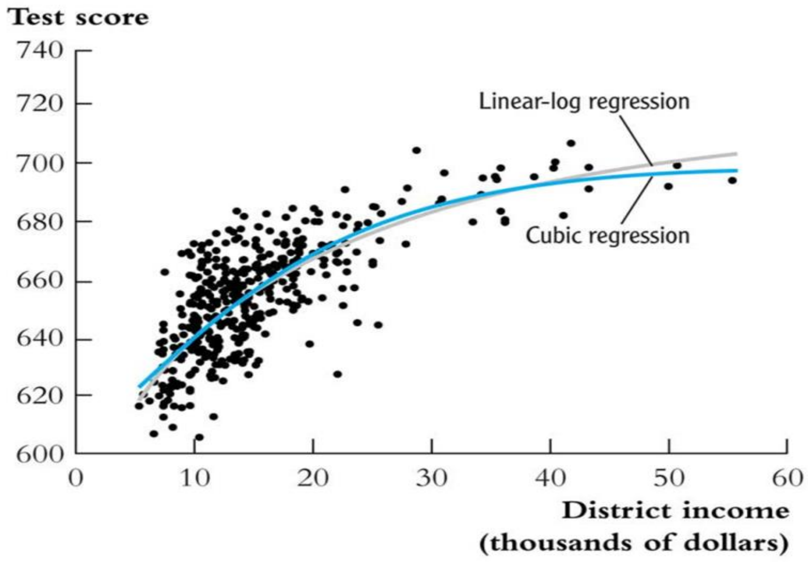
EC 3303: Econometrics I

Nonlinear Regression Functions (Part 1)



Kelvin Seah

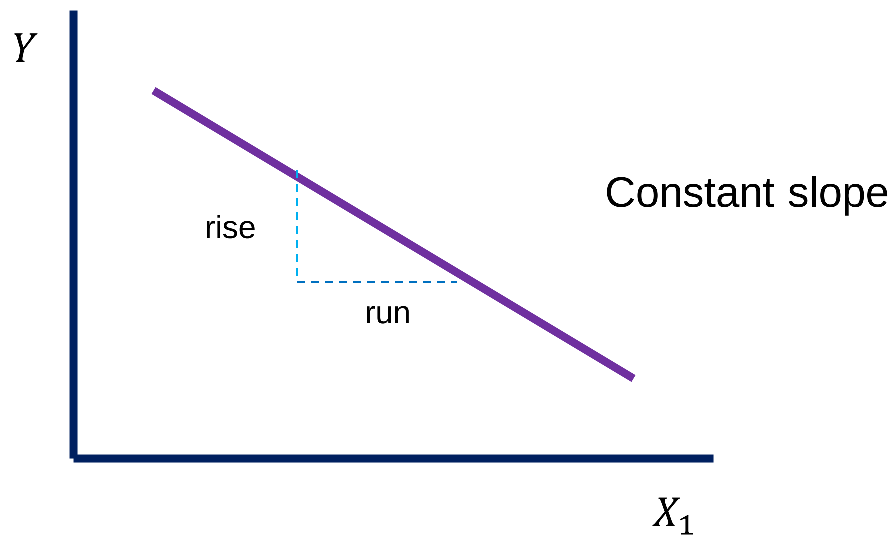
AY 2022/2023, Semester 2

Outline

1. Nonlinear Regression Functions
2. Polynomial Regression Models
3. Logarithmic Regression Models
4. Interactions Between Independent Variables

Introduction

- So far, population regression function is assumed to be ***linear***.
- i.e. a straight-line relationship exists between Y & X_1.
  - ***slope*** of the population regression function is assumed to be ***constant.***

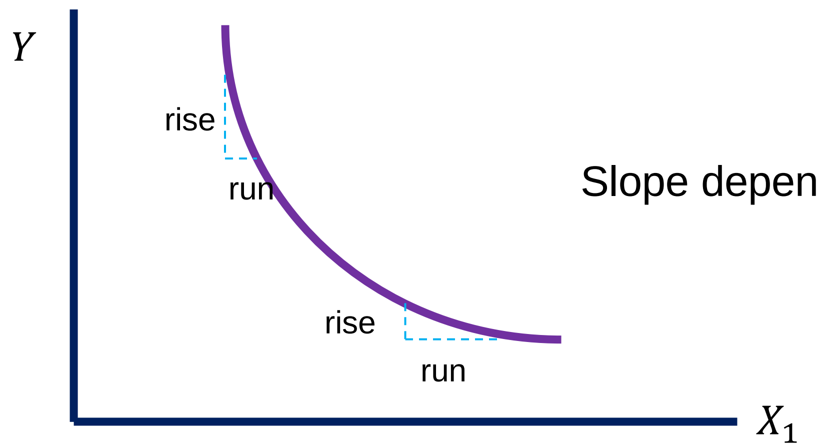


- But what if the effect on Y of a change in X_1 depends on the value of X_1 itself?

- Or what if the effect on Y of a change in X_1 depends on the value of another independent variable?

- If so, the population regression function is ***nonlinear.***

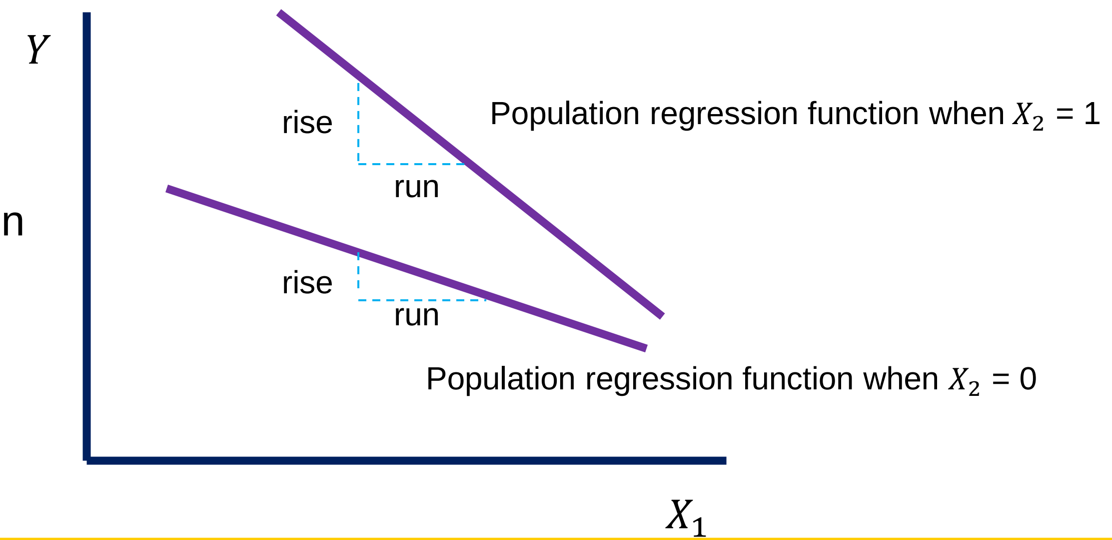
- What if the effect on Y of a change in X_1 depends on the value of X_1 itself?
  - Could happen if reducing class size (i.e. *STR*) has a greater effect when the class size is small compared to when it is large.
  - If so, test score (Y) is a nonlinear function of *STR* (X_1), where this function is steeper when the value of X_1 is small.



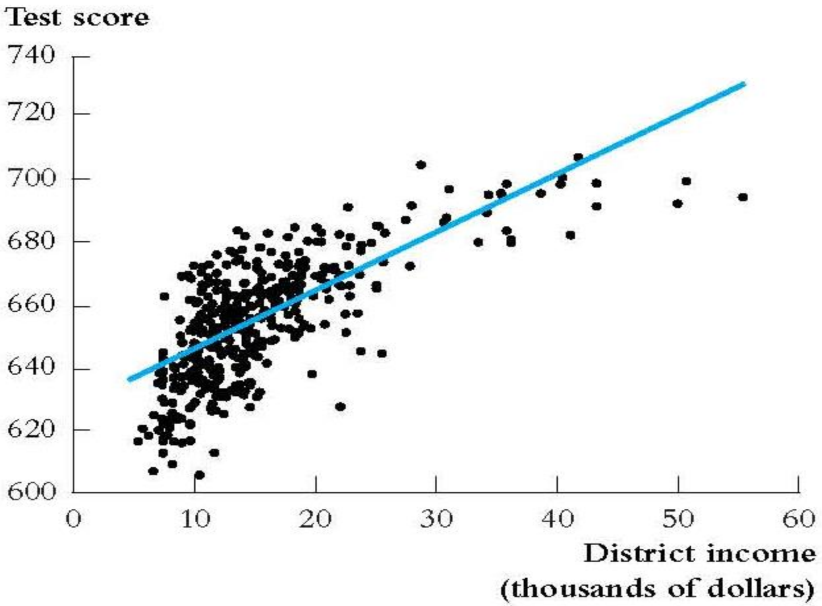
Slope depends on the value of X_1

- What if the effect on Y of a change in X_1 depends on the value of another independent variable?
  - could happen if ELL students learn better under more individual attention.
  - if so, reducing class size (i.e. *STR*) may have a greater effect in districts with many ELL students than in districts with few.
  - here, the effect on test scores (Y) of a change in *STR* (X_1) depends on the percentage of ELL in the district (X_2).

Slope depends on the value of X_2



A General Strategy for Modelling Nonlinear Regression Functions



- relationship is not well summarized by a straight line.
- it is better summarized by a curve – or a nonlinear function.
- slope of the nonlinear function $f(X)$ is not constant, but depends on X (district income).

- We can model the relationship as

$$TestScore_i = \beta_0 + \beta_1 Income_i + \beta_2 {Income_i}^2 + u_i \quad (1)$$

where $Income_i$ is the income in the i^{th} district; ${Income_i}^2$ is the square of the income in the i^{th} district; u_i is the error term; β_0, β_1, β_2 are unknown population coefficients.

- (1) is simply a multiple regression model with 2 regressors.

- You can create the second regressor by squaring the variable $Income$.

- Since (1) is a multiple regression model, can estimate and test β_0, β_1, β_2 using the OLS methods learnt earlier!

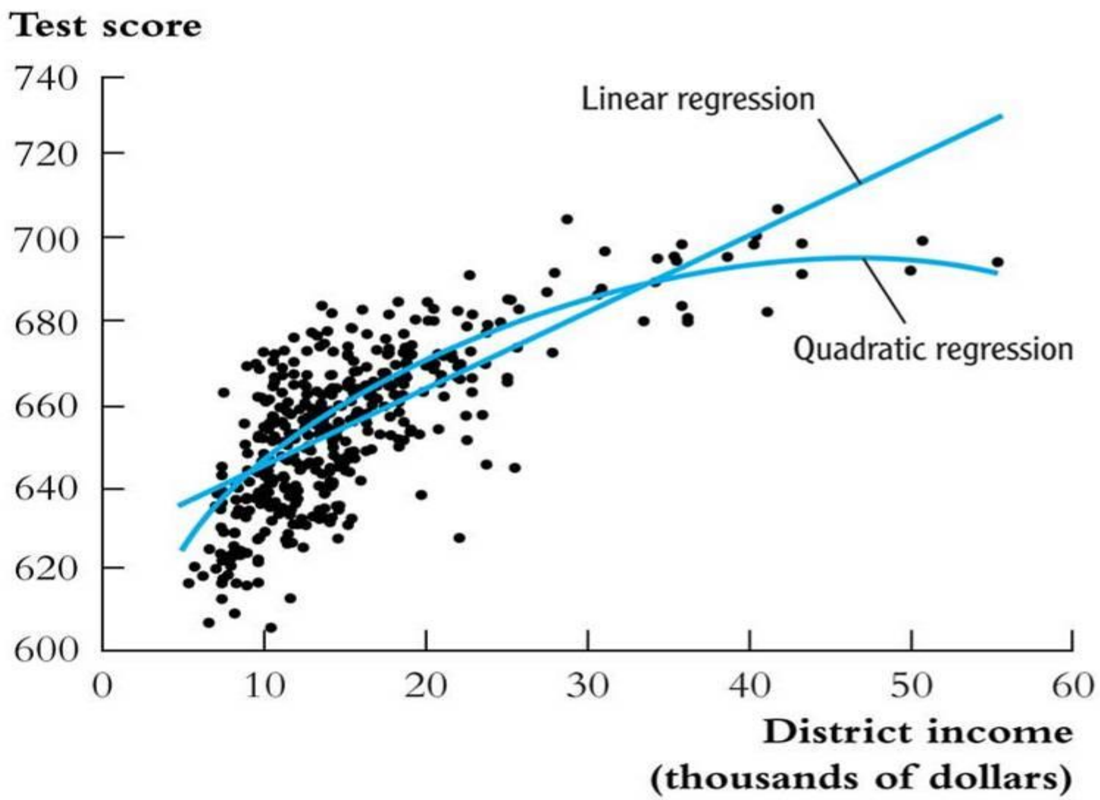
Estimation of the quadratic specification in STATA

```
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
------------------------------------------------------------------------------
```

$$\widehat{TestScore} = \underset{(2.9)}{607.3} + \underset{(0.27)}{3.85}Income - \underset{(0.0048)}{0.0423}Income^2, \bar{R}^2 = 0.554$$

(a) Plot the predicted values

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



- can go one step further beyond the visual comparison

$$TestScore_i = \beta_0 + \beta_1 Income_i + \beta_2 Income_i{}^2 + u_i \quad (1)$$

If the relationship is linear, $Income^2$ does not enter the population regression model, accordingly:

$$\beta_2 = 0$$

If the relationship is quadratic:

$$\beta_2 \neq 0$$

To see if the quadratic model fits better than the linear model,

$$H_0: \beta_2 = 0 \quad \text{vs} \quad H_1: \beta_2 \neq 0$$

$$\widehat{TestScore} = \underset{(2.9)}{607.3} + \underset{(0.27)}{3.85}Income - \underset{(0.0048)}{0.0423}Income^2, \bar{R}^2 = 0.554$$

1) $H_0: \beta_2 = 0;\ \ H_1: \beta_2 \neq 0$

2) Compute the t-statistic: $t^{act} = \frac{-0.0423-0}{0.0048} = -8.81$

3) Calculate the p-value: $2\Phi(-|t^{act}|) = 2\Phi(-8.81) \approx 0.00$

Alternatively, since $|t^{act}| = 8.81$ >1.96 (the 5% two-sided critical value), reject H_0 at the 5% significance level.

- Conclude: quadratic model fits the data better than the linear model.

Effect on Y of a Change in X in Nonlinear Specifications

linear regression model:

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

- β_1 is the expected effect on Y of a unit change in X_1 .

When the population regression function is nonlinear, the expected effect on Y of a unit change in X is more complicated…

- Population regression models (nonlinear or linear) are of the form:

$$Y_i = f(X_{1i}, X_{2i}, \dots, X_{ki}) + u_i, \quad i = 1, \dots., n \qquad (2)$$

where

Y: dependent variable

$X_1, X_2, \dots, X_k$: independent variables

u_i: error term

E.g:

In the quadratic regression model (1), there was a single independent variable. So X_1 is income and

$$Testscore_i = f(Income_i) + u_i \text{ or}$$

where $f(Income_i) = \beta_0 + \beta_1 Income_i + \beta_2 {Income_i}^2$

- Population regression models (nonlinear or linear) are of the form:

$$Y_i = f(X_{1i}, X_{2i}, \dots, X_{ki}) + u_i \qquad (2)$$

- If the population regression function is linear, then

$$f(X_{1i}, X_{2i}, \dots, X_{ki}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

- So (2) allows for both linear & nonlinear regression functions.

Effect on Y of a Change in X_1

$$Y_i = f(X_{1i}, X_{2i}, \dots, X_{ki}) + u_i \quad (2)$$

Consider changing X_1 by ΔX_1, while holding $X_2, X_3, \dots, X_k$ constant:

- Population regression function, ***before*** the change:

$$Y = f(X_1, X_2, \dots, X_k) \quad (3)$$

- Population regression function, ***after*** the change:

$$Y + \Delta Y = f(X_1 + \Delta X_1, X_2, \dots, X_k) \quad (4)$$

So the change in Y, on average (i.e. the expected change in Y), when X_1 changes by ΔX_1 is

$(4) - (3)$:

$$\Delta Y = f(X_1 + \Delta X_1, X_2, \dots, X_k) - f(X_1, X_2, \dots, X_k) \quad (5)$$

- Because the true population regression function f is unknown, the true effect on Y of a change in X_1 is also unknown.

- To estimate the true effect on Y of a change in X_1:

1) estimate the population regression function f.

2) denote this estimated function by $\hat{f}$…

The predicted value of Y, when X_1 takes on the value $X_1 + \Delta X_1$ (while $X_2, X_3, \ldots, X_k$ is held constant) is:

$$\hat{f}(X_1 + \Delta X_1, X_2, \ldots, X_k) \quad (6)$$

The predicted value of Y, when X_1 takes on the value X_1 (while $X_2, X_3, \ldots, X_k$ is held constant) is:

$$\hat{f}(X_1, X_2, \ldots, X_k) \quad (7)$$

So the estimated effect on Y when X_1 changes by an amount ΔX_1 is:

$$\Delta \hat{Y} = \hat{f}(X_1 + \Delta X_1, X_2, \ldots, X_k) - \hat{f}(X_1, X_2, \ldots, X_k) \quad (8)$$

- An example of $\hat{f}$ is the estimated quadratic regression function:

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

Example: *TestScore* – *Income* Relation

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

What is the effect on testscores of an increase in the district income by $1000?

- based on the estimated regression function above, the effect depends on the initial district income.

Consider:

1) An increase in district income from 5 to 6 (i.e. from $5,000 to $6,000).

2) An increase in district income from 25 to 26.

3) An increase in district income from 45 to 46.

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

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

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

- Predicted change in *TestScore* for a change in income from \$25,000 per capita to \$26,000 per capita:

$$\Delta\widehat{TestScore} = (607.3 + 3.85 \times 26 - 0.0423 \times 26^2) - (607.3 + 3.85 \times 25 - 0.0423 \times 25^2) = 1.7$$

ΔIncome (by $1000 per capita)	ΔTestScore
from 5 to 6	3.4
from 25 to 26	1.7
from 45 to 46	____

- The effect of a change in income is greater at low than at high income levels.
- What is the effect of a change in income from 65 to 66? *Caution! Don't extrapolate outside the range of the data!*

Interpreting coefficients in nonlinear (polynomial) specifications

$$TestScore_i = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^{\ 2} + u_i$$

- Regression coefficients here are not easy to interpret.

- It makes no sense to think of β_1, for e.g., as being the effect of changing the district's income, holding the square of the district's income constant.

- when dealing with nonlinear polynomial models, always:

1) plot the estimated regression function (in STATA)

2) calculate the estimated effect on Y associated with a change in X, using the "before and after" method.

Modelling a Nonlinear Regression Function of a Single Independent Variable

- We saw how to model the relationship between Y & X as a quadratic function.

- In general, Y & X could be related in other ways.

There are two complementary approaches to model nonlinear relationships:

1. ***Using Polynomials in X***

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

2. ***Using Logarithmic transformations***

- Y and/or X is transformed by taking its logarithm.

- gives a "percentages" interpretation.

1. Polynomials in X

- Let r denote the highest power of X that is included in the regression. A polynomial regression model of degree r is:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 {X_i}^2 + \cdots + \beta_r {X_i}^r + u_i \quad (9)$$

- this is just the linear multiple regression model – except that the regressors are now powers of the single independent variable X.

- estimation, hypothesis testing, etc. proceeds as with the multiple regression model.

- Let's see how to model a cubic regression function ($r = 3$)

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 {X_i}^2 + \beta_3 {X_i}^3 + u_i \quad (10)$$

unknown coefficients β_0, β_1, β_2, β_3 are estimated using an OLS regression of Y on a $constant$, X, X^2, and X^3.

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
------------------------------------------------------------------------------
```

$$\widehat{TestScore} = \underset{(5.1)}{600.1} + \underset{(0.71)}{5.02}Income - \underset{(0.029)}{0.096}Income^2 + \underset{(0.00035)}{0.00069}Income^3$$

Test the null hypothesis that the population regression function is linear, against the alternative that it is either a quadratic or a cubic, that is, it is a polynomial of degree up to 3:

H_0: population coefficients on $Income^2$ & $Income^3 = 0$

H_1: at least one of these population coefficients is non-zero.

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

Reject hypothesis that the population regression is linear at the 1% level in favour of the alternative that it is a polynomial of degree up to 3.

$$\widehat{TestScore} = 600.1 + 5.02Income - 0.096Income^2 + 0.00069Income^3$$
$$(5.1) \quad (0.71) \quad (0.029) \quad (0.00035)$$

Test the null hypothesis that the population regression function is a quadratic, against the alternative that it is a cubic:

H_0: population coefficient on $Income^3 = 0$

H_1: population coefficient on $Income^3 \neq 0$

t-statistic testing $\beta_{Income^3} = 0$ is $\frac{0.00069-0}{0.00035} = 1.98$

Since $|t^{act}| = 1.98 > 1.96$, reject the null in favour of the alternative at the 5% significance level.

Interpretation of Coefficients in Polynomial Regression Models

- Coefficients in polynomial regressions have complicated interpretations.

- To interpret polynomial regressions:

1) Plot the estimated regression function to see if it fits the data well (i.e. ***plot*** the ***predicted values of Y*** as a function of the ***independent variable X***).

2) Calculate the estimated effect on Y associated with a change in X ***for different values of X.***

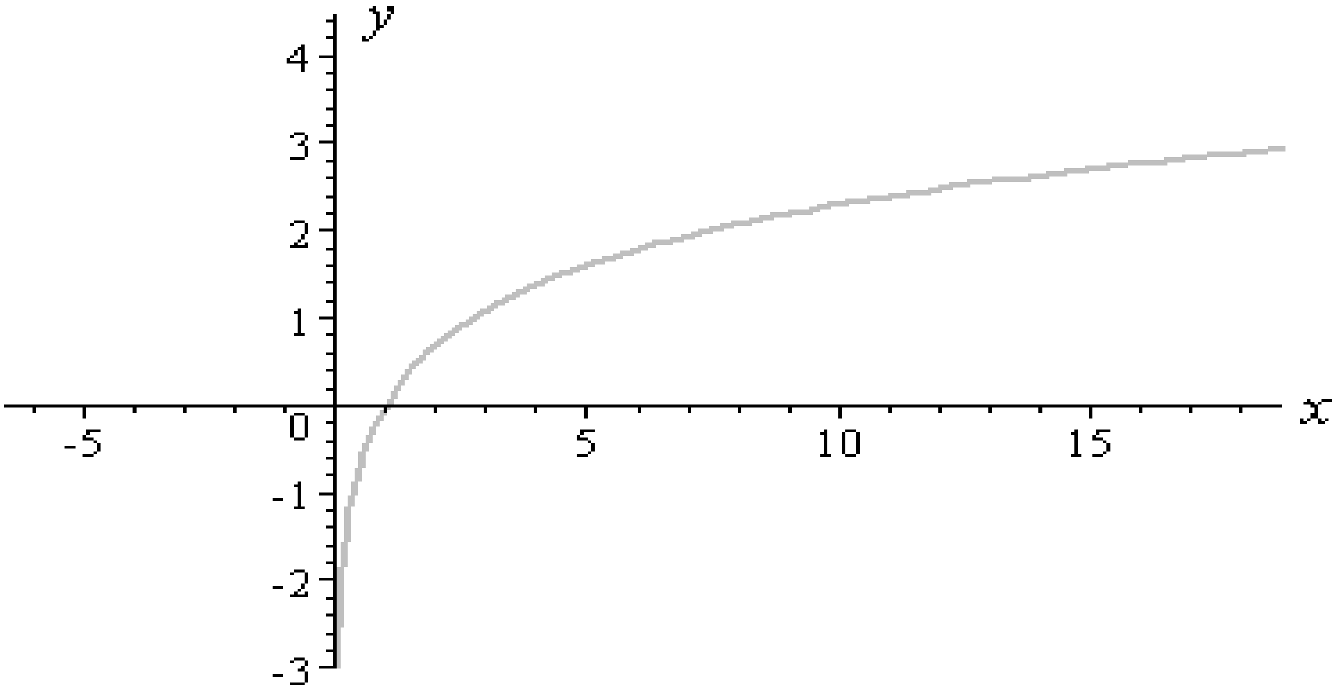
- apply the general "before and after" rule learnt earlier: "calculate the change in Y for a given change in X."

2. Logarithmic Functions of *Y* and/or *X*

- Another way to model a nonlinear relationship is to use the ***natural log***.
- Logarithmic transformations permit modeling relationships in percentage terms.
- Recall micro?
  - price elasticity of demand measures how much quantity demanded changes in percentage terms due to a 1% change in price.
  - So the relationship between quantity demanded and price is linear when both are measured in terms of ***percentages.***
- In fact, many relationships in Economics are naturally expressed in percentage terms.
  - How do wages change with years of service?

Logarithms & Percentages

$$\ln(X) = \text{natural log of } X$$



The link between the log & percentages is based on the fact that when ΔX is small,

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

E.g:

$X=100, \Delta X=1$

$$\ln(100+1)-\ln(100)=0.00995\cong\frac{1}{100}=0.01$$

$X=100, \Delta X=5$

$$\ln(100+5)-\ln(100)=0.04879\cong\frac{5}{100}=0.05$$

Three Log Regression Models

Case	Population Regression Model
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	$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$

- Interpretation of the slope coefficient differs in each case.

1. Linear-Log Population Regression Function

Case 1: X is in logs, Y is not

$$Y_i = \beta_0 + \beta_1 ln(X_i) + u_i \qquad (11)$$

population regression function is:

$$Y = \beta_0 + \beta_1 ln(X) \qquad (12)$$

consider changing X by ΔX. value of the population regression function becomes:

$$Y + \Delta Y = \beta_0 + \beta_1 ln(X + \Delta X) \qquad (13)$$

(13)-*(12)*: $\Delta Y = \beta_1[ln(X + \Delta X) - ln(X)] \cong \beta_1 \frac{\Delta X}{X}$

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

- If X changes by 1%, then $\frac{\Delta X}{X} = 0.01$

- So a 1% change in X, in this model, is associated with a change of Y of $0.01\beta_1$

  - If X changes by 1%, $\Delta Y \cong \beta_1\, 0.01$

E.g. $Testscore$ vs $ln(Income)$

- Instead of a polynomial specification, we could use the linear-log specification to model the relationship between test scores & district income.

1) First, create a new variable $\ln(Income)$

2) The model is now linear in $\ln(Income)$, so the linear-log model can be estimated by OLS:

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

- a 1% increase in income is associated with an increase in testscores of $0.01 \times 36.42 = 0.36$ points

E.g.

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

What is the estimated change in Testscore associated with a change in district income from $10,000 to $11,000?

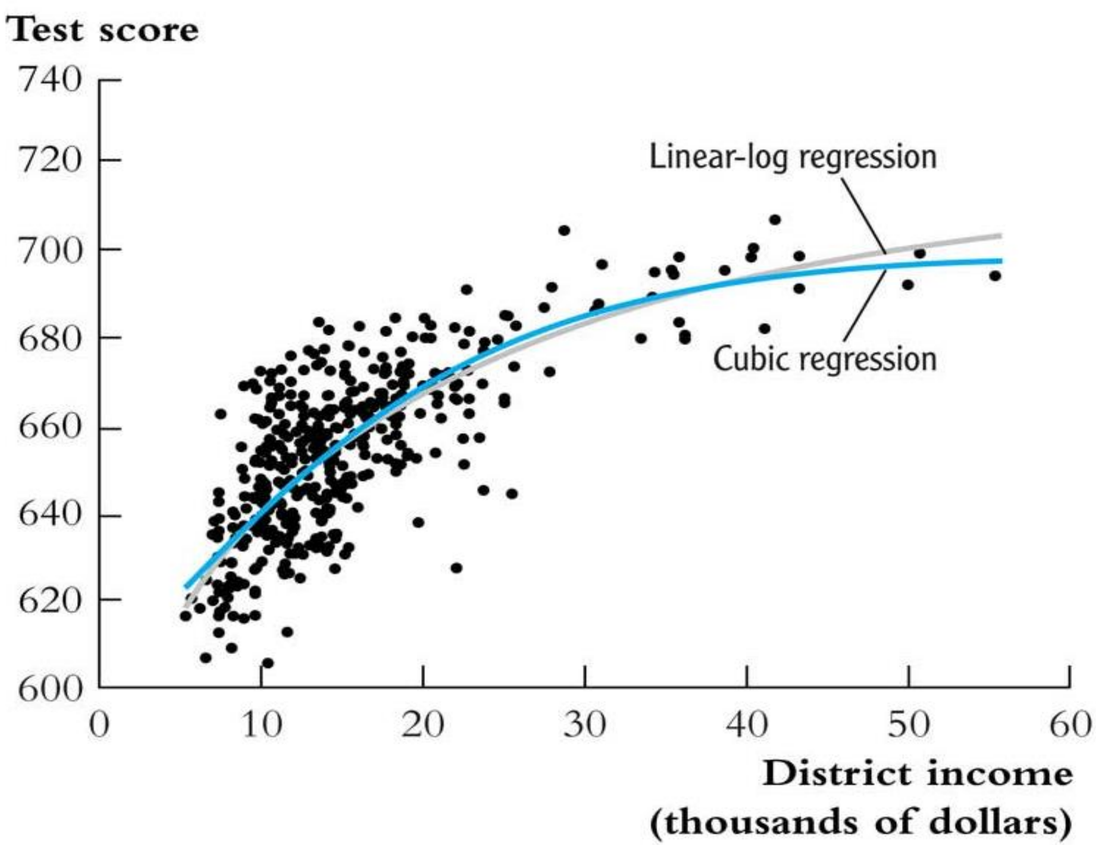
1) predicted testscore when income = $11,000

$$[557.8 + 36.42\ln(11)]$$

2) predicted testscore when income = $10,000

$$[557.8 + 36.42\ln(10)]$$

$$\Delta\widehat{TestScore} = [557.8 + 36.42\ln(11)] - [557.8 + 36.42\ln(10)] = 3.47$$


Test score
740
720
700
680
660
640
620
600
Linear-log regression
Cubic regression
0
10
20
30
40
50
60
District income
(thousands of dollars)

- The two (cubic & linear-log) estimated regression functions are quite similar.
- A way to choose between the two is to use the $\bar{R}^2$
  - $\bar{R}^2$ measures how well the regression function **fits** the data. We can use $\bar{R}^2$ to compare between specifications here because the dependent variable in both regressions are the same – testscores in "level" form.
  - $\bar{R}^2$ of the linear-log regression = 0.561
  - $\bar{R}^2$ of the cubic regression = 0.555
- linear-log specification fits the data (slightly) better.

2. Log-linear Population Regression Function

Case 2: Y is in logs, X is not

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i \qquad (14)$$

population regression function is:

$$\ln(Y) = \beta_0 + \beta_1 X \qquad (15)$$

consider changing X by ΔX. value of the population regression function becomes:

$$\ln(Y + \Delta Y) = \beta_0 + \beta_1 (X + \Delta X) \qquad (16)$$

(16)-(15): $\ln(Y + \Delta Y) - \ln(Y) = \beta_1 \Delta X$

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

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

- if X changes by one unit ($\Delta X = 1$), then $\frac{\Delta Y}{Y} = \beta_1$.

- Since $\frac{\Delta Y}{Y} \times 100\%$ is the percentage change in Y, so if X changes by one unit, then the percentage change in Y is $(\beta_1 \times 100)\%$

- Estimation of a log-linear model proceeds in the same way as with a linear-log model.

E.g. $ln(TestScore)$ vs $Income$

$$\ln(\widehat{TestScore}) = \underset{(0.003)}{6.439} + \underset{(0.00018)}{0.00284}Income$$

- according to the estimated regression, an increase in the district income by \$1,000 is associated with an increase in test scores of $0.00284 \times 100\% = 0.28\%$.

E.g. $ln(TestScore)$ vs $Income$

$$\ln(\widehat{TestScore}) = \underset{(0.003)}{6.439} + \underset{(0.00018)}{0.00284} Income$$

- an increase in the district income by $1,000 is associated with an increase in test scores of $0.00284 \times 100\% = 0.28\%$.
- What is the estimated change in testscore associated with a change in district income from $10,000 to $11,000?
  - Testscore increases by 0.28%.
  - Do not use exponential function to calculate change in predicted testscore (page 323 TB).

3. Log-log Population Regression Function

Case 3: Both Y and X are in logs

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i \qquad (17)$$

population regression function is:

$$\ln(Y) = \beta_0 + \beta_1 \ln(X) \qquad (18)$$

consider changing X by ΔX. value of the population regression function becomes:

$$ln(Y + \Delta Y) = \beta_0 + \beta_1 ln(X + \Delta X) \qquad (19)$$

(19)-(18): $l\,n(Y + \Delta Y) - \ln(Y) = \beta_1[ln(X + \Delta X) - ln(X)]$

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

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

- Since $\frac{\Delta Y}{Y} \times 100\%$ is the percentage change in Y and $\frac{\Delta X}{X} \times 100\%$ is the percentage change in X, so

$$\frac{\Delta Y}{Y} \times 100\% = \beta_1 \frac{\Delta X}{X} \times 100\%$$

- So a 1% change in X is associated with a β_1% change in Y.

E.g. $ln(TestScore)$ vs $ln(Income)$

- Let's use a log-log specification to model the relationship between test scores & district income.

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

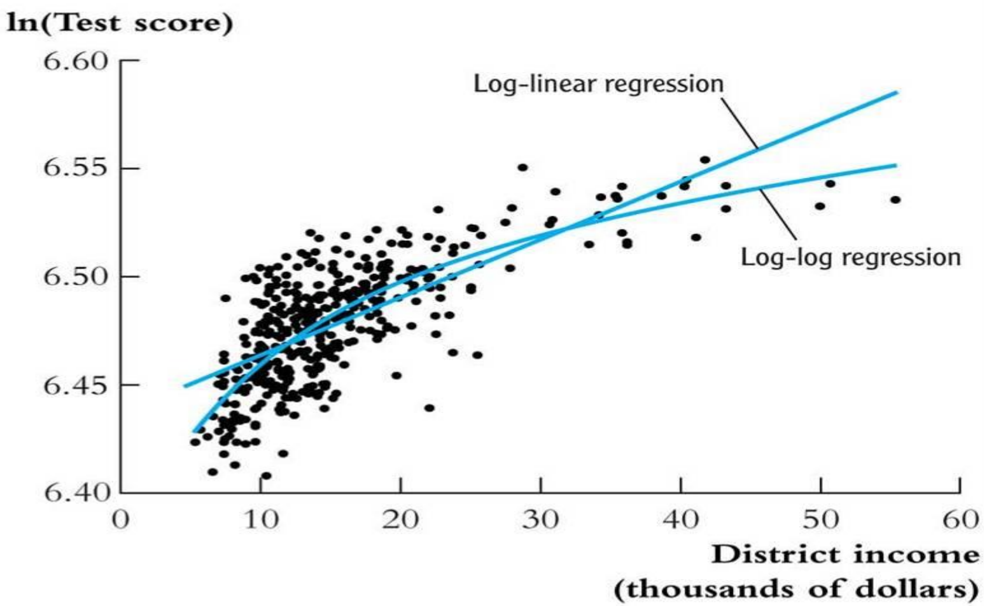
according to this estimated regression, a 1% increase in income is associated with a 0.0554% increase in test scores.

E.g. $ln(TestScore)$ vs $ln(Income)$

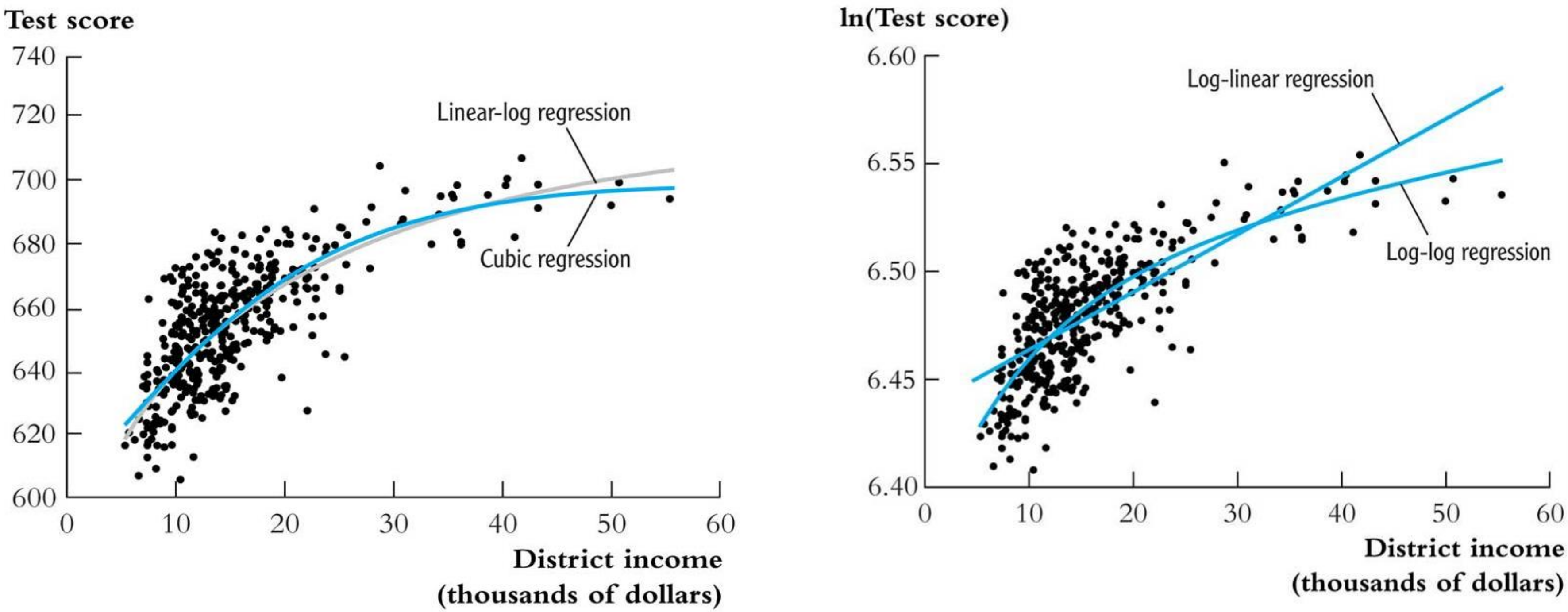
$$\ln(\widehat{TestScore}) = \underset{(0.006)}{6.336} + \underset{(0.0021)}{0.0554}\ln(Income)$$

- a 1% increase in income is associated with a 0.0554% increase in test scores.
- What is the estimated change in testscore associated with a change in district income from \$10,000 to \$11,000?
  - Here, change in district income = 10%
  - So change in testscore is $(10 \times 0.0554)\% = 0.554\%$.
  - Do not use exponential function to calculate change in predicted testscore.

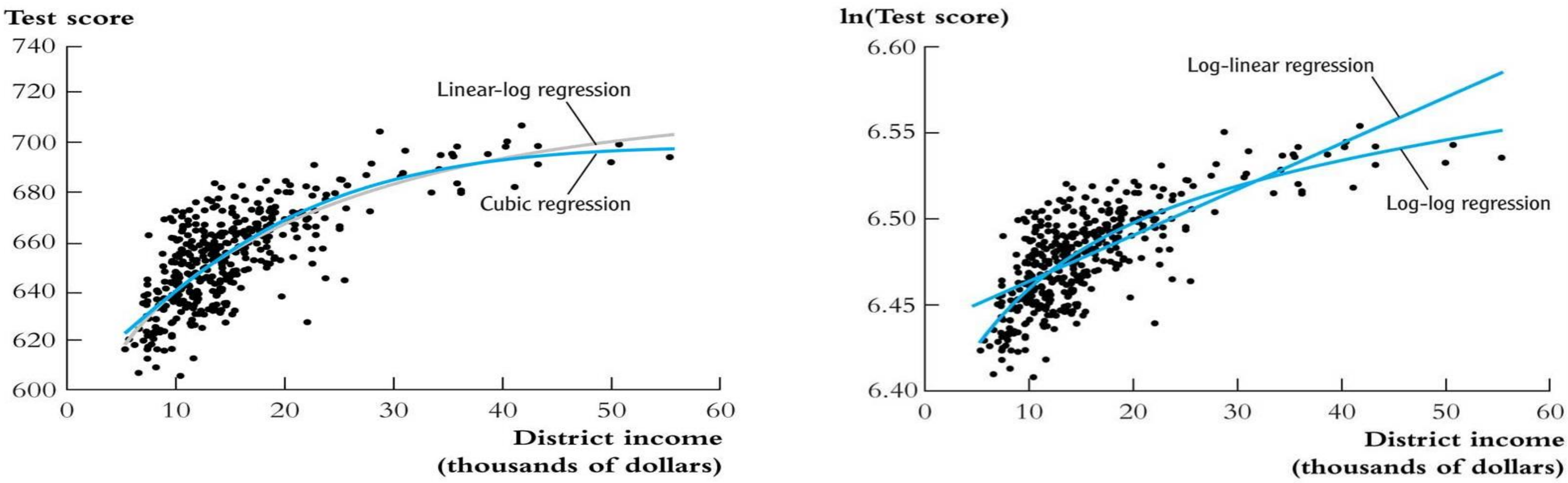
- Which specification fits the data better? log-log or log-linear?



- can use $\bar{R}^2$ to compare.
  - $\bar{R}^2$ of the log-log regression = 0.557
  - $\bar{R}^2$ of the log-linear regression = 0.497
- log-log specification fits better than the log-linear specification.



Can we also use the $\bar{R}^2$ to compare between the log-log & linear-log specifications to see which fits the data best?



No!

<u>Why?</u>

- Because their dependent variables are different [one is in levels "Y", the other is in logs "$\ln(Y)$"]
- Recall $\bar{R}^2$.
- Because the dependent variables in the log-log & linear-log specifications are different, it makes no sense to compare their $\bar{R}^2$s.

- How then do we decide whether the log-log or linear-log specification fits better in practice?

1) Use economic theory

2) Ask if it makes sense to specify Y in logs

E.g.

- relationship between wages & years of service.

- might make sense to use a log-linear specification in this application.

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.

    - The variable being transformed must be positive.

- Hypothesis tests and confidence intervals are implemented and interpreted "as usual".

- The interpretation of β_1 differs in each case.

Case	Population Regression Model	Interpretation of β_1
I. linear-log	$Y_i = \beta_0 + \beta_1 ln(X_i) + u_i$	A 1% change in X is associated with a change in Y of $0.01\beta_1$
II. log-linear	$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$	A change in X by one unit ($\Delta X = 1$) is associated with a $100\beta_1$% change in Y
III. log-log	$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$	A 1% change in X is associated with a β_1% change in Y

Next Lecture

- Which degree polynomial should you use in practice?