Georgetown University
Spring 2016
Advanced Applied Econometrics
(PPOL 754-20)

Andrew H. McCallum, Ph.D.

March 17, 2016

Disclaimer

Any opinions and conclusions expressed during this course are solely the responsibility of the author and do not necessarily represent the views of the Board of Governors or any other person associated with the Federal Reserve System.

Regression: What You Need to Know

Our regression agenda:

1. Three reasons to love
2. The CEF is all you need
3. The long and short of regression anatomy

The CEF

- The *Conditional Expectation Function* (CEF) for a dependent variable, Y_i given a $\mathrm{k} \times 1$ vector of covariates, X_i (with elements x_{ki}) is written $E\left[Y_i \mid X_i\right]$ and is a function of X_i
- Because X_i is random, the CEF is random. For dummy D_i , the CEF takes on two values, $E\left[Y_i \mid D_i = 1\right]$ and $E\left[Y_i \mid D_i = 0\right]$
- For a specific value of X_i, say $X_i = 42$, we write $E\left[Y_i \mid X_i = 42\right]$
- For continuous Y_i with conditional density $f_y\left(\cdot \mid X_i = x\right)$, the CEF is

$$E\left[Y_i \mid X_i = x\right] = \int t f_y\left(t \mid X_i = x\right) dt$$

  If Y_i is discrete, $E\left[Y_i \mid X_i = x\right]$ equals the $\sum_t t f y\left(t \mid X_i = x\right)$
- The CEF residual is uncorrelated with any function of of X_i. Write $\varepsilon_i \equiv Y_i - E\left[Y_i \mid X_i\right]$. Then for any function, $h\left(X_i\right)$:

$$E\left[\varepsilon_i h\left(X_i\right)\right] = E\left[\left(Y_i - E\left[Y_i \mid X_i\right]\right) h\left(X_i\right)\right] = 0$$

  (The LIE proves it)

CEF for log weekly earnings

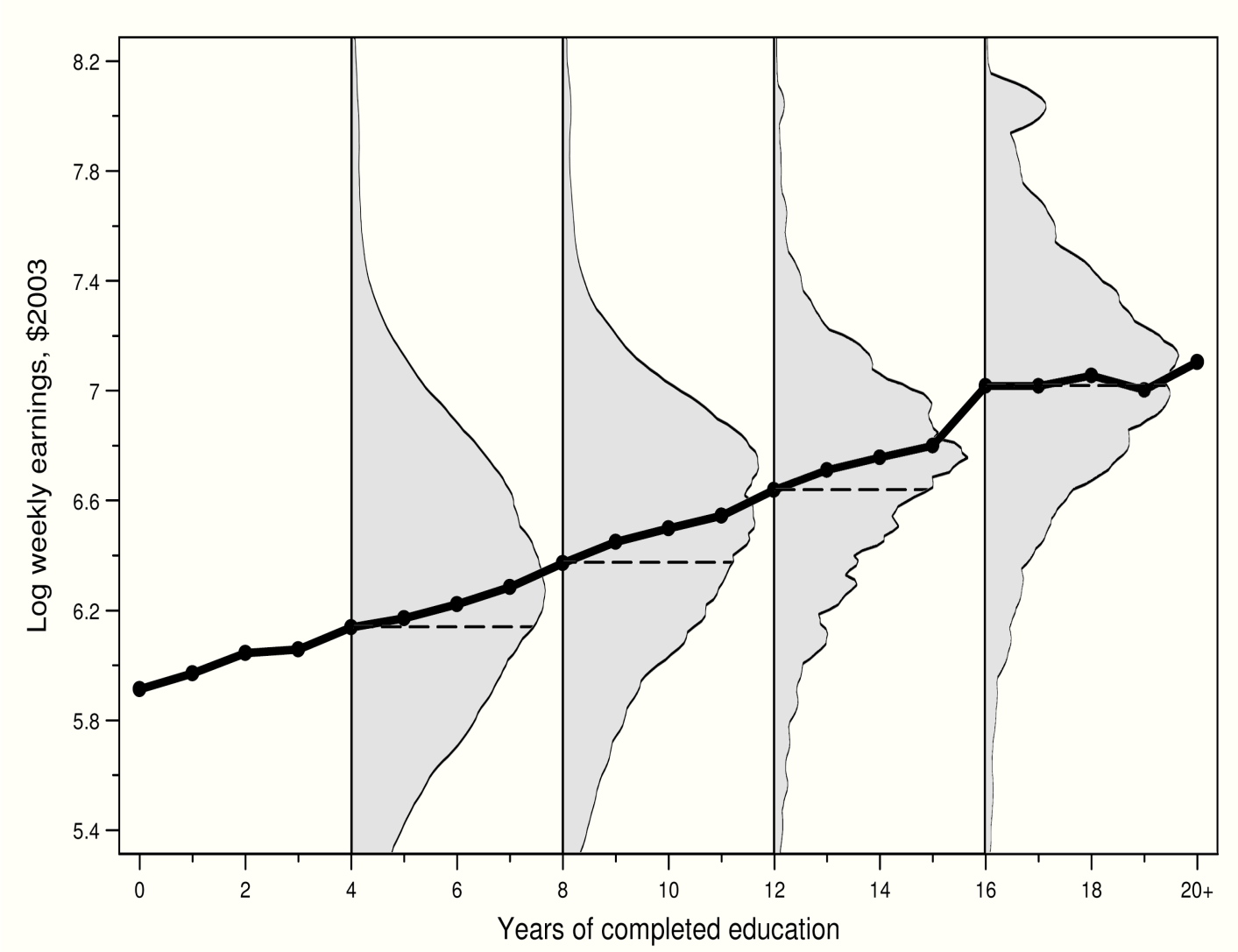

Figure 3.1.1: Raw data and the CEF of average log weekly wages given schooling. The sample includes white men aged 40-49. The data are from the 1980 IPUMS5 percent sample.

Proof of Law of Iterated Expectations

A powerful tool is the ability to always "iterate expectations" (break variables into pieces)

$$E[Y_i] = E\{E[Y_i \mid X_i]\}$$

- Proof of law of iterated expectations (for continuous case):

$$\begin{aligned}
E\{E[Y_i \mid X_i]\} &= \int E[Y_i \mid X_i = u]\, g_x(u)\, du \\
&= \int \left[\int t f_y(t \mid X_i = u)\ dt \right] g_x(u)\, du \\
&= \iint t f_y(t \mid X_i = u)\, g_x(u)\, du dt \\
&= \int t \left[\int f_y(t \mid X_i = u)\, g_x(u)\, du \right] dt = \int t \left[\int f_{xy}(u,t)\, du \right] dt \\
&= \int t g_y(t)\ dt
\end{aligned}$$

CEF Decomposition Property

You can always write (decompose) Y_i as

$$Y_i = E\left[Y_i \mid X_i\right] + \varepsilon_i$$

where

(i) $E\left[\varepsilon_i \mid X_i\right] = 0$ (i.e. ε_i is mean-independent of X_i)

(ii) ε_i is uncorrelated with any function $h\left(X_i\right)$ of X_i

In other words, any Y_i is decomposable into:

- A piece explained by X_i (the CEF)
- A piece orthogonal (uncorrelated) to any function of X_i, $h\left(X_i\right)$

Proof of CEF-Decomposition Property

Prove that under (i) and (ii) we can always write

$$Y_i = E\left[Y_i \mid X_i\right] + \varepsilon_i$$

If the CEF has additive errors, then we want that

$$\begin{aligned} E\left[\varepsilon_i \mid X_i\right] &= E\left[Y_i - E\left[Y_i \mid X_i\right] \mid X_i\right] \\ &= E\left[Y_i \mid X_i\right] - E\left[Y_i \mid X_i\right] \\ &= 0 \end{aligned}$$

If (i) holds, then (ii) is true

$$\begin{aligned} E\left[h\left(X_i\right)\varepsilon_i\right] &= E\left\{E\left[h\left(X_i\right)\varepsilon_i \mid X_i\right]\right\} \ by\ LIE \\ &= E\left\{h\left(X_i\right)E\left[\varepsilon_i \mid X_i\right]\right\} \ by\ CE \\ &= E\left\{h\left(X_i\right)0\right\} \ by\ (i) \\ &= 0 \end{aligned}$$

CEF Prediction Property

- Let $m(X_i)$ be any function of X_i. The CEF solves:

$$E[Y_i \mid X_i] = \underset{m(X_i)}{\arg\min}\, E\left[(Y_i - m(X_i))^2\right]$$

- In other words, the CEF is the Minimum Mean Squared Error (MMSE) predictor of Y_i given X_i
- The CEF is the **BEST** function we can find for minimizing distance.

Proof of CEF-Prediction Property

- Begin with a "trick" of adding and subtracting

$$\left(Y_i - m\left(X_i\right)\right)^2 = \left(\left(Y_i - E\left[Y_i \mid X_i\right]\right) + \left(E\left[Y_i \mid X_i\right] - m\left(X_i\right)\right)\right)^2$$

- Now expand the terms:

$$= \left(Y_i - E\left[Y_i \mid X_i\right]\right)^2 + 2\left(E\left[Y_i \mid X_i\right] - m\left(X_i\right)\right)\left(Y_i - E\left[Y_i \mid X_i\right]\right) + \left(E\left[Y_i \mid X_i\right] - m\left(X_i\right)\right)^2$$

Try to choose the $m\left(X_i\right)$ that makes this smallest

1st term doesn't involve $m\left(X_i\right)$

2nd term is $h\left(X_i\right)\varepsilon_i$, where $h\left(X_i\right) \equiv 2\left(E\left[Y_i \mid X_i\right] - m\left(X_i\right)\right)$ so its expected value is zero by (ii).

3rd term is minimized when $m\left(X_i\right)$ is the CEF

ANOVA (Analysis of Variance)

- ANOVA result:

$$V(Y_i) = V(E[Y_i \mid X_i]) + E[V(Y_i \mid X_i)]$$

- This is a way of decomposing the variance of the outcome of interest (e.g. wages) into
  - Variance explained by covariates (1st term)
    - In wages example: variance in wages explained by worker characteristics (e.g. schooling)
  - Variance unexplained by covariates (2nd term)
    - In wages example: residual inequality

Hence, CEF decomposition property allows us to write the variance of Y_i as the variance of the CEF plus the variance of ε_i.

Proof of ANOVA

Apply the CEF-decomposition property

$$\begin{aligned} Y_i &= E[Y_i \mid X_i] + \varepsilon_i \\ V[Y_i] &= V[E[Y_i \mid X_i]] + V[\varepsilon_i] + COV[\varepsilon_i, E[Y_i \mid X_i]] \ by\ Def. \\ V[Y_i] &= V[E[Y_i \mid X_i]] + V[\varepsilon_i] \ by\ E[\varepsilon_i \mid X_i] = 0 \end{aligned}$$

1. $E[\varepsilon_i \mid X_i] = 0$ (for each X_i), so $V[\varepsilon_i] = E\left[\varepsilon_i^2\right]$
2. By the LIE:
$$E\left[\varepsilon_i^2\right] = E\left[E\left[\varepsilon_i^2 \mid X_i\right]\right]$$
3. Finally, $E\left[E\left[\varepsilon_i^2 \mid X_i\right]\right] = E[V[Y_i \mid X_i]]$, since:
$$\varepsilon_i \equiv Y_i - E[Y_i \mid X_i]$$

Population Regression

- Define *population regression* ("regression," for short) as the solution to the population least squares problem. Specifically, the $\mathrm{k} \times 1$ regression coefficient vector β is defined by solving

$$\beta = \arg\min_{b} E\left[\left(Y_i - X_i'b\right)^2\right]$$

- Using the first-order condition,

$$E\left[X_i\left(Y_i - X_i'b\right)\right] = 0,$$

  the solution for b can be written

$$\beta = E\left[X_i X_i'\right]^{-1} E\left[X_i Y_i\right]$$

- By construction, $E\left[X_i\left(Y_i - X_i'\beta\right)\right] = 0$. In other words, the population residual, defined as $Y_i - X_i'\beta = e_i$, is uncorrelated with the regressors, X_i
- The error term has no life of its own: e_i owes its meaning and existence to β

Regression anatomy lesson

- Bivariate reg recap: the slope coefficient is $\beta_1 = \frac{\text{Cov}(Y_i, x_i)}{V(x_i)}$, and the intercept is $\alpha = E[Y_i] - \beta_1 E[X_i]$
- With more than one non-constant regressor, the k-th non-constant slope coefficient is:

$$\beta_k = \frac{\text{Cov}(Y_i, \tilde{x}_{ki})}{V(\tilde{x}_{ki})}$$

  where $\tilde{x}_{ki}$ is the residual from a regression of x_{ki} on all other covariates
- The anatomy formula shows us that each coefficient in a multivariate regression is the bivariate slope coeffcient for the corresponding regressor, after "partialing out" other variables in the model.
- Verify the regression-anatomy formula by substituting

$$Y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \ldots + \beta_{\text{k}} x_{\text{k}i} + e_i$$

  in the numerator of (2) and work through to find that
  $\text{Cov}(Y_i, \tilde{x}_{ki}) = \beta_k V(\tilde{x}_{ki})$

Regression Anatomy Proof

- Substitute

$$Y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \ldots + \beta_{\mathrm{k}} x_{\mathrm{ki}} + e_i$$

into the numerator ($\mathrm{Cov}\left(Y_i, \tilde{x}_{ki}\right)$) and note that

1. $E\left[X_i\left(Y_i - X_i'\beta\right)\right] = 0$, so the population residual, $e_i = Y_i - X_i'\beta$ is uncorrelated with X_i
2. $\tilde{x}_{ki}$ is uncorrelated with e_i. Why?
   - Because it's a linear combination of the x_{ji}s
3. $\tilde{x}_{ki}$ is uncorrelated with x_{ji} for $j \neq k$. Why?
   - It's a residual from a regression on all the other covariates in the model
4. For the same reason, the covariance of $\tilde{x}_{ki}$ with x_{ki} is just the variance of $\tilde{x}_{ki}$
5. Plugging all these back into the numerator, we get: $\mathrm{Cov}\left(Y_i, \tilde{x}_{ki}\right) = \beta_k V\left(\tilde{x}_{ki}\right)$
6. And therefore: $\beta_k = \frac{\mathrm{Cov}(Y_i, \tilde{x}_{ki})}{V(\tilde{x}_{ki})}$

- In words: each coefficient in a multivariate regression is the bivariate slope coefficient for the corresponding regressor, after "partialling out" all the other variables in the model.

Partialling out

Scalar, mean-zero random variables y_i, x_{1i}, and x_{2i}:

$$(\beta, \gamma) = \arg\min_{b,c} E\left[(y_i - bx_{1i} - cx_{2i})^2\right]$$

$$\text{FOC}_\gamma : -2E\left[x_{2i}\left(y_i - bx_{1i} - \gamma x_{2i}\right)\right] = 0$$

$$\text{IFT} : \quad \gamma(b) = \frac{E\left[x_{2i}\left(y_i - bx_{1i}\right)\right]}{E\left[x_{2i}^2\right]}$$

Plug $\gamma(b)$ back in (sometimes called "concentrating out" γ):

$$\begin{aligned}
\beta &= \arg\min_b E\left[\left(y_i - bx_{1i} - \frac{E\left[x_{2i}\left(y_i - bx_{1i}\right)\right]}{E\left[x_{2i}^2\right]} x_{2i}\right)^2\right] \\
&= \arg\min_b E\left[\left(\left(y_i - \frac{E\left[x_{2i}y_i\right]}{E\left[x_{2i}^2\right]} x_{2i}\right) - b\left(x_{1i} - \frac{E\left[x_{2i}x_{1i}\right]}{E\left[x_{2i}^2\right]} x_{2i}\right)\right)^2\right]
\end{aligned}$$

A bivariate regression! But of what on what?

Partialling out (cont.)

- Special case of the **Frisch-Waugh (sometimes -Lovell) theorem:** If $x_i = [x'_{1i}, x'_{2i}]$, $\tilde{x}_{1i}$ is the residual (vector) from regressing (each component of) x_{1i} on x_{2i}, and $\tilde{y}_i$ is the residual from regressing y_i on x_{2i}, then all three are equivalent:
  1. The component β_1 of $\beta = [\beta'_1, \beta'_2]$ from regressing y_i on x_i
  2. $\tilde{\beta}_1$ from regressing y_i on $\tilde{x}_i$
  3. $\bar{\beta}_1$ from regressing $\tilde{y}_i$ on $\tilde{x}_i$
- Partialling out x_{2i} from y_i is unnecessary! Why? Back to our example:

$$\begin{aligned} y_i &= \beta x_{1i} + y x_{2i} + e_i \\ \tilde{y}_i &= \beta \tilde{x}_{1i} + \tilde{e}_i \\ y_i &= \beta \tilde{x}_{1i} + \tilde{e}_i + y_i - \tilde{y}_i \\ y_i &= \beta \tilde{x}_{1i} + \left(\tilde{e}_i + \frac{E[x_{2i} y_i]}{E[x_{2i}^2]} x_{2i} \right) \end{aligned}$$

  Why must the last line be a *regression* (and not just an *equation*)?

Linear CEF Theorem

- Linear CEF Theorem: Suppose the CEF is linear, then it is the population regression
- Proof:
  - Suppose that $E[Y_i \mid X_i] = X_i'\beta^*$ for a $k \cdot 1$ vector of coefficients β^*
  - From CEF decomposition: $E[X_i(Y_i - E[Y_i \mid X_i])] = 0$
  - Substitute: $E[Y_i \mid X_i] = X_i'\beta^*$
  - And get $\beta^* = E\left[X_i X_i'\right]^{-1} E[X_i Y_i] = \beta$

The Best Linear Predictor Theorem

- The Best Linear Predictor Theorem: The function $X_i'\beta$ is the best linear predictor of Y_i given X_i in a MMSE sense
- Proof: $\beta = E\left[X_i X_i'\right]^{-1} E\left[X_i Y_i\right]$ solves the least squares problem (by construction)
- Interpretation: just as the CEF, is the best (MMSE) predictor of Y_i given X_i in the class of all functions of X_i, the population regression function is the best we can do in the class of linear functions

The Regression-CEF Theorem

- Regression-CEF Theorem (Regression-justification III): The function $X_i'\beta$ provides the MMSE linear approximation to $E[Y_i \mid X_i]$, that is
  $\beta = \arg\min_b E\left\{E\left([Y_i \mid X_i] - X_i'b\right)^2\right\}$

Proof of Regression-CEF Theorem

- We can write:

$$\begin{aligned}\left(Y_i - X_i'b\right)^2 &= \left\{\left(Y_i - E\left[Y_i \mid X_i\right]\right) + \left(E\left[Y_i \mid X_i\right] - X_i'b\right)\right\}^2 \\ &= \left(Y_i - E\left[Y_i \mid X_i\right]\right)^2 + \left(E\left[Y_i \mid X_i\right] - X_i'b\right)^2 \\ &+ 2\left(Y_i - E\left[Y_i \mid X_i\right]\right)\left(E\left[Y_i \mid X_i\right] - X_i'b\right).\end{aligned}$$

  - 1st term does not involve b
  - 2nd term is minimized by β
  - 3rd term has expectation zero by CEF decomposition

How to Interpret last two regression justifications?

- The last two theorems show us two ways to view regression when CEF is nonlinear:
  - Regression provides the best linear predictor for the dependent variable in the same way that the CEF is the best unrestricted predictor of the dependent variable
  - Even if the CEF is nonlinear, regression provides the best linear approximation to it.

1. Three reasons to love

1. Regression solves the population least squares problem and is therefore the BLP of Y_i given X_i
2. If the CEF is linear, regression is it
3. Regression gives the best linear approximation to the CEF
   - The first is true by definition; the second follows immediately from CEF-orthgonality. Let's prove the third — it's my favorite!

Theorem

The Regression-CEF Theorem (MHE 3.1.6) The population regression function $X_i'\beta$ provides the MMSE linear approximation to $E[Y_i \mid X_i]$, that is,

$$\beta = \arg\min_b E\left\{\left(E[Y_i \mid X_i] - X_i'b\right)^2\right\}$$

Figure 3.1.2 illustrates the theorem

Approximately linear CEF

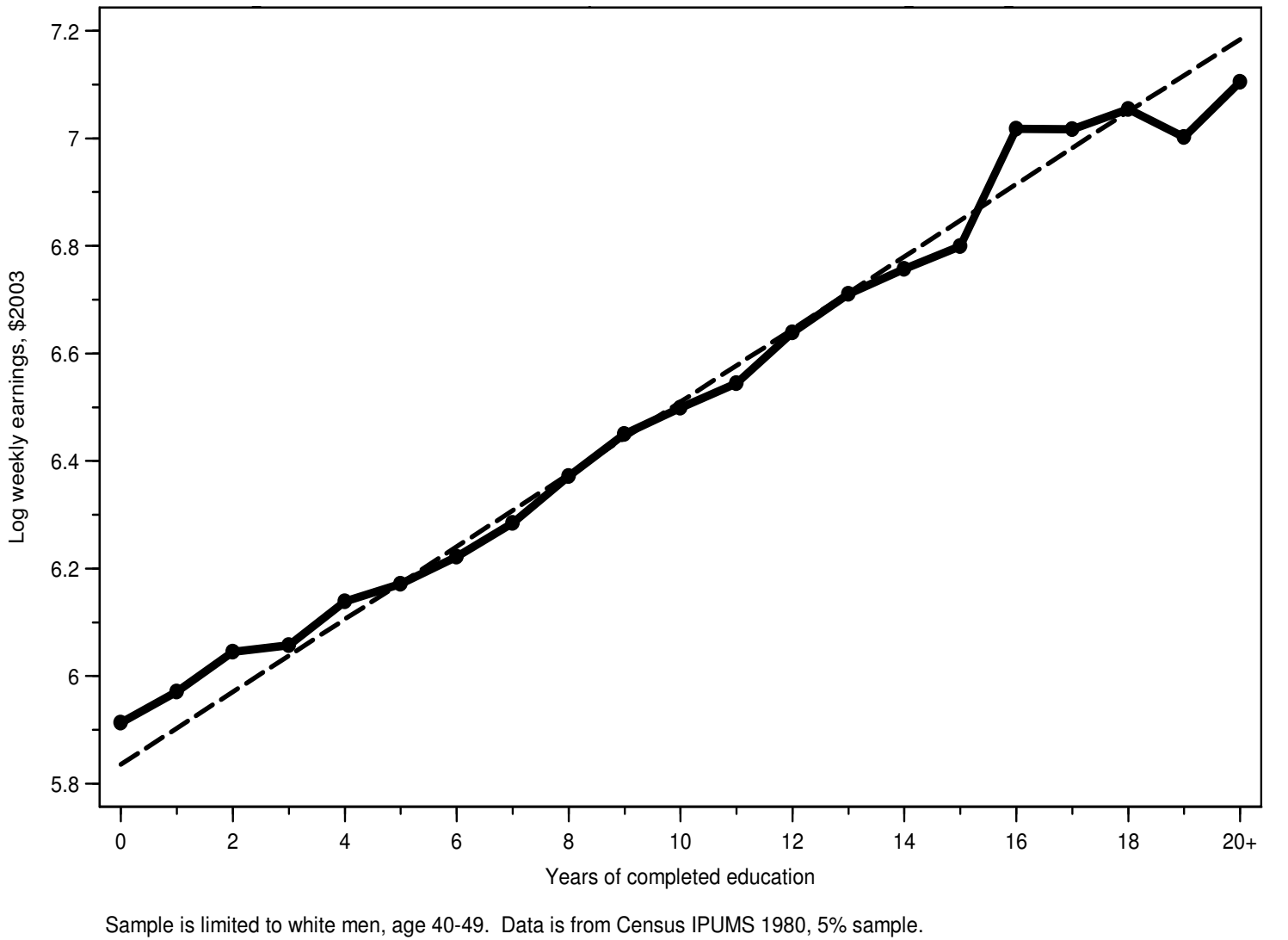

Figure 3.1.2: Regression threads the CEF of average weekly wages given schooling

2. The CEF is all you need

- The regression-CEF theorem implies we can use $E[Y_i \mid X_i]$ as a dependent variable instead of Y_i (but watch the weighting!)
- Another way to see this:

$$\beta = E\left[X_i X_i'\right]^{-1} E\left[X_i Y_i\right] = E\left[X_i X_i'\right]^{-1} E\left[X_i E\left(Y_i \mid X_i\right)\right] \qquad (1)$$

  The CEF or grouped-data version of the regression formula is useful when working on a project that precludes the analysis of micro data
- To illustrate, we can estimate the schooling coe cient in a wage equation using 21 conditional means, the sample CEF of earnings given schooling
- As Figure 3.1.3 shows, grouped data weighted by the number of individuals at each schooling level produces coefficients *identical* to that generated by the underlying micro data

Micro data or group CEF give same estimates

A - Individual-level data

```
. regress earnings school, robust

      Source |       SS       df       MS        Number of obs =  409435
-------------+------------------------------     F(  1,409433) =49118.25
       Model | 22631.4793      1  22631.4793     Prob > F      =  0.0000
    Residual |  188648.31 409433  .460755019     R-squared     =  0.1071
-------------+------------------------------     Adj R-squared =  0.1071
       Total | 211279.789 409434   .51602893     Root MSE      =  .67879
-------------+----------------------------------------------------------
             |               Robust                   Old Fashioned
    earnings |      Coef.   Std. Err.      t            Std. Err.       t
-------------+----------------------------------------------------------
      school |   .0674387   .0003447   195.63           .0003043   221.63
      const. |   5.835761   .0045507  1282.39           .0040043  1457.38
------------------------------------------------------------------------
```

B - Means by years of schooling

```
. regress average_earnings school [aweight=count], robust
(sum of wgt is   4.0944e+05)

      Source |       SS       df       MS        Number of obs =      21
-------------+------------------------------     F(  1,    19) =  540.31
       Model |  1.16077332     1  1.16077332     Prob > F      =  0.0000
    Residual |  .040818796    19  .002148358     R-squared     =  0.9660
-------------+------------------------------     Adj R-squared =  0.9642
       Total |  1.20159212    20  .060079606     Root MSE      =  .04635
-------------+----------------------------------------------------------
     average |               Robust                   Old Fashioned
   _earnings |      Coef.   Std. Err.      t            Std. Err.       t
-------------+----------------------------------------------------------
      school |   .0674387   .0040352    16.71          .0029013     23.24
      const. |   5.835761   .0399452   146.09          .0381792    152.85
------------------------------------------------------------------------
```

Figure 3.1.3: Micro-data and grouped-data estimates of returns to schooling. Source: 1980 Census - IPUMS, 5 percent sample. Sample is limited to white men, age 40-49. Derived from Stata regression output. Old-fashioned standard errors are the default reported. Robust standard errors are heteroscedasticity-consistent. Panel A uses individual-level data. Panel B uses earnings averaged by years of schooling.

The bivariate regression is handy to keep in mind

Scalar random variables x_i and y_i:

$$(\alpha, \beta) = \arg\min_{a,b} E\left[(y_i - a - bx_i)^2\right]$$

FOC:
$$-2E\left[(y_i - \alpha - \beta x_i)\right] = 0$$
$$-2E\left[(y_i - \alpha - \beta x_i)\, x_i\right] = 0$$

or

$$\begin{aligned} \alpha &= E\left[y_i\right] - \beta E\left[x_i\right] \\ \beta E\left[x_i^2\right] &= E\left[y_i x_i\right] - \alpha E\left[x_i\right] \end{aligned}$$

Substituting:

$$\beta E\left[x_i^2\right] = E\left[y_i x_i\right] - E\left[y_i\right] E\left[x_i\right] + \beta E\left[x_i\right]^2$$

$$\beta = \frac{E\left[y_i x_i\right] - E\left[y_i\right] E\left[x_i\right]}{E\left[x_i^2\right] - E\left[x_i\right]^2} = \frac{\mathrm{Cov}\left(y_i, x_i\right)}{Var\left(x_i\right)}$$

From population to sample

- Regression is a **feature of data:** just like expectation, correlation, etc.
- It's a function of population second moments: so easy to estimate!

$$\widehat{\beta} = E_n\left[x_i x_i'\right]^{-1} E_n\left[x_i y_i\right]$$

- A more matrix-y way to write $\widehat{\beta}$:

$$\begin{aligned} E_n\left[x_i x_i'\right]^{-1} E_n\left[x_i y_i\right] &= \left(\frac{1}{n}\sum_i x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_i x_i y_i\right) \\ &= \left(X'X\right)^{-1} X'Y \end{aligned}$$

here

$$X = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}, Y = \begin{bmatrix} y_1' \\ \vdots \\ y_n' \end{bmatrix}$$

Under homoskedasticity

- The default option in Stata assumes homoskedasticity: $E\left[\varepsilon_i^2 \mid X_i\right] = \sigma^2$
- Under this assumption:
  $E\left[X_i X_i' \varepsilon_i^2\right] = E\left(X_i X_i' E\left[\varepsilon_i^2 \mid X_i\right]\right) = \sigma^2 E\left[X_i X_i'\right]$
- So asymptotic covariance matrix of $\widehat{\beta}$ is:

$$E\left[X_i X_i'\right]^{-1} E\left[X_i X_i' \varepsilon_i^2\right] E\left[X_i X_i'\right]^{-1} =$$
$$= E\left[X_i X_i'\right]^{-1} \sigma^2 E\left[X_i X_i'\right]^{-1} E\left[X_i X_i\right]^{-1} = E\left[X_i X_i'\right]^{-1} \sigma^2.$$