is easy to understand (it is only valid if the errors are homoskedastic, but it might help intuition).*

The homoskedasticity-only *F*-statistic

When the errors are homoskedastic, there is a simple formula for computing the "homoskedasticity-only" *F*-statistic:

- Run two regressions, one under the null hypothesis (the "restricted" regression) and one under the alternative hypothesis (the "unrestricted" regression).
- Compare the fits of the regressions – the R^2's – if the "unrestricted" model fits sufficiently better, reject the null

The "restricted" and "unrestricted" regressions

Example: are the coefficients on *STR* and *Expn* zero?

Unrestricted population regression (under H_1):

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

Restricted population regression (that is, under H_0):

$$TestScore_i = \beta_0 + \beta_3 PctEL_i + u_i \qquad (why?)$$

- The number of restrictions under H_0 is $q = 2$ (*why*?).
- The fit will be better (R^2 will be higher) in the unrestricted regression (*why*?)

By how much must the R^2 increase for the coefficients on *Expn* and *PctEL* to be judged statistically significant?

Simple formula for the homoskedasticity-only F-statistic:

$$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n - k_{unrestricted} - 1)}$$

where:

$R^2_{restricted}$= the R^2 for the restricted regression

$R^2_{unrestricted}$ = the R^2 for the unrestricted regression

q = the number of restrictions under the null

$k_{unrestricted}$ = the number of regressors in the unrestricted regression.

- The bigger the difference between the restricted and unrestricted R^2's – the greater the improvement in fit by adding the variables in question – the larger is the homoskedasticity-only F.

Example:

Restricted regression:

$$\widehat{TestScore} = 644.7 - 0.671 PctEL, \quad R^2_{restricted} = 0.4149$$
$$\quad (1.0) \quad (0.032)$$

Unrestricted regression:

$$\widehat{TestScore} = 649.6 - 0.29STR + 3.87Expn - 0.656PctEL$$
$$(15.5) \quad (0.48) \qquad\qquad (1.59) \qquad\qquad (0.032)$$

$$R^2_{unrestricted} = 0.4366, \; k_{unrestricted} = 3, \;\; q = 2$$

so

$$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n - k_{unrestricted} - 1)}$$

$$= \frac{(.4366 - .4149)/2}{(1 - .4366)/(420 - 3 - 1)} = \mathbf{8.01}$$

Note: Heteroskedasticity-robust F = **5.43**…

The homoskedasticity-only F-statistic – summary

$$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n - k_{unrestricted} - 1)}$$

- The homoskedasticity-only *F*-statistic rejects when adding the two variables increased the R^2 by “enough” – that is, when adding the two variables improves the fit of the regression by “enough”
- If the errors are homoskedastic, then the homoskedasticity-only *F*-statistic has a large-sample distribution that is χ^2_q/q.
- But if the errors are heteroskedastic, the large-sample distribution is a mess and is not χ^2_q/q

Digression: The *F* distribution

Your regression printouts might refer to the "*F*" distribution.

If the four multiple regression LS assumptions hold ***and***:

5. u_i is homoskedastic, that is, var($u|X_1,\ldots,X_k$) does not depend on *X*'s
6. $u_1,\ldots,u_n$ are normally distributed

then the homoskedasticity-only *F*-statistic has the "$F_{q,n-k-1}$" distribution, where q = the number of restrictions and k = the number of regressors under the alternative (the unrestricted model).

- **The *F* distribution is to the χ^2_q/q distribution what the t_{n-1} distribution is to the *N*(0,1) distribution**

The $F_{q,n-k-1}$ distribution:

- The *F* distribution is tabulated many places
- As $n \to \infty$, the $F_{q,n\text{-}k-1}$ distribution asymptotes to the χ_q^2/q distribution:

The $F_{q,\infty}$ and χ_q^2/q distributions are the same.

- For q not too big and $n \geq 100$, the $F_{q,n-k-1}$ distribution and the χ_q^2/q distribution are essentially identical.
- Many regression packages (including STATA) compute *p*-values of *F*-statistics using the *F* distribution
- You will encounter the *F* distribution in published empirical work.

Testing Single Restrictions on Multiple Coefficients

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,\ \ i = 1,\ldots,n$$

Consider the null and alternative hypothesis,

$$H_0: \beta_1 = \beta_2 \quad \text{vs.}\ \ H_1: \beta_1 \neq \beta_2$$

This null imposes a *single* restriction ($q = 1$) on *multiple* coefficients – it is not a joint hypothesis with multiple restrictions (compare with $\beta_1 = 0$ and $\beta_2 = 0$).

Testing single restrictions on multiple coefficients, ctd.

Here are two methods for testing single restrictions on multiple coefficients:

1. ***Rearrange ("transform") the regression***

   Rearrange the regressors so that the restriction becomes a restriction on a single coefficient in an equivalent regression; or,

2. ***Perform the test directly***

   Some software, including STATA, lets you test restrictions using multiple coefficients directly

Method 1: Rearrange ("transform") the regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

$$H_0\text{: } \beta_1 = \beta_2 \quad \text{vs. } H_1\text{: } \beta_1 \neq \beta_2$$

Add and subtract $\beta_2 X_{1i}$:

$$Y_i = \beta_0 + (\beta_1 - \beta_2)\, X_{1i} + \beta_2(X_{1i} + X_{2i}) + u_i$$

or

$$Y_i = \beta_0 + \gamma_1\, X_{1i} + \beta_2 W_i + u_i$$

where

$$\gamma_1 = \beta_1 - \beta_2$$

$$W_i = X_{1i} + X_{2i}$$

Rearrange the regression, ctd.

(a) Original system:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

$$H_0: \beta_1 = \beta_2 \quad \text{vs.} \quad H_1: \beta_1 \neq \beta_2$$

(b) Rearranged ("transformed") system:

$$Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i$$

where $\gamma_1 = \beta_1 - \beta_2$ and $W_i = X_{1i} + X_{2i}$

so

$$H_0: \gamma_1 = 0 \quad \text{vs.} \quad H_1: \gamma_1 \neq 0$$

The testing problem is now a simple one:

test whether $\gamma_1 = 0$ in specification (b).

Method 2: Perform the test directly

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

$$H_0\text{: } \beta_1 = \beta_2 \quad \text{vs. } H_1\text{: } \beta_1 \neq \beta_2$$

Example:

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

In STATA, to test $\beta_1 = \beta_2$ vs. $\beta_1 \neq \beta_2$ (two-sided):

```
regress testscore str expn pctel, r
test str=expn
```

The details of implementing this method are software-specific.

Confidence Sets for Multiple Coefficients

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i,\ \ i = 1,\ldots,n$$

What is a *joint* confidence set for β_1 and β_2?

A 95% ***joint confidence set*** is:

- A set-valued function of the data that contains the true parameter(s) in 95% of hypothetical repeated samples.
- The set of parameter values that cannot be rejected at the 5% significance level.
- You can find a 95% confidence set as the set of (β_1, β_2) that cannot be rejected at the 5% level using an F-test (*why not just combine the two 95% confidence intervals?*).

Joint confidence sets ctd.

Let $F(\beta_{1,0},\beta_{2,0})$ be the (heteroskedasticity-robust) F-statistic testing the hypothesis that $\beta_1 = \beta_{1,0}$ and $\beta_2 = \beta_{2,0}$:

95% confidence set = $\{\beta_{1,0}, \beta_{2,0}\text{: } F(\beta_{1,0}, \beta_{2,0}) < 3.00\}$

- 3.00 is the 5% critical value of the $F_{2,\infty}$ distribution
- This set has coverage rate 95% because the test on which it is based (the test it "inverts") has size of 5%

  5% of the time, the test incorrectly rejects the null when the null is true, so 95% of the time it does not; therefore the confidence set constructed as the nonrejected values contains the true value 95% of the time (in 95% of all samples).

The confidence set based on the F-statistic is an ellipse

$$\{\beta_1,\ \beta_2\text{:}\ \ F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2}t_1t_2}{1-\hat{\rho}_{t_1,t_2}^2}\right) \le 3.00\}$$

Now

$$F = \frac{1}{2(1-\hat{\rho}_{t_1,t_2}^2)}\times\left[t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2}t_1t_2\right]$$

$$= \frac{1}{2(1-\hat{\rho}_{t_1,t_2}^2)}\times$$

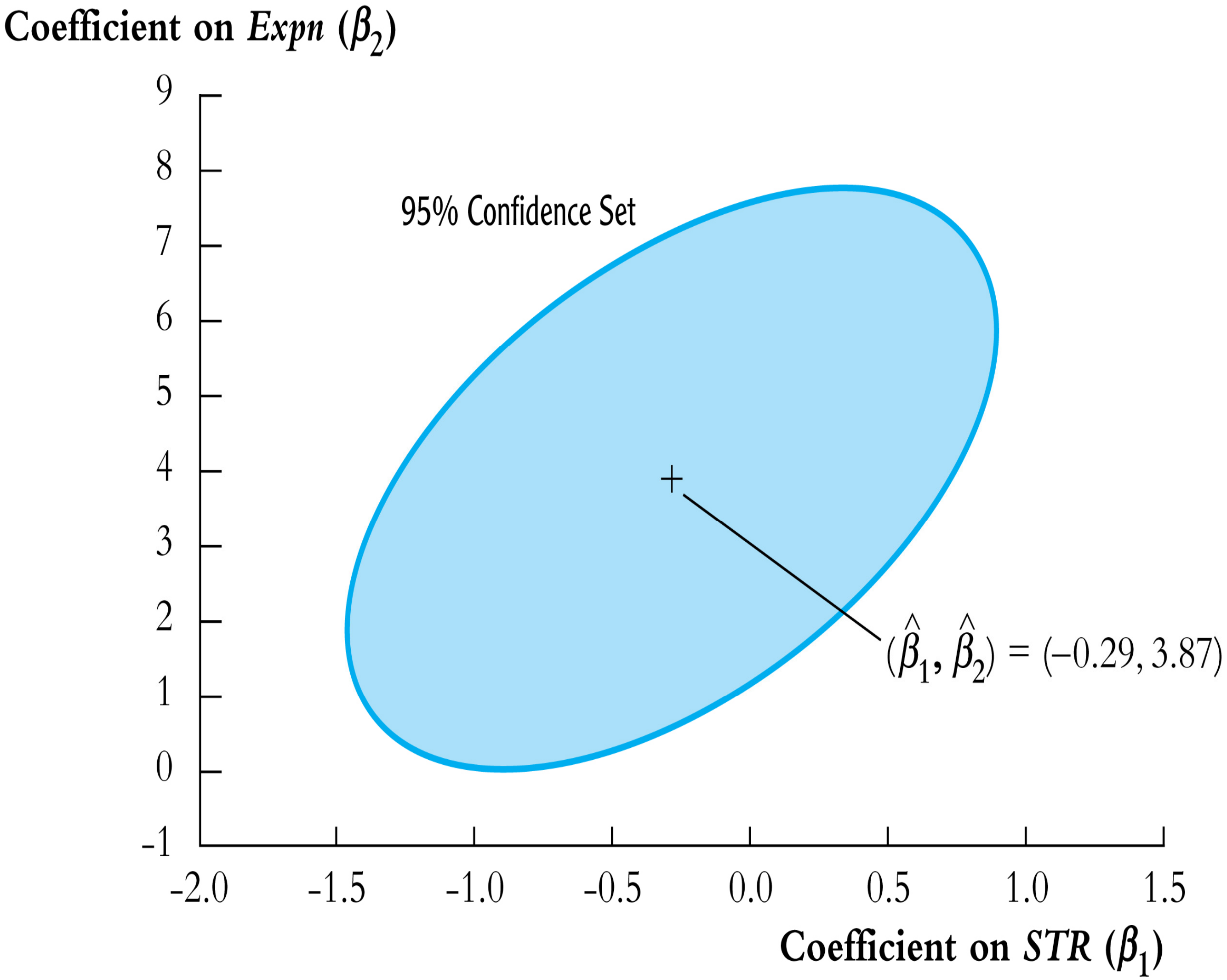
$$\left[\left(\frac{\hat{\beta}_2-\beta_{2,0}}{SE(\hat{\beta}_2)}\right)^2 + \left(\frac{\hat{\beta}_1-\beta_{1,0}}{SE(\hat{\beta}_1)}\right)^2 + 2\hat{\rho}_{t_1,t_2}\left(\frac{\hat{\beta}_1-\beta_{1,0}}{SE(\hat{\beta}_1)}\right)\left(\frac{\hat{\beta}_2-\beta_{2,0}}{SE(\hat{\beta}_2)}\right)\right]$$

This is a quadratic form in $\beta_{1,0}$ and $\beta_{2,0}$ – thus the boundary of the set F = 3.00 is an ellipse.

Confidence set based on inverting the F-statistic

FIGURE 7.1 95% Confidence Set for Coefficients on *STR* and *Expn* from Equation (7.6)

The 95% confidence set for the coefficients on *STR* (β_1) and *Expn* (β_2) is an ellipse. The ellipse contains the pairs of values of β_1 and β_2 that cannot be rejected using the *F*-statistic at the 5% significance level.



Nonlinear Functions of a Single Independent Variable

We’ll look at two complementary approaches:

1. Polynomials in *X*

   The population regression function is approximated by a quadratic, cubic, or higher-degree polynomial

2. 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:

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

- 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

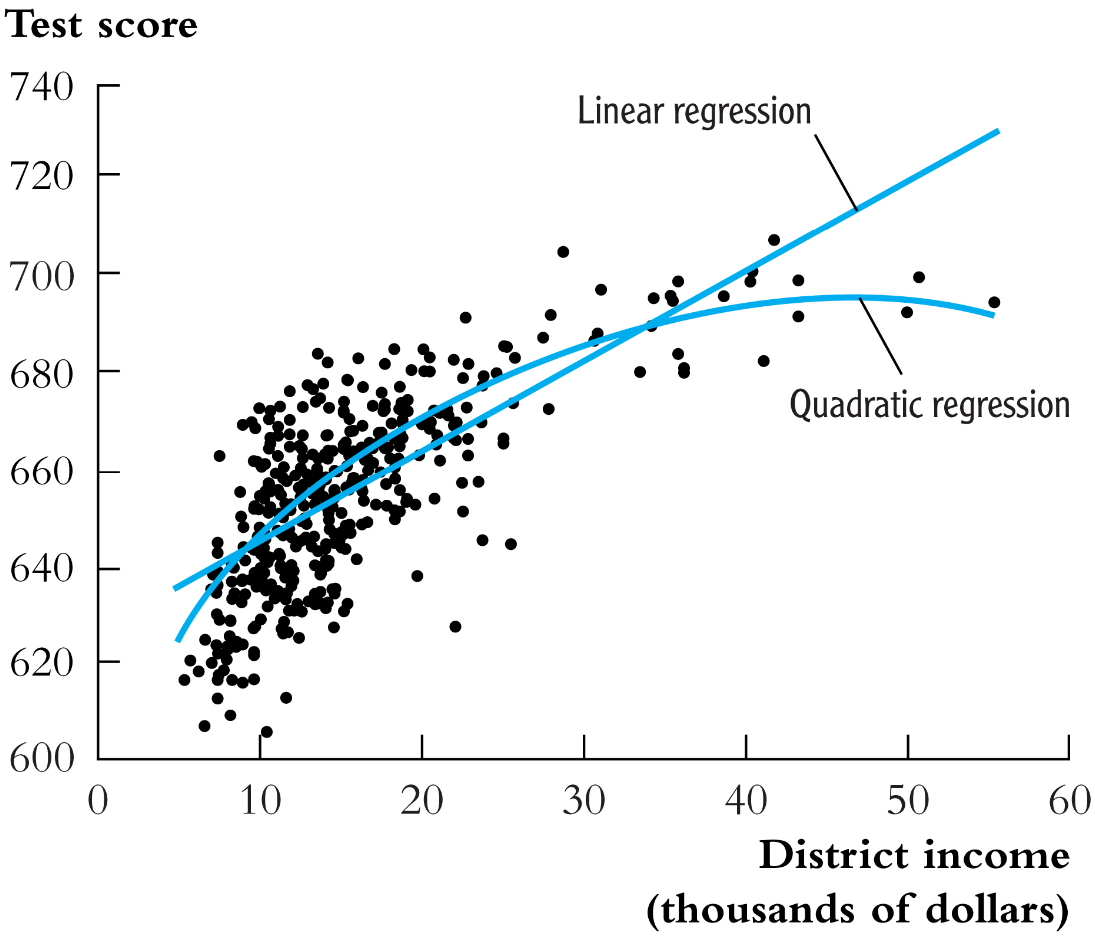
```
generate avginc2 = avginc*avginc;        Create a new regressor
reg testscr avginc avginc2, r;

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

$$\widehat{TestScore} = 607.3 + 3.85Income_i - 0.0423(Income_i)^2$$
$$\qquad\qquad(2.9)\quad(0.27)\qquad\qquad\qquad(0.0048)$$



***Interpreting the estimated regression function, ctd*:**

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

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

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

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

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

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 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, 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.

Summary: polynomial regression functions

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

- 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.

Here's why: $\ln(x+\Delta x) - \ln(x) = \ln\left(1+\frac{\Delta x}{x}\right) \cong \frac{\Delta x}{x}$

(calculus: $\frac{d\ln(x)}{dx} = \frac{1}{x}$)

Numerically:

$\ln(1.01) = .00995 \cong .01$;

$\ln(1.10) = .0953 \cong .10$ (sort of)

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*."

I. Linear-log population regression function

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

Now change X: $$Y + \Delta Y = \beta_0 + \beta_1 \ln(X + \Delta X) \qquad \text{(a)}$$

Subtract (a) – (b): $$\Delta Y = \beta_1[\ln(X + \Delta X) - \ln(X)]$$

now $$\ln(X + \Delta X) - \ln(X) \cong \frac{\Delta X}{X},$$

so $$\Delta Y \cong \beta_1 \frac{\Delta X}{X}$$

or $$\beta_1 \cong \frac{\Delta Y}{\Delta X / X} \quad \text{(small } \Delta X\text{)}$$

Linear-log case, continued

$$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$$

for small ΔX,

$$\beta_1 \cong \frac{\Delta Y}{\Delta X / X}$$

Now $100 \times \frac{\Delta X}{X}$ = percentage change in X, so ***a 1% increase in X (multiplying X by* 1.01*) is associated with a* .01β_1 *change in Y*.**

(1% increase in $X \Rightarrow .01$ increase in $\ln(X)$

$\Rightarrow .01\beta_1$ increase in Y)

Example: TestScore vs. ln(Income)

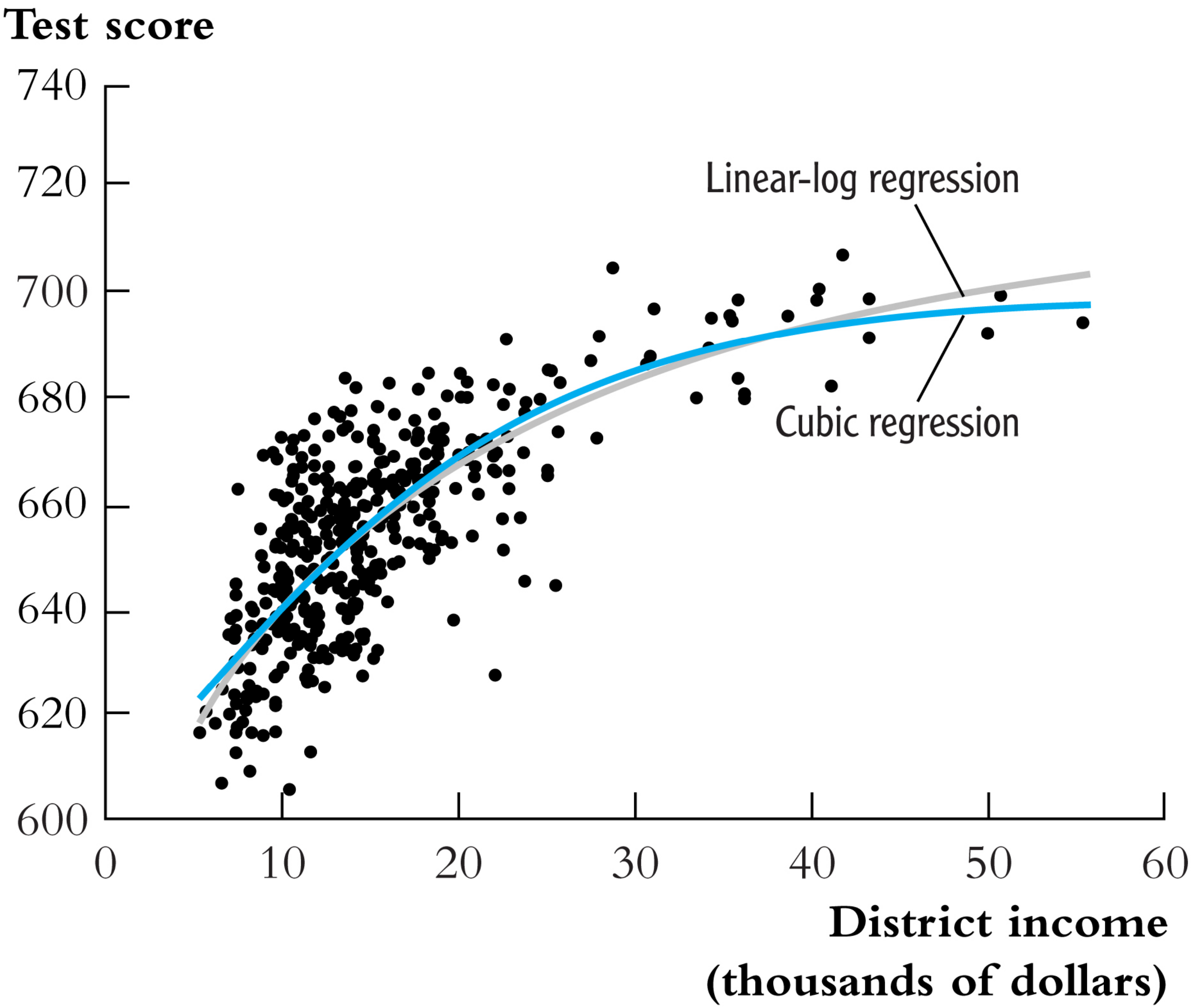
- First defining the new regressor, ln(*Income*)
- The model is now linear in ln(*Income*), so the linear-log model can be estimated by OLS:

$$\widehat{TestScore} = \underset{(3.8)}{557.8} + \underset{(1.40)}{36.42}\times\ln(Income_i)$$

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



II. Log-linear population regression function

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

Now change X: $\ln(Y + \Delta Y) = \beta_0 + \beta_1(X + \Delta X) \qquad \text{(a)}$

Subtract (a) – (b): $\ln(Y + \Delta Y) - \ln(Y) = \beta_1 \Delta X$

so $$\frac{\Delta Y}{Y} \cong \beta_1 \Delta X$$

or $$\beta_1 \cong \frac{\Delta Y / Y}{\Delta X} \text{ (small } \Delta X\text{)}$$

Log-linear case, continued

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$$

for small ΔX, $\beta_1 \cong \frac{\Delta Y / Y}{\Delta X}$

- Now $100 \times \frac{\Delta Y}{Y}$ = percentage change in Y, so ***a change in X by one unit* (ΔX = 1) *is associated with a* 100β_1*% change in Y*.**
- 1 unit increase in $X \Rightarrow \beta_1$ increase in $\ln(Y)$

  $\Rightarrow 100\beta_1\%$ increase in Y

III. Log-log population regression function

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i \qquad \text{(b)}$$

Now change X: $$\ln(Y + \Delta Y) = \beta_0 + \beta_1 \ln(X + \Delta X) \qquad \text{(a)}$$

Subtract: $$\ln(Y + \Delta Y) - \ln(Y) = \beta_1[\ln(X + \Delta X) - \ln(X)]$$

so $$\frac{\Delta Y}{Y} \cong \beta_1 \frac{\Delta X}{X}$$

or $$\beta_1 \cong \frac{\Delta Y / Y}{\Delta X / X} \text{ (small } \Delta X\text{)}$$

Log-log case, continued

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$$

for small ΔX,

$$\beta_1 \cong \frac{\Delta Y / Y}{\Delta X / X}$$

Now $100 \times \frac{\Delta Y}{Y}$ = percentage change in Y, and $100 \times \frac{\Delta X}{X}$ = percentage change in X, so ***a 1% change in X is associated with a β_1% change in Y***.

- ***In the log-log specification, β_1 has the interpretation of an elasticity***.

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

- First defining a new dependent variable, ln(*TestScore*), ***and*** the new regressor, ln(*Income*)
- The model is now a linear regression of ln(*TestScore*) against ln(*Income*), which can be estimated by OLS:

$$\widehat{\ln(TestScore)} = \underset{(0.006)}{6.336} + \underset{(0.0021)}{0.0554}\times\ln(Income_i)$$

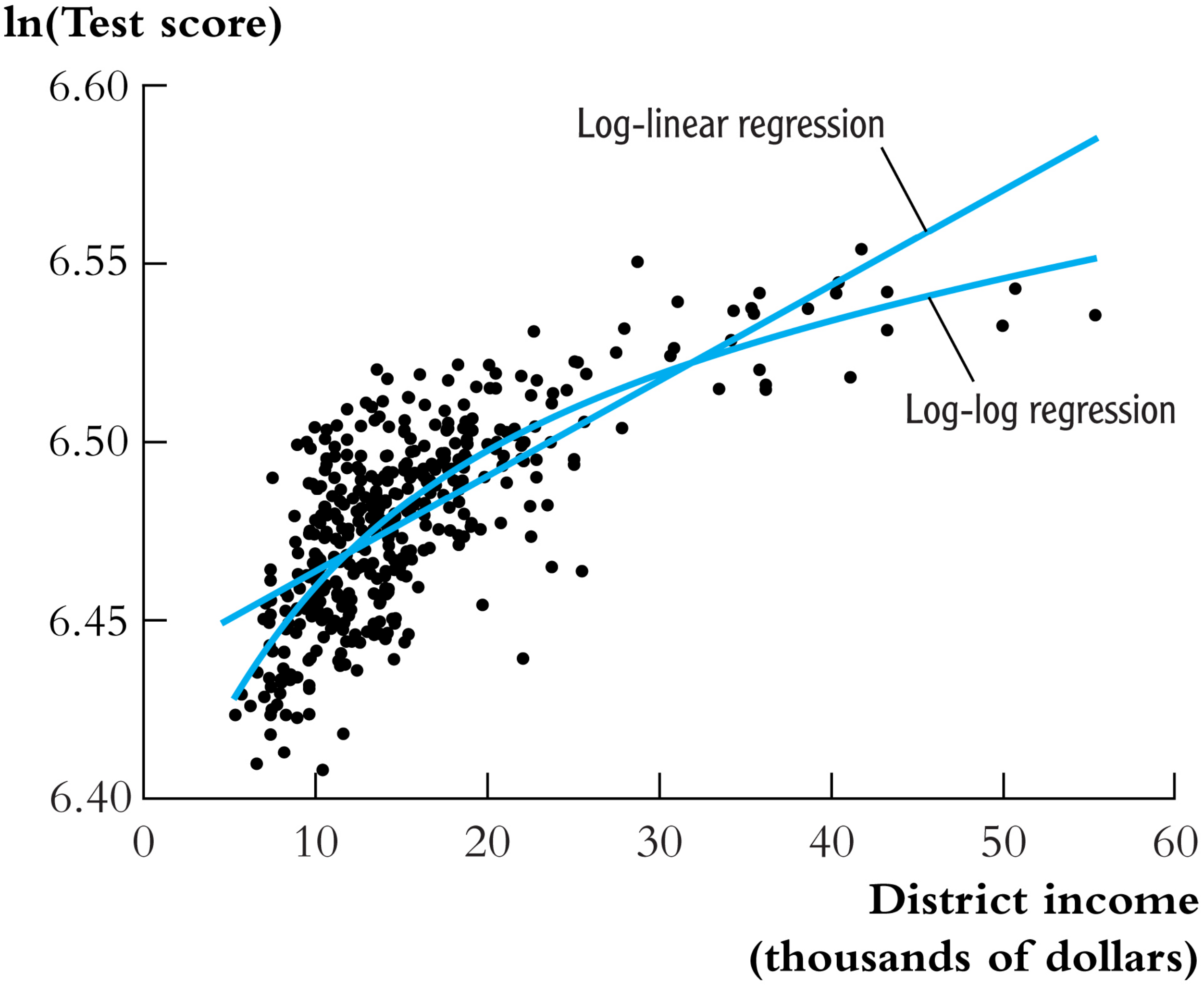
An 1% increase in *Income* is associated with an increase of .0554% in *TestScore* (*Income* up by a factor of 1.01, *TestScore* up by a factor of 1.000554)

Example: ln(TestScore) vs. ln(Income), ctd.

$$\widehat{\ln(TestScore)} = \underset{(0.006)}{6.336} + \underset{(0.0021)}{0.0554}\times\ln(Income_i)$$

- For example, suppose income increases from \$10,000 to \$11,000, or by 10%. Then *TestScore* increases by approximately .0554×10% = .554%. If *TestScore* = 650, this corresponds to an increase of .00554×650 = 3.6 points.
- How does this compare to the log-linear model?

The log-linear and log-log specifications:



- *Note vertical axis*
- *Neither seems to fit as well as the cubic or linear-log*

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.
- Choice of specification should be guided by judgment (which interpretation makes the most sense in your application?), tests, and plotting predicted values