Lecture 2: Functional Forms and Random Samples

Pierre Biscaye

Fall 2022

Linear Models

- Linear Regression Models can model non-linear relationships

$$y_i = \beta_0 + \beta_1 f(x_{1i}) + \beta_2 g(x_{2i}) + ... + \epsilon_i \tag{1}$$

- Linear regression estimates the β parameters
- Restriction is that β parameters enter additively

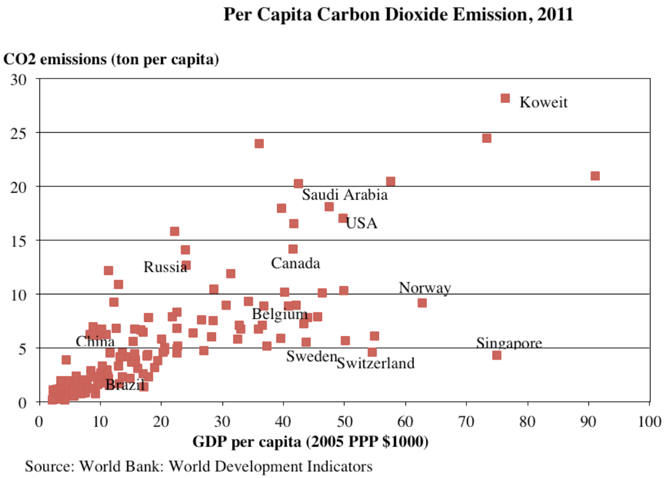
Interpretations

- when $f(x)$ is not linear in x, the interpretation changes
- Some key concepts
  1. Proportional changes: $\frac{x_1 - x_0}{x_0} = \frac{\Delta x}{x}$
  2. Percentage changes: $\frac{\Delta x}{x_0} * 100$
  3. Elasticity (η): $\frac{\Delta z}{z} / \frac{\Delta x}{x}$
     - $\eta < 1$ indicates a relationship is *inelastic*. $\eta > 1$ indicates a relationship is *elastic*

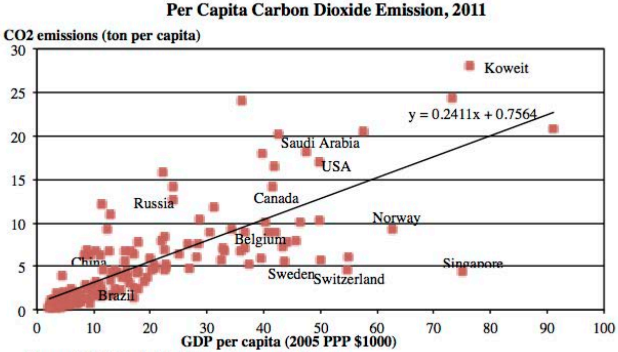
Simple case: $f(x)$ is linear

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{2}$$

CO2 and GDP


Per Capita Carbon Dioxide Emission, 2011
CO2 emissions (ton per capita)
30
25
20
15
10
5
0
Koweit
Saudi Arabia
USA
Russia
Canada
Norway
Belgium
China
Singapore
Sweden
Switzerland
Brazil
0
10
20
30
40
50
60
70
80
90
100
GDP per capita (2005 PPP $1000)
Source: World Bank: World Development Indicators

CO2 and GDP



Source: World Bank: World Development Indicators

Interpretations of the regression line

$$\frac{CO_{2i}}{Pop_i} = 0.75 + 0.24\frac{GDP_i}{Pop_i} + \epsilon_i \tag{3}$$

- $\beta_1 = 0.24$ is the *slope* parameter
  - Interpretation: as x changes, how much does y change on average? β_1 estimates $\frac{\partial y}{\partial x}$
  - 1 additional unit of $\frac{GDP}{Pop}$ is associated with 0.24 additional units of $\frac{CO_2}{Pop}$
  - Units: $\frac{GDP}{Pop}$ is in units of $1000 2005 PPP/Cap; $\frac{CO_2}{Pop}$ is in units of tons/Cap
  - Units are essential for interpretation (and estimation)

Interpretation of β_0

- $\beta_0 = 0.75$ is the intercept parameter
- Structural interpretation: value of y when all x variables are 0
- Often, β_0 will not have a meaningful interpretation
- Units of β_0 are in units of the y variable

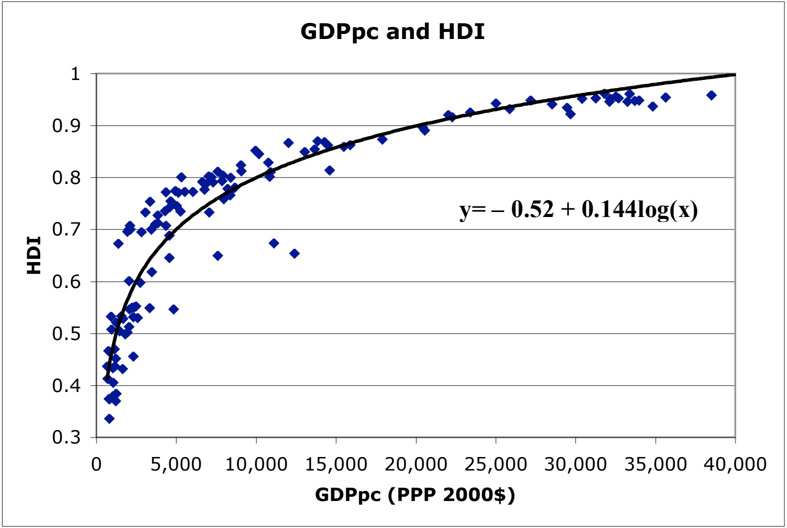
What if $f(x)$ is not linear?

Common case: linear-logarithmic relationship

$$y_i = \beta_0 + \beta_1 \log x_i + \epsilon_i \tag{4}$$

This will change the interpretation of our coefficients

HDI and GDP


GDPpc and HDI
HDI
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
y= – 0.52 + 0.144log(x)
0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
40,000
GDPpc (PPP 2000$)

Change the statistical model

- Linear-logarithmic model

$$HDI_i = \beta_0 + \beta_1 log(\frac{GDP_i}{Pop_i}) + \epsilon_i \tag{5}$$

- In this case

$$HDI_i = -0.52 + 0.14 * log(\frac{GDP_i}{Pop_i}) + \epsilon_i \tag{6}$$

Interpretation in linear-log model: math

- How do we interpret β_1?
- Remember that we are interested in the partial derivative $\frac{\partial y}{\partial x}$
- Recall: $\frac{\partial log(x)}{\partial(x)} = \frac{1}{x}$

$$\frac{\partial HDI}{\partial(\frac{GDP}{Pop})} = \frac{0.14}{\frac{GDP}{Pop}} \tag{7}$$

For small changes (represented by Δ), we can rearrange to write

$$\Delta HDI \approx 0.14 \frac{\Delta(\frac{GDP}{Pop})}{\frac{GDP}{Pop}} \tag{8}$$

Interpretation in linear-log model: HDI

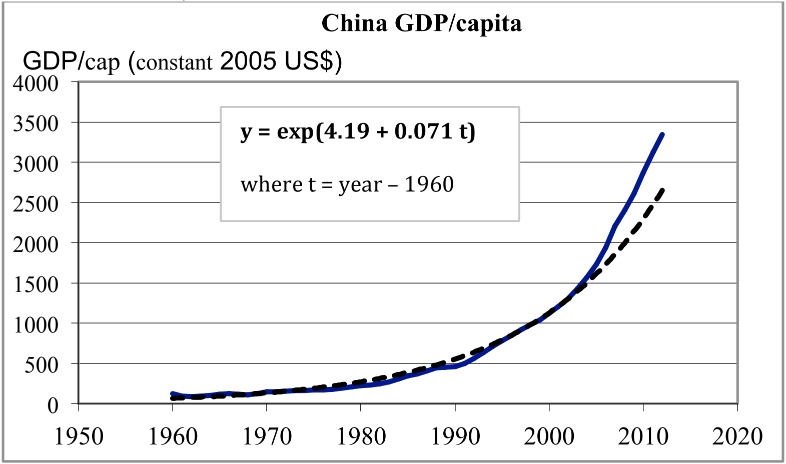
$$\Delta HDI = 0.14 \frac{\Delta(\frac{GDP}{Pop})}{\frac{GDP}{Pop}} \tag{9}$$

- A linear-logarithmic model identifies the change in y for a *proportional* change in x
- Suppose GDP/capita changes by 10% (proportional change of 0.10): $\frac{\Delta(\frac{GDP}{Pop})}{\frac{GDP}{Pop}} = 0.1$
- A 10% increase in GDP per capita would be associated with a $0.14 * 0.1 = 0.014$ *unit* increase in HDI.
- If plugging in the percentage change instead of the proportional change, need to divide by 100: $0.14 * 10/100 = 0.014$

Summary: linear-log estimates

- y is linear and x is measured in logarithms
- A 1 *percent* increase in x is associated with a $\beta_1/100$ *unit* increase in y
- When you see logarithms, think about *percent* changes

The opposite relationship: GDP growth in China



Exponential Model

$$y_t = e^{\beta_0 + \beta_1 t + \epsilon_t} \tag{10}$$

- Not linear in t: is this a problem?

Take logs: log-linear model

$$y_t = e^{\beta_0 + \beta_1 t + \epsilon_t} \tag{11}$$

$$ln(y_t) = \beta_0 + \beta_1 t + \epsilon_t \tag{12}$$

$$ln(\frac{GDP_t}{Capita_t}) = 4.19 + 0.07t \tag{13}$$

Interpretations of log-linear: math

Partial Derivative interpretation: how does y change for a marginal change in x?

$$
\begin{aligned}
log(y) &= \beta_0 + \beta_1 x + \epsilon \\
\frac{\partial log(y)}{\partial x} &= \beta_1 \\
\frac{\partial log(y)}{\partial y}\frac{\partial y}{\partial x} &= \beta_1 \\
\frac{1}{y}\frac{\partial y}{\partial x} &= \beta_1
\end{aligned}
$$

For small changes, we can rearrange to get

$$
\frac{\Delta y}{y} = \beta_1 \Delta x \tag{14}
$$

What is the proportional change in y for a unit change in x?

Interpretations of log-linear: GDP growth

$$ln(\frac{GDP_t}{Capita_t}) = 4.19 + 0.07t \tag{15}$$

- So, one year of time is associated with a 0.07 proportional (or 7%) increase in GDP/capita

Summary: log-linear estimates

- y is measured in logarithms and x is linear
- A 1 *unit* increase in x is associated with a $\beta_1 * 100$ *percent* increase in y
- When you see logarithms, think about *percent* changes
- Multiplying or dividing by 100 confusing? Think about where the proportional change is happening, and that's where you plug in a percentage change divided by 100.

Third basic case: log-log models

$$\log y_i = \beta_0 + \beta_1 \log x_i + \epsilon_i \tag{16}$$

$$\log \mathit{food}_i = \beta_0 + \beta_1 \log \mathit{income}_i + \epsilon_i \tag{17}$$

Interpreting log-log models: math

Again, start with the partial derivative

$$\begin{aligned}
\log \textit{food} &= \beta_0 + \beta_1 \log \textit{income} + \epsilon \\
\frac{\partial \log \textit{food}}{\partial \textit{income}} &= \beta_1 \frac{\partial \log \textit{income}}{\partial \textit{income}} \\
\frac{1}{\textit{food}} \frac{\partial \textit{food}}{\partial \textit{income}} &= \beta_1 \frac{1}{\textit{income}} \\
\frac{\Delta \textit{food}}{\textit{food}} &= \beta_1 \frac{\Delta \textit{income}}{\textit{income}}
\end{aligned}$$

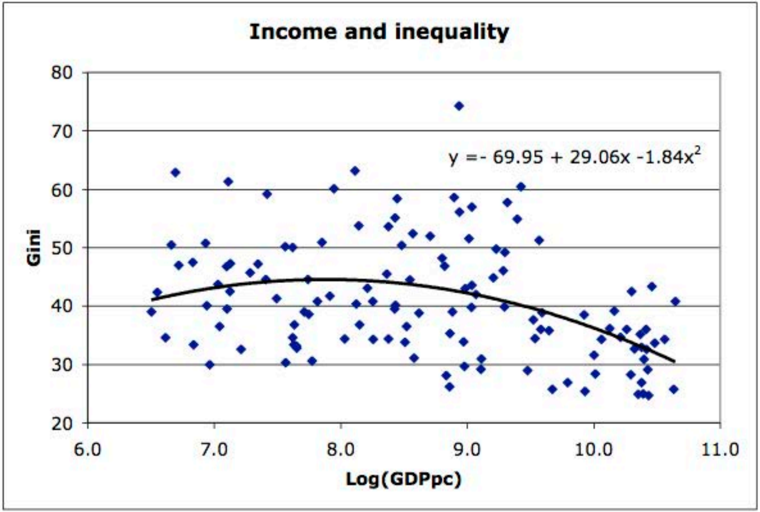
Interpreting log-log models

$$\frac{\Delta food}{food} = \beta_1 \frac{\Delta income}{income} \tag{18}$$

$$\beta_1 = \frac{\frac{\Delta food}{food}}{\frac{\Delta income}{income}} \tag{19}$$

- Observe that this is an elasticity: so a 1% increase in income is associated with a β_1% increase in food consumption
- Helpful summary of interpretations featuring logarithms in Wooldridge table 2.3

Other functional forms: Kuznet's Inverted U-Hypothesis



Accommodating non-monotonic models

- Many non-monotonic models can be accommodated in the linear regression model using polynomials
- e.g. quadratic

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \tag{20}$$

$$Gini_i = \beta_0 + \beta_1 ln(GDP_i) + \beta_2 ln(GDP_i)^2 + \epsilon_i \tag{21}$$

$$Gini_i = -70 + 29 * ln(GDP_i) - 1.84 * ln(GDP_i)^2 \tag{22}$$

Interpreting quadratic models: math

Looking at marginal changes/partial derivatives

$$\begin{aligned}
y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \\
\frac{\partial y}{\partial x} &= \beta_1 + 2\beta_2 x \\
\Delta y &= \Delta x(\beta_1 + 2\beta_2 x)
\end{aligned}$$

The marginal effect of x depends on its starting value.

Interpreting quadratic models: Gini

$$Gini_i = -70 + 29ln(GDP_i) - 1.84 * ln(GDP_i)^2$$
$$\Delta Gini = \Delta ln(GDP)(29 - 3.6ln(GDP))$$

- Need to plug in a value of $ln(GDP)$ and estimate the effect on *Gini* of an increase in $ln(GDP)$ *at that point*
- Will often interpret quadratic models at the mean of x
- Observe that β_1 and β_2 have opposite signs: the effect of an increase in x will be positive for some x and negative for others
- Could try to find the turning point

$$\frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x = 0$$
$$-\frac{\beta_1}{2\beta_2} = x$$

Interpreting quadratic models: turning point

$$-\frac{\beta_1}{2\beta_2} = x$$
$$-\frac{29}{2 * -1.84} = ln(GDP)$$

- The turning point is at $ln(GDP) = \frac{29}{2*1.84} \approx 8$
- Or, inequality increases with GDP until about $e^8 \approx \$3000/capita$ and decreases thereafter

Example: Interpreting Marginal Effects

- Suppose you have collected data on years of education and monthly wages and estimated the model

$$ln(wage_i) = \beta_0 + \beta_1 Ed_i + \epsilon_i \tag{23}$$
$$ln(wage_i) = 5.9 + 0.14 Ed_i \tag{24}$$

- How to interpret these coefficients?

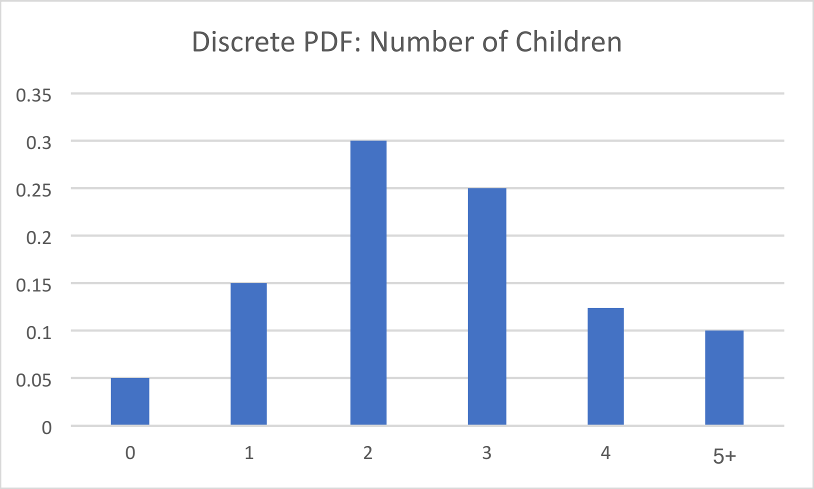
Populations and Samples, statistically

- The *Population* is every unit that could be part of your dataset
  - E.g., all UC Berkeley undergraduates, all households in Tanzania, etc.
  - Units in the population have some distribution of any variable of interest
- The *Sample* is everyone who is in your dataset
  - Since not everyone in the population is in the sample, the characteristics of the sample are different from the population
  - Any random sample from the population won't have exactly the same characteristics but will approximate them, and we rely on this for inference
- Characteristics of units drawn randomly from a population are *random variables*
  - Can be described in terms of probabilities

Characterizing Random Variables

- 2 important types of random variables
  1. Discrete variables
  2. Continuous variables
- Both types of variables can be characterized by a *probability density function* or pdf ($f(x)$)
- The pdf tells you about the probability of sampling a unit with particular characteristics

Discrete RVs



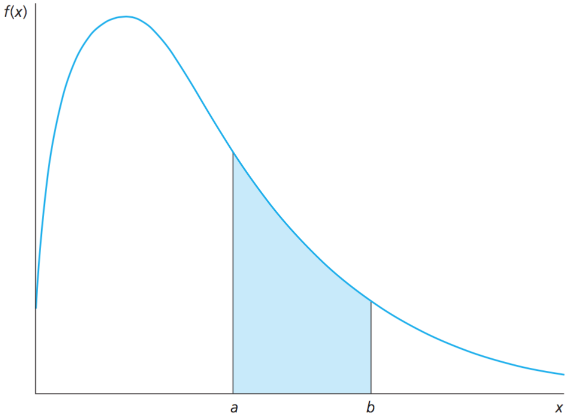
PDF: $f(x) = Pr(X = x)$

Continuous RVs

- With Continuous Random Variables X could take *any* value
- So $f(x) = Pr(X = x) = 0$ for *any* x.
- Instead, Continuous PDFs characterize the probability that X is between two values - say a and b

$$PDF = Pr(a < X < b) = \int_a^b f(x)dx \tag{25}$$

The probability that X lies between the points a and b.



Properties of pdfs

1. $f(x) \geq 0$ for all values of x
2. $\sum_x f(x) = \int_x f(x) dx = 1$

Cumulative Distribution Functions (CDFs)

- *CDFs* $F(x)$ characterize $Pr(X \leq x)$
- for a positive, discrete random variable

$$F(x) = \sum_{min(X)}^{x} f(x) \tag{26}$$

- for a continuous random variable

$$F(x) = \int_{-\infty}^{x} f(x)dx \tag{27}$$