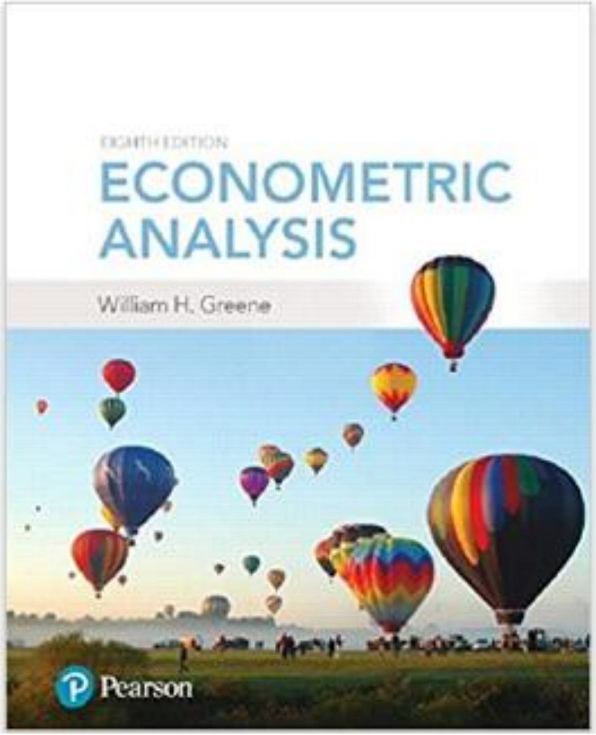
Econometrics I

Professor William Greene

Stern School of Business

Department of Economics



Econometrics I

Part 2 – Projection and Regression

Statistical Relationship

- **Objective**: Characterize the 'relationship' between a variable of interest and a set of 'related' variables
- **Context:** An inverse demand equation,
  - $P = \alpha + \beta Q + \gamma Y$, Y = income. P and Q are two random variables with a joint distribution, f(P,Q). We are interested in studying the 'relationship' between P and Q.
  - By 'relationship' we mean (usually) covariation.

Bivariate Distribution - Model for a Relationship Between Two Variables

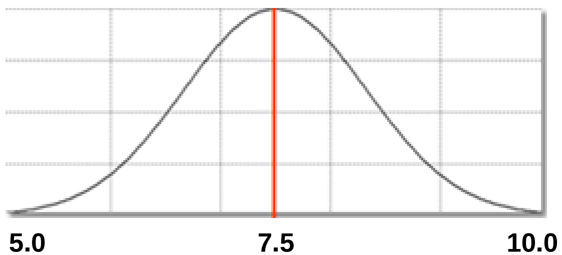
- We might posit a bivariate distribution for P and Q, f(P,Q)
- How does variation in P arise?
  - With variation in Q, and
  - Random variation in its distribution.
- There exists a conditional distribution f(P|Q) and a conditional mean function, E[P|Q]. Variation in P arises because of
  - Variation in the conditional mean,
  - Variation around the conditional mean,
  - (Possibly) variation in a covariate, Y which shifts the conditional distribution

Conditional Moments

- The conditional mean function is the ***regression function***.
  - P = E[P|Q] + (P - E[P|Q]) = E[P|Q] + ε
  - E[ε|Q] = 0 = E[ε]. Proof: (The Law of iterated expectations)
- Variance of the conditional random variable = conditional variance, or the ***scedastic function.***
- A "trivial relationship" may be written as P = h(Q) + ε, where the random variable ε = P-h(Q) has zero mean by construction. Looks like a regression "model" of sorts.
- An extension: Can we carry Y as a parameter in the bivariate distribution? Examine *E*[P|Q,Y]

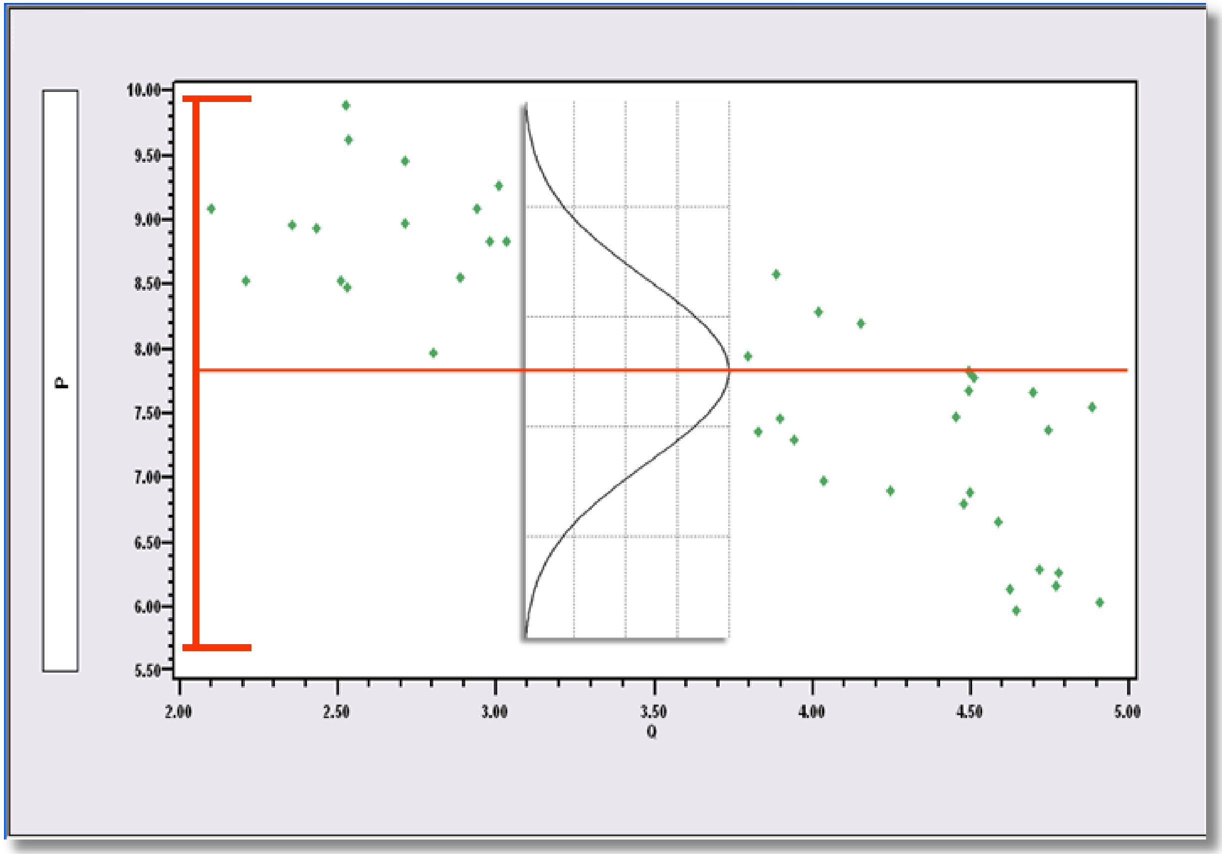
Sample Data (Experiment)

Y	Q	P
2	4.87922	7.54372
1	3.82786	7.34581
2	3.47715	8.06425
1	2.80233	7.95544
1	4.24447	6.89802
2	4.69255	7.65647
1	4.62286	6.13175
1	2.52893	8.4732
2	4.49625	7.81212
1	3.93907	7.28257
1	3.89569	7.45552
1	4.58395	6.65612
1	2.88468	8.54341
1	2.20953	8.52388
1	4.47329	6.79659
1	4.76754	6.15842
2	2.97926	8.81925
1	3.44583	7.31662
2	2.53235	9.60803
2	3.79481	7.93217
2	3.14991	8.35497
1	4.03218	6.96767
1	2.35632	8.95624
2	2.52448	9.88523
2	3.03155	8.82016
1	4.90302	6.03125
2	3.00654	9.25203
2	4.01524	8.28128
1	3.69082	7.57176
2	2.711	8.96197

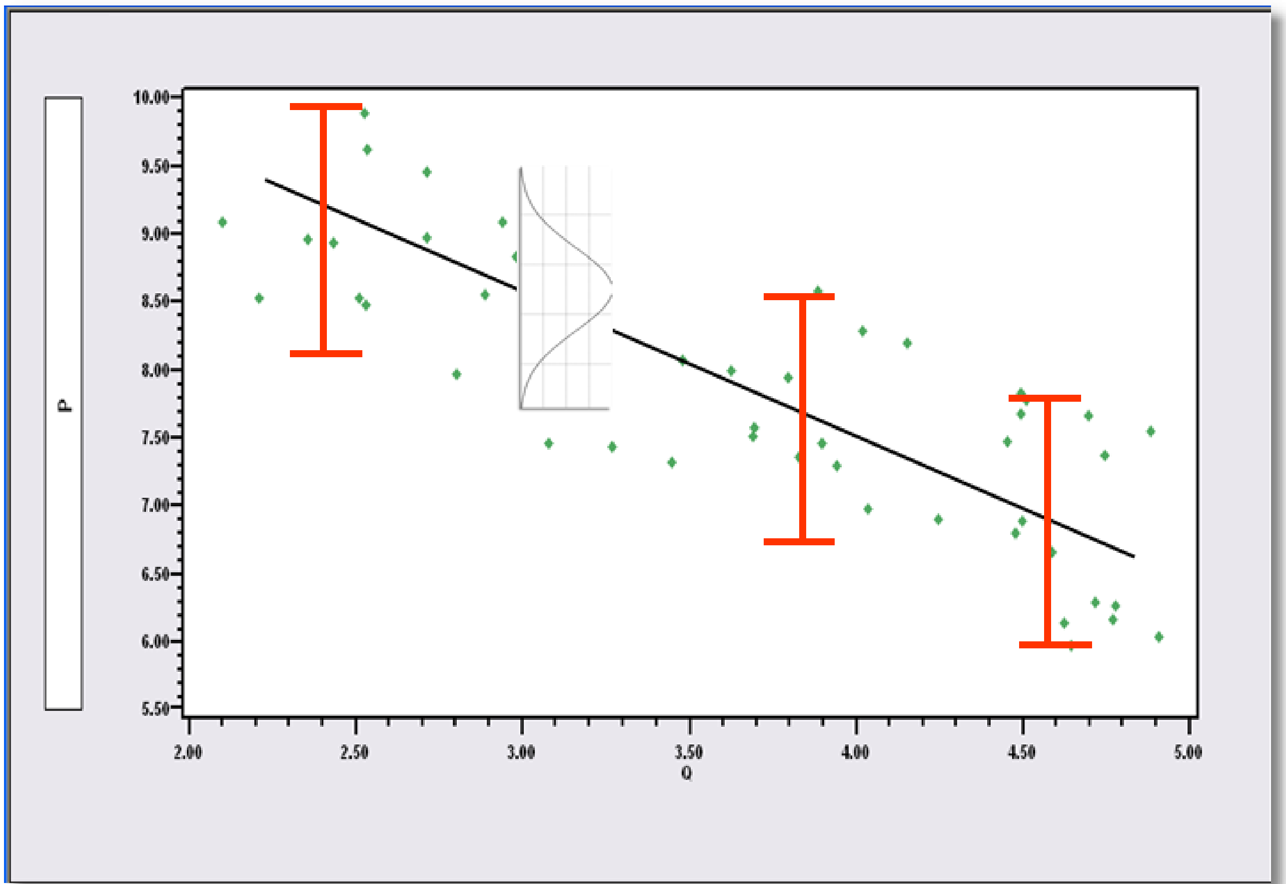


Distribution of P

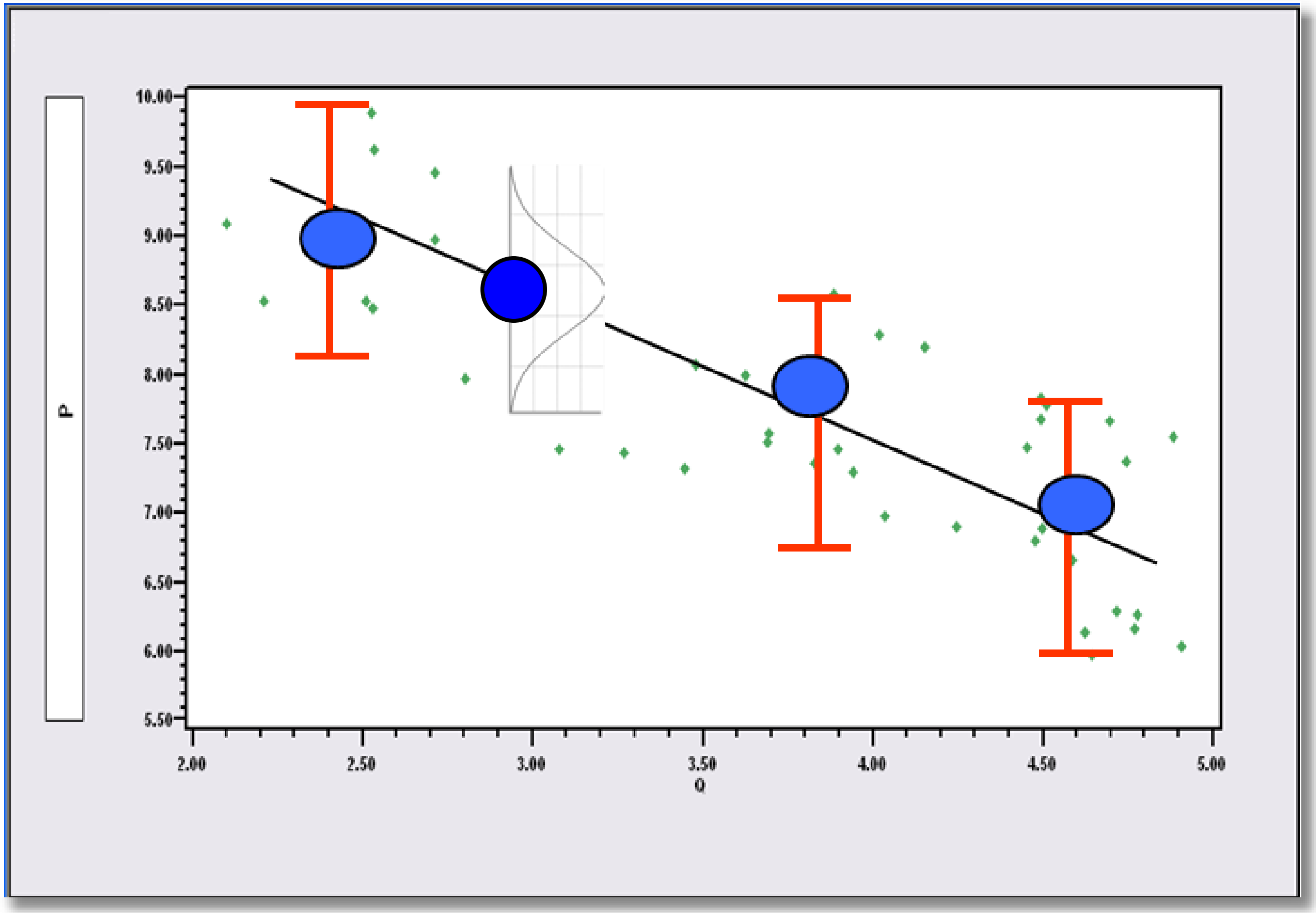
50 Observations on P and Q
Showing Variation of P Around E[P]



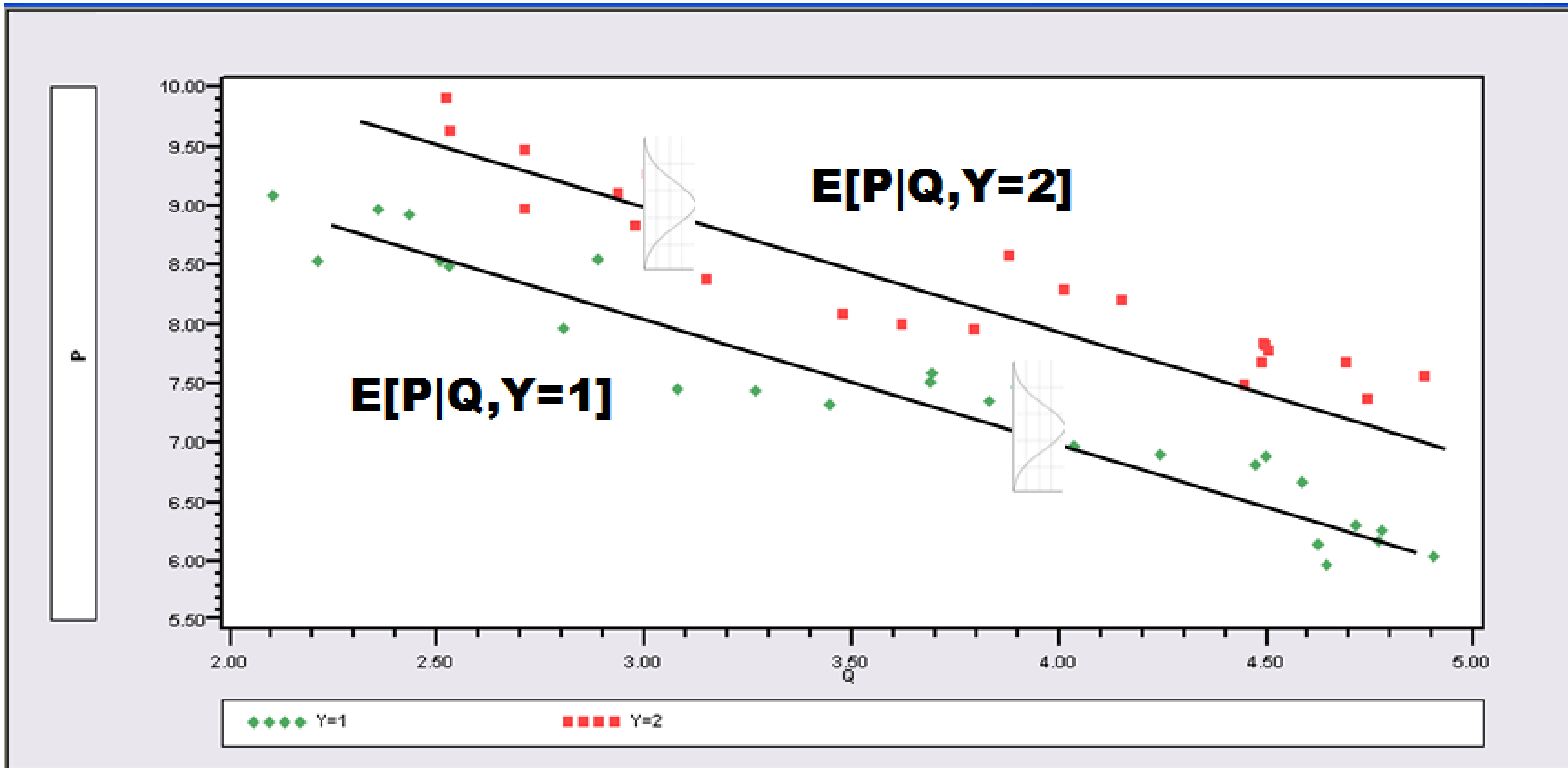
Variation Around E[P|Q] (Conditioning Reduces Variation)



Means of P for Given Group Means of Q



Another Conditioning Variable



Conditional Mean Functions

- No requirement that they be "linear" (we will discuss what we mean by linear)
- Conditional Mean function: h(X) is the function that minimizes $E_{X,Y}[Y – h(X)]^2$
- No restrictions on conditional variances at this point.

Projections and Regressions

- We explore the difference between the linear projection and the conditional mean function
- y and x are two random variables that have a bivariate distribution, f(x,y).
- Suppose there exists a *linear* function such that
- $y = \alpha + \beta x + \varepsilon$ where $E(\varepsilon|x) = 0 \Rightarrow Cov(x,\varepsilon) = 0$

Then,

$$Cov(x,y) = Cov(x,\alpha) + \beta Cov(x,x) + Cov(x,\varepsilon)$$

$$= 0 + \beta\ Var(x) + 0$$

so, $\boxed{\beta = Cov(x,y) / Var(x)}$

and $E(y) = \alpha + \beta E(x) + E(\varepsilon)$

but $E(\varepsilon) = E(\varepsilon|x) = E(0) = 0$ (Law of iterated expectations)

so $E(y) = \alpha + \beta E(x) + 0$

so, $\boxed{\alpha = E[y] - \beta E[x].}$

Regression and Projection

Does this mean $E[y|x] = \alpha + \beta x$?

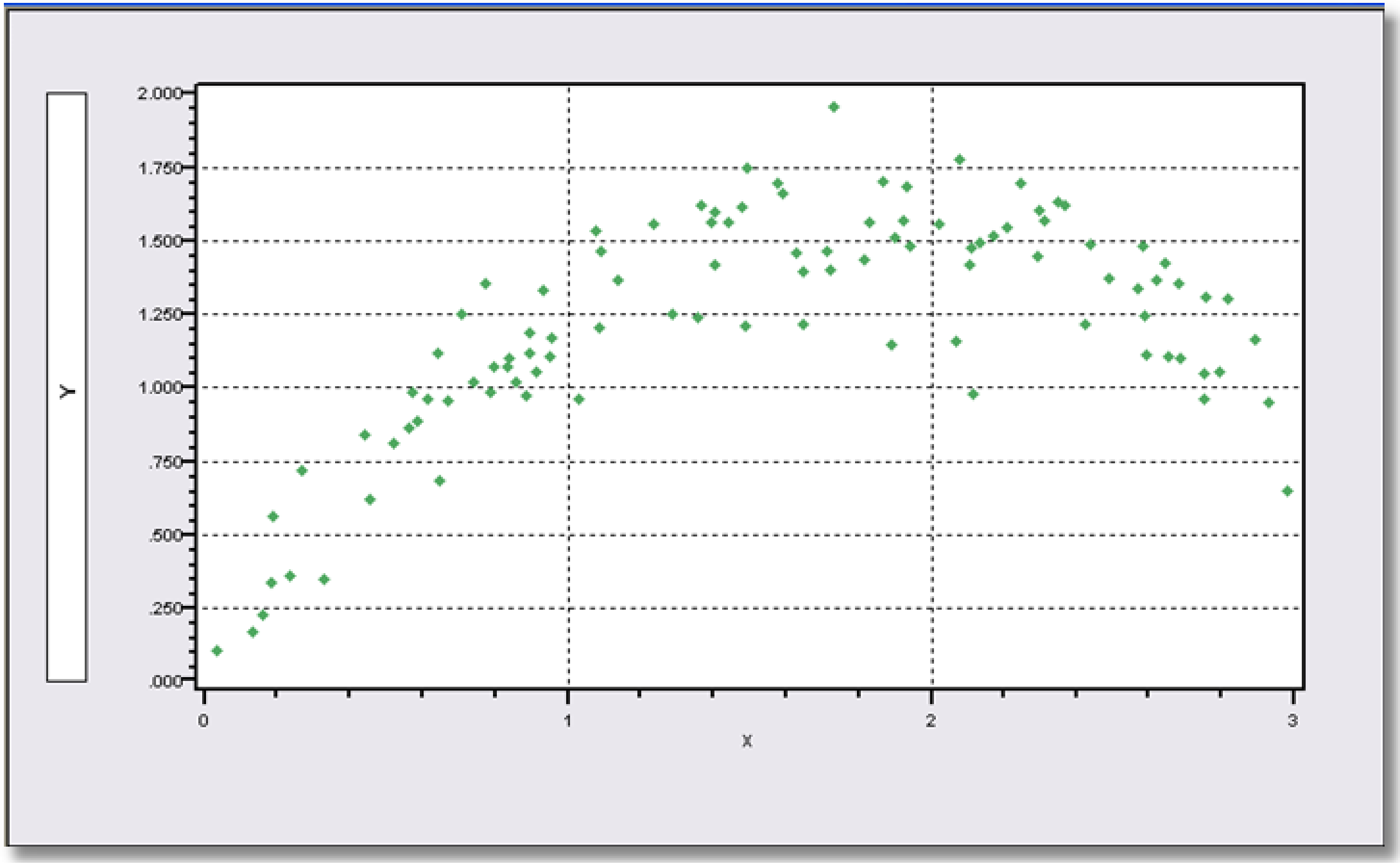
- No. This is *the* ***linear projection*** of y on x
- It is true in every bivariate distribution, whether or not $E[y|x]$ is linear in x.
- y can generally be written $y = \alpha + \beta x + \varepsilon$

  where $\varepsilon \perp x$, $\beta = Cov(x,y) / Var(x)$ etc.

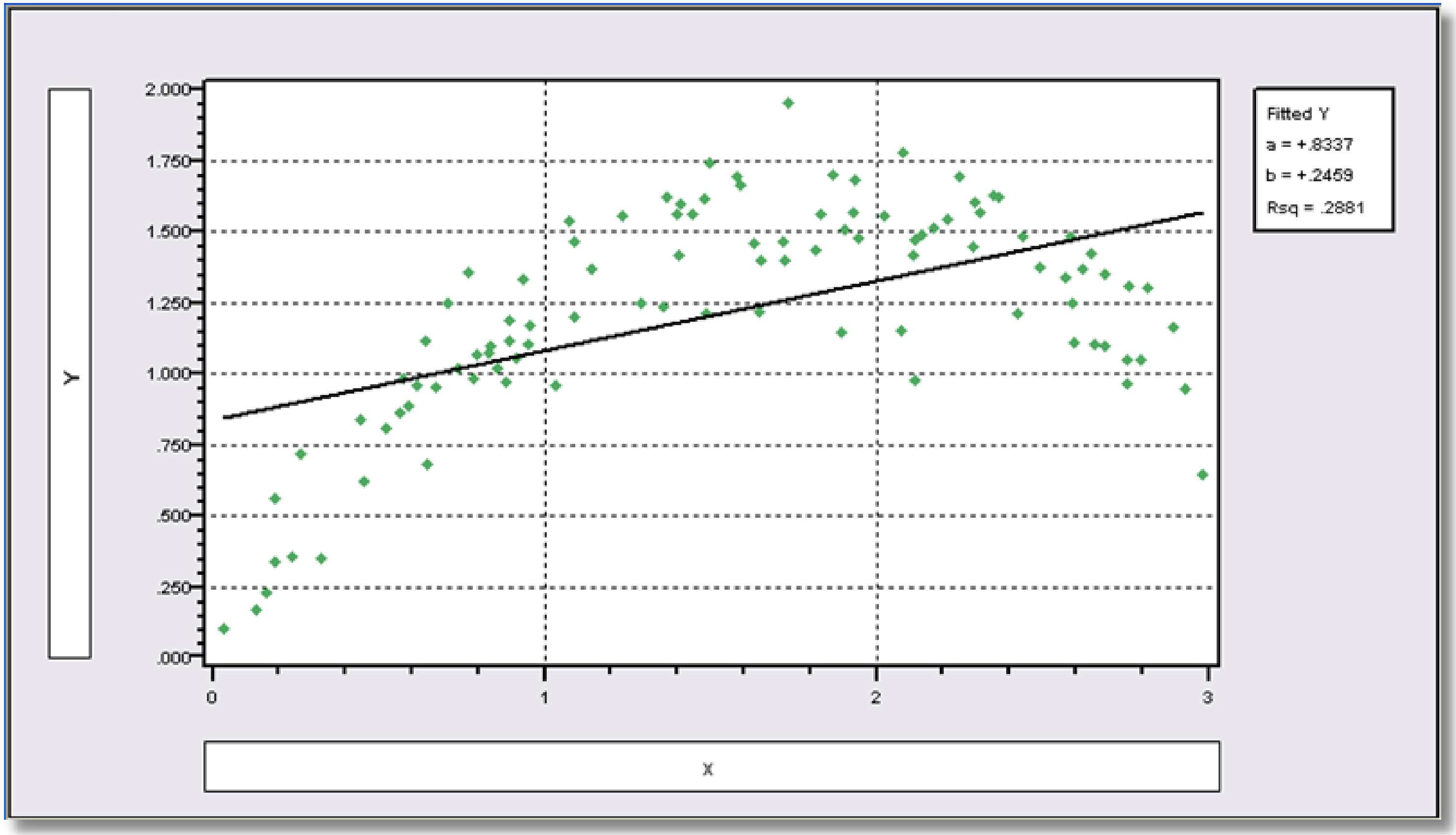
The conditional mean function is h(x) such that $y = h(x) + v$ where $E[v|h(x)] = 0$. But, h(x) does not have to be linear.

The implication: What is the result of "linearly regressing y on ," for example using least squares?

Data from a Bivariate Population



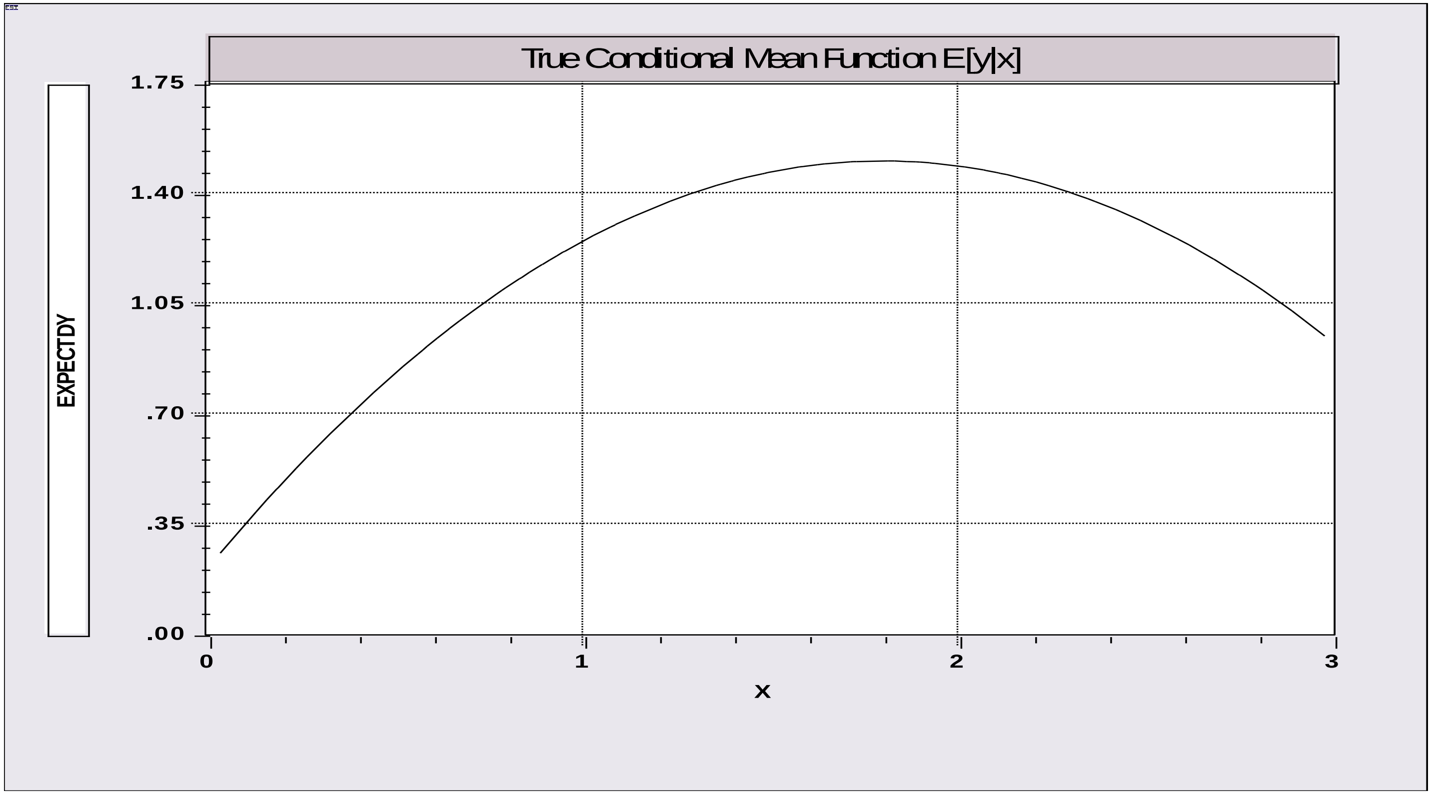
The Linear Projection Computed by Least Squares



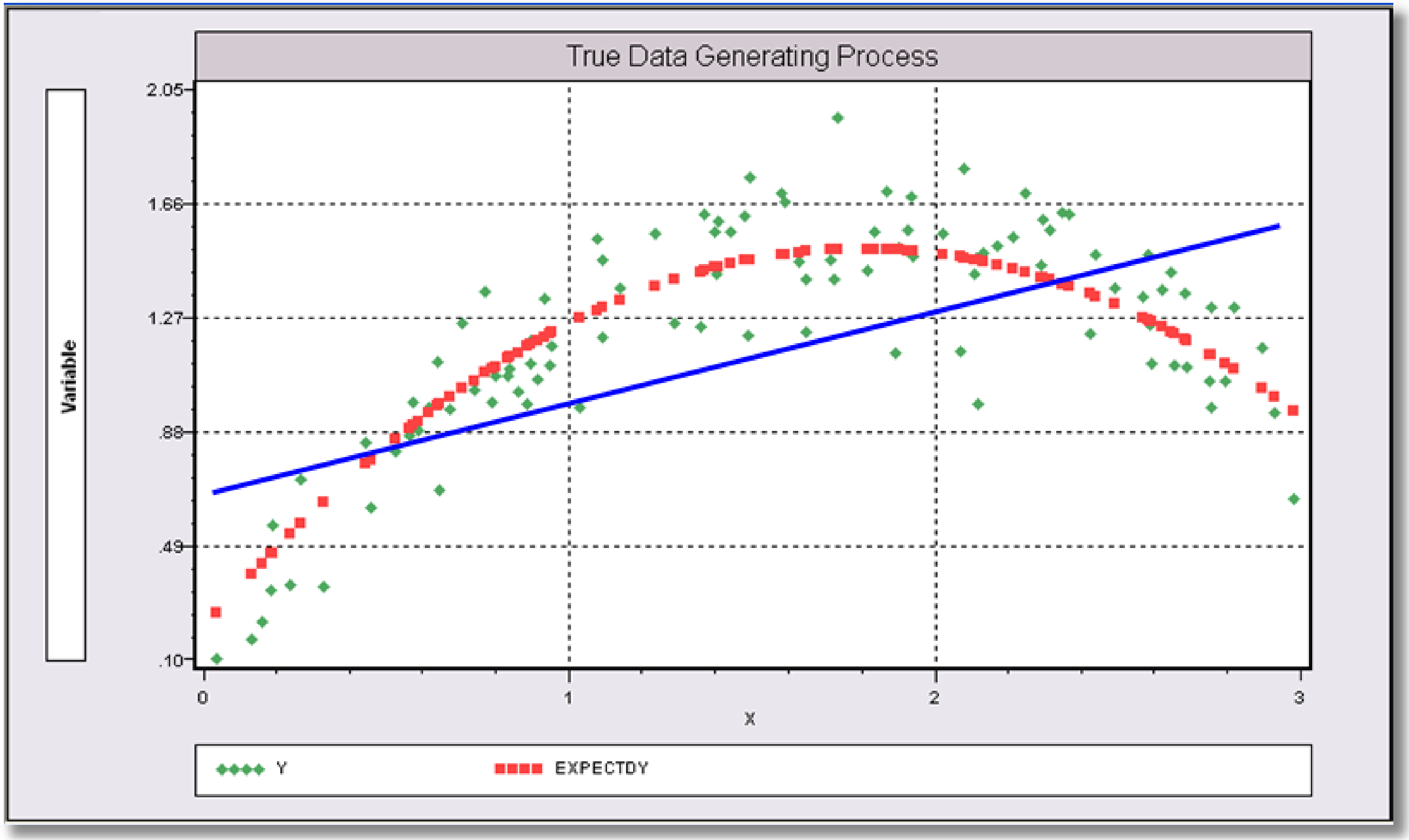
Linear Least Squares Projection

```
----------------------------------------------------------------------
Ordinary     least squares regression ............
LHS=Y        Mean                 =       1.21632
             Standard deviation   =        .37592
             Number of observs.   =           100
Model size   Parameters           =             2
             Degrees of freedom   =            98
Residuals    Sum of squares       =       9.95949
             Standard error of e  =        .31879
Fit          R-squared            =        .28812
             Adjusted R-squared   =        .28086
--------+-------------------------------------------------------------
Variable| Coefficient    Standard Error  t-ratio  P[|T|>t]   Mean of X
--------+-------------------------------------------------------------
Constant|     .83368***       .06861      12.150   .0000
       X|     .24591***       .03905       6.298   .0000      1.55603
--------+-------------------------------------------------------------
```

The True Conditional Mean Function



The True Data Generating Mechanism



What does least squares "estimate?"

Journal of Economic Growth, 5: 5–32 (March 2000)

Inequality and Growth in a Panel of Countries

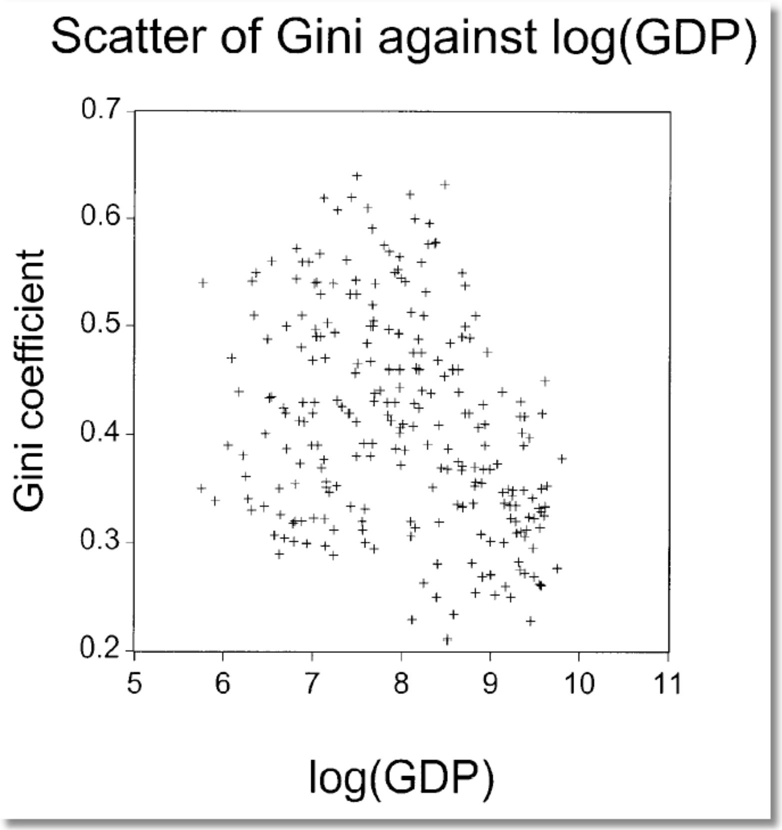
ROBERT J. BARRO

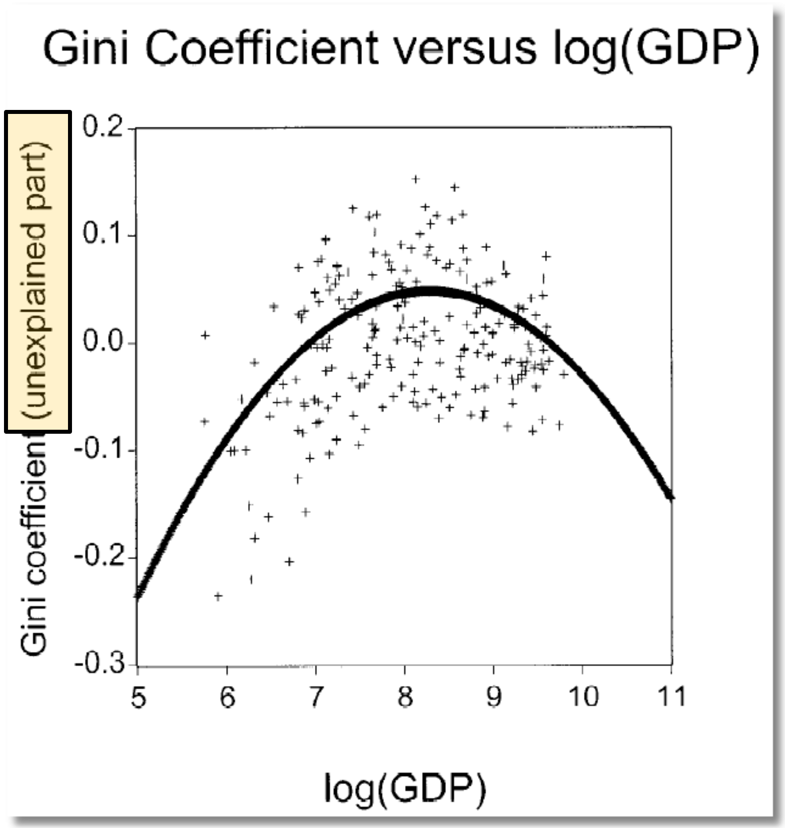
Littauer Center, Department of Economics, Harvard University, Cambridge, MA 02138

Evidence from a broad panel of countries shows little overall relation between income inequality and rates of growth and investment. For growth, higher inequality tends to retard growth in poor countries and encourage growth in richer places. The Kuznets curve—whereby inequality first increases and later decreases during the process of economic development—emerges as a clear empirical regularity. However, this relation does not explain the bulk of variations in inequality across countries or over time.

Keywords: inequality, growth, Kuznets curve, Gini coefficient

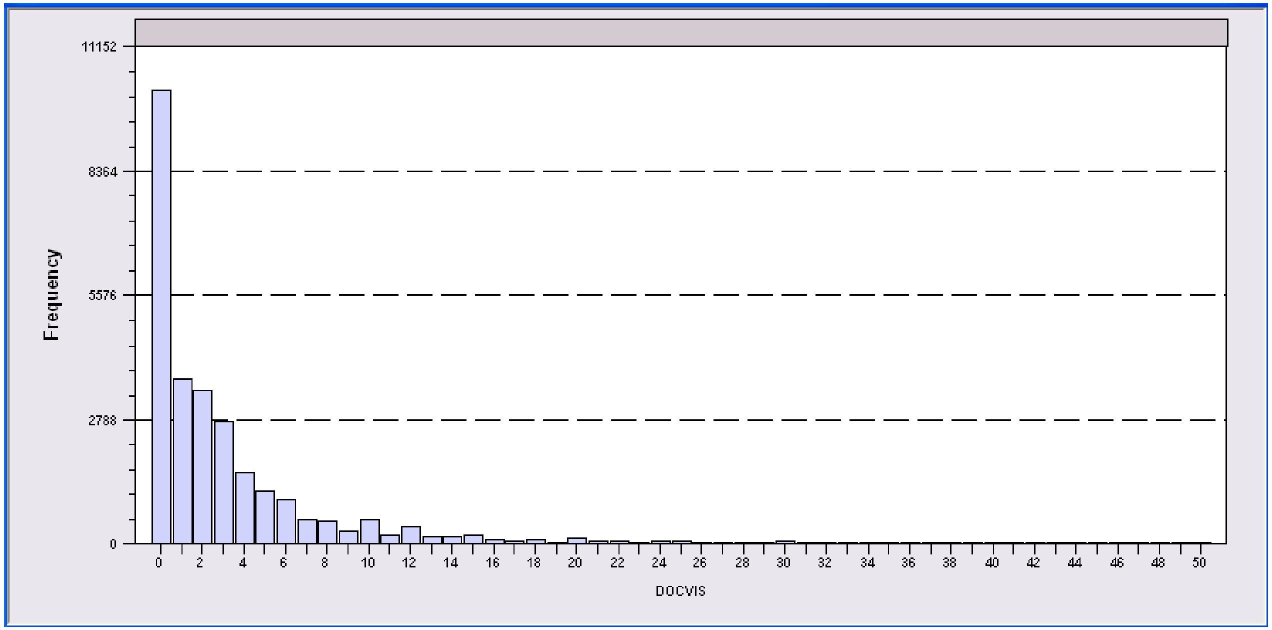
JEL classification: O4, I3


Scatter of Gini against log(GDP)
Gini coefficient
log(GDP)

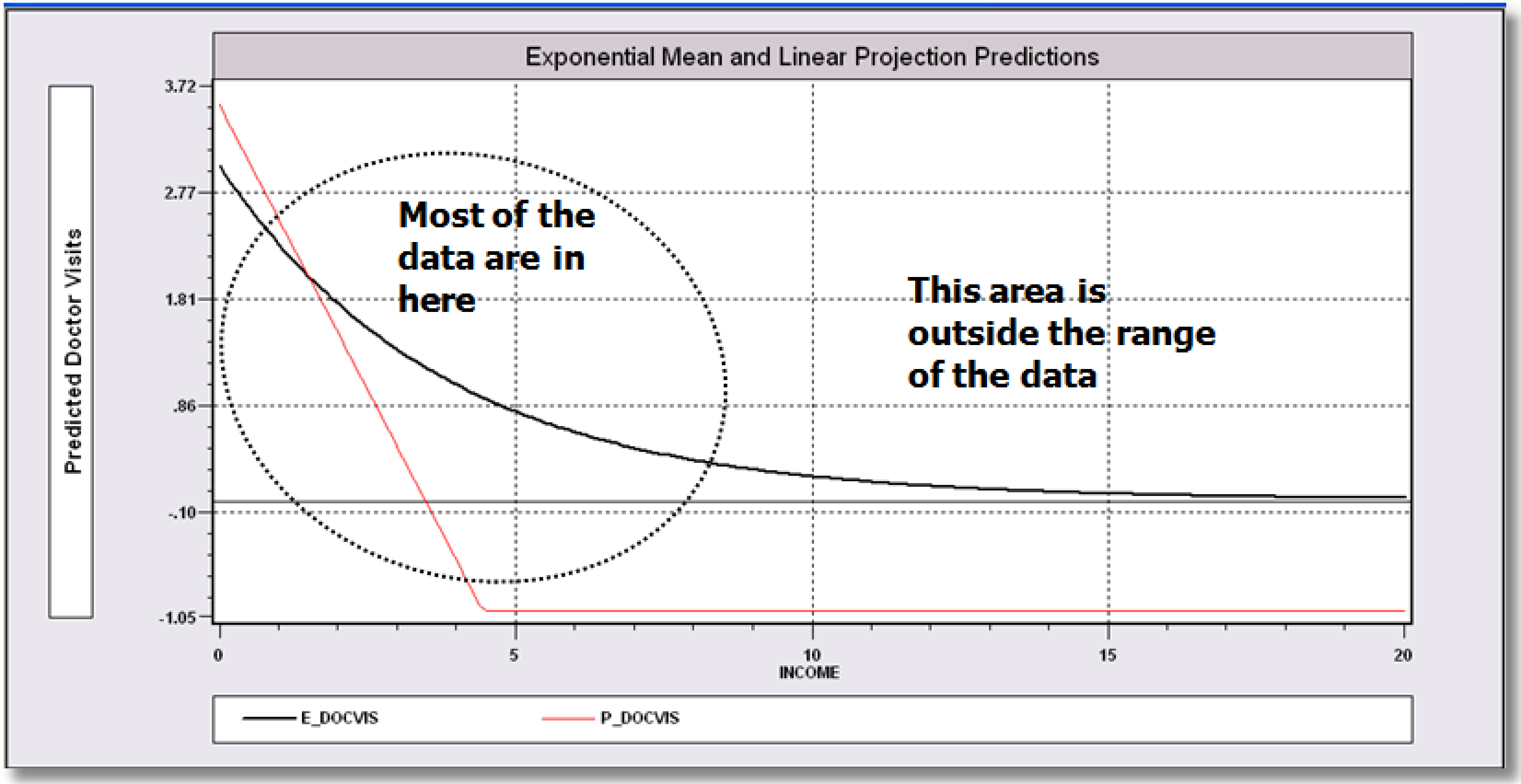

Gini Coefficient versus log(GDP)
(unexplained part)
Gini coefficient
log(GDP)

Application: Doctor Visits

- German Individual Health Care data: n=27,236
- A model for number of visits to the doctor:
  - True E[v|income] = **exp**(1.413 - .747*income)
  - Linear regression: g*(income)=3.918 – 2.087*income



Conditional Mean and Projection



The linear projection somewhat resembles the conditional mean. Notice the problem with the linear approach. Negative predictions.

```
-----------------------------------------------------------------------------
Poisson Regression
Dependent variable                   DOCVIS
Log likelihood function    -108023.08869
Restricted log likelihood -108662.13583
Chi squared [  1](P= .000)   1278.09429
Significance level                 .00000
McFadden Pseudo R-squared       .0058810
Estimation based on N =   27326, K =   2
Inf.Cr.AIC  = 216050.2 AIC/N =    7.906
Chi- squared =270220.31368  RsqP= .0275
G  - squared =163007.59656  RsqD= .0078
Overdispersion tests: g=mu(i)  : 22.805
Overdispersion tests: g=mu(i)^2: 23.248
--------+--------------------------------------------------------------------
        |                Standard            Prob.      95% Confidence
  DOCVIS|  Coefficient     Error       z    |z|>Z*         Interval
--------+--------------------------------------------------------------------
Constant|    1.41304***    .00795   177.84  .0000    1.39747     1.42862
  INCOME|   -.74694***     .02167   -34.47  .0000   -.78941     -.70447
--------+--------------------------------------------------------------------
```

For the Poisson model, E[v|income]=exp(1.41304 - .74694 income)

```
----------------------------------------------------------------------------
Ordinary     least squares regression ............
LHS=DOCVIS   Mean                 =       3.18352
             Standard deviation   =       5.68969
----------   No. of observations  =         27326  DegFreedom   Mean square
Regression   Sum of Squares       =       3721.68           1    3721.67505
Residual     Sum of Squares       =       880859.       27324      32.23755
Total        Sum of Squares       =       884581.       27325      32.37258
----------   Standard error of e  =       5.67781  Root MSE         5.67760
Fit          R-squared            =        .00421  R-bar squared     .00417
Model test   F[  1, 27324]        =     115.44533  Prob F > F*       .00000
--------+-------------------------------------------------------------------
        |                Standard            Prob.      95% Confidence
  DOCVIS|  Coefficient     Error        z    |z|>Z*        Interval
--------+-------------------------------------------------------------------
Constant|    3.91834***     .07653    51.20  .0000     3.76834    4.06833
  INCOME|   -2.08673***     .19421   -10.74  .0000    -2.46738   -1.70608
--------+-------------------------------------------------------------------
```

For the Poisson model, E[v|income]=exp(1.41304 - .74694 income)

Mean income is 0.351235.

The slope is -.74694 * exp(1.41304 - .74694 income(.351235))

```
-----------------------------------------------------------------------
Partial Effects  Analysis for Exponential Regression Function
-----------------------------------------------------------------------
Effects on function with respect to INCOME
Results are computed at sample means of all variables
Partial effects for continuous INCOME   computed by differentiation
Effect is computed as derivative       = df(.)/dx
-----------------------------------------------------------------------
df/dINCOME           Partial    Standard
(Delta method)       Effect      Error      |t|  95% Confidence Interval
-----------------------------------------------------------------------
PE.Func(means)      -2.35903     .06786    34.76    -2.49203    -2.22603
```

Representing the Relationship

- Conditional mean function is : $E[y \mid x] = g(x)$
- The linear projection (linear regression?)

$$g^*(x) = \gamma_0 + \gamma_1 (x - E[x])$$

$$\gamma_0 = E[y],\ \ \gamma = \frac{Cov[x,y]}{Var[x]}$$

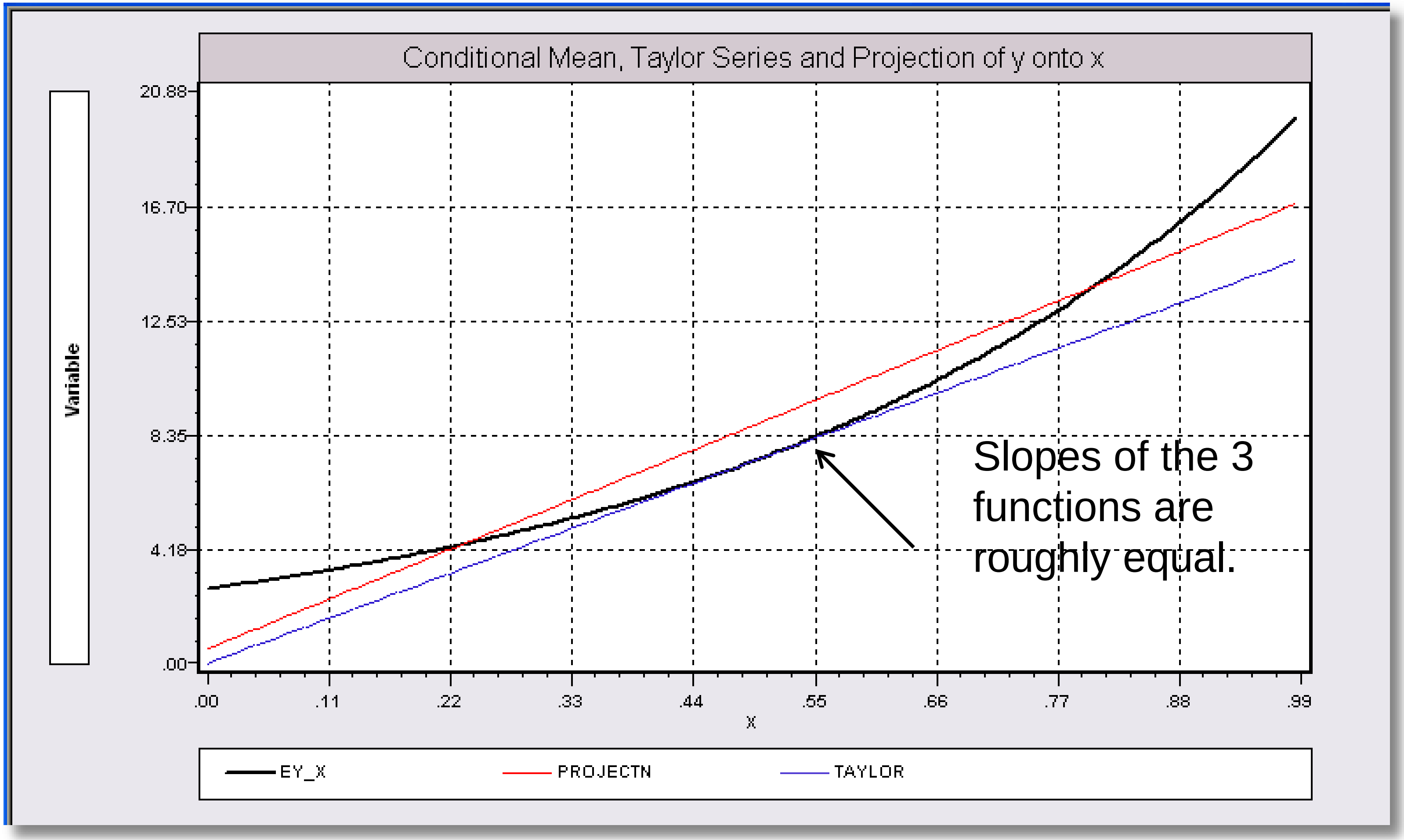
- Linear approximation to the nonlinear conditional mean function: Linear Taylor series evaluated at x^0

$$\hat{g}(x) = g(x^0) + \left[\frac{dg(x)}{dx} \mid \left(x = x^0\right)\right](x - x^0)$$

$$= \delta_0 + \delta_1 (x - x^0)$$

- We will use the projection very often. We will rarely use the Taylor series.

Representations of y
Does $y = \beta_0 + \beta_1 x + \varepsilon$?



Summary

- **Regression function**: $E[y|x] = g(x)$
- **Projection**: $g^*(y|x) = a + bx$ where $b = Cov(x,y)/Var(x)$ and $a = E[y]-bE[x]$
  Projection will equal $E[y|x]$ if $E[y|x]$ is linear.
- $y = E[y|x] + e$
  $y = a + bx + u$

The Linear Regression Model

- The **model** is $y = f(x_1, x_2, \ldots, x_K, \beta_1, \beta_2, \ldots \beta_K) + \varepsilon$

  = **a multiple regression** model (multiple as opposed to multivariate). Emphasis on the "multiple" aspect of multiple regression. Important examples:
- Form of the model – E[y|**x**] = a linear function of **x**. (Regressand vs. regressors)
- Note the presumption that there exists a relationship defined by the model.
- '**Dependent' and 'independent' variables**.
  - Independent of what? Think in terms of autonomous variation.
  - Can y just 'change?' What 'causes' the change?
  - Very careful on the issue of causality. Cause vs. association. Modeling causality in econometrics…

Model Assumptions: Generalities

- **Linearity** means linear in the parameters. We'll return to this issue shortly.
- **Identifiability**. It is not possible in the context of the model for two different sets of parameters to produce the same value of E[y|**x**] for **all** **x** vectors. (It is possible for some **x**.)
- **Conditional expected value of the deviation** of an observation from the conditional mean function is zero
- **Form of the variance** of the random variable around the conditional mean is specified
- Nature of the process by which **x** is observed is not specified. The assumptions are conditioned on the observed **x**.
- Assumptions about a specific probability distribution to be made later.

Linearity of the Model

- $f(x_1,x_2,\ldots,x_K,\beta_1,\beta_2,\ldots\beta_K) = x_1\beta_1 + x_2\beta_2 + \ldots + x_K\beta_K$
- **Notation:** $x_1\beta_1 + x_2\beta_2 + \ldots + x_K\beta_K = \mathbf{x}'\boldsymbol{\beta}$.
  - Boldface letter indicates a column vector. "x" denotes a variable, a function of a variable, or a function of a set of variables.
  - There are K "variables" on the right hand side of the conditional mean "function."
  - The first "variable" is usually a constant term. (Wisdom: Models should have a constant term unless the theory says they should not.)
- $E[y|\mathbf{x}] = \beta_1{*}1 + \beta_2{*}x_2 + \ldots + \beta_K{*}x_K$.

  ($\beta_1{*}1$ = the intercept term).

Linearity

- Simple linear model, $E[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$
- Quadratic model: $E[y|\mathbf{x}] = \alpha + \beta_1 x + \beta_2 x^2$
- Loglinear model, $E[\ln y|\ln \mathbf{x}] = \alpha + \Sigma_k \ln x_k \beta_k$
- Semilog, $E[y|\mathbf{x}] = \alpha + \Sigma_k \ln x_k \beta_k$
- Translog: $E[\ln y|\ln \mathbf{x}] = \alpha + \Sigma_k \ln x_k \beta_k + \Sigma_k \Sigma_l \delta_{kl} \ln x_k \ln x_l$

All are "linear." An infinite number of variations.

Linearity

- **Linearity** means ***linear in the parameters***, not in the variables
- $E[y|\mathbf{x}] = \beta_1 f_1(\ldots) + \beta_2 f_2(\ldots) + \ldots + \beta_K f_K(\ldots)$.

  $f_k()$ may be any function of data.
- Examples:
  - Logs and levels in economics
  - Time trends, and time trends in loglinear models – rates of growth
  - Dummy variables
  - Quadratics, power functions, log-quadratic, trig functions, interactions and so on.

Uniqueness of the Conditional Mean

The conditional mean relationship must hold for any set of N observations, i = 1,…,n*.* Assume, that n ≥ *K* (justified later)

$E[y_1|\mathbf{x}] = \mathbf{x_1}'\beta$

$E[y_2|\mathbf{x}] = \mathbf{x_2}'\beta$

…

$E[y_n|\mathbf{x}] = \mathbf{x_n}'\beta$

All n observations at once: $E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\beta = \mathbf{E}_\beta$.

Uniqueness of E[y|X]

Now, suppose there is a $\gamma \neq \beta$ that produces the same expected value,

$$E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\gamma = \mathbf{E}_\gamma.$$

Let $\delta = \beta - \gamma$. Then,

$$\mathbf{X}\delta = \mathbf{X}\beta - \mathbf{X}\gamma = \mathbf{E}_\beta - \mathbf{E}_\gamma = \mathbf{0}.$$

Is this possible? **X** is an n$\times$K matrix (n rows, K columns). What does $\mathbf{X}\delta = \mathbf{0}$ mean? We assume this is not possible. This is the '***full rank'*** assumption – it is an 'identifiability' assumption. Ultimately, it will imply that we can 'estimate' β. (We have yet to develop this.) This requires $n \geq K$.

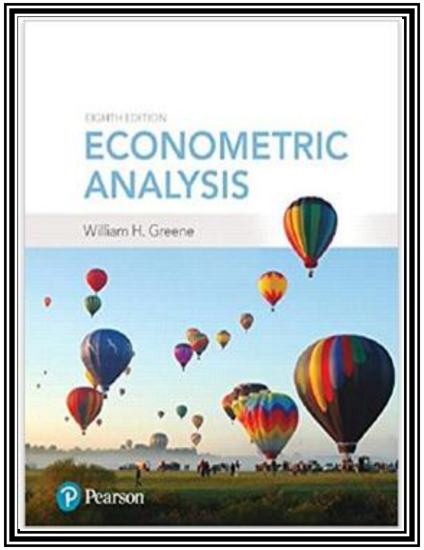
Without uniqueness, neither $\mathbf{X}\beta$ or $\mathbf{X}\gamma$ are E[y|**X**]

Linear Dependence

- Example: (2.5) from your text:
  x = [1 , Nonlabor income, Labor income, Total income]
- More formal statement of the uniqueness condition:
  No linear dependencies: No variable x_k may be written as a linear function of the other variables in the model. An ***identification condition***. Theory does not rule it out, but it makes estimation impossible. E.g.,
  $y = \beta_1 + \beta_2 NI + \beta_3 S + \beta_4 T + \varepsilon$, where $T = NI + S$.
  $y = \beta_1 + (\beta_2 + a)NI + (\beta_3 + a)S + (\beta_4 - a)T + \varepsilon$ for any a,
  $= \gamma_1 + \gamma_2 NI + \gamma_3 S + \gamma_4 T + \varepsilon$.
- What do we estimate if we 'regress' y on (1,NI,S,T)?
- Note, the model does not rule out **nonlinear dependence**. Having x and x^2 in the same equation is no problem.

An Enduring Art Mystery

The Persistence of Econometrics Greene, 2017



Graphics show relative sizes of the two works.

The Persistence of Memory. Salvador Dali, 1931

Why do larger paintings command higher prices?

An Unidentified (But Valid) Theory of Art Appreciation

(Not a Monet)

Enhanced Monet Area Effect Model: Height and Width Effects

$$\text{Log(Price)} = \alpha + \beta_1 \log \text{Area} + \beta_2 \log \text{Aspect Ratio} + \beta_3 \log \text{Height} + \beta_4 \text{Signature} + \varepsilon$$

$$= \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

(Aspect Ratio = Width/Height). This is a perfectly respectable theory of art prices. However, it is not possible to learn about the parameters from data on prices, areas, aspect ratios, heights and signatures.

$$x_3 = (1/2)(x_1 - x_2)$$

Notation

Define column vectors of N observations on y **and the** K **variables.**

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nK} \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$= \mathbf{X}\beta + \varepsilon$$

The assumption means that the rank of the matrix X is K**.**
No linear dependencies => FULL COLUMN RANK of the matrix X.

Expected Values of Deviations from the Conditional Mean

Observed *y* will equal E[y|**x**] + random variation.

y = E[*y*|**x**] + ε (disturbance)

- Is there any ***information*** about ε in **x**? That is, does movement in **x** provide useful information about movement in ε? If so, then we have not fully specified the conditional mean, and this function we are calling '*E*[y|**x**]' is not the conditional mean (regression)
- There may be information about ε in other variables. But, not in **x**. If E[ε|**x**] $\neq$ 0 then it follows that Cov[ε,**x**] $\neq$ 0. This violates the (as yet still not fully defined) 'independence' assumption

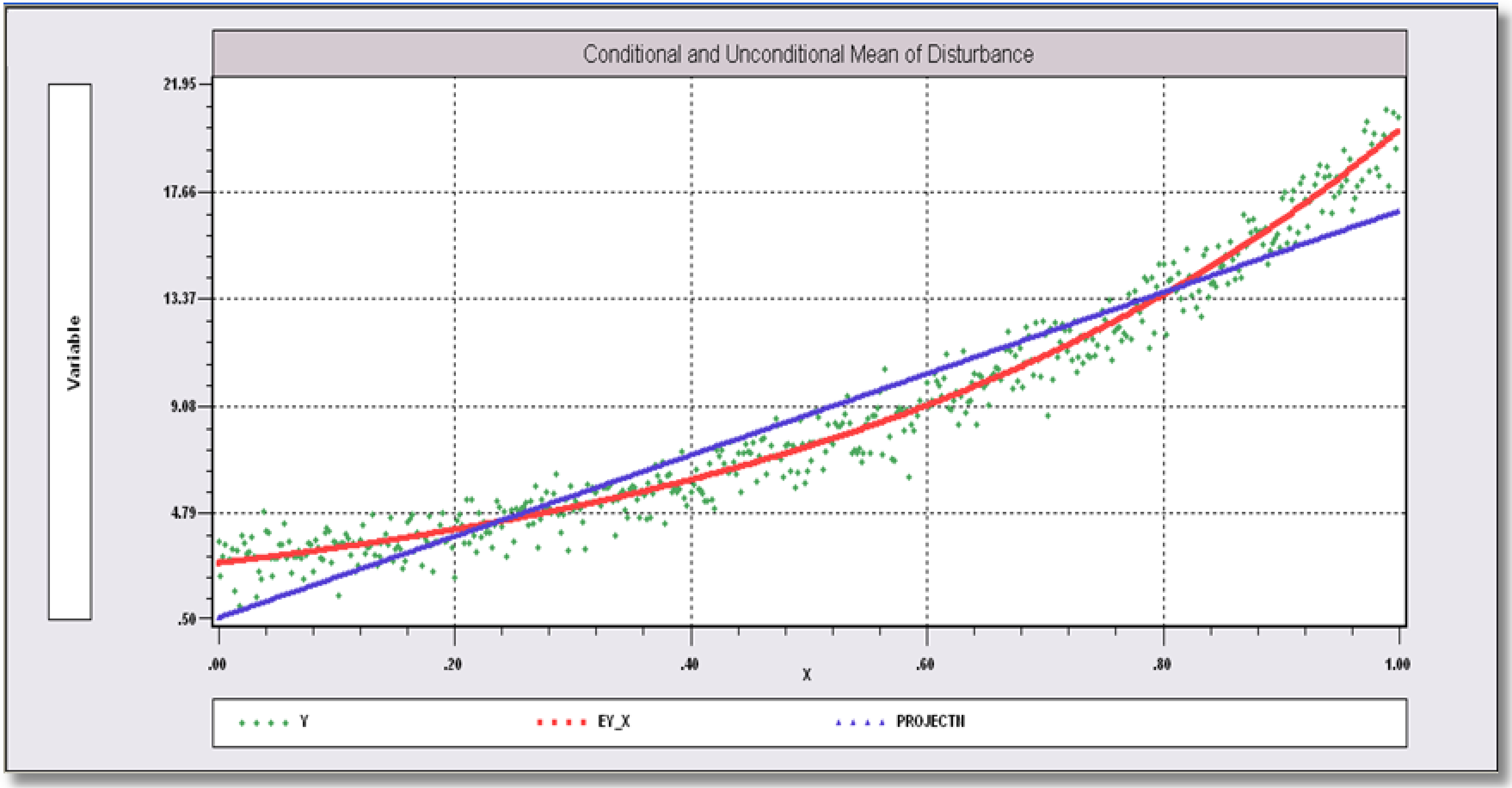
Zero Conditional Mean of ε

- $E[\varepsilon|\text{all data in } \mathbf{X}] = 0$

- $E[\boldsymbol{\varepsilon}|\mathbf{X}] = \mathbf{0}$ is stronger than $E[\varepsilon_i \mid \mathbf{x}_i] = 0$
  - The second says that knowledge of $\mathbf{x}_i$ provides no information about the mean of ε_i. The first says that *no* $\mathbf{x}_j$ provides information about the expected value of ε_i, not the i^{th} observation and not any other observation either.
  - "No information" is the same as no correlation. Proof: $\text{Cov}[\mathbf{X},\boldsymbol{\varepsilon}] = \text{Cov}[\mathbf{X},E[\boldsymbol{\varepsilon}|\mathbf{X}]] = \mathbf{0}$

The Difference Between E[ε |**x**]=0 and E[**ε**]=0

With respect to ——, E[**ε**|**x**] ≠ 0, but E_x[E[**ε**|**x**]] = E[**ε**] = 0



Conditional Homoscedasticity and Nonautocorrelation

Disturbances provide no information about each other, whether in the presence of **X** or not.

- $\text{Var}[\varepsilon|\mathbf{X}] = \sigma^2\mathbf{I}$.
- Does this imply that $\text{Var}[\varepsilon] = \sigma^2\mathbf{I}$? Yes:
  Proof: $\text{Var}[\varepsilon] = E[\text{Var}[\varepsilon|\mathbf{X}]] + \text{Var}[E[\varepsilon|\mathbf{X}]]$.

Insert the pieces above. What does this mean? It is an additional assumption, part of the model. We'll change it later. For now, it is a useful simplification

Normal Distribution of ε

- Used to facilitate finite sample derivations of certain test statistics.
- Temporary. We'll return to this later. For now, we only assume ε are i.i.d. with zero conditional mean and constant conditional variance.

The Linear Model

- $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$, n observations, K columns in **X**, including a column of ones.
  - Standard assumptions about **X**
  - Standard assumptions about **ε|X**
  - **E[ε|X]=0, E[ε]=0 and Cov[ε,x]=0**

- Regression?
  - If E[**y**|**X**] = **X**β then E[y|**x**] is also the projection.

Cornwell and Rupert Panel Data

Cornwell and Rupert Returns to Schooling Data, 595 Individuals, 7 Years
Variables in the file are

EXP	= work experience
WKS	= weeks worked
OCC	= occupation, 1 if blue collar,
IND	= 1 if manufacturing industry
SOUTH	= 1 if resides in south
SMSA	= 1 if resides in a city (SMSA)
MS	= 1 if married
FEM	= 1 if female
UNION	= 1 if wage set by union contract
ED	= years of education
LWAGE	= log of wage = dependent variable in regressions

These data were analyzed in Cornwell, C. and Rupert, P., "Efficient Estimation with Panel Data: An Empirical Comparison of Instrumental Variable Estimators," Journal of Applied Econometrics, 3, 1988, pp. 149-155. See Baltagi, page 122 for further analysis. The data were downloaded from the website for Baltagi's text.

Regression Specification: Quadratic Effect of Experience

```
----------------------------------------------------------------------------
Ordinary    least squares regression ............
LHS=LWAGE   Mean                 =         6.67635
            Standard deviation   =          .46151
----------  No. of observations  =            4165  DegFreedom   Mean square
Regression  Sum of Squares       =         370.955         10      37.09546
Residual    Sum of Squares       =         515.950       4154        .12421
Total       Sum of Squares       =         886.905       4164        .21299
----------  Standard error of e  =          .35243  Root MSE          .35196
Fit         R-squared            =          .41826  R-bar squared     .41686
Model test  F[ 10,  4154]        =       298.66153  Prob F > F*       .00000
--------+-------------------------------------------------------------------
        |                 Standard            Prob.      95% Confidence
   LWAGE|  Coefficient     Error       z     |z|>Z*         Interval
--------+-------------------------------------------------------------------
Constant|    5.24547***    .07170    73.15  .0000     5.10493    5.38600
      ED|     .05654***    .00261    21.64  .0000      .05142     .06166
     EXP|     .04045***    .00217    18.61  .0000      .03619     .04471
 EXP*EXP|    -.00068***  .4783D-04  -14.24  .0000     -.00077    -.00059
     WKS|     .00449***    .00109     4.12  .0000      .00235     .00662
     OCC|    -.14053***    .01472    -9.54  .0000     -.16939    -.11167
   SOUTH|    -.07210***    .01249    -5.77  .0000     -.09658    -.04762
    SMSA|     .13901***    .01207    11.51  .0000      .11534     .16267
      MS|     .06736***    .02063     3.26  .0011      .02692     .10779
     FEM|    -.38922***    .02518   -15.46  .0000     -.43857    -.33987
   UNION|     .09015***    .01289     6.99  .0000      .06488     .11542
--------+-------------------------------------------------------------------
nnnnn.D-xx or D+xx => multiply by 10 to -xx or +xx.
***, **, * ==>  Significance at 1%, 5%, 10% level.
----------------------------------------------------------------------------
```

Model Implication:
Effect of Experience and Male vs. Female