ECOM20001
Econometrics 1

Lecture Note 8
Nonlinear Regression

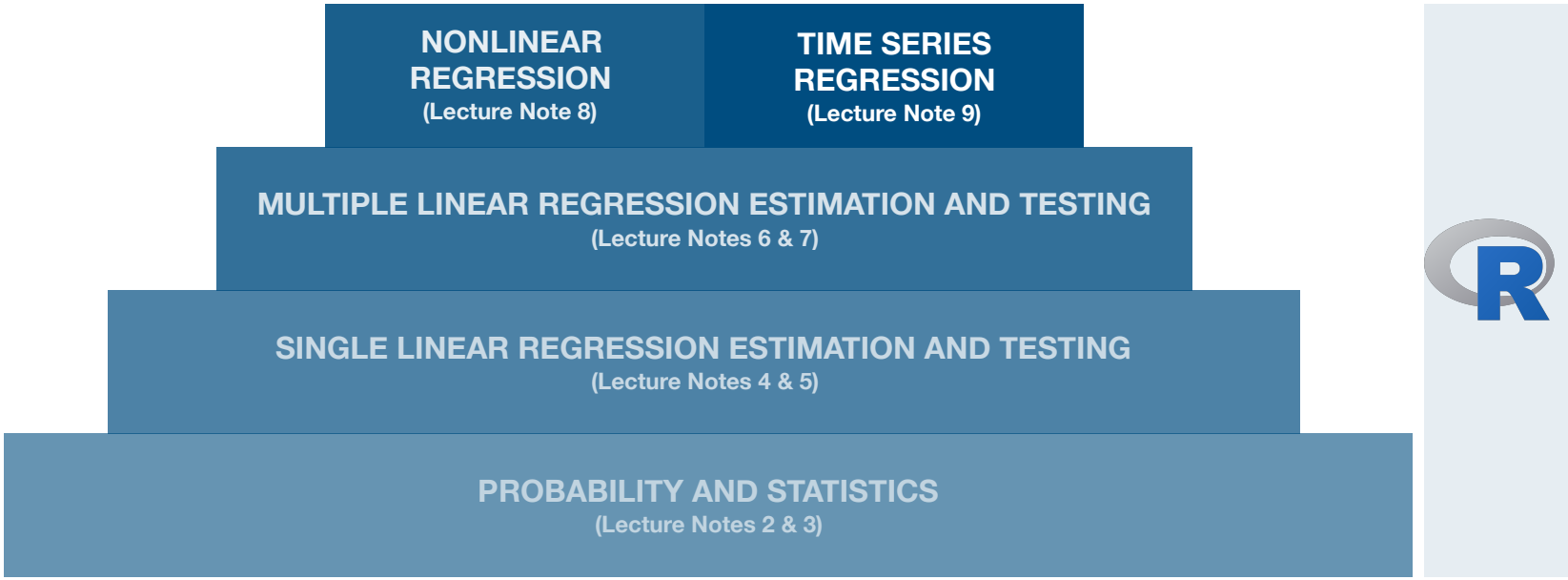
A/Prof David Byrne
Department of Economics
University of Melbourne

Stock and Watson: Chapter 8

Summary of Key Concepts

- Economics of immigration motivating example
- Modeling nonlinear regression functions
- Quadratic regression
- General framework for estimating and testing nonlinear regression models
- Standard errors for estimated nonlinear effects
- Polynomial regression functions
- Logarithmic regression functions: log-linear, linear-log, log-log models
- Interactions between independent variables: binary-binary, continuous-binary, continuous-continuous models
- Polynomial or logarthmic models with interactions
- Difference-in-Differences

Building our Econometric Toolkit



Economics of Immigration

- There is a substantial amount of empirical research in economics around the economics of immigration
- One of the fundamental research questions this work addresses is the following:

  How many years, since first arrival in a country, does it take for an immigrant to catch up to a similar native resident in terms of earnings?

- Let's break this sentence down in terms of econometrics
  - Variable of interest (X_{1i}): "years, since first arrival in a country"
  - Outcome variable (Y_i): "earnings"
  - Controls ($X_{2i}, \ldots, X_{ki}$): "for an immigrant to catch up to a similar native resident"

Economics of Immigration

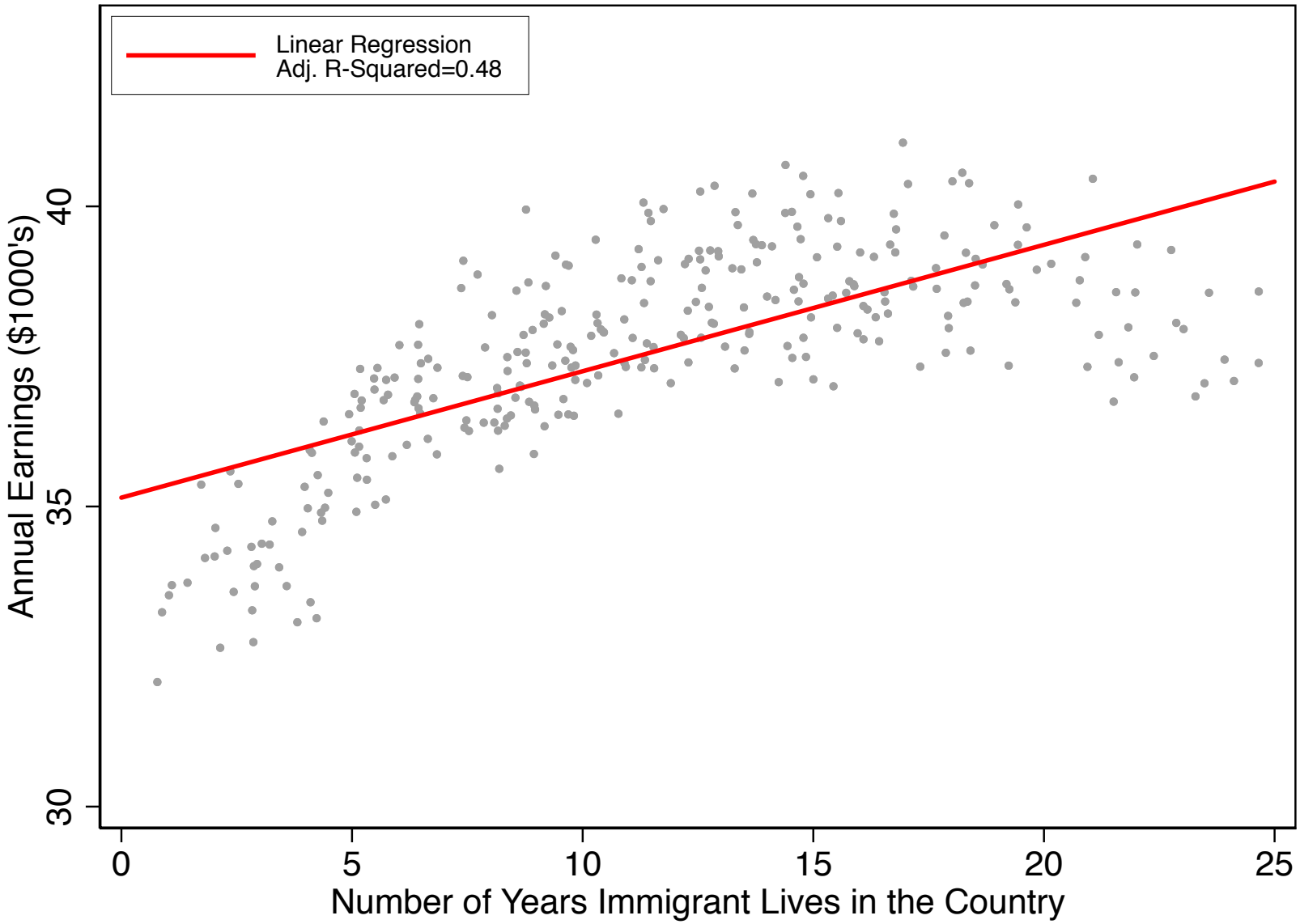
- To begin investigating this question empirically, let's consider a dataset that contains $i = 1, \ldots, n$ individuals
- The data consists of both immigrants and natives, and has the following two variables:
  - $Earnings_i$: annual earnings in \$1000's
  - $Immigrant_i$: dummy=1 individual i is an immigrant, 0 if native
  - $Nyears_i$: number of years an immigrant i has been in the country; set to 0 if individual i is native
  - Age_i: age of individual i
  - $Male_i$: dummy=1 if individual i is male, 0 otherwise
  - $Educ_i$: dummy=1 if individual i has an undergrad degree, 0 otherwise

Economics of Immigration

- We could begin investigating how immigrant earnings "catch up" over time to native earnings by looking at a scatter plot with $Earnings_i$ and $Nyears_i$ among the immigrants in the sample (e.g., all data points where $Immigrant_i = 1$)
- Also plot the estimated regression function from the single linear regression among immigrants:

$$Earnings_i = \beta_0 + \beta_1 Nyears_i + u_i$$

Economics of Immigration



Shortcomings of the Simple Linear Regression

- The estimated OLS regression line indeed quantifies the positive relationship between $Earnings_i$ and $Nyears_i$
- However, a peculiar feature of the graph is that the data points are below the OLS line when $Nyears_i$ is very small and very large
- There appears to be a curvature in the relationship between $Earnings_i$ and $Nyears_i$ that our simple linear regression is not capturing
- In short, $Earnings_i$ and $Nyears_i$ has a nonlinear relationship

Modeling Nonlinear Regression Functions

- We now lay out a general strategy for modeling nonlinear regression functions
  - Estimating a nonlinear relationship
  - Plotting a nonlinear relationship
  - Testing a nonlinear relationship
  - Computing the effect of X on Y when there is a nonlinear relationship
- Throughout, we will work through our example of modeling the relationship $Earnings_i$ and $Nyears_i$ among immigrants to illustrate the main concepts

Quadratic Regression

- To improve our model's ability to predict the $Earnings_i$ and $Nyears_i$ relationship, we want the fitted curve to be steep when $Nyears_i$ is small, and flat when $Nyears_i$ is large
- A quadratic population regression model can model this:

$$Earnings_i = \beta_0 + \beta_1 Nyears_i + \beta_2 Nyears_i^2 + u_i$$

which has the population regression function

$$E(Earnings_i|Nyears_i) = \beta_0 + \beta_1 Nyears_i + \beta_2 Nyears_i^2$$

Quadratic Regression

- Modeling a nonlinear relationship might sound complicated
- However, notice that the quadratic regression is just a multiple linear regression that adds $Nyears_i^2$ as a regressor to the simple linear regression

$$Earnings_i = \beta_0 + \beta_1 Nyears_i + u_i$$

- So to model the nonlinear relationship, we take two steps:
  1. construct $Nyears_i^2$ variable in your dataset by multiplying:

$$Nyears_i^2 = Nyears_i \times Nyears_i$$

  2. Run the multiple linear regression

$$Earnings_i = \beta_0 + \beta_1 Nyears_i + \beta_2 Nyears_i^2 + u_i$$

Comparing Model Fit

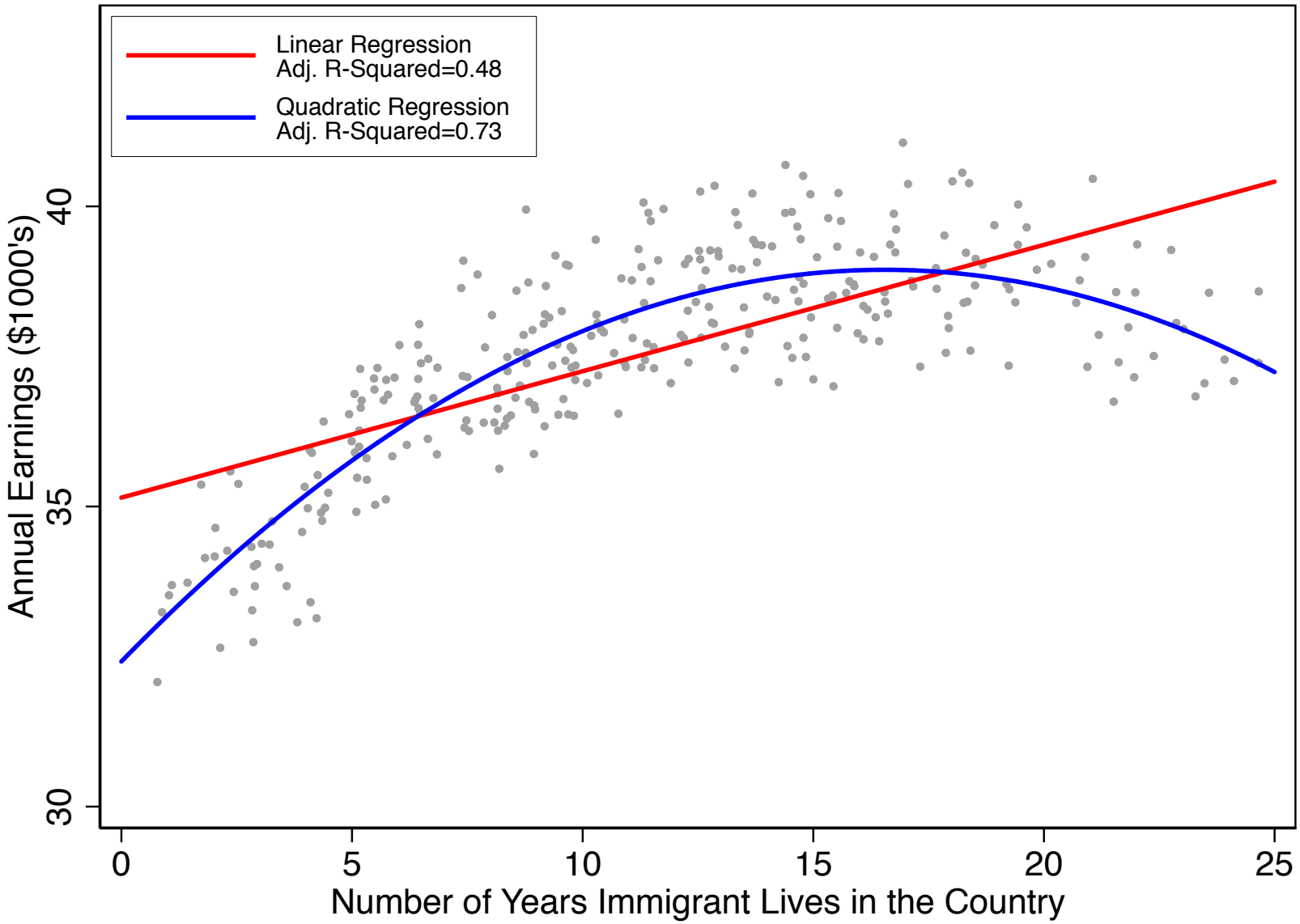
- Using our, here are regression results for the linear and quadratic relationships between $Earnings_i$ and $Nyears_i$ based on the subsample immigrants ($Immigrant_i = 1$):

$$\widehat{Earnings}_i = \underset{(0.19)}{35.15} + \underset{(0.02)}{0.21}\, Nyears_i, \bar{R}^2 = 0.48$$

$$\widehat{Earnings}_i = \underset{(0.19)}{32.42} + \underset{(0.03)}{0.79}\, Nyears_i - \underset{(0.001)}{0.02}\, Nyears_i^2, \bar{R}^2 = 0.73$$

- There are a number of notable differences between the models, the first of which is the large jump in the $\bar{R}^2$
- The quadratic regression model fits the data much better

Economics of Immigration



Testing Nonlinear Relationships

$$\widehat{Earnings}_i = \underset{(0.19)}{32.42} + \underset{(0.03)}{0.79}\, Nyears_i - \underset{(0.001)}{0.02}\, Nyears_i^2, \bar{R}^2 = 0.73$$

- We can also formally test whether there is a nonlinear relationship between $\widehat{Earnings}_i$ and $Nyears_i$ by testing $H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$
- That is, under the null there is a linear relationship between $\widehat{Earnings}_i$ and $Nyears_i$
- Doing so yields a t-statistic of $t = (-0.02 - 0)/0.001 = -20$ and a p-value$<$0.0001. Reject the null of a linear relationship.

General Nonlinear Regression Framework

- Let's put aside the immigration example for a moment and consider the general regression function:

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

where $f(X_{1i}, X_{2i}, \ldots, X_{ki})$ is a general population nonlinear regression function

- Both the linear and quadratic regression functions are particular cases of a nonlinear regression function
- For instance, with linear regression:

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

- The conditional expectation of Y given the $X_{1i}, X_{2i}, \ldots, X_{ki}$ is also a nonlinear function:

$$E[Y_i | X_{1i}, X_{2i}, \ldots, X_{ki}] = f(X_{1i}, X_{2i}, \ldots, X_{ki})$$

General Nonlinear Regression Framework

- There are two general classes of nonlinear models we could consider for $f(X_{1i}, X_{2i}, \ldots, X_{ki})$

1. Population regression function is a nonlinear function of the regressors X's, but a linear function of the parameters β's.
   Example:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \beta_3 X_{1i} X_{2i} + \beta_4 X_{2i}^2 + \beta_5 \sqrt{X_{3i}} + u_i$$

2. Population regression function is a nonlinear or linear function of the regressors X's, and a nonlinear function of the parameters β's.
   Example:

$$Y_i = \beta_0 + \beta_1 X_{1i} + e^{\beta_2} X_{2i}^2 + \ln(\beta_3) X_{3i} + \beta_4 X_{1i}^2 + u_i$$

- We only consider the first class of regression models that are nonlinear in the X's and linear in the parameters β's

The Effect on Y for a Change in X in Nonlinear Specifications

- So far, with linear regressions, the expected change in Y from a ΔX_1 change in X_1 has simply been $\beta_1 \times \Delta X_1$
- How do we compute partial effects with nonlinear regression functions?
- Remember that we are interested in the change in X_1, ΔX_1 holding $X_2, \ldots, X_k$ constant
- The predicted effect of Y from such a change in a nonlinear regression model is:

$$\Delta Y = f(X_1 + \Delta X_1, X_2, \ldots, X_k) - f(X_1, X_2, \ldots, X_k)$$

The Effect on Y for a Change in X in Nonlinear Specifications

- While we do not know the true population regression function, $f(X_1, X_2, \ldots, X_k)$, we can use a random sample to estimate it using OLS, like we did with the quadratic regression model
- We will denote the OLS estimated nonlinear regression model:

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

- The estimated predicted change in Y from a ΔX_1 change is then given by:

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

- This formula for estimating ΔY from a ΔX_1 change in X_1 holding $X_2, \ldots, X_k$ fixed always works no matter the shape of the regression function $f()$ or the size of ΔX_1

Application: Economics of Immigration

- Let's return to our estimated quadratic regression model:

$$\widehat{Earnings} = \underset{(0.19)}{32.42} + \underset{(0.03)}{0.79}\, Nyears - \underset{(0.001)}{0.02}\, Nyears^2, \bar{R}^2 = 0.73$$

- Let's consider changing *Nyears* from 5 to 6:

$$\Delta\widehat{Earnings} = \left(32.42 + 0.79 \times 6 - 0.02 \times 6^2\right) - \left(32.42 + 0.79 \times 5 - 0.02 \times 5^2\right)$$

$$\rightarrow \Delta\widehat{Earnings} = 0.57 \text{ (or \$570/year)}$$

- Now let's consider changing *Nyears* from 15 to 16:

$$\Delta\widehat{Earnings} = \left(32.42 + 0.79 \times 16 - 0.02 \times 16^2\right) - \left(32.42 + 0.79 \times 15 - 0.02 \times 15^2\right)$$

$$\rightarrow \Delta\widehat{Earnings} = 0.17 \text{ (or \$170/year)}$$

- Notice how the initial level of *Nyears* (e.g. 5 vs 15) shapes the impact of a $\Delta Nyears = 1$ on $\widehat{Earnings}$. There's a diminishing effect for higher values of *Nyears*

Standard Errors of Estimated Effects

- Because the estimated regression function $\hat{f}$ varies from one sample to the next due to random sampling, the predicted value of $\Delta\hat{Y}$ also varies and therefore has a standard error (just like OLS regression coefficients)
- Recall with linear regressions, the estimated effect of a ΔX_1 change in X_1 holding all other regressors fixed was

$$\Delta\hat{Y} = \hat{\beta}_1 \Delta X_1$$

and the standard error was simply

$$SE(\Delta\hat{Y}) = SE(\hat{\beta}_1)\Delta X_1$$

- With nonlinear regression models, we need to take a slightly different approach

Standard Errors of Estimated Effects

- We can compute $SE(\Delta \hat{Y})$ by using methods for jointly testing linear restrictions involving multiple regressors from the previous lecture
- Returning to our example for testing the effect of changing *Nyears* from 5 to 6 on *Earnings*, the general formula (e.g., written in terms of $\hat{\beta}$'s symbols) from our quadratic regression model for $\Delta \hat{Y}$ was:

$$\Delta \hat{Y} = \left(\hat{\beta}_0 + \hat{\beta}_1 \times 6 + \hat{\beta}_2 \times 6^2\right) - \left(\hat{\beta}_0 + \hat{\beta}_1 \times 5 + \hat{\beta}_2 \times 5^2\right)$$

$$\rightarrow \Delta \hat{Y} = \hat{\beta}_1 + 11\hat{\beta}_2$$

- The standard error of the predicted change is therefore:

$$SE(\Delta \hat{Y}) = SE(\hat{\beta}_1 + 11\hat{\beta}_2)$$

Standard Errors of Estimated Effects

- Let F be the F-statistic for testing the hypothesis that

$$\beta_1 + 11\beta_2 = 0$$

  You can compute F using a statistics program like R
- Using F from the test, the standard error can be computed as:

$$SE(\Delta\hat{Y}) = \frac{|\Delta\hat{Y}|}{\sqrt{F}}$$

  See footnote 2 on page 311 of the text for the derivation
- The 95% confidence interval for $\Delta\hat{Y}$ can then be computed as

$$[\Delta\hat{Y} - 1.96SE(\Delta\hat{Y}), \Delta\hat{Y} + 1.96SE(\Delta\hat{Y})]$$

Nonlinear Partial Effects and Standard Errors in 5 Steps

- This 5-step method for computing $SE(\Delta\hat{Y})$ works for any nonlinear regression function that we consider:

1. Choose a level of the variable of interest, X_1
   - in our example: $Nyears = 5$
2. Specify a change in the variable of interest, ΔX
   - in our example: $\Delta Nyears = 1$, so a change in *Nyears* from 5 to 6
3. Write the formula for $\Delta\hat{Y}$ from the nonlinear regression from a change in X_1 to $X_1 + \Delta$ holding $X_2, \ldots X_k$ fixed,

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

   in general form in terms of the estimated $\hat{\beta}$. Plug in the $\hat{\beta}$'s you estimated to compute the partial effect $\Delta\hat{Y}$
   - in our example: $\Delta\hat{Y} = \hat{\beta}_1 + 11\hat{\beta}_2$

Nonlinear Partial Effects and Standard Errors in 5 Steps

4. Using the estimated model, run the joint test of the null that the general formula for $\Delta\hat{Y} = 0$ at the estimated $\hat{\beta}'s$ values and obtain the corresponding F-statistic for the test, F
   - in our example: test $\Delta\hat{Y} = \hat{\beta}_1 + 11\hat{\beta}_2 = 0$
5. Compute the standard error and 95% CI for $\Delta\hat{Y}$:

$$SE(\Delta\hat{Y}) = \frac{|\Delta\hat{Y}|}{\sqrt{F}}$$

$$[\Delta\hat{Y} - 1.96SE(\Delta\hat{Y}), \Delta\hat{Y} + 1.96SE(\Delta\hat{Y})]$$

- This approach always works for computing $SE(\Delta\hat{Y})$ no matter the shape of the nonlinear regression function $f()$ or the size of ΔX_1, where X_1 enters $f()$ as a quadratic/polynomial function

Standard Errors of Estimated Effects Application

- From our example of changing *Nyears* from 5 to 6 with corresponding $\Delta\hat{Y} = 0.57$, we obtain:

Steps 1-3. Predicted change in $\Delta\hat{Y}$ from *Nyears* changing from 5 to 6:

$$\begin{aligned}\Delta\hat{Y} &= \left(\hat{\beta}_0 + \hat{\beta}_1 \times 6 + \hat{\beta}_2 \times 6^2\right) - \left(\hat{\beta}_0 + \hat{\beta}_1 \times 5 + \hat{\beta}_2 \times 5^2\right) \\ &= \hat{\beta}_1 + 11\hat{\beta}_2\end{aligned}$$

Step 4. $F = 810.99$ for joint test of $\hat{\beta}_1 + 11\hat{\beta}_2 = 0$

Steps 5-6. Compute standard error and 95% CI

$$SE(\Delta\hat{Y}) = 0.57/\sqrt{810.99} = 0.02$$

$$95\%\ \text{CI} = [0.57 - 1.96 * 0.02, 0.57 + 1.96 * 0.02] = [0.53, 0.61]$$

- Interpreting the results in words: holding other factors fixed, the predicted effect from changing *Nyears* from 5 to 6 years on earnings is \$570/year, with a 95% CI of [\$530, \$610]

Interpreting Coefficients in Nonlinear Models

- In all of our analyses of single and multiple linear regression models, we focused on directly interpreting regression coefficients from the model because the coefficients had a natural interpretation
- With nonlinear models, it is generally not useful to interpret individual regression coefficients
- For instance, in our regression

$$Earnings_i = \beta_0 + \beta_1 Nyears_i + \beta_2 Nyears_i^2 + u_i$$

  β_1 does not correspond to the corresponding change in *Earnings* from a one-unit change in *Nyears*.
  - As we have seen, the change depends on the level of *Nyears* because of the nonlinearity in the relationship
- In interpreting nonlinear regression models, it is best to graph the nonlinear regression function (like we did in the example), and calculate $\Delta \hat{Y}$ for a given ΔX_1 for different levels of X_1

Summarizing a General Approach to Modeling Nonlinear Relationships Using Multiple Linear Regression

1. Investigate scatter plots for your dependent variable Y, variable of interest X_1 (and possibly key control variables too)
2. Specify different noninear functions and estimate their parameters by OLS
3. Determine whether the nonlinear model improves upon the linear model using $\bar{R}^2$; test the null hypothesis of no nonlinear relationship against the alternative that it is nonlinear
4. Plot the estimated nonlinear regression function
5. Estimate the effect on Y of a change in X_1, $\Delta\hat{Y}$, for different X_1 values; compute $SE(\Delta\hat{Y})$ and the 95% CI for $\Delta\hat{Y}$ using an appropriate F statistic, which depends on the nonlinear regression model, level of X_1 and size of ΔX_1

Nonlinear Functions of a Single Independent Variable

- We will now work through two types of nonlinear regression models of the following form:

$$Y_i = f(X_i) + u_i$$

- That is, nonlinear regressions with a single independent variable, X_{i1}
- We focus on these models to simplify things as we consider different types of regression functions for $f()$
- Once we understand the single independent variable model, it is straightforward to extend to extend the model to one with multiple independent variables:

$$Y_{i1} = f(X_{1,i}, X_{2i}, \ldots, X_{ki}) + u_i$$

- The models for $f()$ we consider are: polynomials and logarithms

Polynomial Regression Functions

- The polynomial regression model of degree r is:

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

- Varying r yields some familiar functions
  - $r = 2$ is a quadratic regression model
  - $r = 3$ is a cubic regression model
  - $r = 4$ is a quartic regression model

  and so on.
- If our dataset contains (Y_i, X_i), we can construct all the higher order regressors in the regression model using X_i:
  - $X_i^2 = X_i \times X_i$
  - $X_i^3 = X_i \times X_i \times X_i$
  - $X_i^4 = X_i \times X_i \times X_i \times X_i$

  and so on

Polynomial Regression Functions

- We estimate a polynomial regression function using OLS as these regressions correspond to a multiple linear regression
- Similarly, all of our techniques for assessing model fit, model testing, and computing confidence intervals from multiple linear regression directly apply to polynomial regressions
- An important joint test to conduct with polynomial regressions is whether the population regression function is linear:
  - $H_0 : \beta_2 = 0, \beta_3 = 0, \ldots, \beta_r = 0$
  - H_1: at least one of $\beta_j \neq 0$, $j = 2, \ldots, r$
- This test imposes $q = r - 1$ restrictions under the null, and can be implemented using the F-statistic as discussed in Lecture Note 7.

Polynomial Regression Functions

- Which degree polynomial should I use?
  - In other words, what value should I choose for r?
- There is a tradeoff to consider when you increase r, and hence add additional regressors to the nonlinear regression model
  - flexibility: as you make r higher, you make your regression function model flexible, thereby allowing you to fit more complicated curves to data
    - a polynomial of degree r can fit up to $r - 1$ bends (or inflection points) in its graph
  - precision: how as you increase r, you have more regressors meaning lower degrees of freedom and less precise regression coefficient estimates with larger standard errors
    - a common problem with making r too high in practise is you run into imperfect mulitcollinearity problems (see textbook p 251)

Polynomial Regression Functions

- Sequential hypothesis testing is often used in practice to determine r in polynomial regressions:
  1. Pick a max value for r and estimate the polynomial regression model of degree r
  2. Using the t-statitic to test $\beta_r = 0$. If you reject this hypothesis, then X^r belongs in the regression, and use a polynomial of degree r and STOP.
  3. If you fail to reject $\beta_r = 0$, eliminate X^r from the regression, and estimate a polynomial of degree $r - 1$
  4. Using the t-statitic to test $\beta_{r-1} = 0$. If you reject this hypothesis, then X^{r-1} belongs in the regression, and use a polynomial of degree $r - 1$ and STOP.

Polynomial Regression Functions (cont.)

5. If you fail to reject $\beta_{r-1} = 0$, eliminate X^{r-1} from the regression, and estimate a polynomial of degree $r - 2$ ...
6. Continue this procedure until the coefficient on the highest power in your polynomial regression is statistically significant, then STOP.

- Typically we start with a maximum r value of $r = 4$ or $r = 5$ and work backwards from there

Polynomial Regression Functions: Application

- Polynomial regression for $r = 4$:

$$\widehat{Earnings}_i = \underset{(0.19)}{32.42} + \underset{(0.19)}{0.88}\,Nyears_i - \underset{(0.03)}{0.04}\,Nyears_i^2 + \underset{(0.005)}{0.0017}\,Nyears_i^3 - \underset{(0.0001)}{0.00003}\,Nyears_i^4,\ \bar{R}^2 = 0.73$$

  $\rightarrow$ p-value for the test of $\beta_4 = 0$ is 0.79

- Polynomial regression for $r = 3$:

$$\widehat{Earnings}_i = \underset{(0.19)}{32.42} + \underset{(0.09)}{0.84}\,Nyears_i - \underset{(0.01)}{0.03}\,Nyears_i^2 + \underset{(0.001)}{0.0002}\,Nyears_i^3,\ \bar{R}^2 = 0.73$$

  $\rightarrow$ p-value for the test of $\beta_3 = 0$ is 0.58

- Polynomial regression for $r = 2$:

$$\widehat{Earnings}_i = \underset{(0.19)}{32.42} + \underset{(0.03)}{0.79}\,Nyears_i - \underset{(0.001)}{0.02}\,Nyears_i^2,\ \bar{R}^2 = 0.73$$

  $\rightarrow$ p-value for the test of $\beta_2 = 0$ is < 0.001 STOP

Polynomial Regression Functions: Application

- In our application, notice how the standard errors on regression coefficients fall as r falls
- Also notice how $\bar{R}^2$ is unchanged across regressions; higher order terms in the polynomial for $r = 3$ and $r = 4$ do little to improve model fit
- The significant digits are mixed with the polynomials, which is common. This is because regressors like $Nyears^3$ and $Nyears^4$ are huge relative to $Nyears$ and $Nyears^2$, which causes their regression coefficients to be very small
- Interpretation: to interpet the results from polynomials, graph the estimated function at the chosen value of r, and evaluate the impact of ΔX on Y for different levels of X
- Re-emphasising: do not interpet individual regression coefficients as they are individually meaningless in nonlinear models

Logarithms

- We can also use $\ln(x)$, logarithmic function or logarithms, to model nonlinear relationships
- The logarithmic function is closely related to the exponential function, e^x, which is e to the power of x, where e is the constant 2.71828...
- The natural logarithm (or just logarithm) is the inverse of the exponential function:

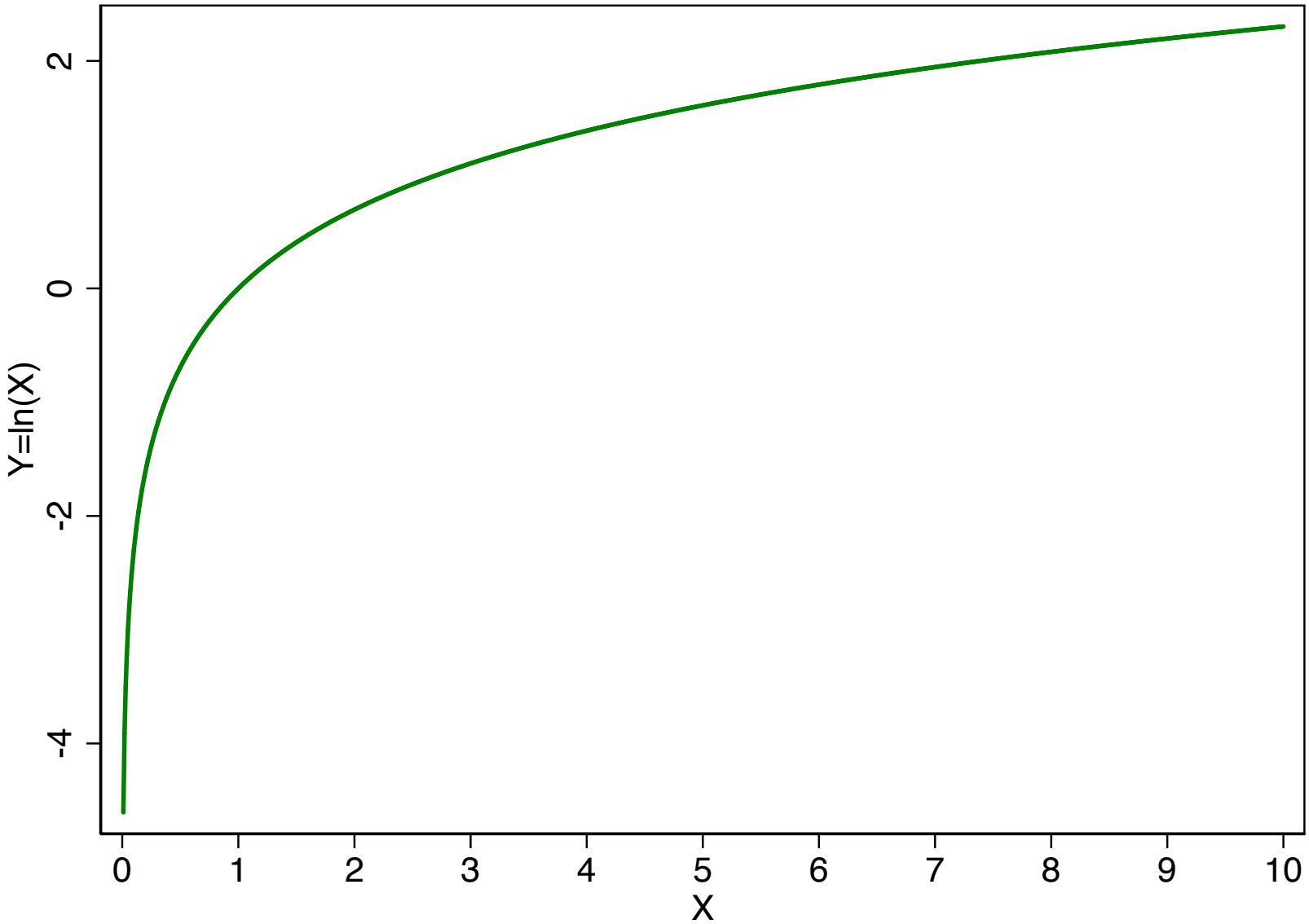
$$x = \ln(e^x)$$

- $\ln(x)$ is defined for $x > 0$, with $\ln(x) < 0$ if $x < 1$; $\ln(x) = 0$ if $x = 1$; and $\ln(x) > 0$ if $x > 1$
- The slope of the $\ln(x)$ function is:

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

That is, the slope of $\ln(x)$ is always positive, is large for small values of x, and small for large values of x

Logarithms



Logarithms properties

- For 2 numbers, a and x, the following properties hold:
  - $\ln(1/x) = -\ln(x)$
  - $\ln(ax) = \ln(a) + \ln(x)$
  - $\ln(x/a) = \ln(x) - \ln(a)$
  - $\ln(x^a) = a\ln(x)$

Logarithms

- Logarithms also have a very close connection with percentage changes and elasticities, which makes them extreme useful for econometric analysis
- For small changes in x, Δx:

$$\ln(x + \Delta x) - \ln(x) \approx \frac{\Delta x}{x}$$

  where $\approx$ means "approximately equal to", and where $\frac{\Delta x}{x}$ is the percentage change in x divided by 100
- Example: if $x = 100$ and $\Delta x = 1$, then $\Delta x/x = 0.01$, or a 1% change. Similarly, $\ln(101) - \ln(100) = 0.00995$ or a 0.995% change, which is very close to 1%

Logarithmic Econometric Models

- There are three key logarithmic models often used in econometric practice

1. Linear-Log model:

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

2. Log-Linear model:

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

3. Log-Log model:

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

- We consider these models' use and interpretation in turn.
- Each of these models are estimated via OLS, are tested using standard methods for single and multiple linear regression, and confidence intervals for regression coefficients are computed by inverting t-statistics as as previously discussed

Linear-Log Model

- Model where X is in logarithms and Y is not:

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

- Estimation: If we have a dataset with (X_i, Y_i) for $i = 1, \ldots,$ we can estimate this model by constructing a new variable in the dataset $W_i = \ln(X_i)$, and then regress Y_i on W_i
- Interpretation: A 1% change in X is associated with a change in Y of $\Delta Y = 0.01\beta_1$

Linear-Log Model

- Deriving the partial effect in the Linear-Log Model, ΔY:

$$E[Y|X] = \beta_0 + \beta_1 \ln(X)$$
$$E[Y|X + \Delta X] = \beta_0 + \beta_1 \ln(X + \Delta X)$$

which implies

$$\begin{aligned}
\Delta Y &= E[Y|X + \Delta X] - E[Y|X] \\
&= (\beta_0 + \beta_1 \ln(X + \Delta X)) - (\beta_0 + \beta_1 \ln(X)) \\
&= \beta_1 (\ln(X + \Delta X) - \ln(X)) \\
&\approx \beta_1 \frac{\Delta X}{X} = 0.01\beta_1 \text{ (if } X \text{ changes by 1\%)}
\end{aligned}$$

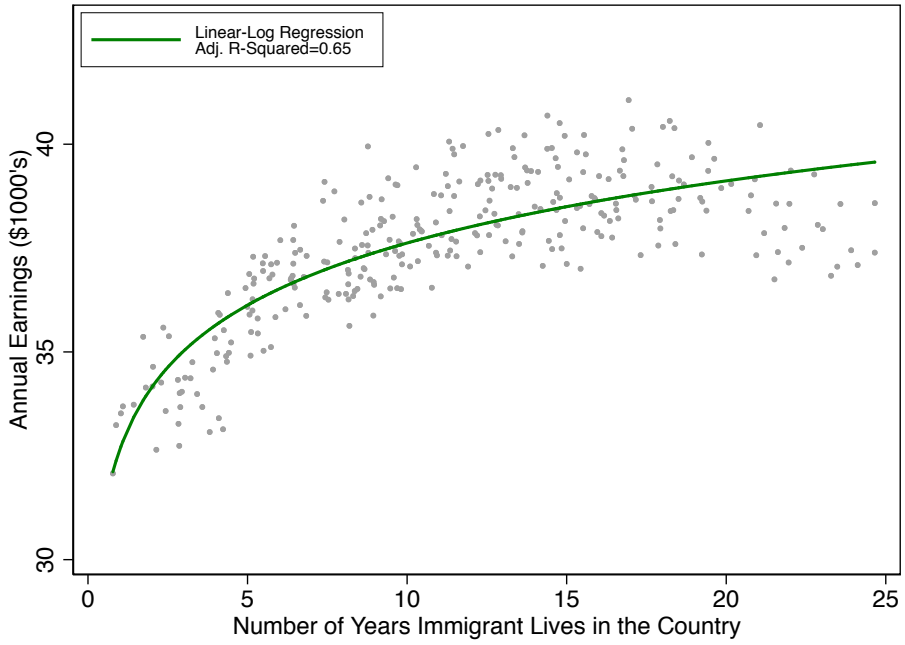
Linear-Log Model Application

- Applying the linear-log model to the immigration data:

$$\widehat{Earnings}_i = \underset{(0.22)}{32.65} + \underset{(0.10)}{2.16}\ln(Nyears_i), \bar{R}^2 = 0.65$$

- Interpretation: a 1% increase in $Nyears_i$ is associated with a $0.01 \times 2.16 = 0.0216$. Recalling Earnings is in \$1000's, this partial effect is equivalent to a \$1000 $\times$ 0.0216=\$21.60/year increase in earnings from a 1% increase in $Nyears_i$

- Linear-Log Model graphically:



Log-Linear Model

- Model where Y is in logarithms and X is not:

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

- Estimation: If we have a dataset with (X_i, Y_i) for $i = 1, \ldots,$ we can estimate this model by constructing a new variable in the dataset $Z_i = \ln(Y_i)$, and then regress Z_i on X_i
- Interpretation: A $\Delta X = 1$ unit change in X is associated with a $\Delta Y = (100\ \beta_1)\%$ change in Y.

Log-Linear Model

- Deriving the partial effect in the Log-Linear Model, $\Delta \ln(Y)$:

$$E[\ln(Y)|X] = \ln(Y) = \beta_0 + \beta_1 X$$
$$E[\ln(Y)|X + \Delta X] = \ln(Y + \Delta Y) = \beta_0 + \beta_1(X + \Delta X)$$

  which implies:

$$\begin{aligned}\Delta \ln(Y) &= E[\ln(Y)|X + \Delta X] - E[\ln(Y)|X] \\ &= \ln(Y + \Delta Y) - \ln(Y) \\ &= (\beta_0 + \beta_1(X + \Delta X)) - (\beta_0 + \beta_1 X) = \beta_1 \Delta X\end{aligned}$$

- Further, if ΔY is small, then $\ln(Y + \Delta Y) - \ln(Y) \approx \frac{\Delta Y}{Y}$.
- So, given the above, if $\Delta X = 1$, then $\frac{\Delta Y}{Y} = \beta_1$. That is, $\Delta X = 1$ is associated with a $100 \times \beta_1 = 100 \times \frac{\Delta Y}{Y}$ percent change in Y

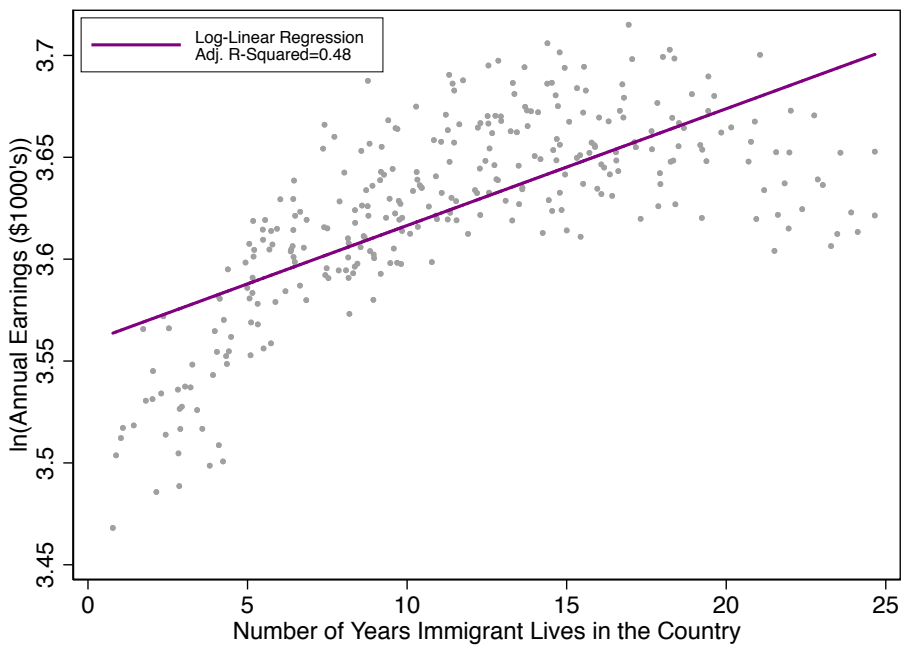
Log-Linear Model Application

- Applying the log-linear model to the immigration data:

$$\widehat{\ln(Earnings_i)} = \underset{(0.005)}{3.60} + \underset{(0.0004)}{0.005}\, Nyears_i,\ \bar{R}^2 = 0.48$$

- Interpretation: a $\Delta X = 1$ year increase in $Nyears_i$ is associated with a $100 \times 0.005 = 0.5\%$ increase in earnings

- Log-Linear Model graphically (note: y-axis is $\ln(Earnings_i)$):



Log-Log Model

- Model where both X and Y are in logarithms:

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

- Estimation: If we have a dataset with (X_i, Y_i) for $i = 1, \ldots,$ we can estimate this model by constructing a new variable in the dataset $Z_i = \ln(Y_i)$, and another new variable $W_i = \ln(X_i)$ then regress Z_i on W_i
- Interpretation: A 1% change in X_i is associated with a β_1% change in Y. That is, β_1 is an elasticity.

Log-Log Model

- Deriving the partial effect in the Log-Log Model, $\Delta \ln(Y)$:

$$E[\ln(Y)|X] = \ln(Y) = \beta_0 + \beta_1 \ln(X)$$
$$E[\ln(Y)|X + \Delta X] = \ln(Y + \Delta Y) = \beta_0 + \beta_1 \ln(X + \Delta X)$$

  which implies:

$$\begin{aligned}\Delta \ln(Y) &= E[\ln(Y)|X + \Delta X] - E[\ln(Y)|X] \\ &= \ln(Y + \Delta Y) - \ln(Y) \\ &= \beta_1 (\ln(X + \Delta X) - \ln(X))\end{aligned}$$

- If ΔX is small, then $\ln(X + \Delta X) - \ln(X) = \frac{\Delta X}{X}$ and $\ln(Y + \Delta Y) - \ln(Y) \approx \frac{\Delta Y}{Y}$ and the equation is approximated by:

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

  or

$$\beta_1 = \frac{\Delta Y/Y}{\Delta X/X} = \frac{100 \times \Delta Y/Y}{100 \times \Delta X/X} = \frac{\text{percent change in } Y}{\text{percent change in } X}$$

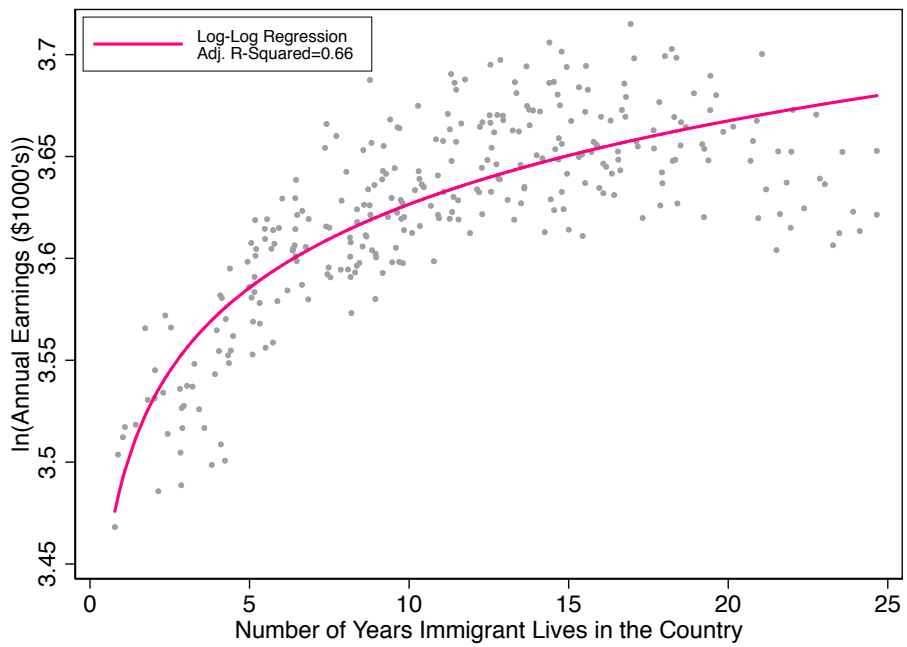
  which is the definition of the elasticity of Y with respect to X

Log-Log Model Application

- Applying the log-log model to the immigration data:

$$\widehat{\ln(Earnings_i)} = \underset{(0.006)}{3.49} + \underset{(0.003)}{0.059} \ln(Nyears_i), \bar{R}^2 = 0.66$$

- Interpretation: a 1% change in $Nyears_i$ is associated with a 0.059% change in $Earnings_i$.
- Log-Log Model graphically (note: y-axis is $\ln(Earnings_i)$):



Summarizing Logarithmic Specifications

1. Linear-Log model:

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

   1% change in X_i is associated with a $0.01\beta_1$ change in Y

2. Log-Linear model:

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

   1 unit change in X_i is associated with a $100\beta_1\%$ change in Y

3. Log-Log model:

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

   1% change in X_i is associated with a $\beta_1\%$ change in Y.
   Equivalently, β_1 is the elasticity of Y with respect to X

Comparing Logarithmic Specifications

- We can compare log-linear and log-log models in terms of model fit since they have the same dependent variable, $\ln(Y_i)$
  - in our example, the log-log model with $\bar{R}^2 = 0.66$ has a better fit than the log-linear model with $\bar{R}^2 = 0.48$
- We can also compare linear regressions with linear-log models because they have the same dependent variable, Y_i
  - in our example, the linear-log model with $\bar{R}^2 = 0.65$ has a better fit than the linear regression model with $\bar{R}^2 = 0.48$
- We cannot compare (linear-log vs. log-linear) nor (linear-log vs. log-log) because they have different dependent variables: Y_i for linear-log, and $\ln(Y_i)$ for log-linear and log-log

Which Model to Use?

- There is no clear-cut rule for deciding whether to use a polynomial regression or a particular type of logarithmic regression. It depends on the context and data at hand.
- When deciding on which specification to use, and whether to have a $\ln(Y_i)$ dependent variable, you want to ask yourself is it clearer to discuss the regression findings and dependent variable in terms of percent changes (and elasticities) or in terms of levels?
  - if percentage changes and elasticities, then log-linear or log-log
  - if levels, then linear-log or polynomials
- Once the best way to discuss the empirics are in terms of percent changes or levels, then look at model fit with linear-log vs log-log or linear-log vs polynomial to determine which is the better fitting model

Adding Controls to a Nonlinear Regression Function

- To hold other factors fixed in estimating a non-linear relationship (to avoid omitted variable bias with your variable of interest), you can add additional control variables to your nonlinear regression model
- For instance, returning to our immigration example, you could estimate the following regression:

$$\begin{aligned} Earnings_i = &\beta_0 + \beta_1 Immigrant_i + \beta_2 Nyears_i + \beta_3 Nyears_i^2 \\ &+ \beta_4 Male_i + \beta_5 Age_i + u_i \end{aligned}$$

  where the coefficients of interest are β_2 and β_3
- If you want added flexibility in your controls, you can include nonlinear functions there too. For example:

$$\begin{aligned} Earnings_i = &\beta_0 + \beta_1 Immigrant_i + \beta_2 Nyears_i + \beta_3 Nyears_i^2 \\ &+ \beta_4 Male_i + \beta_5 Age_i + \beta_6 Age_i^2 + u_i \end{aligned}$$

Adding Controls to a Nonlinear Regression Function

- Alternatively, you could use log-linear model for modeling how immigrants catch-up to natives over time in earnings:

$$\begin{aligned}\log(Earnings_i) =& \beta_0 + \beta_1 Immigrant_i + \beta_2 Nyears_i \\ &+ \beta_3 Male_i + \beta_4 Age_i + u_i\end{aligned}$$

  where the key coefficient of interest is β_2, which has the interpretation that 1 additional year in the country yields a $100\beta_2\%$ change in $Earnings_i$

Adding Controls to a Nonlinear Regression Function

Combining Logarithms and Polynomials with Controls

- You can even combine logarithmic and quadratic models:

$$\begin{aligned}\log(Earnings_i) = &\beta_0 + \beta_1 Immigrant_i + \beta_2 Nyears_i + \beta_3 Nyears_i^2 \\ &+ \beta_4 Male_i + \beta_5 Age_i + u_i\end{aligned}$$

- Polynomials combined with log-linear regressions are popular in earnings-immigration regressions because it is natural to interpret results in terms adding one year to an immigrant's time in a country and earnings growth in terms of percentages

Adding Controls to a Nonlinear Regression Function

Earnings and the Number of Years Since Immigration

	Dep. Var: $Earnings_i$			Dep. Var: $\ln(Earnings_i)$		
	(1)	(2)	(3)	(4)	(5)	(6)
$Immigrant_i$	-10.088**	-10.179**	-10.187**	-0.269**	-0.271**	-0.271**
	(0.176)	(0.123)	(0.120)	(0.005)	(0.004)	(0.003)
$Nyears_i$	0.799**	0.825**	0.826**	0.022**	0.023**	0.023**
	(0.030)	(0.021)	(0.021)	(0.001)	(0.001)	(0.001)
$Nyears_i^2$	-0.024**	-0.026**	-0.026**	-0.001**	-0.001**	-0.001**
	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)
$Male_i$		0.979**	0.982**		0.025**	0.025**
		(0.043)	(0.042)		(0.001)	(0.001)
Age		0.028**	0.053**		0.001**	0.001**
		(0.002)	(0.011)		(0.000)	(0.000)
Age^2			-0.000*			-0.000**
			(0.000)			(0.000)
Constant	42.023**	40.383**	39.910**	3.738**	3.696**	3.682**
	(0.048)	(0.087)	(0.211)	(0.001)	(0.002)	(0.005)
R-Squared	0.926	0.968	0.969	0.926	0.969	0.969
Observations	549	549	549	549	549	549

Notes: Heteroskedasticity robust standard errors are reported in parentheses. ** $p < 0.01$,* $p < 0.05$

Implications of Estimated Nonlinear Relationships

- Consider our regression in column (3) of the results table

$$\begin{aligned} Earnings_i = & 39.910 - 10.187 Immigrant_i + 0.826 Nyears_i \\ & - 0.026 Nyears_i^2 + 0.982 Male_i + 0.53 Age_i \\ & - 0.0001 Age_i^2 \end{aligned}$$

- So, if $Immigrant_i = 1$, holding other factors fixed, there is a -10.187 earnings disadvantage from being an immigrant, irrespective of $Nyears_i$
- For instance, we can compute the number of years it takes for an immigrant to make up 50% of the wage gap with a native:

$$0.5 \times 10.187 = 0.826 Nyears - 0.026 Nyears^2$$

- Solving this quadratic equation using the quadratic formula we obtain $Nyears = 8.37$ (based on the first root)

Interactions Between Independent Variables

- Let's return to our dataset and consider the following (different) simple linear regression that relates earnings to age:

$$Earnings_i = \beta_0 + \beta_1 Age_i + u_i$$

- We would expect that on average over someone's working life that they would earning more each year as they get older, suggesting $\beta_1 > 0$
- A key question for policy might be whether this relationship differs for males versus females.
- That is, does the value of β_1 differ for observations where $Male_i = 1$ and where $Male_i = 0$?
- To formally study this type of question with data, we develop a separate class of nonlinear regression models that involve regressor interactions

Interactions Between Independent Variables

- We consider three types of nonlinear regression models that involve interactions between:

1. Binary-Binary interaction models: two binary variables, D_{1i} and D_{2i}
2. Continuous-Binary interaction models: a continuous and binary variable, X_i and D_i
3. Continuous-Continuous interaction models: two continuous variables, X_{1i} and X_{2i}

Note: For now, we will "pause" our analysis of polynomial and logarithmic regressions and come back to them at the end of the lecture note

Interactions Between Independent Variables

- Just like with polynomial and logarithmic regressions, interactions models:
  1. Are estimated using Ordinary Least Squares as these models are also linear in the regression parameters $\beta_1, \beta_2, \ldots, \beta_k$
  2. Yield results with different interpretations, depending on which of the three types of interactions nonlinear regression functions is estimated
- We consider multiple linear regression models using just two regressors to keep things simple; adding additional regressors to the model is straightforward

Model 1: Binary-Binary Interaction Models

- Let's first consider the following regression model with dependent variable Y_i and two binary variables, D_{1i} and D_{2i}:

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

- An example using our dataset:

$$Earnings_i = \beta_0 + \beta_1 Male_i + \beta_2 Educ_i + u_i$$

  - β_1 is the effect of being male ($Male_i = 1$) on earnings, holding education constant
  - β_2 is the effect of having a university degree ($Educ_i = 1$) on earnings, holding gender constant

- Limitation of this regression: the effect of $Educ_i$ on $Earnings_i$ is the same for males ($Male_i = 1$) and females ($Male_i = 0$)
  - males and females may have different returns on education

Binary-Binary Interaction Models

- Let's consider a variant on this regression:

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i$$

  where $(D_{1i} \times D_{2i})$ is called an interaction term or an interacted regressor, and the population regression model is called a binary variable interaction regression model.

- With our application:

$$Earnings_i = \beta_0 + \beta_1 Male_i + \beta_2 Educ_i + \beta_3 (Male_i \times Educ_i) + u_i$$

- Using our general notation ($D_{1i} = Male_i, D_{2i} = Educ_i$), what is the predicted effect of having a university degree on earnings depending on whether you are female or male?

Binary-Binary Interaction Models

- Here, we derive the partial effect of D_{2i} (e.g. $Educ_i$) using the population regression function by taking conditional expectations of Y_i for different D_{2i} values
- Recall from earlier in the lecture note that this is a general approach for deriving partial effects in nonlinear models
- Start with the binary-binary interactions model:

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i$$

- Expected value of Y_i if $D_{1i} = d_1$ and $D_{2i} = 0$:

$$E(Y_i | D_{1i} = d_1, D_{2i} = 0) = \beta_0 + \beta_1 d_1$$

- Expected value of Y_i if $D_{1i} = d_1$ and $D_{2i} = 1$:

$$E(Y_i | D_{1i} = d_1, D_{2i} = 1) = \beta_0 + \beta_1 d_1 + \beta_2 + \beta_3 d_1$$

where d_1 is either 0 or 1 (as recall D_{1i} is dummy variable)

Binary-Binary Interaction Models

- Partial effect of changing $D_{2i} = 0$ to $D_{2i} = 1$ is thus:

$$\begin{aligned}\Delta Y &= E(Y_i|D_{1i} = d_1, D_{2i} = 1) - E(Y_i|D_{1i} = d_1, D_{2i} = 0) \\ &= (\beta_0 + \beta_1 d_1 + \beta_2 + \beta_3 d_1) - (\beta_0 + \beta_1 d_1) \\ &= \beta_2 + \beta_3 d_1\end{aligned}$$

- If $d_1 = 0$ (that is, $Male_i = 0$), the partial effect of $Educ_i$ is:

$$\Delta Y = \beta_2 + \beta_3 0 = \beta_2$$

- If $d_1 = 1$ (that is, $Male_i = 1$), the partial effect of $Educ_i$ is:

$$\Delta Y = \beta_2 + \beta_3 1 = \beta_2 + \beta_3$$

Binary-Binary Interaction Models

- Reproducing the application regression for binary-binary interactions models

$$Earnings_i = \beta_0 + \beta_1 Male_i + \beta_2 Educ_i + \beta_3 (Male_i \times Educ_i) + u_i$$

- These conditional expectations imply that the partial effect of having a university degree D_{2i} on earnings Y is:
  - $\Delta Y = \beta_2$ if female ($Male_i = 0$)
  - $\Delta Y = \beta_2 + \beta_3$ if male ($Male_i = 1$)
- In words, including the interaction term $(Male_i \times Educ_i)$ in the regression implies the effect of having a university degree on earnings depends if person i is male or female, and the parameter β_3 governs this gender-specific education effect

Binary-Binary Interaction Models Application

- Suppose we estimated this regression using our dataset:

$$\widehat{Earnings}_i = \underset{(0.18)}{39.43} + \underset{(0.24)}{0.61}\, Male_i + \underset{(1.83)}{4.93}\, Educ_i - \underset{(1.11)}{2.71}\,(Male_i \times Educ_i); \quad \bar{R}^2 = 0.15$$

- Interpretations (remembering $Earnings_i$ is in \$1000's):
  - Males with no university degree earn $0.61 \times 1000 = \$610$ more per year than females with no university degree
  - Females with university degrees earn $4.93 \times 1000 = \$4,930$ more per year than females with no university degree
  - A university degrees increases earnings by $2.71 \times 1000 = \$2,710$ less for males
- In words, a university degree is associated with:
  - \$4,930 in earnings per year for females
  - \$4,930-\$2,710=\$2,220 in earnings per year for males

Model 2: Continuous-Binary Interaction Models

- Now let's consider the following regression model with dependent variable Y_i, a continuous regressor X_i, and a dummy variable D_i:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i$$

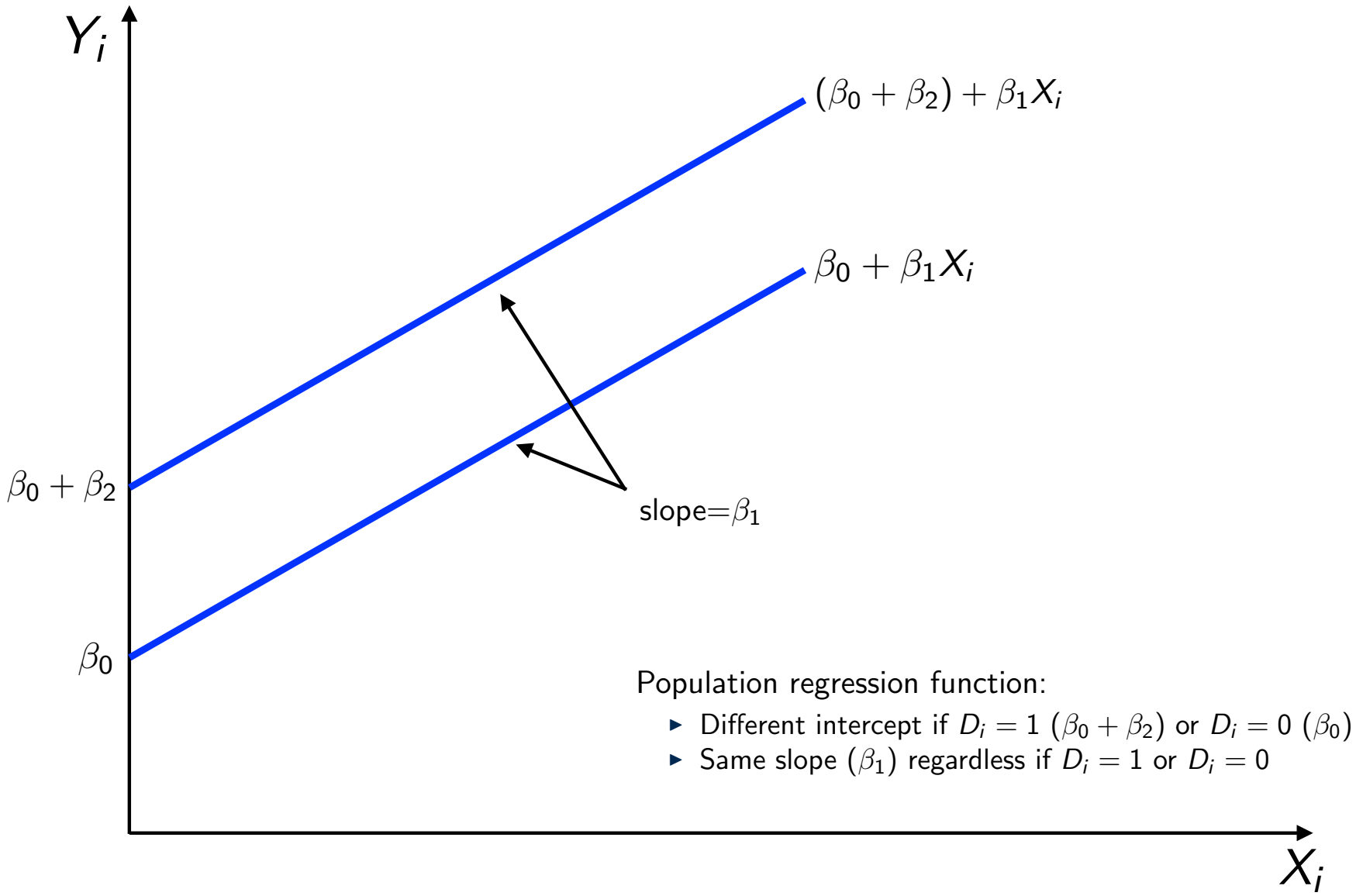
- An example using our dataset (ignoring $Nyears_i^2$):

$$Earnings_i = \beta_0 + \beta_1 Nyears_i + \beta_2 Male_i + u_i$$

  - β_1 is the effect of one additional year of being in the country for an immigrant, holding gender constant
  - β_2 is the effect of being male on earnings, holding number of years in the country constant

- Limitation of this regression: effect of $Nyears_i$ on $Earnings_i$ is the same for males ($Male_i = 1$) and females ($Male_i = 0$)
  - males and females might have different relationships between the number of years in a country and their earnings!

Graphically: $E[Y_i|X_i, D_i] = \beta_0 + \beta_1 X_i + \beta_2 D_i$



Population regression function:

- Different intercept if $D_i = 1$ $(\beta_0 + \beta_2)$ or $D_i = 0$ (β_0)
- Same slope (β_1) regardless if $D_i = 1$ or $D_i = 0$

Continuous-Binary Interaction Models

- To allow for different intercepts and different slopes depending on D_i we add an interaction term to our regression:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i$$

- An example using our dataset :

$$Earnings_i = \beta_0 + \beta_1 Nyears_i + \beta_2 Male_i + \beta_3 (Nyears_i \times Male_i) + u_i$$

(where for the moment we leave out $Nyears_i^2$ as a regressor)

- The inclusion of the interaction term $(Nyears_i \times Male_i)$ allows us to estimate β_3 which tells how having an additional year in the country differs for males and females

Continuous-Binary Interaction Models

- Regression model with interactions:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i$$

- What is the conditional mean of Y_i depending on whether D_i equals 0 or 1?

$$\begin{aligned} E(Y_i|X_i, D_i = 1) &= \beta_0 + \beta_1 X_i + \beta_2 1 + \beta_3 (X_i \times 1) \\ &= \beta_0 + \beta_2 + (\beta_1 + \beta_3) X_i \end{aligned}$$

and

$$\begin{aligned} E(Y_i|X_i, D_i = 0) &= \beta_0 + \beta_1 X_i + \beta_2 0 + \beta_3 (X_i \times 0) \\ &= \beta_0 + \beta_1 X_i \end{aligned}$$

Continuous-Binary Interaction Models

- Using our procedure for computing partial effects based on conditional expectations from before we can then show that:

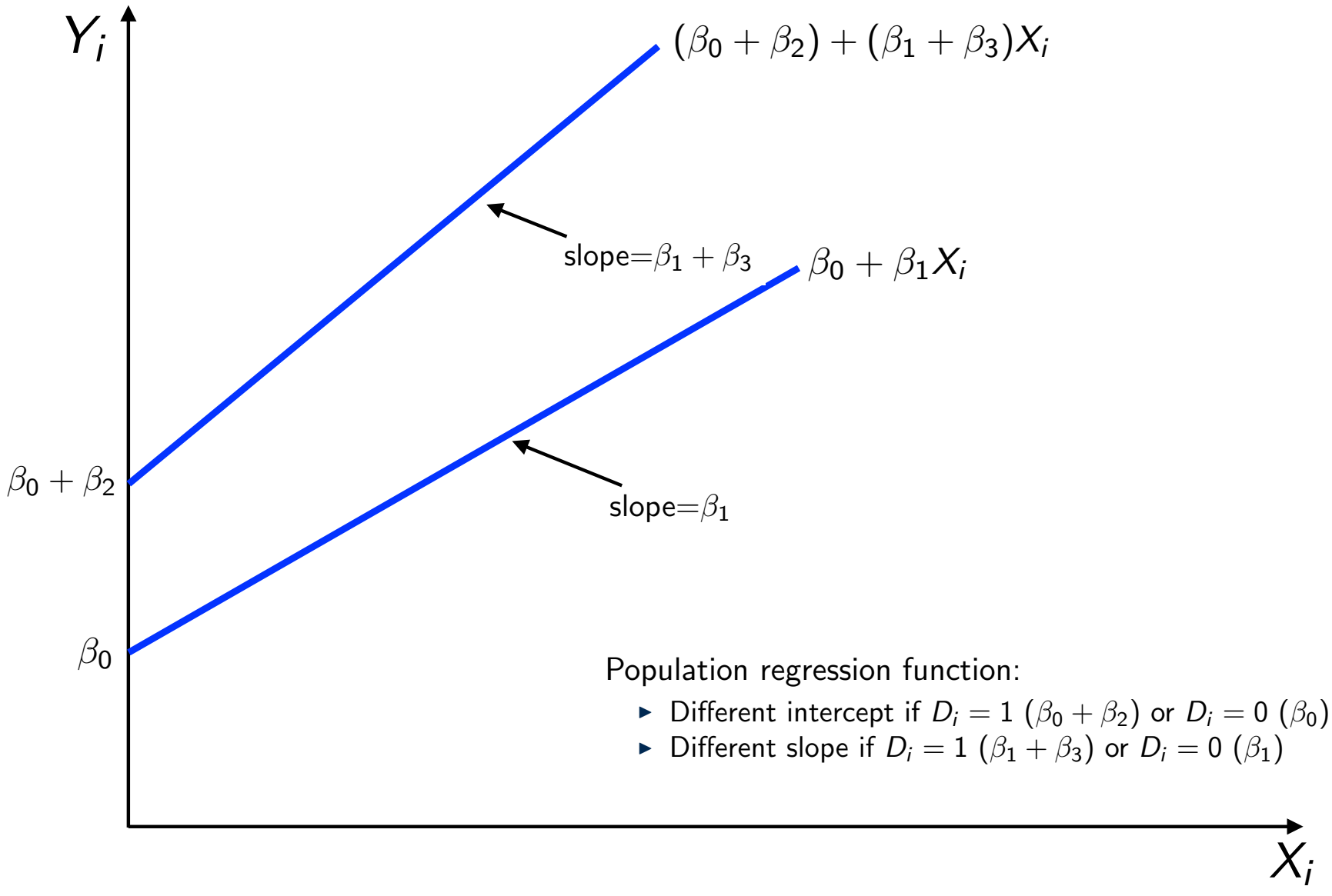
$$\begin{aligned}\Delta Y &= E(Y_i|X_i, D_i = 1) - E(Y_i|X_i, D_i = 0) \\ &= (\beta_0 + \beta_2 + (\beta_1 + \beta_3)X_i) - (\beta_0 + \beta_1 X_i) \\ &= \beta_2 + \beta_3 X_i\end{aligned}$$

- In our continuous-binary interactions example, which recall is:

$$Earnings_i = \beta_0 + \beta_1 Nyears_i + \beta_2 Male_i + \beta_3 (Nyears_i \times Male_i) + u_i,$$

β_3 is the partial effect of an additional year of being in a country (e.g., $\Delta X_i = 1$) for males relative to females

Graphically: $E[Y_i|X_i, D_i] = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i)$



Population regression function:

- Different intercept if $D_i = 1$ $(\beta_0 + \beta_2)$ or $D_i = 0$ (β_0)
- Different slope if $D_i = 1$ $(\beta_1 + \beta_3)$ or $D_i = 0$ (β_1)

Continuous-Binary Interaction Models Application

- Suppose we estimated the following regression model with a different intercept but same slope for males and females:

$$\widehat{Earnings}_i = \underset{(4.94)}{38.75} + \underset{(0.44)}{1.84}\, Nyears_i + \underset{(0.056)}{1.22}\, Male_i; \;\; \bar{R}^2 = 0.11$$

  - 1 extra year in the country is associated with and increase in earnings by $1.84 \times 1000 = \$1,840$ per year for both males and females
  - Being male is associated with an increase in earnings by $1.22 \times 1000 = \$1,220$ per year, holding the number of years in the country fixed

Continuous-Binary Interaction Models Application

- Now suppose we estimated the following regression model with a different intercept and different slope for males and females:

$$\widehat{Earnings}_i = \underset{(5.12)}{39.88} + \underset{(0.60)}{1.52}\, Nyears_i + \underset{(0.12)}{2.12}\, Male_i - \underset{(0.03)}{0.50}\,(Nyears_i \times Male_i);\ \bar{R}^2 = 0.16$$

  - 1 extra year in the country is associated with and increase in earnings by $1.52 \times 1000 - 0.50 \times 1000 = \$1,020$ per year if the immigrant is male (e.g., $Male_i = 1$)
  - 1 extra year in the country is associated with and increase in earnings by $1.52 \times 1000 = \$1,520$ per year if the immigrant is female (e.g., $Male_i = 0$)

Continuous-Binary Interaction Models Application

$$\widehat{Earnings}_i = \underset{(5.12)}{39.88} + \underset{(0.60)}{1.52}\, Nyears_i + \underset{(0.12)}{2.12}\, Male_i - \underset{(0.03)}{0.50}\,(Nyears_i \times Male_i);\ \bar{R}^2 = 0.16$$

- This model also allows for differential immigration earning effects across males and females as a function of $Nyears_i$
- Relative to females, males have an earnings premium of $2.12 - 0.50 \times Nyears_i$, or equivalently, $\$2120 - \$500 \times Nyears_i$
  - $Nyears_i = 1$: the male earnings premium relative to females is $\$2120 - \$500 \times 1 = \$1620$
  - $Nyears_i = 2$: the male earnings premium relative to females is $\$2120 - \$500 \times 2 = \$1120$
  - $Nyears_i = 5$: the male earnings premium relative to females is $\$2120 - \$500 \times 5 = -\$380$
- In words: The gender earnings gap among immigrants declines over time; female immigrants eventually catch and surpass male immigrants the longer they live in a country!

Continuous-Binary Interaction Models

- Another variant on these models to have the same intercept but different slopes:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i \times D_i) + u_i$$

- An example using our dataset (again ignoring $Nyears_i^2$):

$$Earnings_i = \beta_0 + \beta_1 Nyears_i + \beta_2 (Nyears_i \times Male_i) + u_i$$

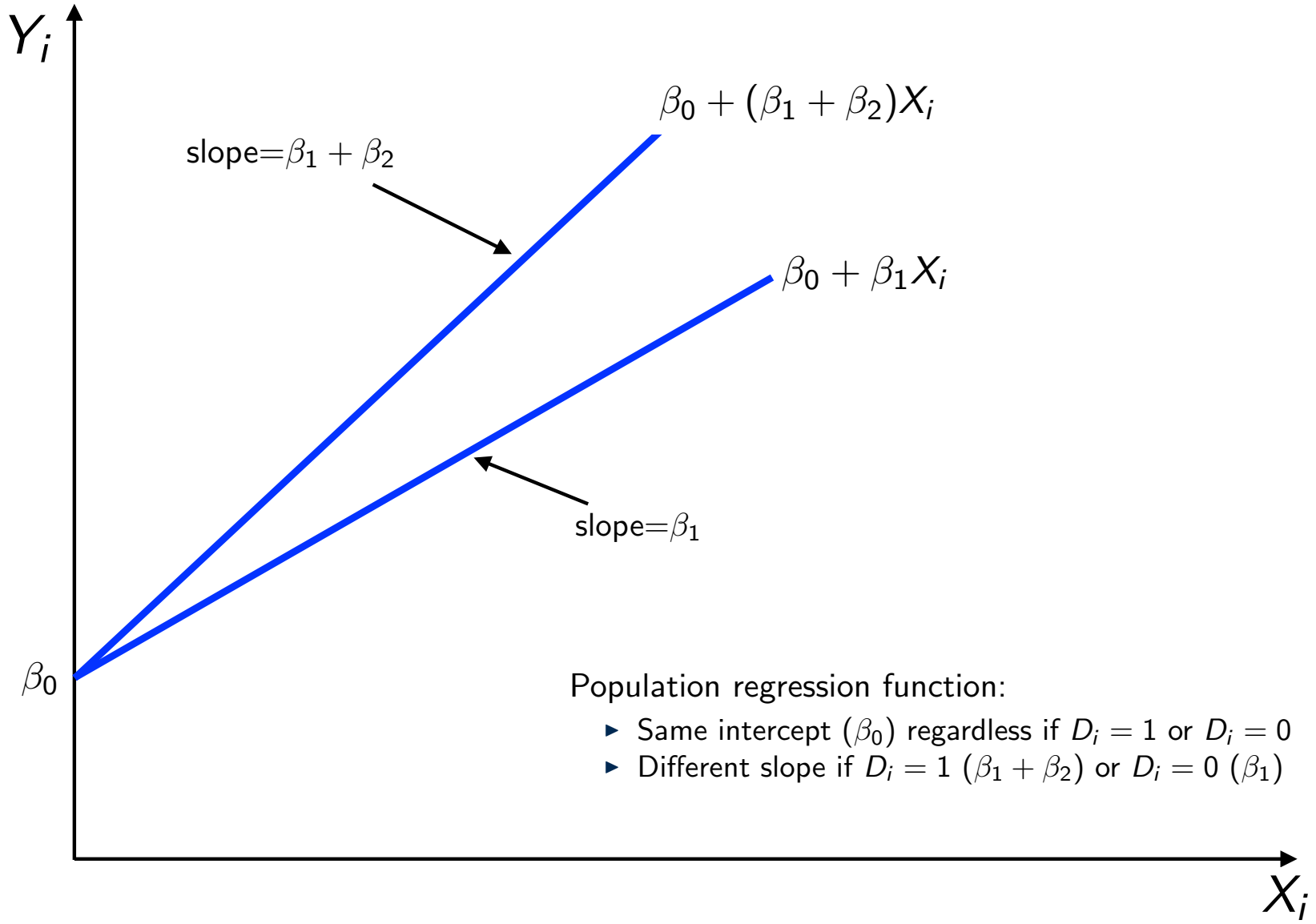
- This model is used very rarely; in practice the focus on:
  - different intercept and same slope model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i$$

  - different intercept and different slope model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i$$

Graphically: $E[Y_i|X_i, D_i] = \beta_0 + \beta_1 X_i + \beta_2(X_i \times D_i)$



Population regression function:

- Same intercept (β_0) regardless if $D_i = 1$ or $D_i = 0$
- Different slope if $D_i = 1$ ($\beta_1 + \beta_2$) or $D_i = 0$ (β_1)

Model 3: Continuous-Continuous Interaction Models

- The final type of nonlinear interactions regressions involves interactions between two continuous regressors, X_{1i} and X_{2i}:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i$$

- An example using our dataset (ignoring $Nyears_i^2$):

$$Earnings_i = \beta_0 + \beta_1 Nyears_i + \beta_2 Age_i + \beta_3 (Nyears_i \times Age_i) + u_i$$

- Interpreting the coefficients in the example:
  - β_1 is the partial effect of 1 more year in the country on earnings
  - β_3 is the change in the partial effect of 1 more year in the country on earnings if someone is 1 more year older

- In other words, β_3 captures different relationships between $Earnings_i$ and $Nyears_i$ for people of different Age_i
  - for instance, we might expect $\beta_3 < 0$, which would mean the effect of $Nyears_i$ on $Earnings_i$ is smaller for older immigrants

Continuous-Continuous Interaction Models

- Continuous-Continuous interactions model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i$$

- In this model, the effect on Y of a change in X_1 given X_2 is:

$$\frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2$$

  where we see the partial effect of X_1 depends on the level of X_2

- This can be seen in our example where the impact on earnings from additional years depends on an immigrants age:

$$\frac{\Delta Earnings}{\Delta Nyears} = \beta_1 + \beta_3 Age_i$$

Continuous-Continuous Interaction Models

- Continuous-Continuous interactions model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i$$

- Similarly, in this model the effect on Y of a change in X_2 given X_1 is:

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

  where we see the partial effect of X_2 depends on the level of X_1

- In our example, this also means the impact on additional years of age on earnings depends on how many years an immigrant has been in the country:

$$\frac{\Delta Earnings}{\Delta Age_i} = \beta_2 + \beta_3 Nyears_i$$

Continuous-Continuous Interaction Models

- Continuous-Continuous interactions model:
- Finally, the effect on Y of simultaneous changes in X_1, ΔX_1, and in X_2, ΔX_2, is:

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

- In our example, if we simultaneously increased age by one year ($\Delta Age_i = 1$) and the number of years in the country by one ($\Delta Nyears_i = 1$), the partial effect on earnings would be:

$$\frac{\Delta Earnings}{\Delta Age_i \Delta Nyears_i} = \underbrace{\beta_1 + \beta_3 Age_i}_{\Delta Nyears_i = 1 \ impact} + \underbrace{\beta_2 + \beta_3 Nyears_i}_{\Delta Age_i = 1 \ impact} + \underbrace{\beta_3}_{interaction \ effect}$$

Continuous-Continuous Interaction Models Application

- Suppose we estimated the following regression model with the same slope for $Nyears_i$ for all ages Age_i:

$$\widehat{Earnings}_i = \underset{(5.58)}{41.02} + \underset{(0.31)}{1.44}\, Nyears_i + \underset{(0.092)}{0.811}\, Age_i; \quad \bar{R}^2 = 0.06$$

  where again recall $Earnings_i$ is in \$1000's, so we continue to multiple the coefficients by 1000 to have interpretations of effects in terms of dollars
  - 1 extra year in the country is associated with an increase in earnings by 1000×1.44=\$1,440 per year, holding age fixed
  - 1 extra year of age is associated with an increase in earnings of 1000×0.811=\$811 per year, holding the number of years in the country fixed
- Interpretations change considerably once we include a $(Nyears_i \times Age_i)$ interaction ...

Continuous-Continuous Interaction Models Application

- Now suppose we estimated the following regression model with the same slope for $Nyears_i$ for all ages Age_i:

$$\widehat{Earnings}_i = \underset{(5.99)}{40.75} + \underset{(0.36)}{1.69}\, Nyears_i + \underset{(0.12)}{1.28}\, Age_i - \underset{(0.01)}{0.02}\,(Nyears_i \times Age_i); \quad \bar{R}^2 = 0.08$$

- Partial effects of $Nyears_i$ as a function of Age_i:
  - 1 extra year in the country is associated with an increase in earnings of: \$1,690 - \$20 $\times$ Age_i per year
  - $Age_i = 25$: 1 extra year in the country is associated with an increase in earnings of \$1,690 - \$20 $\times$ 25=\$1190 per year
  - $Age_i = 40$: 1 extra year in the country is associated with an increase in earnings of \$1,690 - \$20 $\times$ 40=\$890 per year
- In words: older immigrants realise less earnings gain from each additional year in the country

Continuous-Continuous Interaction Models Application

(continued)

$$\widehat{Earnings}_i = \underset{(5.99)}{40.75} + \underset{(0.36)}{1.69}\, Nyears_i + \underset{(0.12)}{1.28}\, Age_i - \underset{(0.01)}{0.02}\,(Nyears_i \times Age_i); \quad \bar{R}^2 = 0.08$$

- Partial effects of Age_i as a function of $Nyears_i$:
  - 1 additional year of age is associated with an increase in earnings of: \$1,280 - \$20 $\times$ $Nyears_i$ per year
  - $Nyears_i = 5$: 1 additional year of age is associated with an increase in earnings of \$1,280 - \$20 $\times$ 5=\$1180 per year
  - $Nyears_i = 15$: 1 additional year of age is associated with an increase in earnings of \$1,280 - \$20 $\times$ 15=\$980 per year
- In words: immigrants who have been in the country longer realize less earnings gains from each additional year of age

Summary of Nonlinear Regression with Interactions

- In summary, we have considered the 3 classes of nonlinear interactive regression models, which differ in their application and interpretation
- Binary-Binary interaction models:

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i$$

- Binary-Continuous interaction models:

$$\begin{aligned}
\text{(diff int., same slope)} \quad & Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i \\
\text{(diff int., diff slope)} \quad & Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i \\
\text{(same int., diff slope)} \quad & Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i \times D_i) + u_i
\end{aligned}$$

- Continuous-Continuous interaction models:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i$$

Polynomials + Logarithms + Interactions

- We can combine polynomial, logarithmic, and interaction nonlinear regression functions
- An example in the immigration literature is:

$$\begin{aligned}\log(Earnings_i) = \beta_0 &+ \beta_1 Immigrant_i + \beta_2 Nyears_i + \beta_3 Nyears_i^2 \\ &+ \beta_4 (Nyears_i \times Male_i) + \beta_5 (Nyears_i^2 \times Male_i) \\ &+ \beta_6 Male_i + \beta_7 Educ_i + \beta_8 Age_i + \beta_9 Age_i^2 + u_i\end{aligned}$$

which is a Log-Linear quadratic model with interactions

Polynomials + Logarithms + Interactions

How to interpret this regression function?

$$\begin{aligned}\log(Earnings_i) = \beta_0 &+ \beta_1 Immigrant_i + \beta_2 Nyears_i + \beta_3 Nyears_i^2 \\ &+ \beta_4(Nyears_i \times Male_i) + \beta_5(Nyears_i^2 \times Male_i) \\ &+ \beta_6 Male_i + \beta_7 Educ_i + \beta_8 Age_i + \beta_9 Age_i^2 + u_i\end{aligned}$$

- Interpretations
  - β_2 and β_3 governs how immigrant earnings catch native earnings as $Nyears_i$ grows for females ($Male_i = 0$), holding fixed gender, education, and age
  - β_2, β_3, β_4, β_5, governs how immigrant earnings catch native earnings as $Nyears_i$ grows for males ($Male_i = 1$), holding fixed gender, education, and age
  - The $\log(Earnings_i)$ dependent variable implies that the β_2, β_3, β_4, β_5 coefficients collectively predict the percentage changes in annual earnings from $\Delta Nyears_i = 1$ for males and females

Polynomials + Logarithms + Interactions

- Interpretations need to account for the levels of $Nyears_i$ because of the quadratic specification
- Interpretations also need to account for male/female status because of the interaction terms with $Male_i$
- Interpretations are in terms of % changes because it is a Log-Linear model with dependent variable $log(Earnings_i)$
- In sum, this model incorporates all of the aspects of nonlinear regression models that we have considered in this lecture

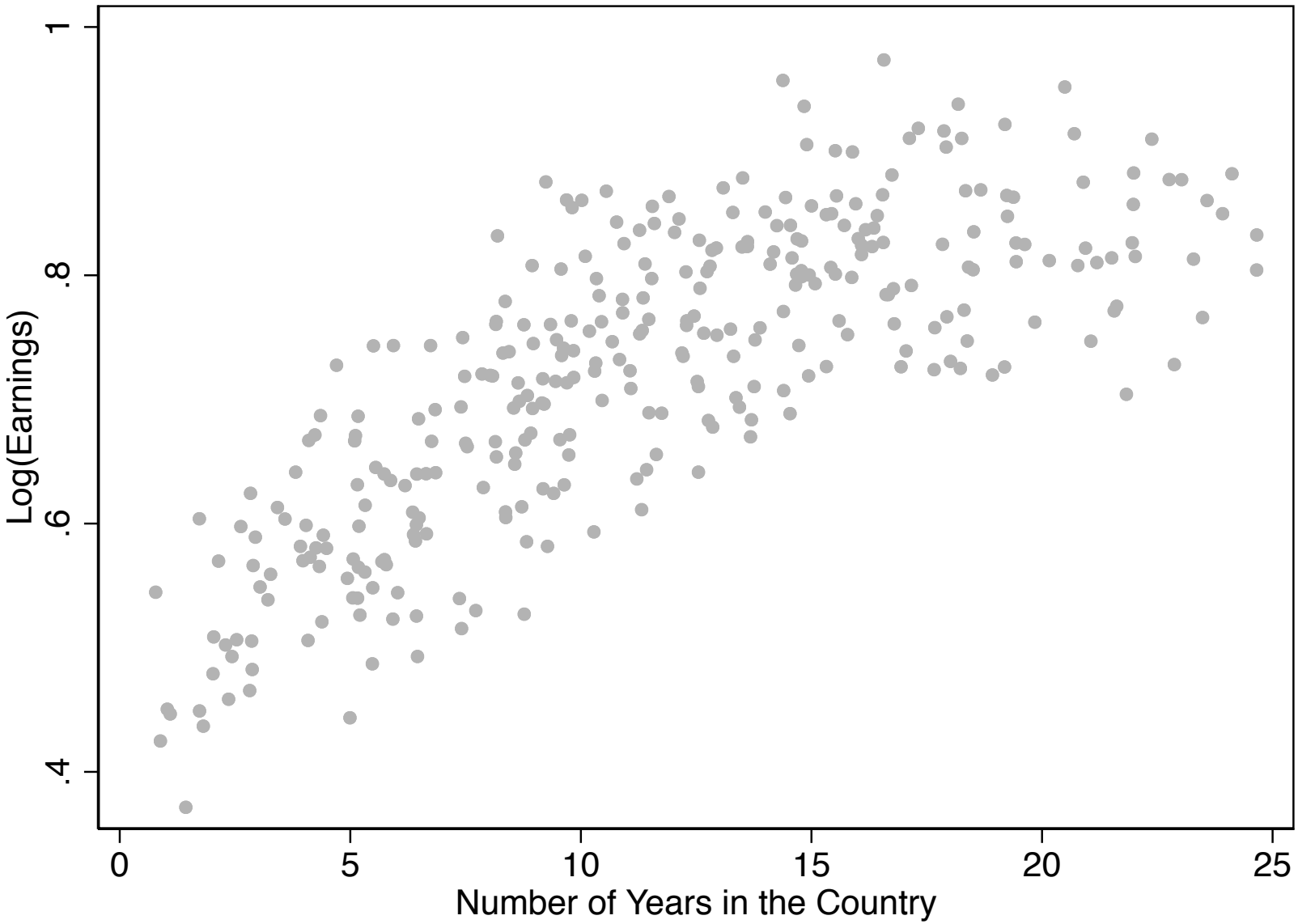
Polynomials + Logarithms + Interactions

Steps for Analysing Such Rich Models

1. Data Visualisation
2. Model Estimation
3. Model Testing
4. Partial Effects Estimation
5. Partial Effects Testing

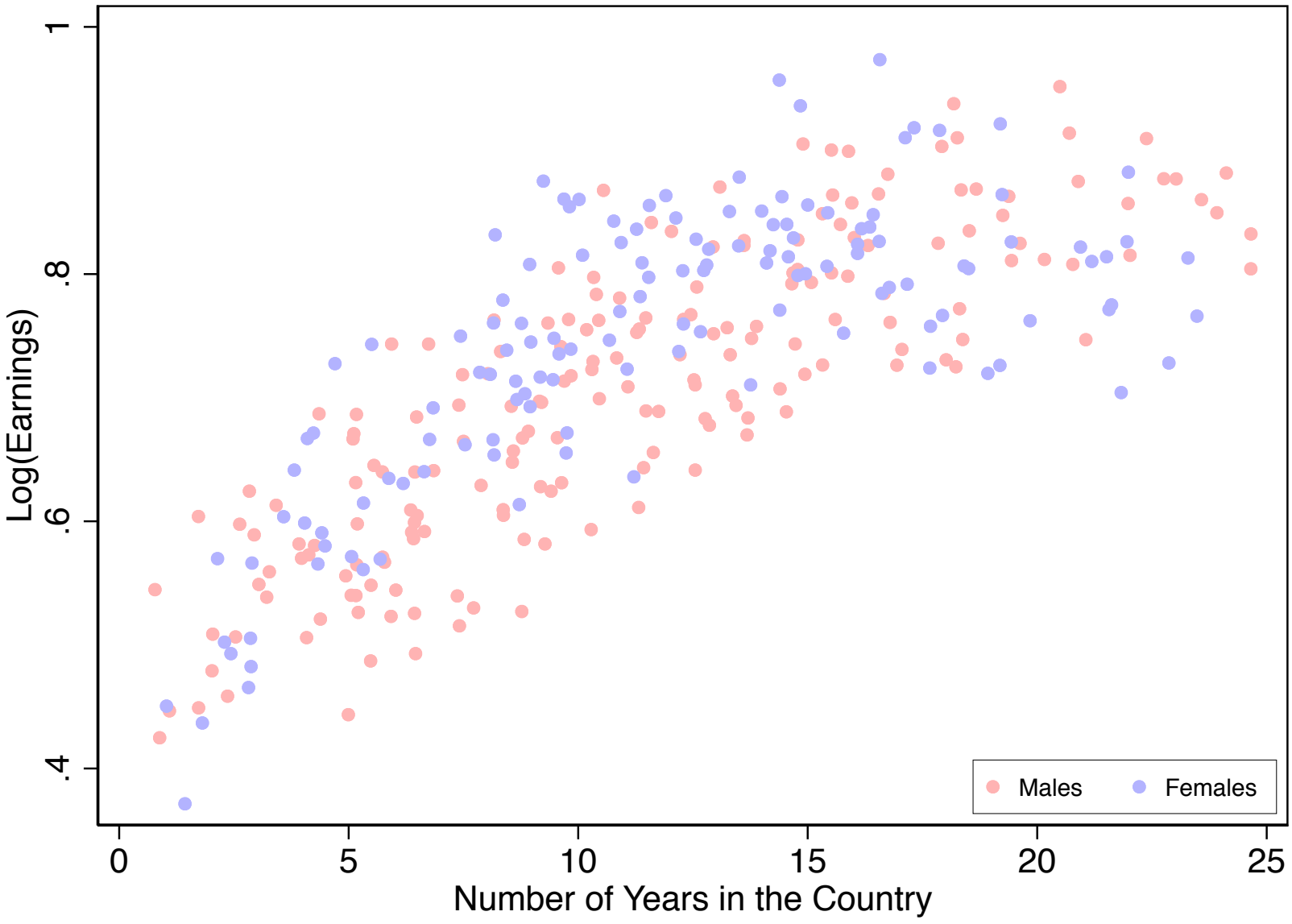
Polynomials + Logarithms + Interactions

Data Visualisation



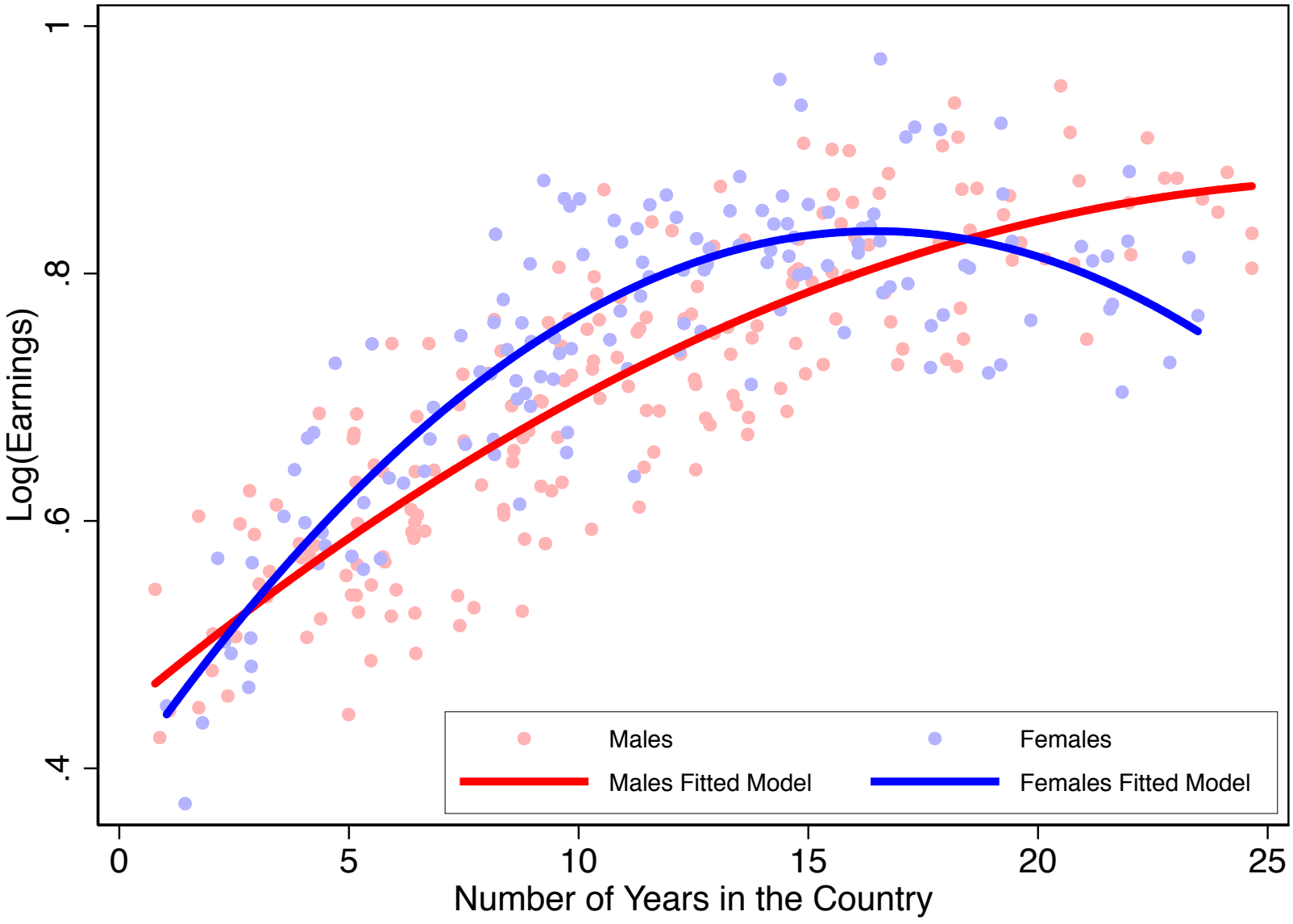
Polynomials + Logarithms + Interactions

Data Visualisation



Polynomials + Logarithms + Interactions

Data Visualisation



Polynomials + Logarithms + Interactions

- **Model Estimation**: suppose we obtained the following regression estimates by multiple linear regression:

$$\begin{aligned}\widehat{\log(Earnings_i)} = & \underset{(0.424)}{1.450} - \underset{(0.002)}{0.010} Immigrant_i + \underset{(0.001)}{0.049} Nyears_i - \underset{(0.0001)}{0.001} Nyears_i^2 \\ & - \underset{(0.0003)}{0.019}(Nyears_i \times Male_i) + \underset{(0.0001)}{0.0005}(Nyears_i^2 \times Male_i) \\ & + \underset{(0.001)}{0.039} Male_i + \underset{(0.001)}{0.021} Educ_i + \underset{(0.017)}{0.022} Age_i - \underset{(0.001)}{0.0002} Age_i^2\end{aligned}$$

- Regression function for females ($Male_i = 0$):

$$\begin{aligned}\widehat{\log(Earnings_i)} = & \ 1.450 - 0.010 Immigrant_i + 0.049 Nyears_i - 0.001 Nyears_i^2 \\ & + 0.021 Educ_i + 0.022 Age_i - 0.0002 Age_i^2\end{aligned}$$

- Regression function for males ($Male_i = 1$):

$$\begin{aligned}\widehat{\log(Earnings_i)} = & \ 1.450 - 0.010 Immigrant_i + 0.030 Nyears_i - 0.0005 Nyears_i^2 \\ & + 0.039 Male_i + 0.021 Educ_i + 0.022 Age_i - 0.0002 Age_i^2\end{aligned}$$

Polynomials + Logarithms + Interactions

- **Model Testing**: Testing the shape of the regression function

$$\widehat{\log(Earnings_i)} = \underset{(0.424)}{1.450} - \underset{(0.002)}{0.010} Immigrant_i + \underset{(0.001)}{0.049} Nyears_i - \underset{(0.0001)}{0.001} Nyears_i^2$$
$$- \underset{(0.0003)}{0.019}(Nyears_i \times Male_i) + \underset{(0.0001)}{0.0005}(Nyears_i^2 \times Male_i)$$
$$+ \underset{(0.001)}{0.039} Male_i + \underset{(0.001)}{0.021} Educ_i + \underset{(0.017)}{0.022} Age_i - \underset{(0.001)}{0.0002} Age_i^2$$

- Does there exist a nonlinear relationship between *Nyears* and $\log(Earnings_i)$? The test depends on the gender dummy.
- Females test for nonlinearity
  - **individual test** that the coefficient on $Nyears^2$ equals 0 (e.g., β_3, with OLS estimate $\hat{\beta}_3 = 0.001$)
  - formally the test is:

$$H_0 : \beta_3 = 0; \quad H_1 : \beta_3 \neq 0$$

  - corresponding p-value from the test is < 0.00001 (based on t-statistic) so we reject the null, implying the model is nonlinear in *Nyears* for females

Polynomials + Logarithms + Interactions

- **Model Testing**: Testing the shape of the regression function

$$\widehat{\log(Earnings_i)} = \underset{(0.424)}{1.450} - \underset{(0.002)}{0.010} Immigrant_i + \underset{(0.001)}{0.049} Nyears_i - \underset{(0.0001)}{0.001} Nyears_i^2$$
$$- \underset{(0.0003)}{0.019} (Nyears_i \times Male_i) + \underset{(0.0001)}{0.0005} (Nyears_i^2 \times Male_i)$$
$$+ \underset{(0.001)}{0.039} Male_i + \underset{(0.001)}{0.021} Educ_i + \underset{(0.017)}{0.022} Age_i - \underset{(0.001)}{0.0002} Age_i^2$$

- Males test for nonlinearity
  - **joint test** that the sum of the coefficients on $Nyears^2$ and $(Nyears_i \times Male_i)$ equals 0 (e.g., $\beta_3 + \beta_5$, with OLS estimates $\hat{\beta}_3 = -0.001$ and $\hat{\beta}_5 = 0.0005$)
  - formally the test is:

$$H_0 : \beta_3 + \beta_5 = 0; \quad H_1 : \beta_3 + \beta_5 \neq 0$$

  - corresponding p-value from the test is < 0.00001 (based on F-statistic) so we reject the null, implying the model is nonlinear in *Nyears* for males

Polynomials + Logarithms + Interactions

- **Model Testing**: Testing gender–related differences in the regression function

$$\begin{aligned}\widehat{\log(Earnings_i)} = & \underset{(0.424)}{1.450} - \underset{(0.002)}{0.010} Immigrant_i + \underset{(0.001)}{0.049} Nyears_i - \underset{(0.0001)}{0.001} Nyears_i^2 \\ & - \underset{(0.0003)}{0.019} (Nyears_i \times Male_i) + \underset{(0.0001)}{0.0005} (Nyears_i^2 \times Male_i) \\ & + \underset{(0.001)}{0.039} Male_i + \underset{(0.001)}{0.021} Educ_i + \underset{(0.017)}{0.022} Age_i - \underset{(0.001)}{0.0002} Age_i^2\end{aligned}$$

- Is the nonlinear relationship between *Nyears* and $\log(Earnings_i)$ the different for males and females?
  - **joint test** that the coefficient on $(Nyears_i \times Male_i)$ equals 0 and the coefficient on $(Nyears_i^2 \times Male_i)$ equals 0 (e.g., $\beta_4 = 0$ and $\beta_5 = 0$, with OLS estimates $\hat{\beta}_4 = -0.019$ and $\hat{\beta}_5 = 0.0005$
  - formally the test is:

    $H_0 : \beta_4 = 0$ and $\beta_5 = 0$; $\quad H_1 :$ at least one of $\beta_4 \neq 0$ or $\beta_5 \neq 0$

  - corresponding p-value from the test is < 0.00001 (based on F-statistic) so we reject the null, implying that the nonlinear relationship between *Nyears* and $\log(Earnings_i)$ is different between males and females

Polynomials + Logarithms + Interactions

- **Partial Effects Estimation**: What is the partial effect of a change in $Nyears_i$, $\Delta Nyears_i$ on $\log(\widehat{Earnings_i})$?
- We now use the general techniques from slides 16-22 of this lecture note for computing nonlinear partial effects and their standard errors
- Recall the general definition of a nonlinear partial effect in Y, ΔY from changing X_1 to $X_1 + \Delta X_1$ holding all other regressors $X_2, \ldots, X_k$ fixed:

$$\Delta Y = f(X_1 + \Delta X_1, X_2, \ldots, X_k) - f(X_1, X_2, \ldots, X_k)$$

- In our example:

$$\begin{aligned}\Delta\log(\widehat{Earnings_i}) = &\hat{f}(Nyears + \Delta Nyears, Immigrant_i, Male_i, Educ_i, Age_i) \\ &- \hat{f}(Nyears, Immigrant_i, Male_i, Educ_i, Age_i)\end{aligned}$$

Polynomials + Logarithms + Interactions

- Deriving the partial effect from our estimated model:

$$\begin{aligned}\Delta\log(\widehat{Earnings}) = & \\ & [1.450 - 0.010 Immigrant + 0.049(Nyears + \Delta Nyears) - 0.001(Nyears + \Delta Nyears)^2 \\ & \quad -0.019((Nyears + \Delta Nyears) \times Male) + 0.0005((Nyears + \Delta Nyears)^2 \times Male) \\ & \quad +0.039 Male + 0.021 Educ + 0.022 Age - 0.0002 Age^2] \\ & -[1.450 - 0.010 Immigrant + 0.049 Nyears + -0.001 Nyears^2 \\ & \quad -0.019(Nyears \times Male) + 0.0005(Nyears^2 \times Male) \\ & \quad +0.039 Male + 0.021 Educ + 0.022 Age - 0.0002 Age^2]\end{aligned}$$

which after cancelling common terms implies:

$$\begin{aligned}\Delta\log(\widehat{Earnings}) = & [0.049(Nyears + \Delta Nyears) - 0.001(Nyears + \Delta Nyears)^2 \\ & \quad -0.019((Nyears + \Delta Nyears) \times Male) + 0.0005((Nyears + \Delta Nyears)^2 \times Male)] \\ & -[0.049 Nyears + -0.001 Nyears^2 \\ & \quad -0.019(Nyears \times Male) + 0.0005(Nyears^2 \times Male)]\end{aligned}$$

Polynomials + Logarithms + Interactions

- Partial effect on $\log(\widehat{Earnings})$ from $\Delta Nyears$ in $Nyears$:

$$\begin{aligned}\Delta\log(\widehat{Earnings}) = &[0.049(Nyears + \Delta Nyears) - 0.001(Nyears + \Delta Nyears)^2 \\ &- 0.019((Nyears + \Delta Nyears) \times Male) + 0.0005((Nyears + \Delta Nyears)^2 \times Male)] \\ &-[0.049 Nyears + -0.001 Nyears^2 \\ &- 0.019(Nyears \times Male) + 0.0005(Nyears^2 \times Male)]\end{aligned}$$

- How large is the partial effect of $\Delta Nyears = 1$ from $Nyears = 5$ to $Nyears + \Delta Nyears = 6$ for males and females?
- Males: if $Nyears_i$ changes from 5 to 6 for a male ($Male_i = 1$) the partial effect on $\log(Earnings_i)$ is

$$\begin{aligned}\Delta\log(\widehat{Earnings}) &= [0.049 \times 6 - 0.001 \times 6^2 - 0.019(6 \times 1) + 0.0005(6^2 \times 1)] \\ &\quad - [0.049 \times 5 - 0.001 \times 5^2 - 0.019(5 \times 1) + 0.0005(5^2 \times 1)] \\ &= 0.162 - 0.137 = 0.025\end{aligned}$$

- Interpretation: increasing $Nyears_i$ from 5 to 6 for a male immigrant leads to a 2.5% increase in their earnings

Polynomials + Logarithms + Interactions

- Females: if $Nyears_i$ changes from 5 to 6 for a female ($Male_i = 0$) the partial effect on $log(Earnings_i)$ is

$$\begin{aligned}\Delta\widehat{\log(Earnings)} &= [0.049 \times 6 - 0.001 \times 6^2 - 0.019(6 \times 0) + 0.0005(6^2 \times 0)] \\ &\quad - [0.049 \times 5 - 0.001 \times 5^2 - 0.019(5 \times 0) + 0.0005(5^2 \times 0)] \\ &= 0.258 - 0.220 = 0.038\end{aligned}$$

- Interpretation: increasing $Nyears_i$ changes from 5 to 6 for a female immigrant leads to a 3.8% increase in their earnings

Polynomials + Logarithms + Interactions

- **Partial Effects Testing**: Are these partial effects for males and females statistically significant? What are their 95% confidence intervals?
- To conduct our test, we continue to use the general approach outlined in slides 23-24 of this lecture note for computing standard errors for nonlinear partial effects

Polynomials + Logarithms + Interactions

- Partial effect for males ($Male_i = 1$) in general terms of the population regression coefficients β from a change in *Nyears* from 5 to 6:

$$\begin{aligned}\Delta\log(\widehat{Earnings}) &= [\hat{\beta}_2 6 + \hat{\beta}_3 6^2 + \hat{\beta}_4(6 \times 1) + \hat{\beta}_5(6^2 \times 1)] \\ &\quad - [\hat{\beta}_2 5 + \hat{\beta}_3 5^2 + \hat{\beta}_4(5 \times 1) + \hat{\beta}_5(5^2 \times 1)] \\ &= \hat{\beta}_2(6-5) + \hat{\beta}_3(36-25) + \hat{\beta}_4(6-5) + \hat{\beta}_5(36-25) \\ &= \hat{\beta}_2 + 11\hat{\beta}_3 + \hat{\beta}_4 + 11\hat{\beta}_5\end{aligned}$$

- Next, we conduct the following joint null hypothesis test:

$$H_0 : \hat{\beta}_2 + 11\hat{\beta}_3 + \hat{\beta}_4 + 11\hat{\beta}_5 = 0; \quad H_1 : \hat{\beta}_2 + 11\hat{\beta}_3 + \hat{\beta}_4 + 11\hat{\beta}_5 \neq 0$$

which yields an F-statistic of $F = 102.99$

Polynomials + Logarithms + Interactions

- We then compute the standard error for the partial effect as:

$$SE(\Delta \log(\widehat{Earnings})) = \frac{|0.025|}{\sqrt{102.99}} = 0.0025$$

  which implies a 95% CI around the 0.025 (2.5%) partial effect on log(earnings) for increasing *Nyears* from 5 to 6 among males of:

  $$[0.025 - 1.96 \times 0.0025, 0.025 + 1.96 \times 0.0025] = [0.020, 0.029]$$

- In other words, the partial effect is 2.5% with a 95% CI of [2.0%, 2.9%]

Polynomials + Logarithms + Interactions

- Partial effect for females ($Male_i = 0$) in general terms of estimated regression coefficients from a change in *Nyears* from 5 to 6:

$$\begin{aligned}\Delta\log(\widehat{Earnings}) &= [\hat{\beta}_2 6 + \hat{\beta}_3 6^2 + \hat{\beta}_4(6 \times 0) + \hat{\beta}_5(6^2 \times 0)] \\ &\quad - [\hat{\beta}_2 5 + \hat{\beta}_3 5^2 + \hat{\beta}_4(5 \times 0) + \hat{\beta}_5(5^2 \times 0)] \\ &= \hat{\beta}_2(6-5) + \hat{\beta}_3(36-25) \\ &= \hat{\beta}_2 + 11\hat{\beta}_3\end{aligned}$$

- Next, we conduct the following joint null hypothesis test:

$$H_0 : \hat{\beta}_2 + 11\hat{\beta}_3 = 0; \quad H_1 : \hat{\beta}_2 + 11\hat{\beta}_3 \neq 0$$

which yields an F-statistic of $F = 141.06$

Polynomials + Logarithms + Interactions

- We then compute the standard error for the partial effect as:

$$SE(\Delta \log(\widehat{Earnings})) = \frac{|0.038|}{\sqrt{141.06}} = 0.0032$$

  which implies a 95% CI around the 0.038 (3.8%) partial effect on log(earnings) for increasing *Nyears* from 5 to 6 among females of:

$$[0.038 - 1.96 \times 0.0032, 0.038 + 1.96 \times 0.0032] = [0.032, 0.044]$$

- In other words, the partial effect is 3.8% with a 95% CI of [3.2%, 4.4%]

Differences-in-Differences

- A final application of non-linear models involves one of the most important tools in applied empirical research: differences-in-differences estimation
- It is a method that is widely used in for program evaluation of major policy interventions
  - tax changes
  - changes in businesses' investment incentives
  - construction of new schools
- Application of the Binary-Binary interactions model, except we require having data over $t = 1, \ldots, T$ periods over time for a set of $i = 1, \ldots, N$ cross-sectional units
  - We get to observe cross-sectional unit i for T periods in a row
  - Example: we get to see the same worker i's $Earnings_{it}$ for 8 years in a row

Differences-in-Differences

General set-up

- Suppose we have observe some outcome Y_{it} $i = 1, \ldots, N$ individuals across $t = 1, \ldots, T$ periods
- Further suppose that a subset $M < N$ individuals are exposed to a "policy" in period $K < T$
  - treatment group: M individuals exposed to the policy
  - control group: $N - M$ individuals not exposed to the policy
- Example: we observe $Earnings_{it}$ for $i = 1, \ldots, 1000$ workers for the years $t = 2000, \ldots, 2010$. 300 workers are given training in 2005.
  - $N = 1000$, $M = 300$
    $\rightarrow$ 300 treatment workers; 700 control workers
  - $T = 2010$, $K = 2005$
    $\rightarrow$ $2000, \ldots, 2004$ pre-policy; $2005, \ldots, 2010$ post-policy
  - Dataset has $1000 \times 11 = 11000$ (i, t) observations

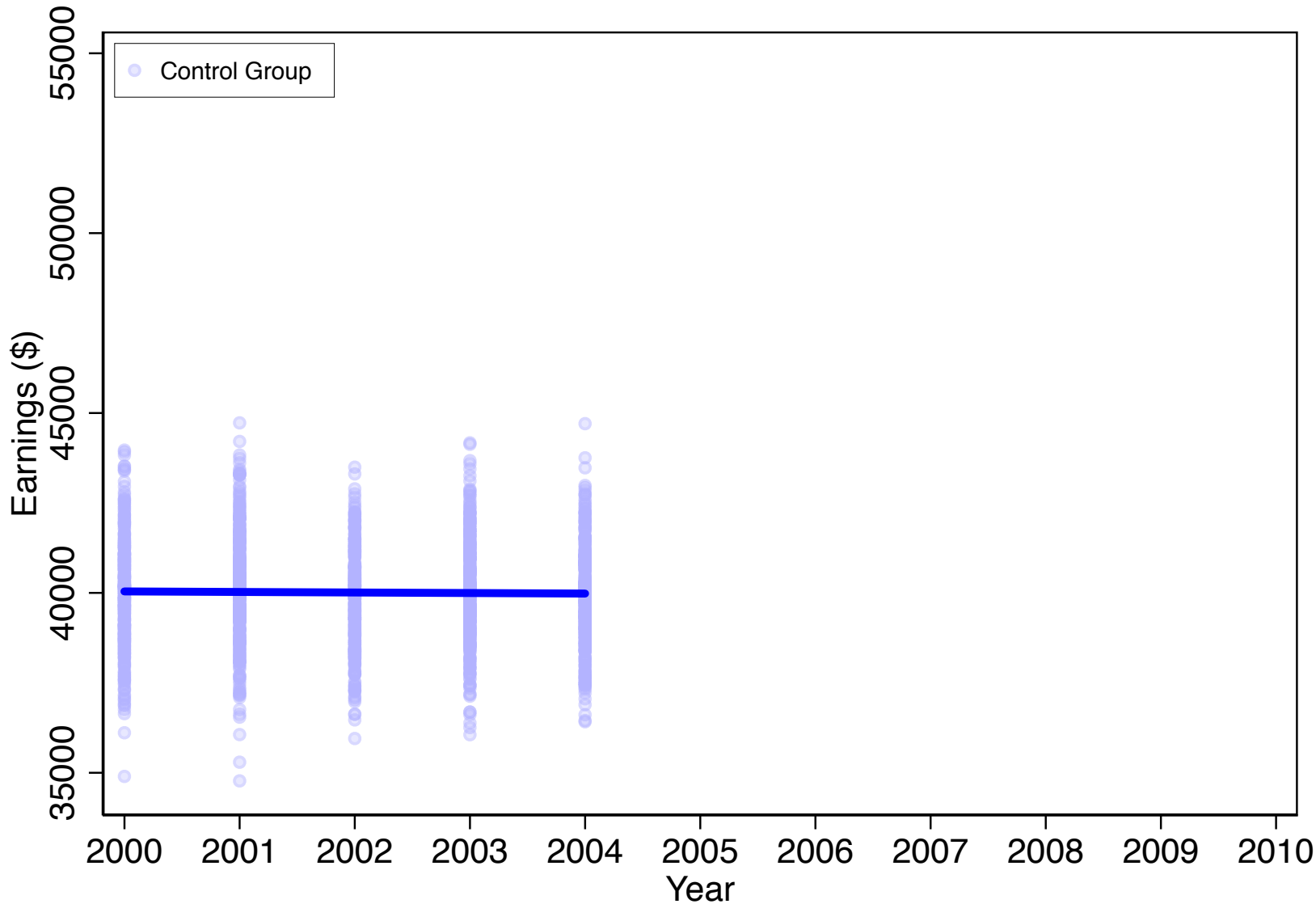
Differences-in-Differences

General set-up

	Before Period K (pre-policy period)	After Period K (post-policy period)
Treatment Group (gets the policy)	No Policy	Policy
Control Group (does not get policy)	No Policy	No Policy

Differences-in-Differences

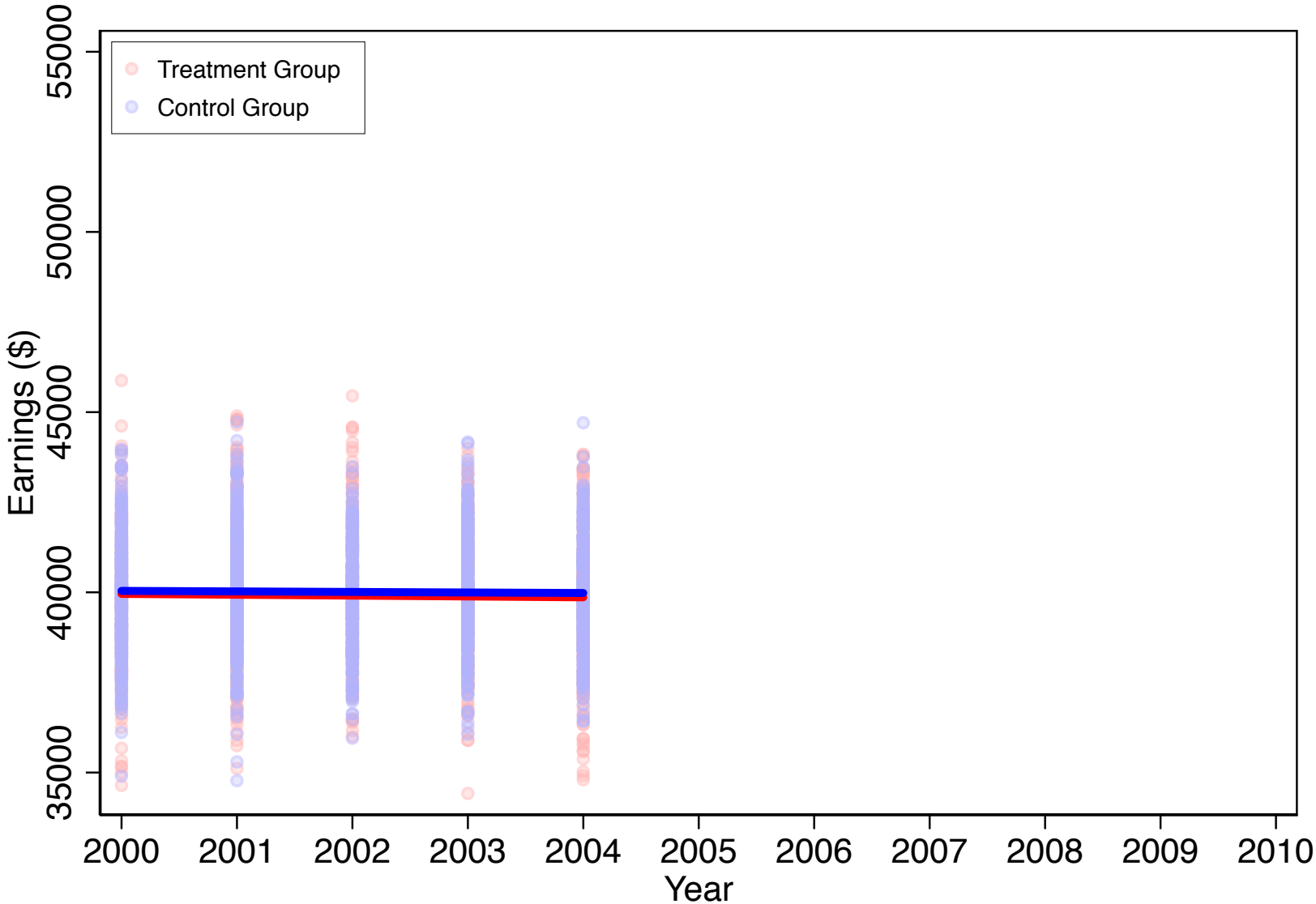
Visualising the data



Note: Job Training Policy Rolled Out in 2005

Differences-in-Differences

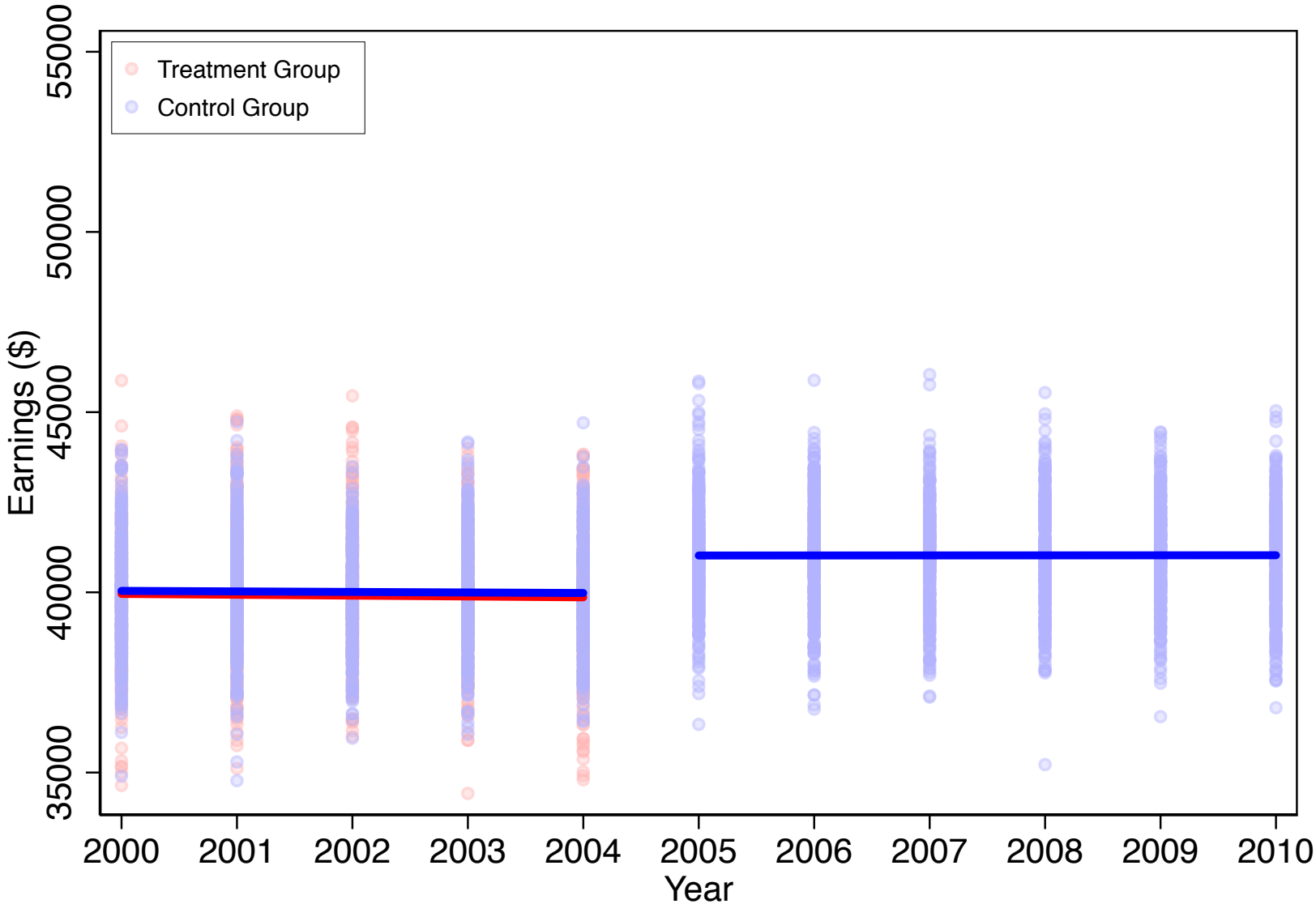
Visualising the data



Note: Job Training Policy Rolled Out in 2005

Differences-in-Differences

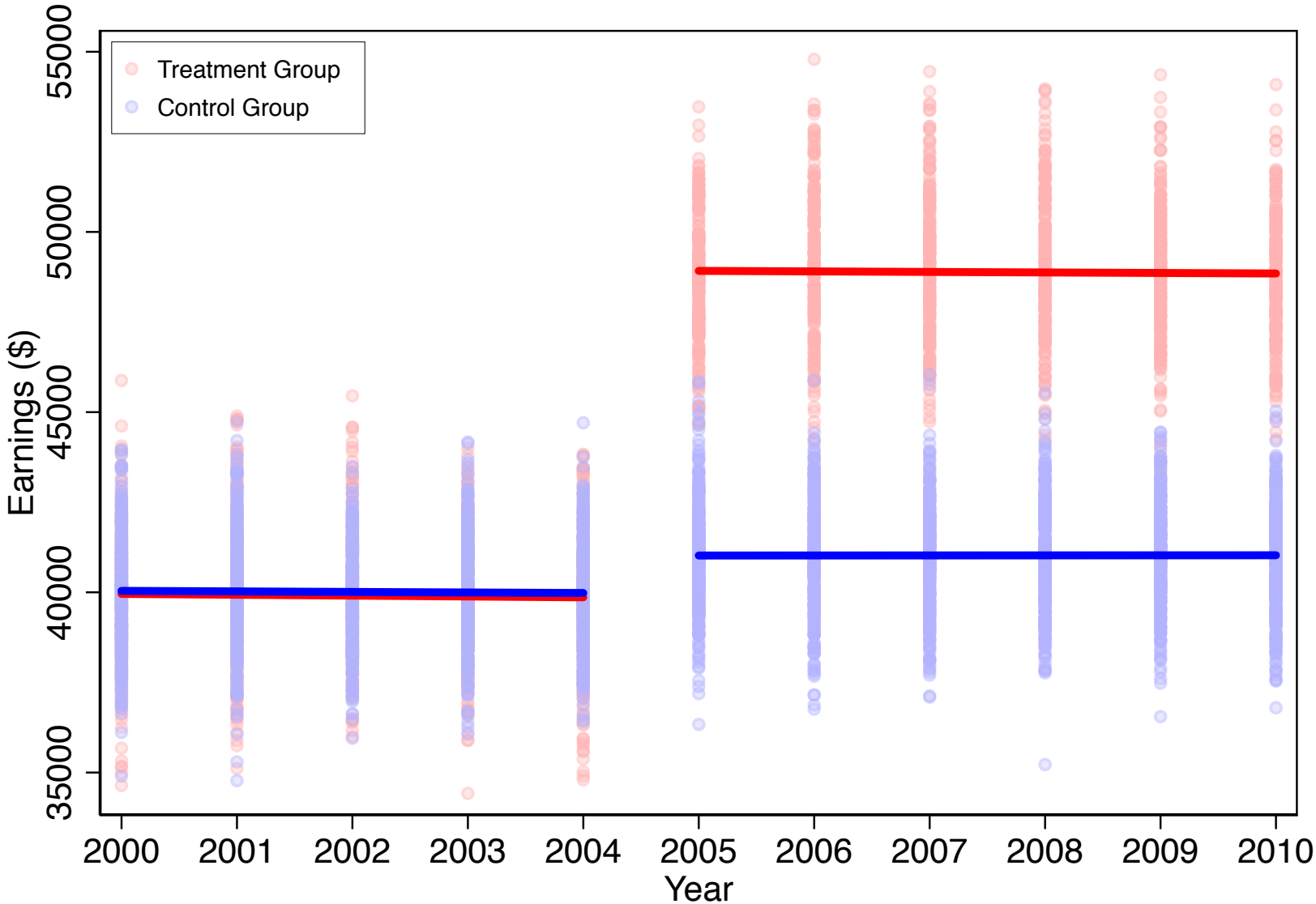
Visualising the data



Note: Job Training Policy Rolled Out in 2005

Differences-in-Differences

Visualising the data



Note: Job Training Policy Rolled Out in 2005

Differences-in-Differences

Regression Model

- Create 2 dummy variables based on the data:
  - $Treat_{it} = 1$ if $i \in N$, and 0 otherwise
    $\rightarrow$ treatment group dummy
  - $Post_{it} = 1$ if $t \geq K$, and 0 otherwise
    $\rightarrow$ post-policy period dummy
- Difference-in-differences regression:

$$Y_{it} = \beta_0 + \beta_1 Treat_{it} + \beta_2 Post_{it} + \beta_3 Treat_{it} \times Post_{it} + u_{it}$$

where

- β_0: mean of Y_{it} for the control group in the pre-policy period
- β_1: controls for pre-policy period differences between treatment and control groups
- β_2: control for post-policy trends in the outcome variable
- β_3: difference-in-difference estimator the policy effect

Differences-in-Differences

Example

- Difference-in-differences estimate of the impact of training from our example:

$$\widehat{Earnings}_{it} = \underset{(38.91)}{40011.71} - \underset{(73.01)}{102.19}\, Treat_{it} + \underset{(52.49)}{1014.25}\, Post_{it} + \underset{(99.67)}{7959.99}\, Treat_{it} \times Post_{it}$$

- Interpretation: job yields a (statistically significant) \$7,959.99 increase in $Earnings_{it}$
  - compares to a \$40,011.71 baseline $Earnings_{it}$ level in the control group during the pre-policy period

Summary of Nonlinear Regression

- We have substantially extended our toolkit for doing econometric analysis using nonlinear regression
  - Polynomial models allows for potentially rich curvature in regression functions
  - Logarithmic models permit different interpretations of results in terms of levels, percent changes, and elasticities
  - Interactions allow for different shifts, slopes, curves, elasticities with regression functions among different subgroups in our datasets (e.g., male vs female, educated vs non-educated, immigrants vs natives, etc.)

Summary of Nonlinear Regression

- All of the nonlinear specification types can be combined:
  - 4 polynomial models: linear, quadratic, cubic, quartic
  - 4 linear/log models: linear-linear, linear-log, log-linear, log-log
  - 3 interactions: binary-binary, binary-continuous, continuous-continuous
  - which implies $4 \times 4 \times 3 = 48$ model-type combinations
- Introduced difference-in-differences estimator, an application of binary-binary interactions, commonly used for policy evaluation in practice by businesses and governments
- In sum, we have a rich class of econometric models that are estimated (OLS) and tested (t-tests, F-tests) using techniques for multiple linear regression