

# **Lecture 1: The Classical Linear Regression Model**

**(Hayashi, pp. 3-14)**

- 1. Fundamental concepts: Data, Models, Estimation Procedures**
- 2. The Classical Linear Regression Model and OLS: Overview**
- 3. The Classical Linear Regression Model: Details**

# 1. Fundamental concepts: Data, Models, Estimation Procedures

**Data** is a set of observations on (economic) quantities.

A **model** is a *family* of probability distributions that could possibly have generated the data. The model is ASSUMED TO BE TRUE.

An **estimation procedure** is a *data-based protocol* for selecting the “best” or “most likely” probability distribution from the family given by the model.

A Trivial Example:

**Data:** annual incomes,  $z_i$ , of  $n = 50,000$  persons in US March 2009 CPS

**Model:** annual incomes follow a lognormal distribution with mean  $\mu$  and variance  $\sigma^2$ .

**Estimation procedure:**  $y_i = \log(z_i)$ ;  $\hat{\mu} = \sum_{i=1}^N y_i / N$ ;  $\sigma^2 = \sum_{i=1}^n (y_i - \hat{\mu})^2 / n$

A more interesting example:

**Data:**

log annual incomes of  $n = 50,000$  workers in US March 2009 CPS  
 $(y_i; i = 1, \dots, 50,000)$  [dependent variable; regressand, LHS variable]

an intercept term: a variable that equals one for every person in the sample:  
 $x_{i1} = 1; i = 1, \dots, 50,000$

age of each person, in years:  $x_{i2}; i = 1, \dots, 50,000$

sex of each person,  $x_{i3}; i = 1, \dots, 50,000$ , where  $x_{i3} = 0$  for men and 1 for women.

50,000 observations on other possible characteristics  $j = 4, \dots, K$ .

Together,  $x_{i1}, x_{i2}, \dots, x_{iK}$  are the independent variables; regressors, RHS variables

Thus:  $i = 1, \dots, n$  indexes observations;  $k = 1, \dots, K$  indexes regressors;

One possible **model** with which to summarize/interpret this data is the **classical linear regression model**; one possible **procedure** for estimating this model is OLS. In the remainder of today's lecture we'll describe the former & preview the latter.

## A couple of notes, before we move on:

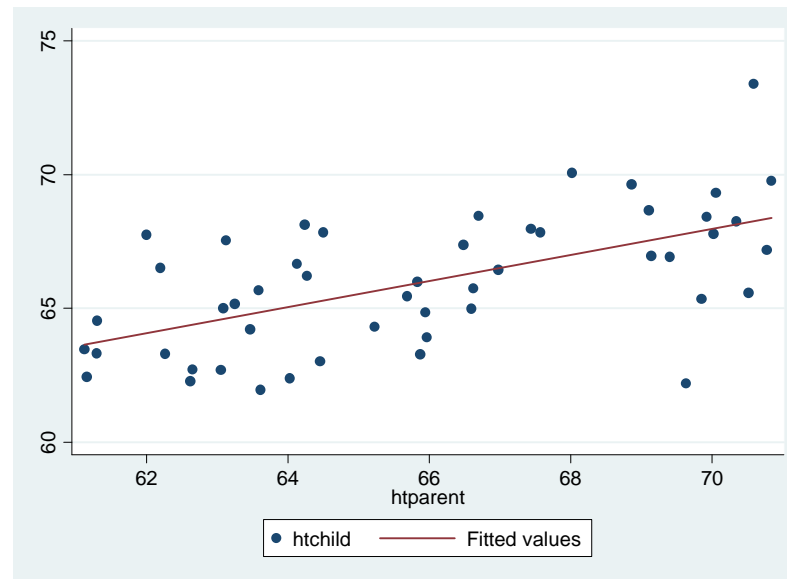
*What is the point* of estimating models like the classical linear regression model?

1. A parsimonious description (in the form of a handful of parameters) of the data.
2. We wish, ultimately, to make inferences about *causality*, in this case about the effects of  $x_1, x_2, \dots, x_K$  on  $y$ . Aside from describing the data more parsimoniously, how does regression help with this?
  - a) The earliest and still most common use of regression in attempting to infer causal relationships is as a way of estimating “partial derivatives” when  $y$  is potentially affected by more than one  $x$ , i.e.  $y_i = f(x_{i1}, x_{i2}, \dots, x_{iK}; \varepsilon_i)$ , where  $\varepsilon_i$  is a random variable. In this context, multiple regression allows us to estimate  $\partial E(y)/\partial x_{ik}$ , helping to ‘parcel out’ the competing effects of all the different  $x$ ’s on  $y$ . NOTE: simply finding  $\partial E(y)/\partial x_{ik} \neq 0$  EVEN WHEN HOLDING LOTS OF OTHER  $X$ ’s CONSTANT does not prove that  $x_k$  has a causal effect on  $y$ .
  - b) Under certain conditions, some simple extensions of the classical linear regression model (e.g. instrumental variables, regression discontinuity models), can make more powerful statements about causality.

## *Why is regression called “regression”?*

It really has nothing to do with what regression actually does; it comes from the first *application* of the technique.

Francis Galton (1886) studied the bivariate relationship between parents’ and children’s height, which had the form  $y_{child} = a + b(y_{parent}) + \varepsilon$ , where  $0 < b < 1$ . :



Because the slope is less than one (and heights are approximated stationary across generations), it follows that parents who are taller than average tend to have children who are shorter than them; the opposite is true for short parents. Thus, across generations, heights ‘regress’ towards the population mean.

## 2. The Classical Linear Regression Model and OLS: Overview

The Classical Linear Regression Model consists of:

- a distribution of an error term,  $\varepsilon_i$
- plus a set of parameters  $\sigma^2$  and  $\beta_1, \beta_2, \dots, \beta_K$ , such that:

1. **Linearity:**  $y_i$  is a linear function of the  $x_i$ 's, plus an error term,  $\varepsilon_i$ , *i.e.*:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{1k} x_{iK} + \varepsilon_i, \text{ or } y_i = \sum_{k=1}^K \beta_k x_{ik} + \varepsilon_i$$

2. **Strict Exogeneity:** The expected value of  $\varepsilon_i$  equals zero for every  $i$ . Thus, it does not depend on  $x_{i1}, x_{i2}, \dots, x_{iK}$  (the observed characteristics of that person/observation), nor on  $x_{j1}, x_{j2}, \dots, x_{jK}, j \neq i$  (the observed characteristics of any other person/observation).

3. **No Multicollinearity:** Consider all our observations on two of our  $x$  variables, say  $x_{i2}$  (height), and on another variable,  $x_{i4}$  (e.g. weight). It cannot be case that these variables are perfectly linearly related, *i.e.* that  $x_{i2} = c + dx_{i4}$ . More generally, we cannot have  $x_{ip} = c + dx_{iq}$  for any  $p \neq q$ . (more realistic examples: experience and tenure; age, cohort and year).

4. **Homoskedasticity:** The variance of  $\varepsilon_i$  equals a constant,  $\sigma^2$  for every  $i$ . Thus, it does not depend on  $x_{i1}, x_{i2}, \dots, x_{iK}$  (the observed characteristics of that person/observation), nor on  $x_{j1}, x_{j2}, \dots, x_{jK}, j \neq i$  (the observed characteristics of any other person/observation).

Ordinary Least Squares (OLS) is an *Estimation Procedure* for the Classical Linear Regression Model, which proceeds as follows:

-define  $b_k$  as the *OLS estimate* of  $\beta_k$ ;  $s^2$  as the OLS estimate of  $\sigma^2$

-define  $\hat{y}_i = \sum_{k=1}^K b_k x_{ik}$  as the “predicted” value of  $y_i$  (it will equal  $E(y | \mathbf{x})$ ).

-define  $e_i = y_i - \hat{y}_i = y_i - \sum_{k=1}^K b_k x_{ik}$  as the *residual* for observation  $i$ .

Then the OLS *procedure* is to:

1. Choose  $b_1, \dots, b_K$  to minimize the sum of squared residuals:

$$\text{Min} \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \sum_{k=1}^K b_k x_{ik})^2, \text{ then}$$

2. Calculate  $s^2$  as:  $s^2 = \sum_{i=1}^n e_i^2 / (n - K)$

In what sense is the *OLS estimation procedure* a “good”/ “correct”/ or “best” way to choose estimates of  $\beta_1, \beta_2, \dots, \beta_K$  and  $\sigma^2$ ?

Much of what econometric theory does is to provide answers to questions like these, for a wide class of

Models (i.e. what we assume is true, or the “maintained hypothesis”)

and

Estimation Techniques (e.g. OLS, LAD, ML, Probit, MNL)

Another big thing we do is study how much/in what sense our estimation techniques deviate from “correctness/bestness” when the model’s maintained assumptions are violated (this is the study of bias, and inconsistency)

### 3. The Classical Linear Regression Model—Details

In matrix notation, define the  $K$ -dimensional column vectors:

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iK} \end{bmatrix}; \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}$$

So we can write an individual observation's dependent variable as  $y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i$ .  
And define  $\mathbf{y}$  ( $n \times 1$ ),  $\boldsymbol{\varepsilon}$  ( $n \times 1$ ), and  $\mathbf{X}$  ( $n \times K$ ) as:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}; \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nK} \end{bmatrix}$$

Note that the rows of  $\mathbf{X}$  correspond to observations, the columns to (independent) variables. (It will in general be much taller than it is wide.)

Putting these together, we can write *Assumption 1 (linearity)* as:

$$\underbrace{\mathbf{y}}_{(n \times 1)} = \underbrace{\mathbf{X}}_{(n \times K)} \underbrace{\boldsymbol{\beta}}_{(K \times 1)} + \underbrace{\boldsymbol{\varepsilon}}_{(n \times 1)}$$

*Notes on the Linearity Assumption:*

Not as restrictive as it seems, because we can:

-transform the variables (e.g. use  $\log y$  instead of  $y$ )

-add higher-order terms (e.g. enter  $x_{i2} = z_i$ ,  $x_{i3} = z_i^2$ , and  $x_{i4} = z_i^3$  as three distinct variables, each with its own coefficient.)

-add interaction terms, to let the marginal effect of one variable vary with the level of another, (e.g. enter  $x_{i2} = z_{i2}$ ,  $x_{i3} = z_{i3}$ , and  $x_{i4} = z_{i2} \cdot z_{i3}$  as three distinct variables, each with its own coefficient.)

Likewise, we can write *Assumption 2 (strict exogeneity)* as:

$$E(\varepsilon_i | \mathbf{X}) = 0, \quad (i = 1, 2, \dots, n).$$

Or, put a different way,

$$E(\varepsilon_i | \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n) = 0, \quad (i = 1, 2, \dots, n).$$

*Notes on Strict Exogeneity:*

First, note that whenever  $\mathbf{X}$  includes a constant, it is immaterial whether we assume  $E(\varepsilon_i | \mathbf{X}) = 0$  or any other constant—what matters is that the expectation *be* constant.

In general, both the  $X$ 's and  $\varepsilon$ 's are random variables, so we might expect the  $X$ 's and  $\varepsilon$ 's to covary both within and between observations. This assumption rules out both within- and between- forms of dependence. Note that  $E(\varepsilon_i | \mathbf{X}) = 0$  also rules out serial dependence in the  $\varepsilon$ 's.

Examples of dependence:

***Within obs:*** Most common is a left-out variable, leading to LOVE (left-out variable error).

Example: in an earnings equation, unobserved ‘ability’ or personality factors are part of the error term. If these are positively correlated with years of schooling, then the OLS regression coefficient doesn’t correctly estimate the (partial) effect of schooling on earnings.

***Across obs:*** Most common is a lagged dependent variable, but could arise from other sources as well.

Note: because independence implies zero correlation (but not conversely), strict exogeneity is violated whenever the correlation between any  $x$  variable and  $\varepsilon$  is nonzero. Put a different way, strict exogeneity requires (among other things):

$$\text{Cov}(\varepsilon_i, x_{ik}) = 0, \text{ for all } k.$$

In matrix notation, *Assumption 3 (no multicollinearity)* is just:

The rank of the  $n \times K$  data matrix,  $\mathbf{X}$ , is  $K$  with probability one.

*Notes on Multicollinearity:*

Recall that the rank of a matrix is the number of linearly independent columns.

So, multicollinearity would occur if, for example,  $x_{ip} = c + dx_{iq}$  for any  $p \neq q$ , as already stated.

*Example of multicollinearity:*

(closely related to *identification* in econometrics)

Barry Chiswick (JPE 1978) studied a cross-section of men in the U.S. Roughly speaking, he proposed that:

$$Y^M/Y^N = a + b(\text{YSM}) + c(\text{Age}) + dX + \varepsilon. \quad (1)$$

He estimated this relationship, estimated  $b > 0$ , and calculated that immigrants would ‘overtake’ natives ( $Y^M/Y^N = 1$ ) within 20 years of arrival.

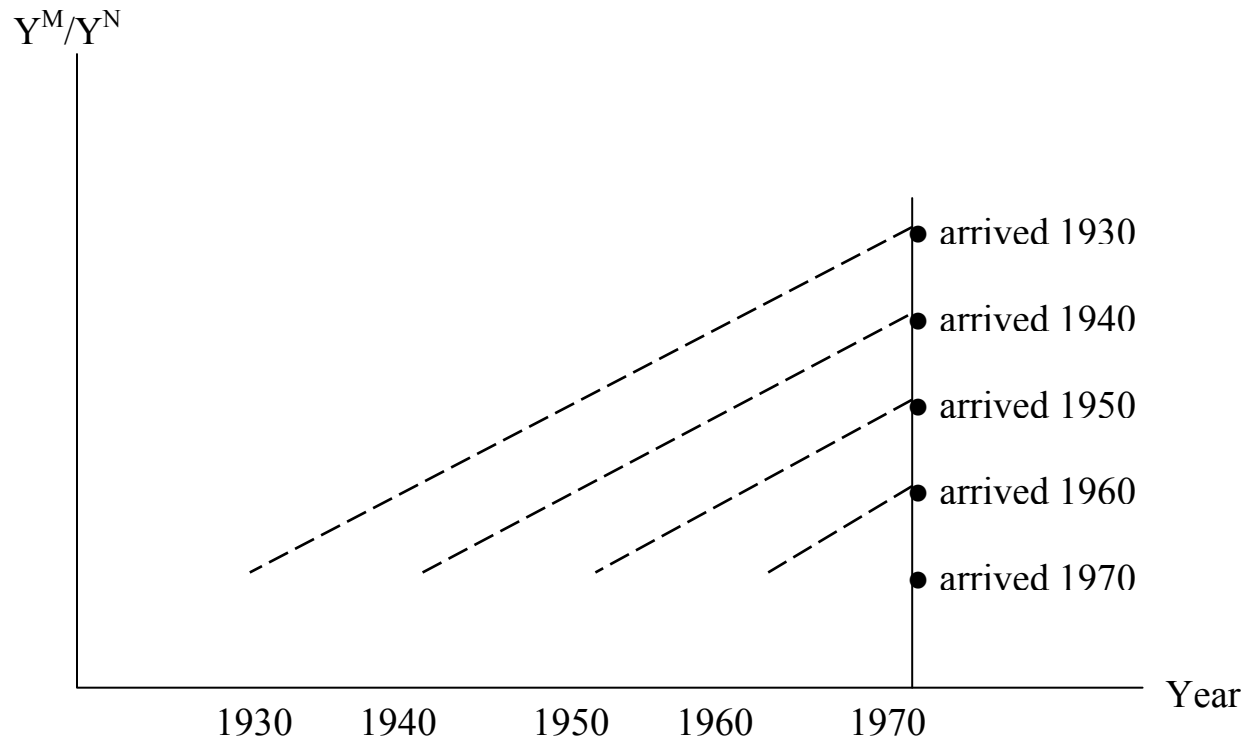
But note that  $\text{YSM} = K - \text{YOA}$ , where  $K$  is the current calendar year, and  $\text{YOA}$  (year of arrival—an ‘arrival cohort’ effect) might also affect immigrants’ relative earnings. (i.e., the model we would like to estimate contains both  $\text{YSM}$  and  $\text{YOA}$ , but that model can’t be estimated/is not identified because the two variables are perfectly collinear).

Substituting the definition of  $\text{YSM}$ , it follows that Chiswick’s model is observationally equivalent to a model where:

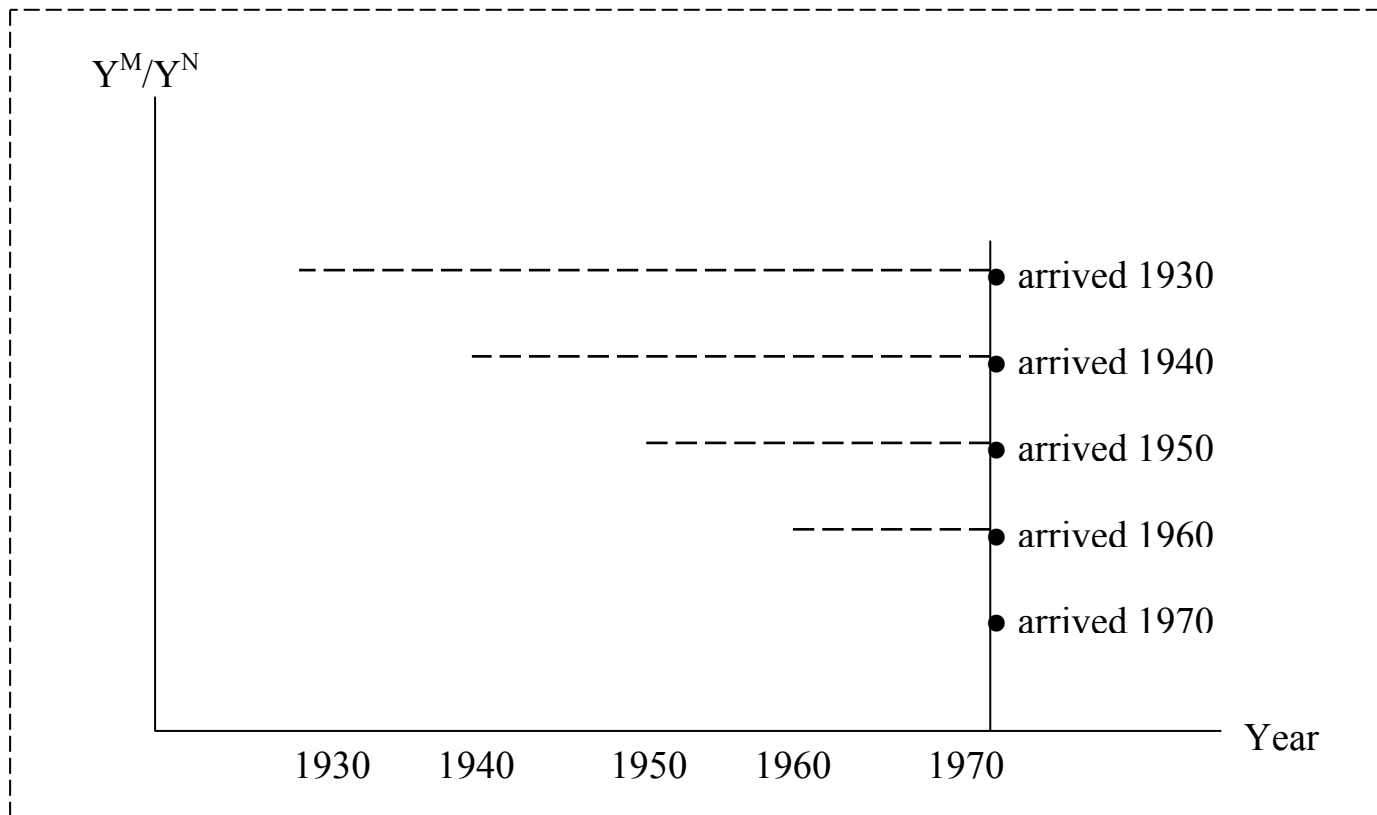
$$Y^M/Y^N = [a + bK] - b(\text{YOA}) + c(\text{Age}) + dX + \varepsilon. \quad (2)$$

So Chiswick’s data is equally consistent with a scenario in which there is strong assimilation but no cohort effects (1), or one with *no* assimilation but *declining* ‘cohort quality’ (2).

## Chiswick's 'assimilation' scenario (1):



An alternative ‘pure cohort effects’ scenario (2) that is equally consistent with the Chiswick’s cross-sectional data:



Borjas (1985) used two cross-sections of Census data to argue that the real situation in the U.S. was closer to scenario 2.

## An earlier example:

Robert M. Yerkes, *Psychological Examining in the U.S. Army*, 1921.  
(as described in Steven Jay Gould, *The Mismeasure of Man*)

“Mental tests” administered to 1.75 million WWI recruits, assigning “mental age” to adults.

Results for European Immigrants:

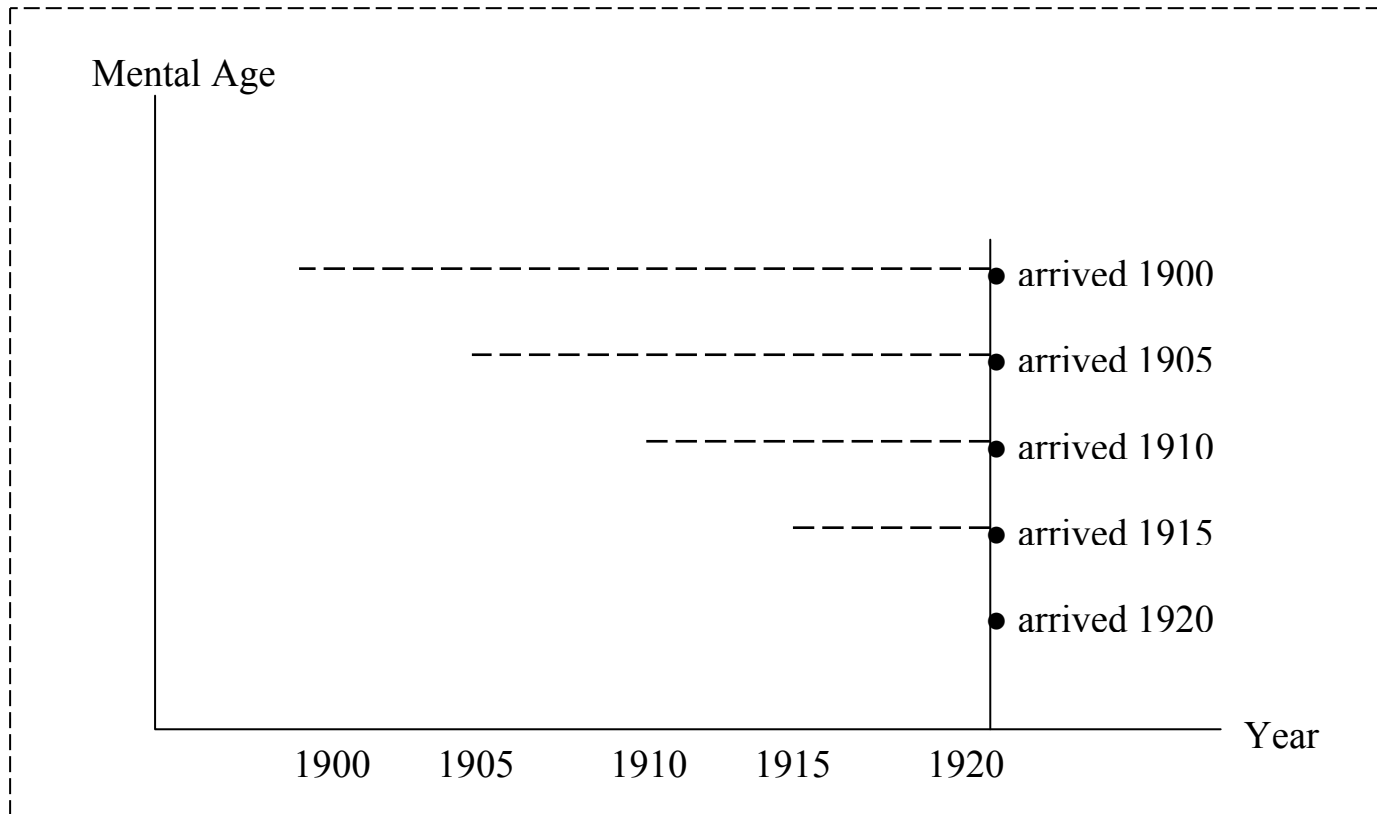
<b>Years of Residence in U.S.</b>	<b>Mean Mental Age</b>
0-5	11.29
6-10	11.70
11-15	12.53
15-20	13.50
20+	13.74

Conclusion: Recent arrivals more likely to come from ‘inferior stock’, such as Russians (11.34), Italians (11.01) and Poles (10.74), in contrast to Northern European immigrants who disproportionately arrived later.

Other possible explanation –assimilation over time—simply dismissed as implausible.

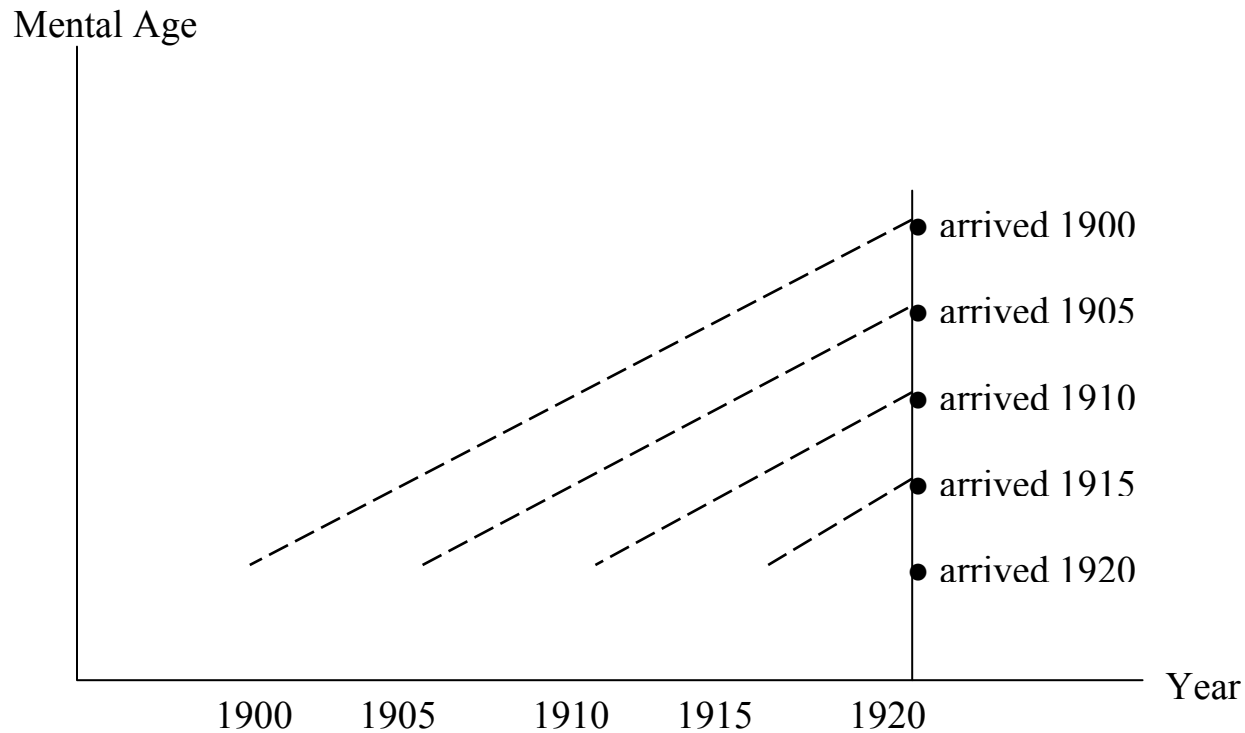
In other words, 1920 cross-section data was interpreted as consistent with:

**Scenario A (Pure Cohort Effects):**



Of course, the same data is also consistent with:

### Scenario B: Pure Assimilation Effects



Lesson: Some things are inherently unknowable from certain data sets. Apparent 'knowledge' in those cases is just the result of an implicit assumption.

**Finally, Assumption 4 (Homoskedasticity) is just:**

$$E(\varepsilon_i^2 \mid \mathbf{X}) = \sigma^2 > 0,$$

and

$$E(\varepsilon_i \varepsilon_j \mid \mathbf{X}) = 0$$

Or more succinctly,

$$E(\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}' \mid \mathbf{X}) = \sigma^2 \mathbf{I}_n.$$

*Notes on Homoskedasticity:*

1. Since  $E(\varepsilon_i | \mathbf{X}) = 0$ , homoskedasticity can also be expressed in terms of variances/covariances, i.e. as

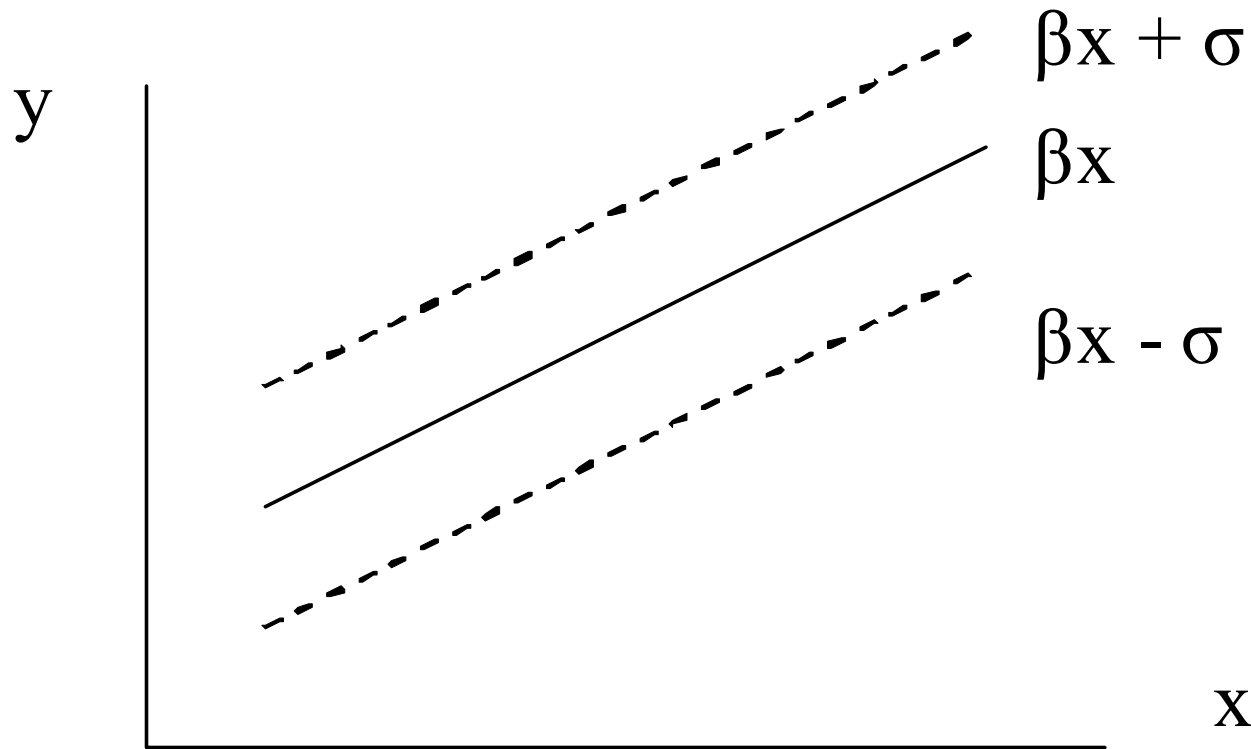
$Var(\varepsilon_i | \mathbf{X}) = \sigma^2 > 0$ ,  $Cov(\varepsilon_i \varepsilon_j | \mathbf{X}) = 0$ . (The latter implies *no serial [or spatial] correlation* of the error term), or:

$$Var(\boldsymbol{\varepsilon} | \mathbf{X}) = \sigma^2 \mathbf{I}_n.$$

2. Simply assuming that our sample of  $(y_i, \mathbf{x}_i)$  is random (iid) across observations is not sufficient to guarantee homoskedasticity as we define it which is *conditional* (on the X's) homoskedasticity. (Thus, Assumption 4 has content even when the sample is randomly drawn from the same joint distribution).

3. It is more traditional (and easier) to think of the Xs in the classical linear regression model as fixed, rather than random. (This would economize on notation a lot, b/c we could drop all the “|X’s” from our definitions of homoskedasticity). But it’s not realistic for most economic applications, and doesn’t highlight the need for  $\varepsilon_i$  to be independent of all current, past, *and* future regressors.

4. Although homoskedasticity is sometimes referred to as “spherical error variance”, there’s another sense in which it assumes the errors are “rectangular”:



The “spread” of the values of  $y$  around the regression line is the same everywhere along the line.

*In the problem set:*

- you'll get familiar with data manipulation in Stata
- you'll work through a 'practical' example of multicollinearity

*Next time:*

One (very well known and widely used) way to estimate the parameters ( $\sigma^2$  and  $\beta_1, \beta_2, \dots, \beta_K$ ) of the classical linear regression model: OLS.