Microeconometrics

Linear model with cross section data

Alex Armand
alex.armand@novasbe.pt

Topics

1. **Linear model with cross section data**
   - Identification and inference using the linear model
   - Identification versus prediction
2. **Violation of orthogonality**
   - IV Estimation of a General Equation
   - Two Stage Least Squares

Topics

1. **Linear model with cross section data**
   - **Identification and inference using the linear model**
   - Identification versus prediction
2. **Violation of orthogonality**
   - IV Estimation of a General Equation
   - Two Stage Least Squares

Linear model setting

Parametric solution to the identification problem for $f(y|\mathrm{x};\beta) \Rightarrow$

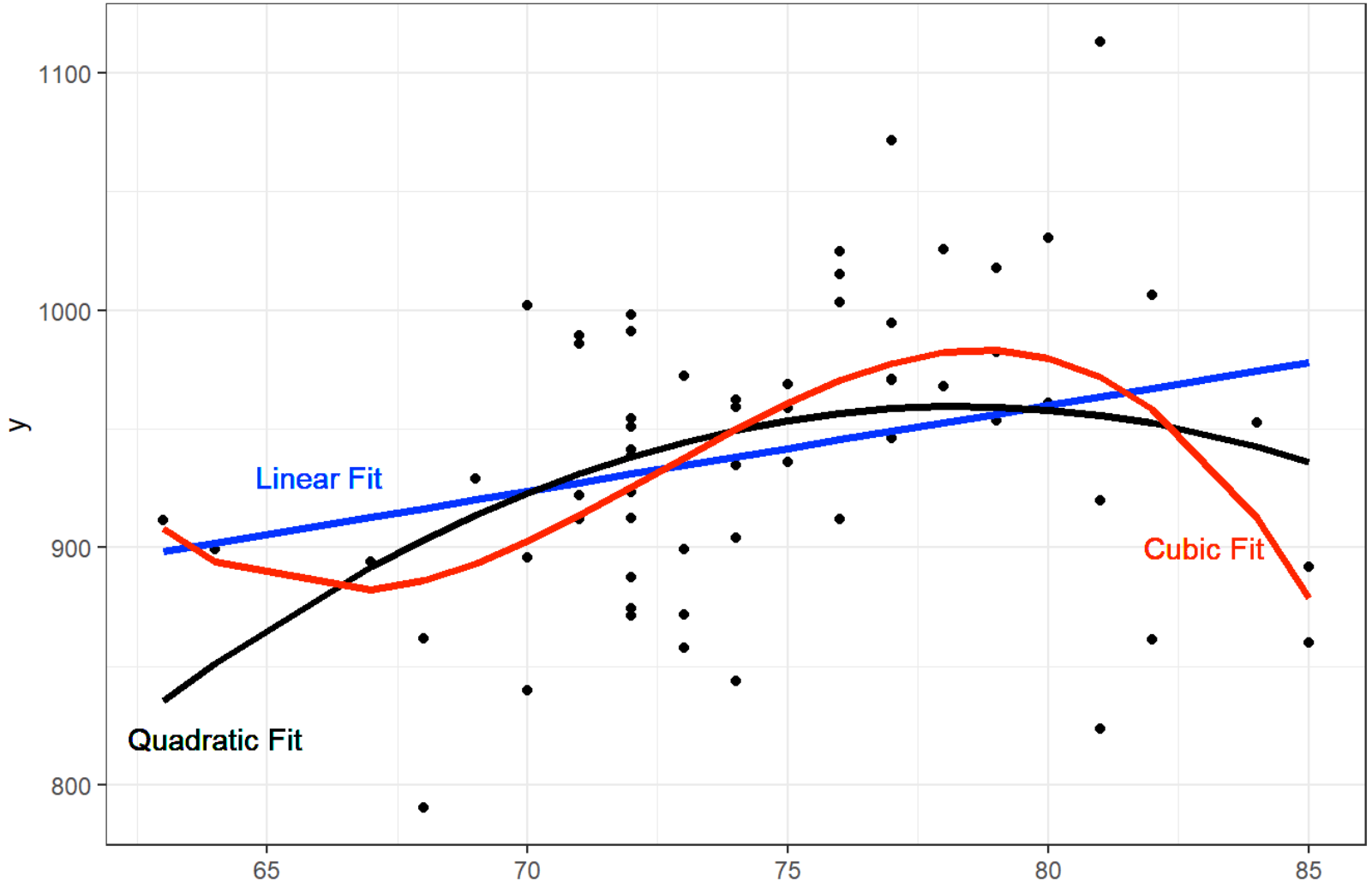
- Focus on the **conditional mean** $E[y|\mathrm{x};\beta]$
- The conditional mean is **linear in parameters** in terms of a (well-defined) population

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + ... + \beta_{K-1} x_{K-1} + u \\ &= \mathrm{x}\beta + u \\ E[y|x] &= \mathrm{x}\beta = f(\mathrm{x}) \end{aligned}$$

  - x is $1 \times K$ and observed
  - β is the $K \times 1$ vector of unknown slope parameters
  - u is an error term

Flexibility versus interpretation

Linear model can be fairly general as x can include nonlinear functions (logarithms, squares, reciprocals and interactions)



Flexibility versus interpretation

Flexibility comes at what cost?

- We are interested in causal relationships $\Rightarrow$ **partial effect**
- How changing x causes a change in the outcome
- For a continuous variable $x_j \Rightarrow$ first derivative

$$\frac{\partial f(x)}{\partial x_j}$$

Average Partial Effect

Averages partial effects are across the sample distribution of x.

$$APE_j = E_x \left[\frac{\partial f(x)}{\partial x_j} \right]$$

Flexibility versus interpretation: an example

Consider the following specifications and the APE w.r.t. *match*:

1. $contribs = \beta_0 + \beta_1 match + \beta_2 income + u$
2. $contribs = \beta_0 + \beta_1 match + \beta_2 income + \beta_3 match \cdot income + u$

- In the first, β_1 captures the APE
- In the second, the APE is equal to $\beta_1 + \beta_3 \cdot income$
- **Flexibility comes at the cost of interpretation** $\Rightarrow$
  - Coefficients on level terms may become essentially meaningless
  - APE becomes functions of observable characteristics

Flexibility versus interpretation: an example

Consider the following specifications and the APE w.r.t. *match*:

1. $contribs = \beta_0 + \beta_1 match + \beta_2 income + u$
2. $contribs = \beta_0 + \beta_1 match + \beta_2 income + \beta_3 match \cdot income + u$

- In the first, β_1 captures the APE
- In the second, the APE is equal to $\beta_1 + \beta_3 \cdot income$
- **Flexibility comes at the cost of interpretation** $\Rightarrow$
  - Coefficients on level terms may become essentially meaningless
  - APE becomes functions of observable characteristics

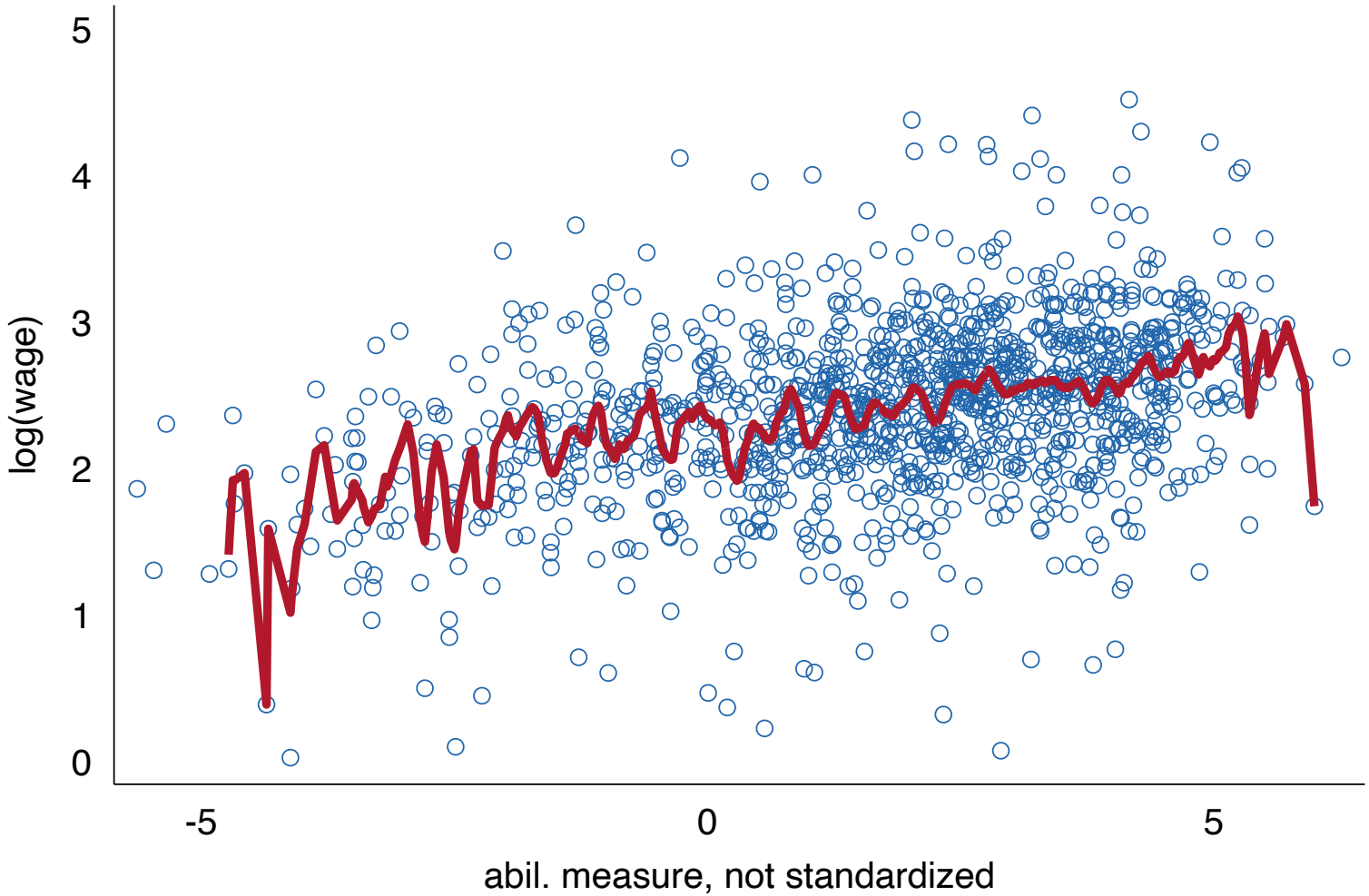
What is flexibility?

We compare parametric versus non-parametric approach:

- Non-parametric approach (kernel regression) performs **local smoothing**
- Linear models perform **global smoothing**
- Let's see an example: relationship between wage and ability

Local smoothing

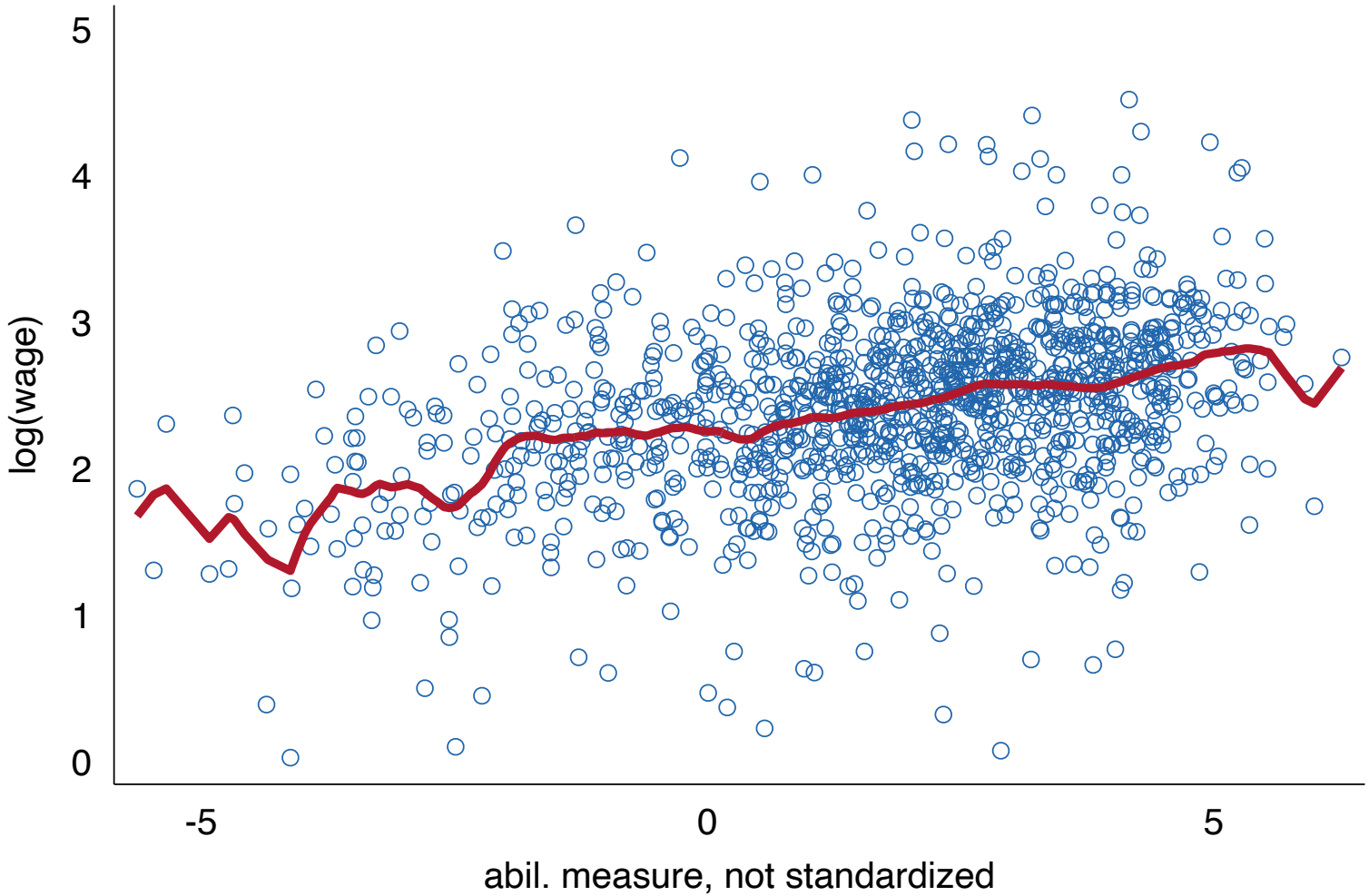
- Choice of bandwidth **0.05** 0.20 0.50 2.00



Local-linear estimates
kernel = epanechnikov bandwidth = .05

Local smoothing

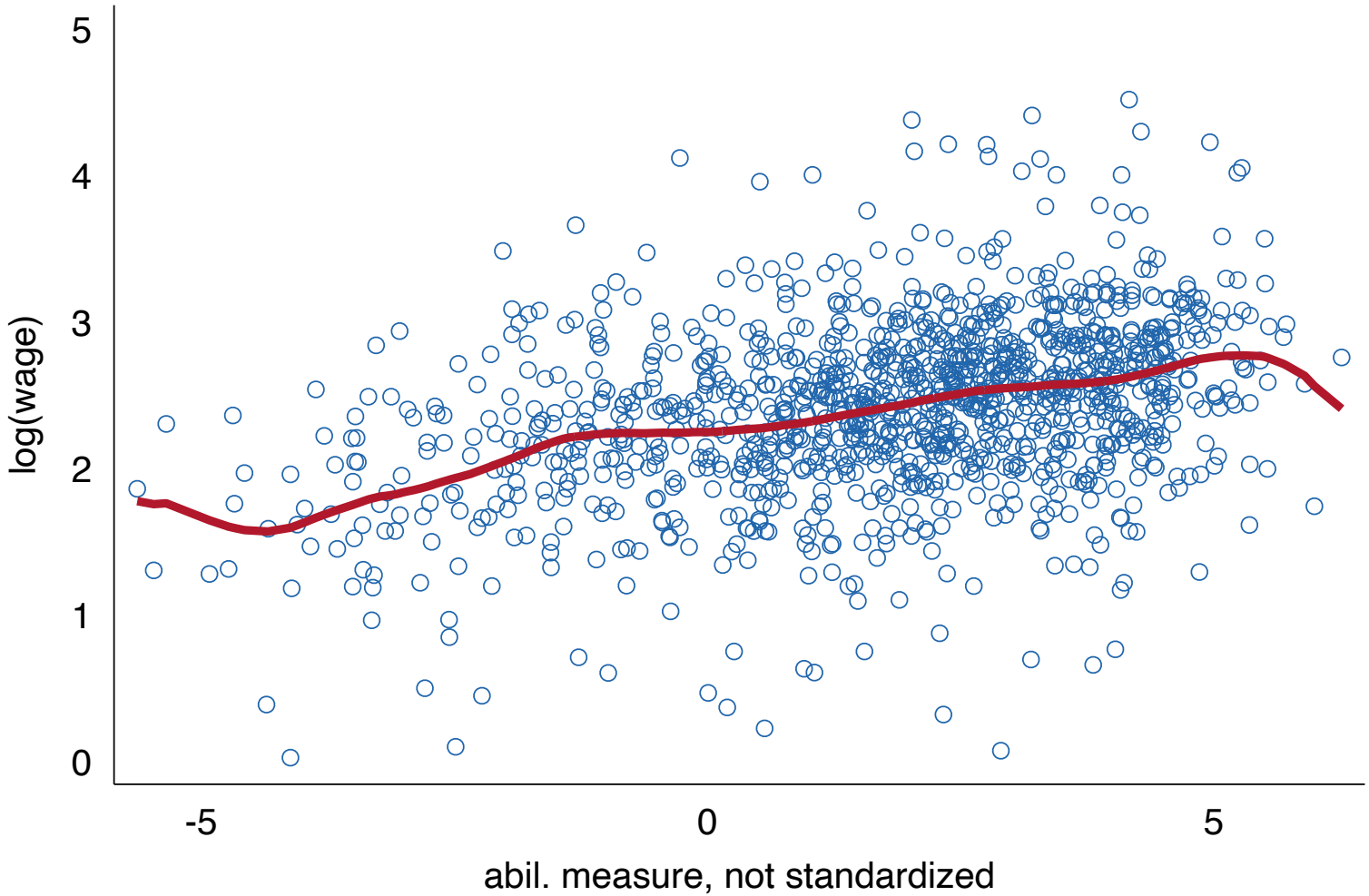
- Choice of bandwidth 0.05 **0.20** 0.50 2.00



Local-linear estimates
kernel = epanechnikov bandwidth = .2

Local smoothing

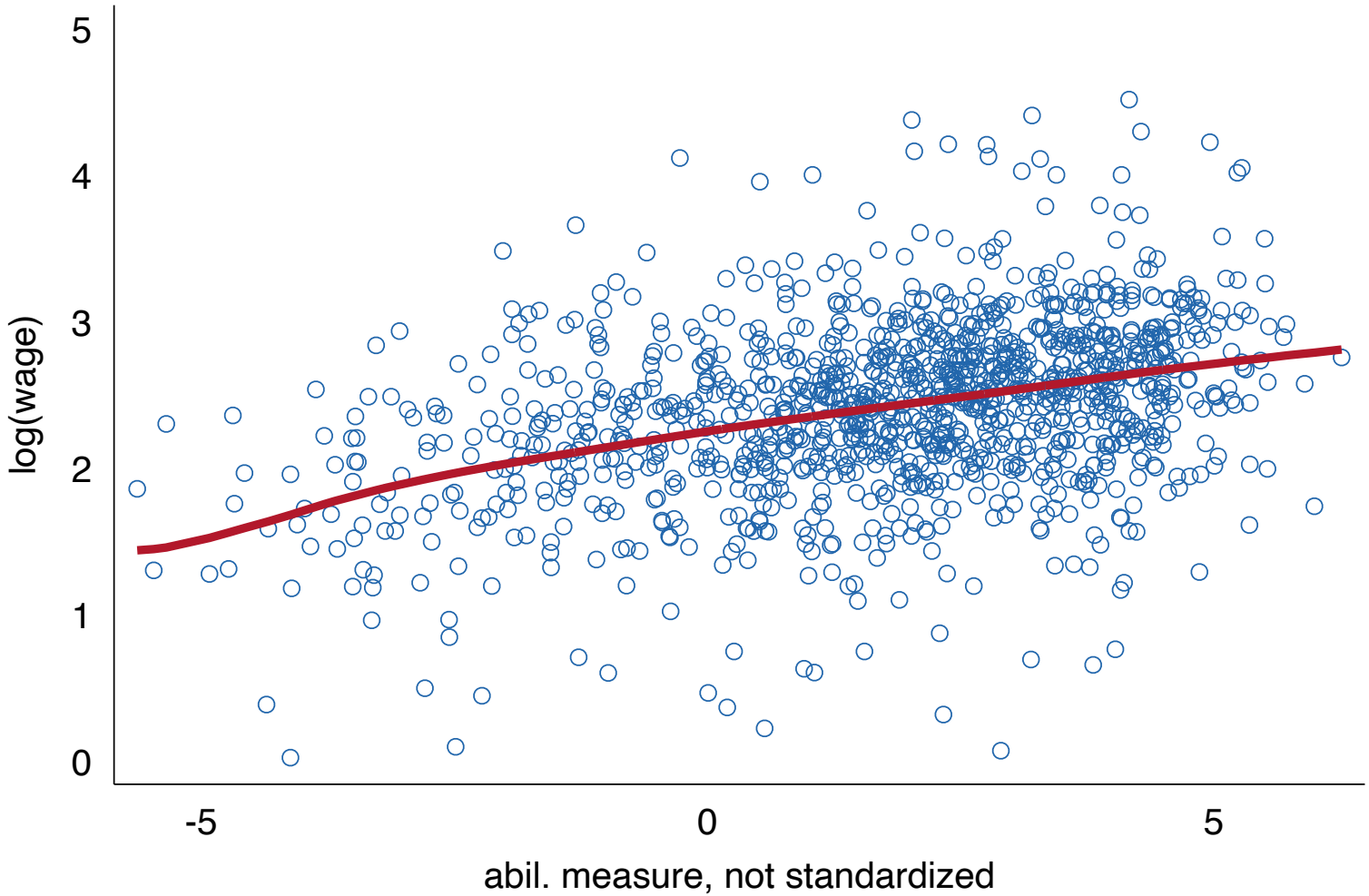
- Choice of bandwidth 0.05 0.20 **0.50** 2.00



Local-linear estimates
kernel = epanechnikov bandwidth = .5

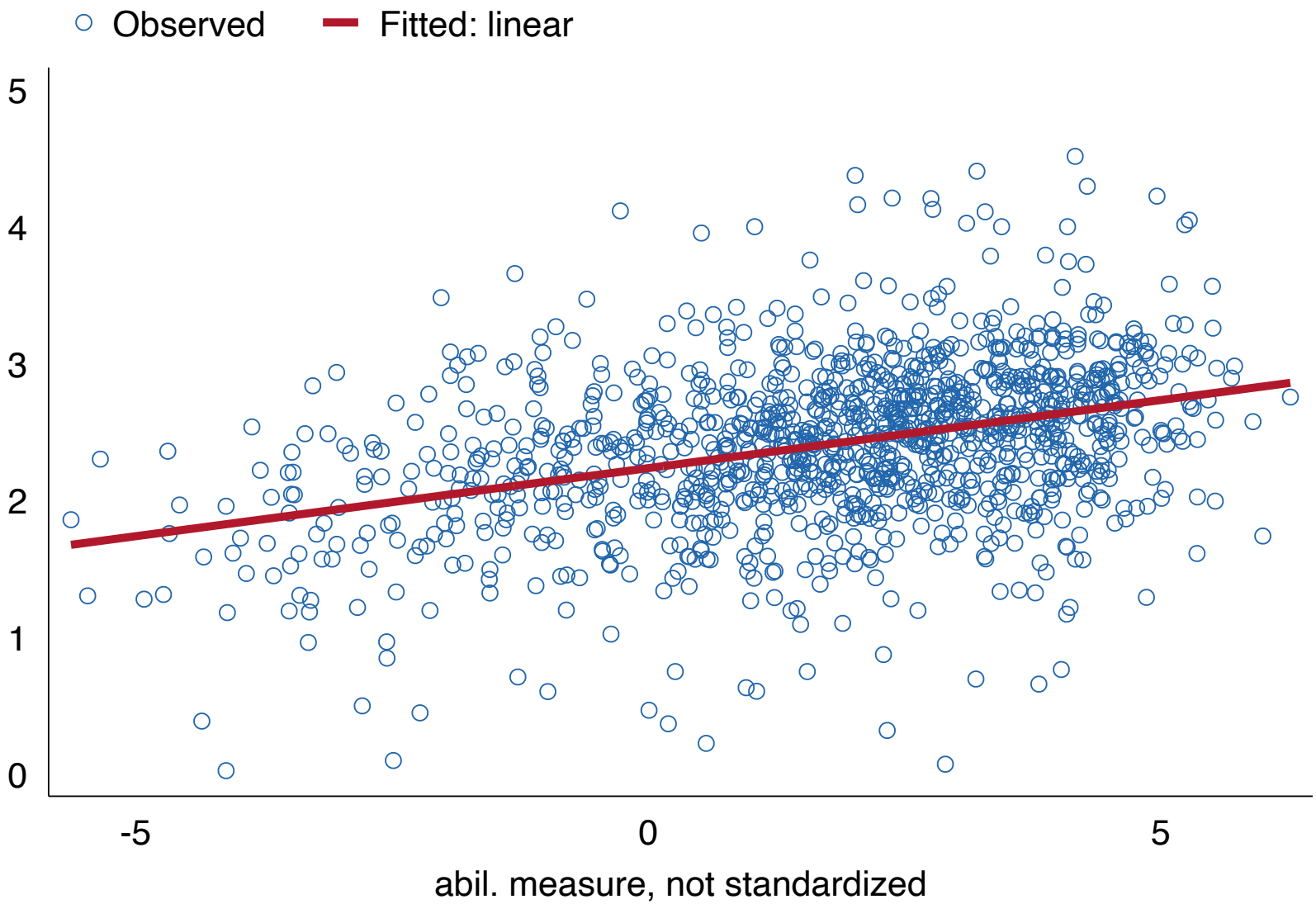
Local smoothing

- Choice of bandwidth 0.05 0.20 0.50 **2.00**

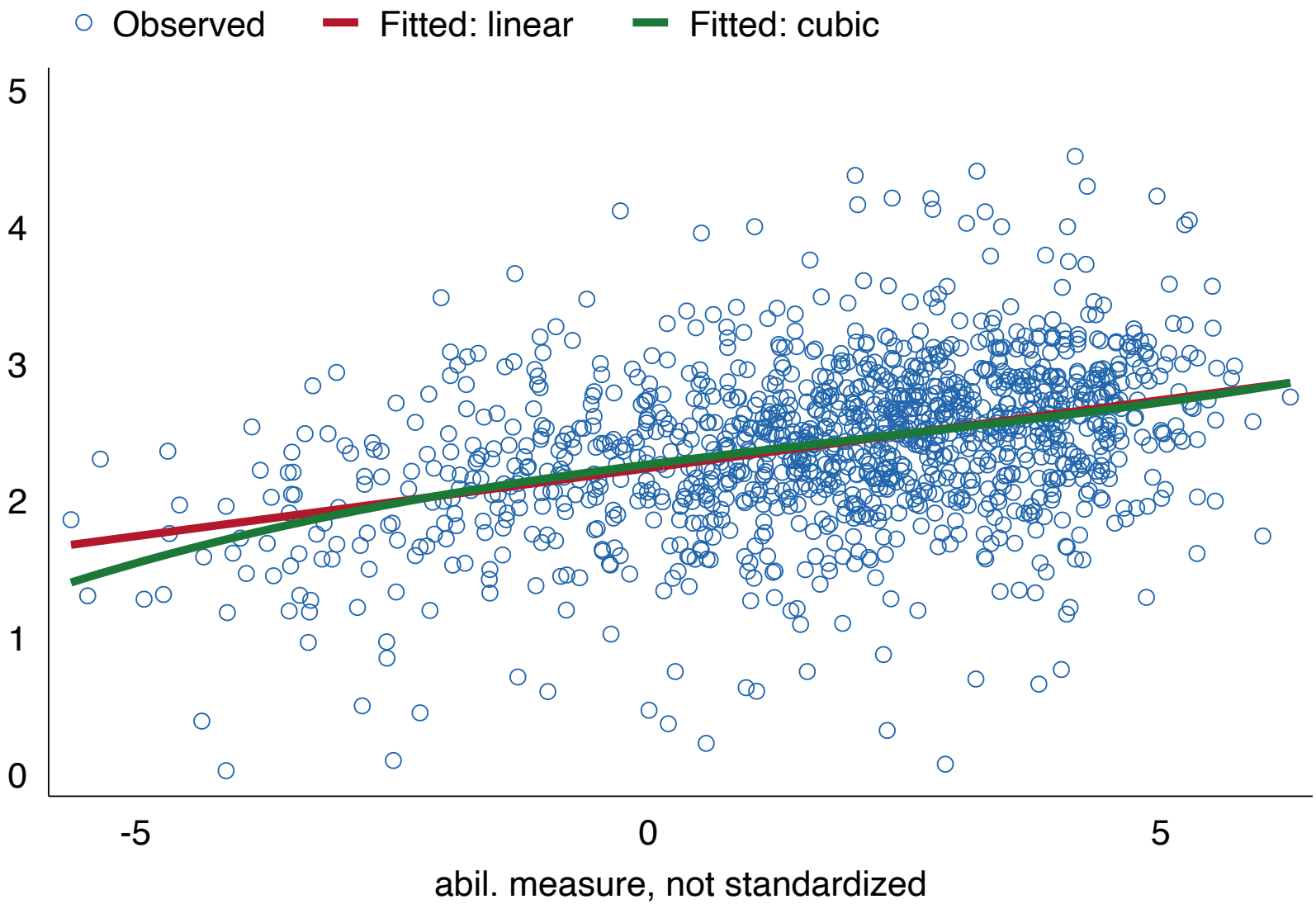


Local-linear estimates
kernel = epanechnikov bandwidth = 2

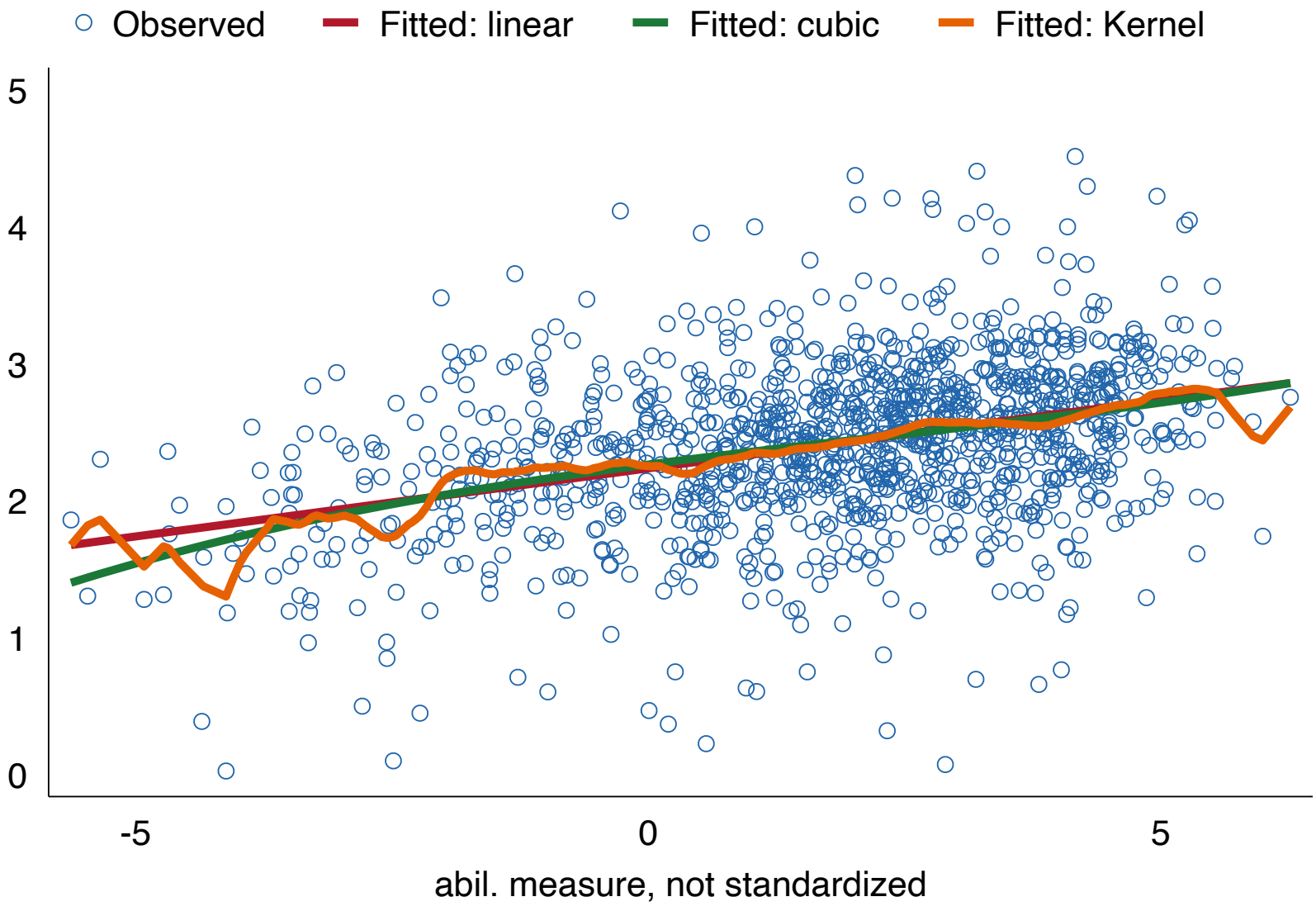
Global smoothing: linear model linear in ability



Global smoothing: linear model cubic in ability



Global versus local smoothing



Ordinary Least Squares (OLS) setting

1. **Population model**

$$y = \mathrm{x}\beta + u \tag{1}$$

   - x is $1 \times K$ (for notational convenience, x_1 is unity)

2. **Random sample**

$$y_i = \mathrm{x}_i\beta + u_i \tag{2}$$

   - For each random draw i: $\{(\mathrm{x}_i, y_i) : i = 1, ..., N\}$

- **Can we identify the parameters β?**
  - Think at the problem as a population problem using equation (1)

Assumption OLS.0 – Linearity

Linearity

$$y = \mathrm{x}\beta + u$$

1. Observable variables enter linearly in the equation
   - We can call $\mathrm{x}\beta$ a *linear index*
2. Error term (unobservable) is separable and additive

Violations

- Model nonlinear in parameters *may* be more appropriate
  - Example: the range of y is restricted, such as binary variables

Identification steps

$$y = \mathrm{x}\beta + u$$

- u is unobserved $\Rightarrow$ β cannot be identified without assumptions
- OLS trick: **think about expected values** using the following steps:
  1. Multiply y by x'

$$\mathrm{x}'y = (\mathrm{x}'\mathrm{x})\beta + \mathrm{x}'u$$

  2. Take the expected value:

$$E(\mathrm{x}'y) = E(\mathrm{x}'\mathrm{x})\beta + E(\mathrm{x}'u)$$

- What assumptions are needed now?

Assumption OLS.1 – Orthogonality

Orthogonality

$$E(\mathbf{x}'u) = 0$$

- When x has an intercept (almost always), orthogonality includes the following K conditions:

$$\begin{aligned} E(u) &= 0 \\ Cov(x_j, u) &= 0 \quad j = 2, ..., K \end{aligned}$$

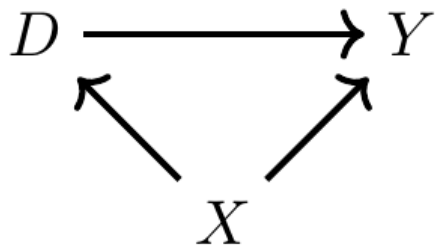
- OLS.1 allows deleting the last term in our derivation

$$E(\mathbf{x}'y) = E(\mathbf{x}'\mathbf{x})\beta + E(\mathbf{x}'u) = E(\mathbf{x}'\mathbf{x})\beta$$

Violations of orthogonality condition

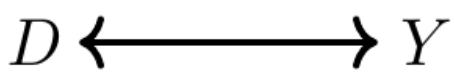
$$Y_i = \alpha + \beta D_i + \epsilon_i$$

1. **Omitted variables**: certain explanatory variables are excluded from the regression model but are correlated with independent variables



2. **Measurement error**: variables are measured with error
   - D is uncorrelated with ϵ, but I observe only $\tilde{D} = D + v$ and v is correlated with ϵ

3. **Simultaneity**: dependent variable causes dependent variables



Assumption OLS.2 – No Perfect Collinearity

$$E(\mathrm{x}'y) \quad = \quad E(\mathrm{x}'\mathrm{x})\beta$$

To identify $\beta \Rightarrow E(\mathrm{x}'\mathrm{x})$ needs to be invertible

No Perfect Collinearity

$$rank\ E(\mathrm{x}'\mathrm{x}) = K$$

where the rank is the number of linearly independent rows or columns in the matrix.

- Violations:
  - Examples?
  - High correlation among regressors often cannot be avoided, but not a violation of assumptions

Assumption OLS.2 – No Perfect Collinearity

$$E(\mathrm{x}'y) \quad = \quad E(\mathrm{x}'\mathrm{x})\beta$$

To identify $\beta \Rightarrow E(\mathrm{x}'\mathrm{x})$ needs to be invertible

No Perfect Collinearity

$$rank\ E(\mathrm{x}'\mathrm{x}) = K$$

where the rank is the number of linearly independent rows or columns in the matrix.

- **Violations:**
  - Examples?
  - High correlation among regressors often cannot be avoided, but not a violation of assumptions

Under OLS.1 and OLS.2, β is identified

$$y = \mathrm{x}\beta + u$$

- Multiply y by x' and take the expected value:

$$\begin{aligned} E(\mathrm{x}'y) &= E(\mathrm{x}'\mathrm{x})\beta + E(\mathrm{x}'u) \\ E(\mathrm{x}'y) &= E(\mathrm{x}'\mathrm{x})\beta \text{ by OLS.1} \\ \beta &= [E(\mathrm{x}'\mathrm{x})]^{-1}E(\mathrm{x}'y) \text{ by OLS.2} \end{aligned}$$

- Intuition:
  1. $E(\mathrm{x}'\mathrm{x})$ is a $K \times K$ matrix of variances and covariances in the population (**variance-covariance matrix**)
  2. $E(\mathrm{x}'y)$ is a $K \times 1$ vector of population covariances

- **Example**: apply this procedure using linear algebra to $y_i = \beta x_i + u_i$?

Under OLS.1 and OLS.2, β is identified

$$y = \mathrm{x}\beta + u$$

- Multiply y by x' and take the expected value:

$$\begin{aligned} E(\mathrm{x}'y) &= E(\mathrm{x}'\mathrm{x})\beta + E(\mathrm{x}'u) \\ E(\mathrm{x}'y) &= E(\mathrm{x}'\mathrm{x})\beta \text{ by OLS.1} \\ \beta &= [E(\mathrm{x}'\mathrm{x})]^{-1}E(\mathrm{x}'y) \text{ by OLS.2} \end{aligned}$$

- Intuition:
  1. $E(\mathrm{x}'\mathrm{x})$ is a $K \times K$ matrix of variances and covariances in the population (**variance-covariance matrix**)
  2. $E(\mathrm{x}'y)$ is a $K \times 1$ vector of population covariances

- **Example**: apply this procedure using linear algebra to $y_i = \beta x_i + u_i$?

Estimation follows from identification

From identification $\Rightarrow$ apply the **random sample correspondent**

- Replace population means (expected values) with sample means
- OLS estimator:

$$\begin{aligned}\hat{\beta} &= \left(N^{-1}\sum_{i=1}^{N} \mathrm{x}_i'\mathrm{x}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N} \mathrm{x}_i' y_i\right) \\ &= (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y}\end{aligned}$$

  - X is $N \times K$ with i^{th} row x_i
  - Y is $N \times 1$ with i^{th} entry y_i

Consistency

Consistency of OLS estimator

Under OLS.1 and OLS.2, **OLS consistently estimates** β, or

$$plim_{N\to\infty}(\hat{\beta}) = \beta$$

$$\begin{aligned} plim_{N\to\infty}(\hat{\beta}) &= plim\left[\left(N^{-1}\sum_{i=1}^{N}\mathrm{x}_i'\mathrm{x}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\mathrm{x}_i'y_i\right)\right] \\ &= \left(plim\ N^{-1}\sum_{i=1}^{N}\mathrm{x}_i'\mathrm{x}_i\right)^{-1} plim\left(N^{-1}\sum_{i=1}^{N}\mathrm{x}_i'y_i\right) \\ &= [E(\mathrm{x}'\mathrm{x})]^{-1}E(\mathrm{x}'y) = \beta \end{aligned}$$

Efficiency or precision of the estimator

Standard errors account for uncertainty in estimated coefficients

From the estimator, replace y with the true population model

$$\begin{aligned}\hat{\beta} &= [E(\mathrm{x}'\mathrm{x})]^{-1}E(\mathrm{x}'y) \\ &= [E(\mathrm{x}'\mathrm{x})]^{-1}E(\mathrm{x}'x)\beta + [E(\mathrm{x}'\mathrm{x})]^{-1}E(\mathrm{x}'u) \\ &= \beta + [E(\mathrm{x}'\mathrm{x})]^{-1}E(\mathrm{x}'u)\end{aligned}$$

- We can therefore write:

$$\hat{\beta} - \beta = [E(\mathrm{x}'\mathrm{x})]^{-1}E(\mathrm{x}'u) \quad (3)$$

- If OLS assumptions are:
  - Not valid $\Rightarrow$ **bias** (does not converge to 0 as $n \to \infty$).
  - Valid $\Rightarrow$ **sampling error** ($\to 0$ as $n \to \infty$).

Asymptotic distribution of OLS

- Start from equation (3) and write the sample correspondent

$$\sqrt{N}(\hat{\beta} - \beta) = \left(N^{-1} \sum_{i=1}^{N} \mathrm{x}_i' \mathrm{x}_i \right)^{-1} \left(N^{-1/2} \sum_{i=1}^{N} \mathrm{x}_i' u_i \right)$$

(Strong) Law of large numbers (LLN)

Let $X_1, X_2, \ldots$ be a random sample of size n – a sequence of independent and identically distributed (i.i.d.) random variables drawn from a distribution with $E[X] = \mu$.

According to the LLN, the sample average converges almost surely to the expected value:

$$Pr\left[\lim_{x \to \infty} \overline{X}_n = \mu \right] = 1$$

Asymptotic distribution of OLS: revision

- Start from equation (3) and write the sample correspondent

$$\sqrt{N}(\hat{\beta} - \beta) = \left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \left(N^{-1/2} \sum_{i=1}^{N} \mathbf{x}_i' u_i \right)$$

(Lindeberg-Lévy) Central limit theorem (CLT)

Let $X_1, X_2, \ldots$ be a random sample of size n – a sequence of independent and identically distributed (i.i.d.) random variables drawn from a distribution with $E[X] = \mu$ and $Var[X] = \sigma^2 < \infty$.

As $n \to \infty$,

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \tag{4}$$

Asymptotic distribution of OLS

- To get the limiting distribution of OLS, start from equation (3)

$$\sqrt{N}(\hat{\beta} - \beta) = \left(N^{-1} \sum_{i=1}^{N} \mathrm{x}_i' \mathrm{x}_i \right)^{-1} \left(N^{-1/2} \sum_{i=1}^{N} \mathrm{x}_i' u_i \right)$$

- The right-hand side is the product of two elements

1. Apply the **LLN** to the first element

$$\mathrm{A} = E(\mathrm{x}_i' \mathrm{x}_i) \tag{5}$$

2. Apply the **CLT** to the second element

$$N^{-1/2} \sum_{i=1}^{N} \mathrm{x}_i' u_i \xrightarrow{d} Normal(0, \mathrm{B}) \tag{6}$$

$$\mathrm{B} = Var(\mathrm{x}_i' u_i) = E(u_i^2 \mathrm{x}_i' \mathrm{x}_i)$$

Asymptotic distribution of OLS

- We can then write

$$\sqrt{N}(\hat{\beta} - \beta) = \mathrm{A}^{-1} \left(N^{-1/2} \sum_{i=1}^{N} \mathrm{x}_i' u_i \right)$$

Asymptotic distribution of OLS estimator

From (6) and (5)

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} Normal(0, \mathrm{A}^{-1}\mathrm{B}\mathrm{A}^{-1})$$

where the variance matrix, $\mathrm{A}^{-1}\mathrm{B}\mathrm{A}^{-1}$, is a **robust sandwich** form.

- The variance matrix contains the variance of each β estimated in the main diagonal $\Rightarrow$ their square roots are the **standard errors**

Homoskedasticity

We can impose further assumptions – one example is **homoskedasticity**

Errors are **homoskedastic**

Sufficient condition is that

$$E(u^2|\mathrm{x}) = \sigma^2$$

- We can then write the variance of B as

$$E(u^2\mathrm{x}'\mathrm{x}) = \sigma^2 E(\mathrm{x}'\mathrm{x})$$

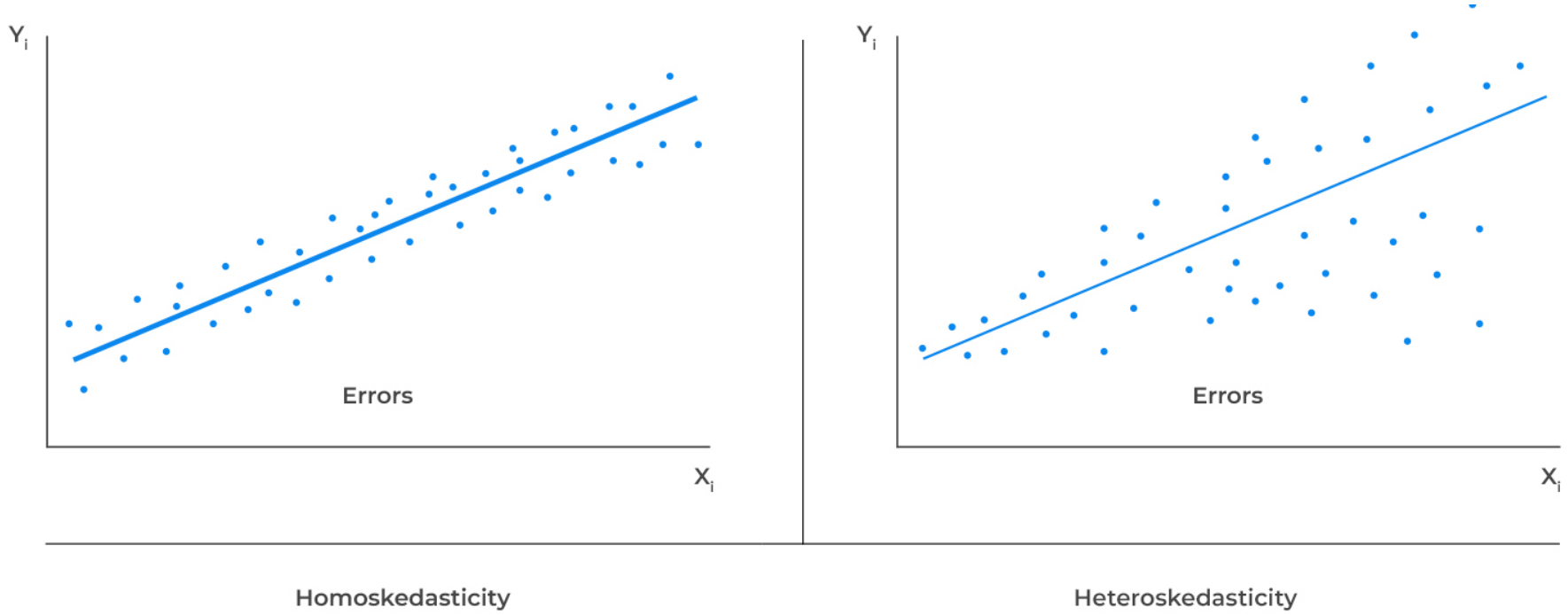
- If we add homoskedasticity, then

$$\mathrm{B} = \sigma^2 \mathrm{A}$$

- (Unrealistic) simplification of the variance of the estimator

Violations of homoskedasticity

Whether homoskedasticity is satisfied is **always an empirical issue**



- Homoskedasticity is often violated!

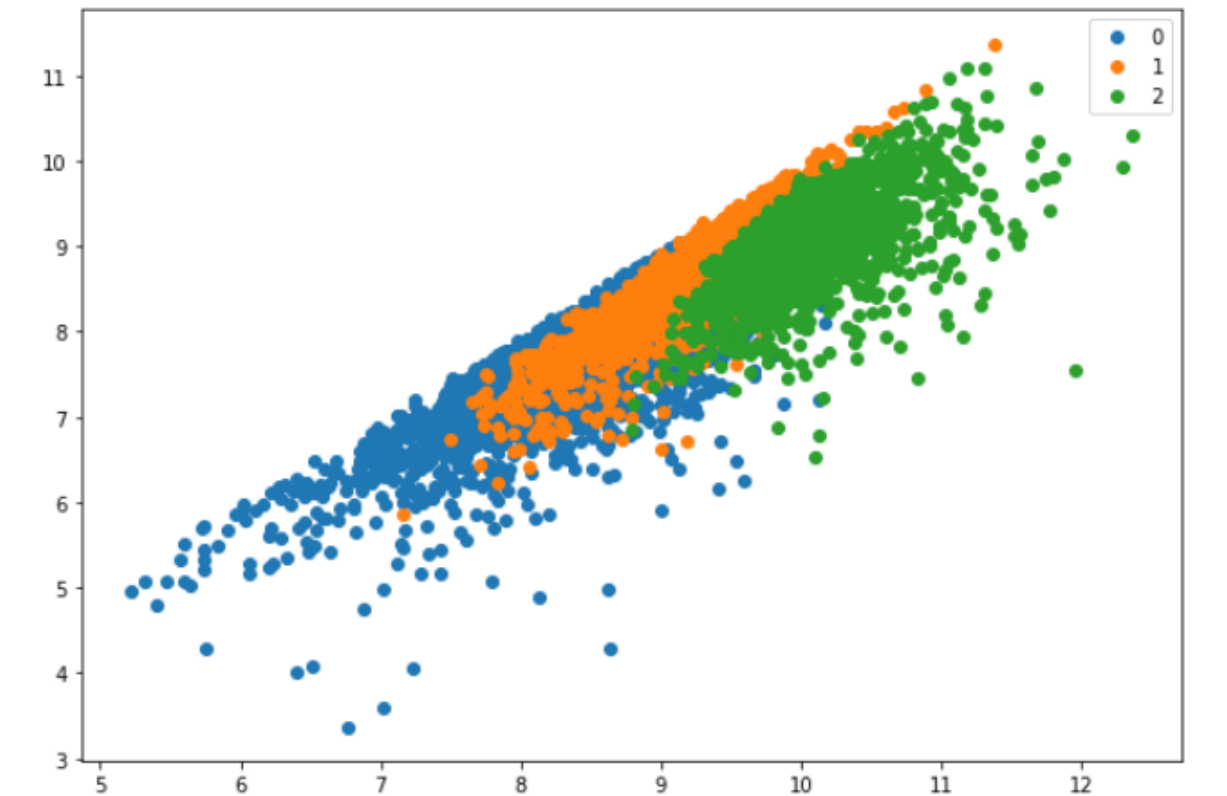
Robust inference

- Steps:
  1. Estimate $\hat{\beta}$ and residuals $\hat{u}$
  2. Apply the estimator for $\widehat{\text{Avar}}(\hat{\beta})$
- Avar($\hat{\beta}$) is estimated with the **sandwich form**:

$$\begin{aligned}
\widehat{\text{Avar}}(\hat{\beta}) &= \hat{\text{A}}^{-1}\hat{\text{B}}\hat{\text{A}}^{-1}/N \\
&= \frac{N}{(N-K)}\left(\sum_{i=1}^{N}\text{x}_i'\text{x}_i\right)^{-1}\left(\sum_{i=1}^{N}\hat{u}_i^2\text{x}_i'\text{x}_i\right)\left(\sum_{i=1}^{N}\text{x}_i'\text{x}_i\right)^{-1}
\end{aligned}$$

Additional corrections: clustering

- Used when observations are grouped into clusters
  - Firms, households, schools, villages
- The observations within each cluster may be correlated
  - Violates the assumption of independence $\Rightarrow$ underestimate s.e.



Normality of the error term

- Do we need to assume **normality** of the error term?

$$u|x_1, ..., x_K \sim Normal(0, \sigma^2)$$

- Normality is *not* needed for large-sample inference
- Normality underlies **exact inference**
  - So why assuming normality?

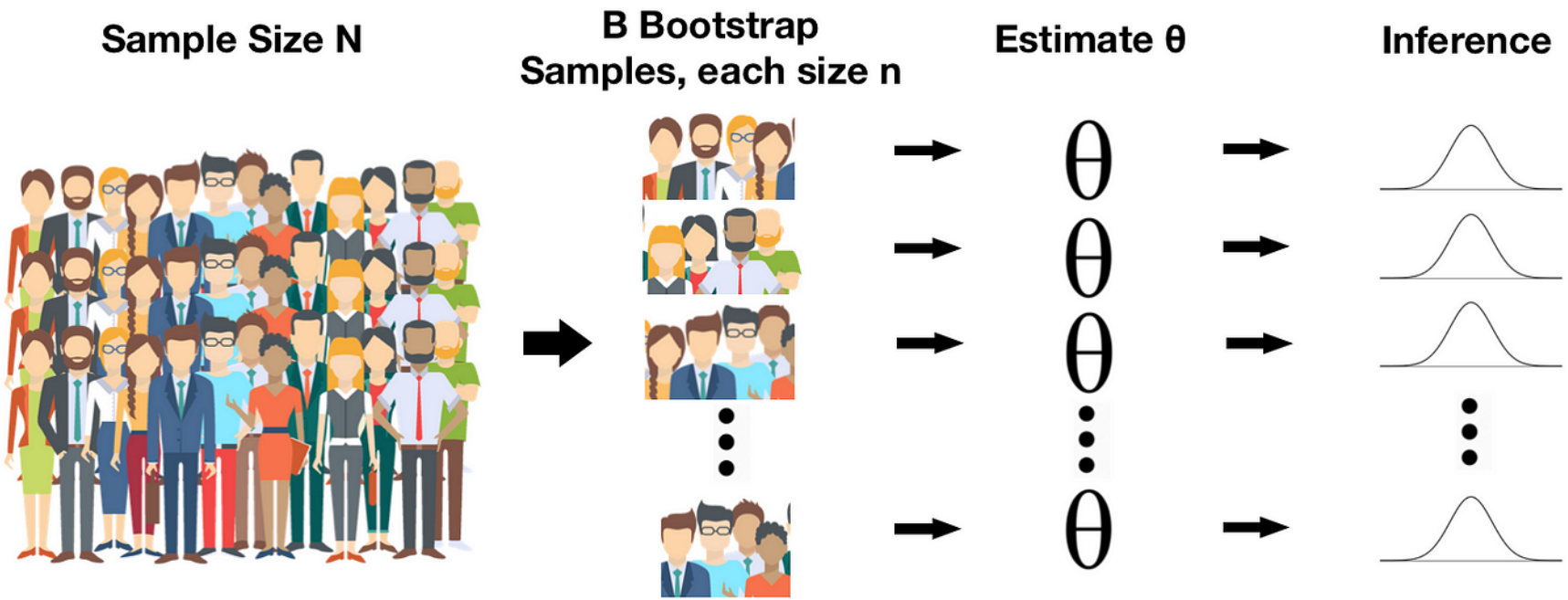
Parametric versus non-parametric s.e.

Why non-parametric? $\Rightarrow$ unsure about the (parametric) formula for s.e.

- $\{(y_1, x_1), ..., (y_N, x_N)\}$ is the sample
- **Bootstrap procedure:**
  1. Obtain B different random samples from this sample using resampling with replacement
  2. Generate estimates for each B: $\hat{\theta}_1, .., \hat{\theta}_B$
  3. Estimate the variance of these estimates $\Rightarrow$ tells us about how much variation there is in the estimates
  4. Square root is called **bootstrap standard error**
  5. Empirical percentiles can be used as confidence intervals (see empirical cdf)

Parametric versus non-parametric s.e.

Why non-parametric? $\Rightarrow$ unsure about the (parametric) formula for s.e.



Causal effects versus forecasting

- If we are interested in forecasting, then we should consider if our model fit the data well
- R^2 is a goodness-of-fit measure

$$\rho^2 = 1 - \sigma_u^2/\sigma_y^2$$

- It depends on the unconditional variance

$$\sigma_y^2 = N^{-1} \sum_i (y_i - \bar{y}_i)^2$$

- It ranges between 0 and 1, with 1 = perfectly fitting the data
- Careful in the use of R^2, we are interested in derivatives!

APPLICATION: 401(k) pension plan

- In the US, a 401(k) plan is a defined-contribution pension plan
  - Retirement contributions are provided by an employer, deducted from the employee's paycheck before taxation and limited to a maximum pre-tax annual contribution.
- The model with constant partial effects is

$$prate = \beta_0 + \beta_1 mrate + \beta_2 age + \beta_3 ltotemp + \beta_4 sole + u$$

  - *prate*: firm participation rate
  - *mrate*: amount the firm contributes for each $ put in by the employee
  - *age*: age of the plan
  - *ltotemp*: log of total firm employment
  - *sole*: dummy variable for the plan being the only retirement option
- Data in 401KPART.DTA.

Describe your data

```
. use 401kpart

. des

Contains data from C:\mitbook1_2e\statafiles\401kpart.dta
  obs:         4,075
 vars:            10                          2 Nov 2005 15:30
 size:       138,550
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
partic          float  %9.0g                  # employees partic. in 401(k)
totemp          float  %9.0g                  # worldwide firm employees
employ          float  %9.0g                  # employees eligible for 401(k)
mrate           float  %9.0g                  plan match rate, per $
prate           float  %9.0g                  partic/employ
age             byte   %9.0g                  age of the plan
sole            byte   %9.0g                  =1 if only pension plan
ltotemp         float  %9.0g                  log(totemp)
agesq           float  %9.0g                  age^2
ltotempsq       float  %9.0g                  ltotemp^2
-------------------------------------------------------------------------------
Sorted by:
```

Describe your data

Check for **variation in your data**

- Look at means and standard deviations
- First check for multicollinearity

```
. sum prate mrate age ltotemp sole

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       prate |      4075    .840607    .1874841   .0036364          1
       mrate |      4075    .463519    .4187388          0          2
         age |      4075   8.186503    9.257011          1         71
     ltotemp |      4075    6.97439    1.539165    4.65396   13.00142
        sole |      4075   .3693252    .4826813          0          1
```

Estimate the linear model with constant effects

```
. reg prate mrate age ltotemp sole, robust

Linear regression                               Number of obs =    4075
                                                F(  4,  4070) =  202.82
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1755
                                                Root MSE      =  .17033

------------------------------------------------------------------------------
             |               Robust
       prate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       mrate |   .1072729   .0060035    17.87   0.000     .0955027    .1190432
         age |      .0037   .0002493    14.84   0.000     .0032113    .0041887
     ltotemp |  -.0281719   .0021148   -13.32   0.000    -.0323181   -.0240257
        sole |   .0177024   .0059192     2.99   0.003     .0060977    .0293072
       _cons |   .9505378   .0149728    63.48   0.000     .9211829    .9798927
------------------------------------------------------------------------------
```

Introducing non-linearities

What happens to interpretions?

```
. gen mrateage = mrate*age

. gen mrateltotemp = mrate*ltotemp

. reg prate mrate age mrateage ltotemp mrateltotemp sole, robust

Linear regression                               Number of obs =    4075
                                                F(  6,  4068) =  156.51
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1940
                                                Root MSE      =  .16845

------------------------------------------------------------------------------
             |               Robust
       prate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       mrate |  -.0014222   .0275289    -0.05   0.959     -.055394    .0525496
         age |   .0066224   .0004247    15.59   0.000     .0057898     .007455
    mrateage |  -.0054106   .0005122   -10.56   0.000    -.0064148   -.0044065
     ltotemp |  -.0390588   .0032932   -11.86   0.000    -.0455153   -.0326023
mrateltotemp |   .0240843   .0044453     5.42   0.000     .0153691    .0327995
        sole |   .0170137   .0058649     2.90   0.004     .0055153    .0285121
       _cons |   1.001494   .0219434    45.64   0.000     .9584733    1.044515
------------------------------------------------------------------------------
```

De-meaning to simplify interpretation

```
. gen mrateage0 = mrate*(age - 8.19)

. gen mrateltotemp0 = mrate*(ltotemp - 6.974)

. reg prate mrate age mrateage0 ltotemp mrateltotemp0 sole, robust

Linear regression                               Number of obs =    4075
                                                F(  6,  4068) =  156.51
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1940
                                                Root MSE      =  .16845

------------------------------------------------------------------------------
             |               Robust
       prate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       mrate |   .1222283   .0066737    18.32   0.000     .1091443    .1353124
         age |   .0066224   .0004247    15.59   0.000     .0057898     .007455
   mrateage0 |  -.0054106   .0005122   -10.56   0.000    -.0064148   -.0044065
     ltotemp |  -.0390588   .0032932   -11.86   0.000    -.0455153   -.0326023
mrateltot~p0 |   .0240843   .0044453     5.42   0.000     .0153691    .0327995
        sole |   .0170137   .0058649     2.90   0.004     .0055153    .0285121
       _cons |   1.001494   .0219434    45.64   0.000     .9584733    1.044515
------------------------------------------------------------------------------
```

OLS with heteroskedastic s.e.

```
. reg prate mrate age ltotemp sole, robust

Linear regression                               Number of obs =    4075
                                                F(  4,  4070) =  202.82
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1755
                                                Root MSE      =  .17033

------------------------------------------------------------------------------
             |               Robust
       prate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       mrate |   .1072729   .0060035    17.87   0.000     .0955027    .1190432
         age |      .0037   .0002493    14.84   0.000     .0032113    .0041887
     ltotemp |  -.0281719   .0021148   -13.32   0.000    -.0323181   -.0240257
        sole |   .0177024   .0059192     2.99   0.003     .0060977    .0293072
       _cons |   .9505378   .0149728    63.48   0.000     .9211829    .9798927
------------------------------------------------------------------------------
```

OLS with bootstrap s.e.: 100 repetitions

```
. bootstrap, rep(100): reg prate mrate age ltotemp sole
(running regress on estimation sample)

Bootstrap replications (100)

Linear regression                                       Number of obs =  4,075
                                                        Replications  =    100
                                                        Wald chi2(4)  = 750.84
                                                        Prob > chi2   = 0.0000
                                                        R-squared     = 0.1755
                                                        Adj R-squared = 0.1747
                                                        Root MSE      = 0.1703

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
       prate | coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       mrate |   .1072729   .0067057    16.00   0.000       .09413    .1204159
         age |       .0037   .0002677    13.82   0.000     .0031754    .0042246
     ltotemp |  -.0281719   .0019182   -14.69   0.000    -.0319315   -.0244123
        sole |   .0177024   .0059831     2.96   0.003     .0059757    .0294291
       _cons |   .9505378    .013466    70.59   0.000     .9241449    .9769307
------------------------------------------------------------------------------
```

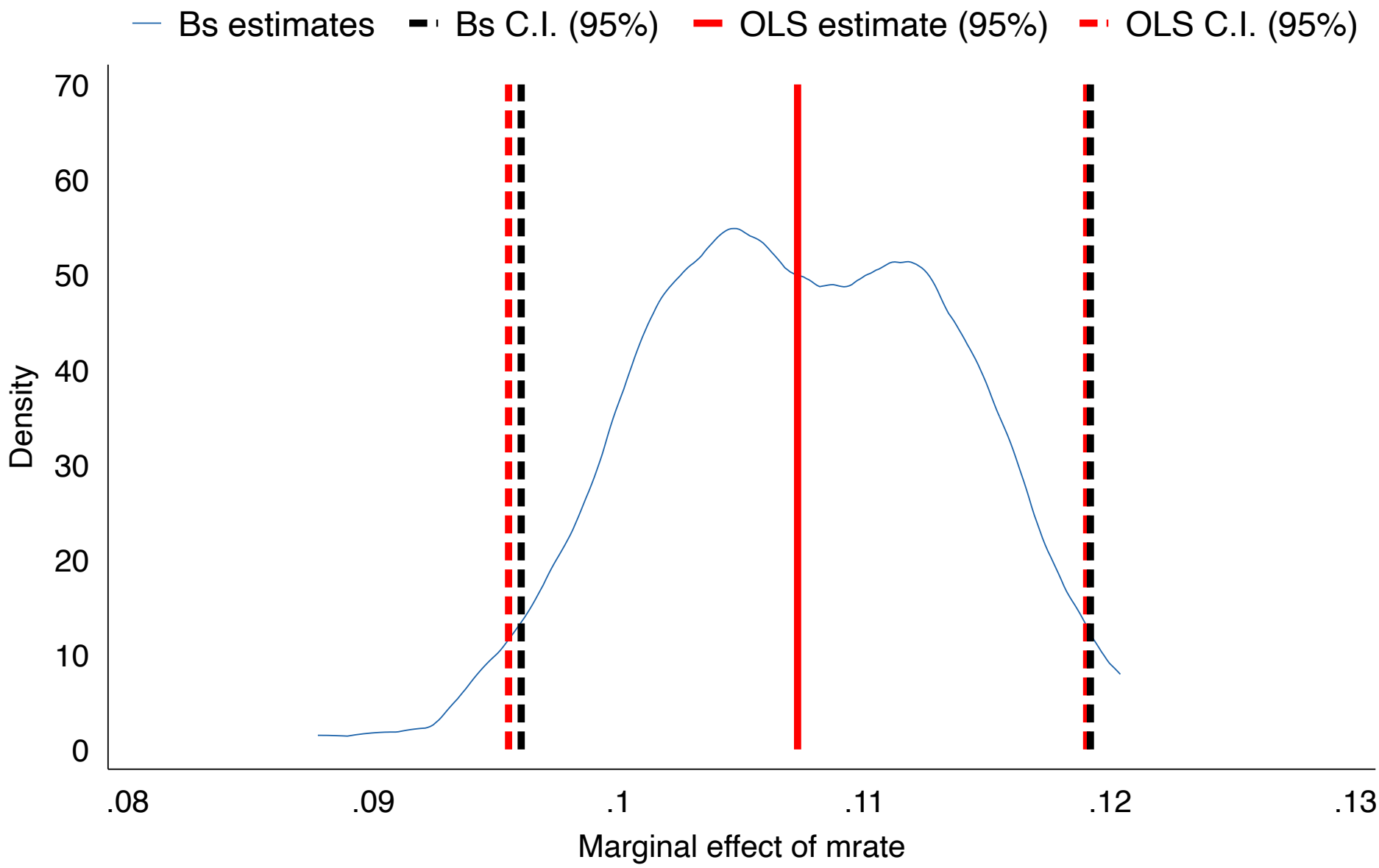
OLS with bootstrap s.e.: 500 repetitions

```
. bootstrap, rep(500): reg prate mrate age ltotemp sole
(running regress on estimation sample)

Bootstrap replications (500)

Linear regression                                      Number of obs =  4,075
                                                       Replications  =    500
                                                       Wald chi2(4)  = 900.30
                                                       Prob > chi2   = 0.0000
                                                       R-squared     = 0.1755
                                                       Adj R-squared = 0.1747
                                                       Root MSE      = 0.1703

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
       prate | coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       mrate |   .1072729   .0058201    18.43   0.000     .0958656    .1186802
         age |       .0037   .0002511    14.74   0.000     .0032078    .0041921
     ltotemp |  -.0281719    .002168   -12.99   0.000     -.032421   -.0239228
        sole |   .0177024   .0059192     2.99   0.003     .0061011    .0293037
       _cons |   .9505378   .0152011    62.53   0.000     .9207442    .9803314
------------------------------------------------------------------------------
```

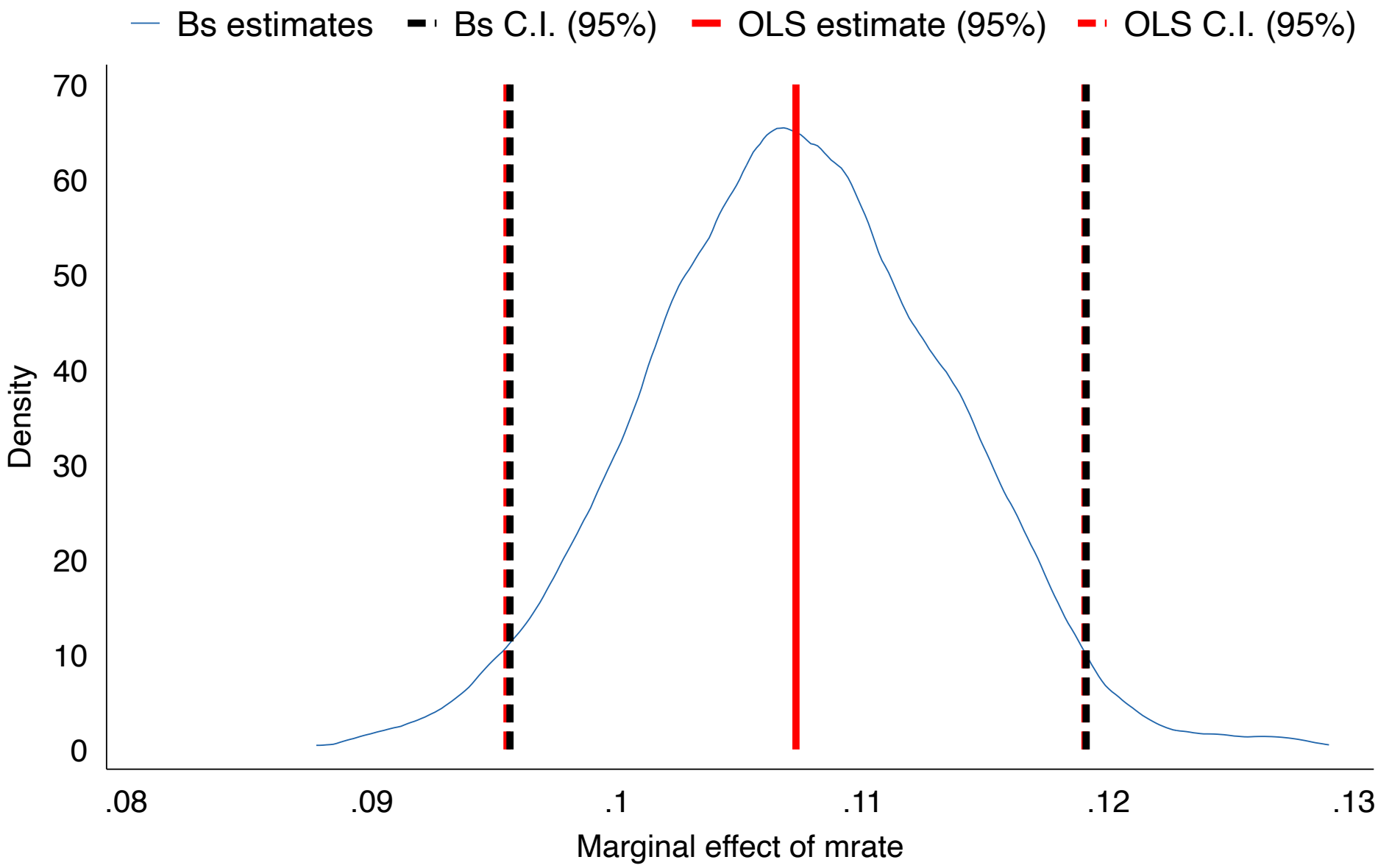
OLS with bootstrap s.e.: 1000 repetitions

```
. bootstrap, rep(1000): reg prate mrate age ltotemp sole
(running regress on estimation sample)

Bootstrap replications (1,000)

Linear regression                                       Number of obs =  4,075
                                                        Replications  =  1,000
                                                        Wald chi2(4)  = 802.62
                                                        Prob > chi2   = 0.0000
                                                        R-squared     = 0.1755
                                                        Adj R-squared = 0.1747
                                                        Root MSE      = 0.1703

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
       prate | coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       mrate |   .1072729   .0060208    17.82   0.000     .0954724    .1190735
         age |      .0037   .0002391    15.47   0.000     .0032313    .0041687
     ltotemp |  -.0281719   .0021217   -13.28   0.000    -.0323303   -.0240134
        sole |   .0177024   .0060789     2.91   0.004     .0057879    .0296169
       _cons |   .9505378   .0150687    63.08   0.000     .9210037    .9800719
------------------------------------------------------------------------------
```

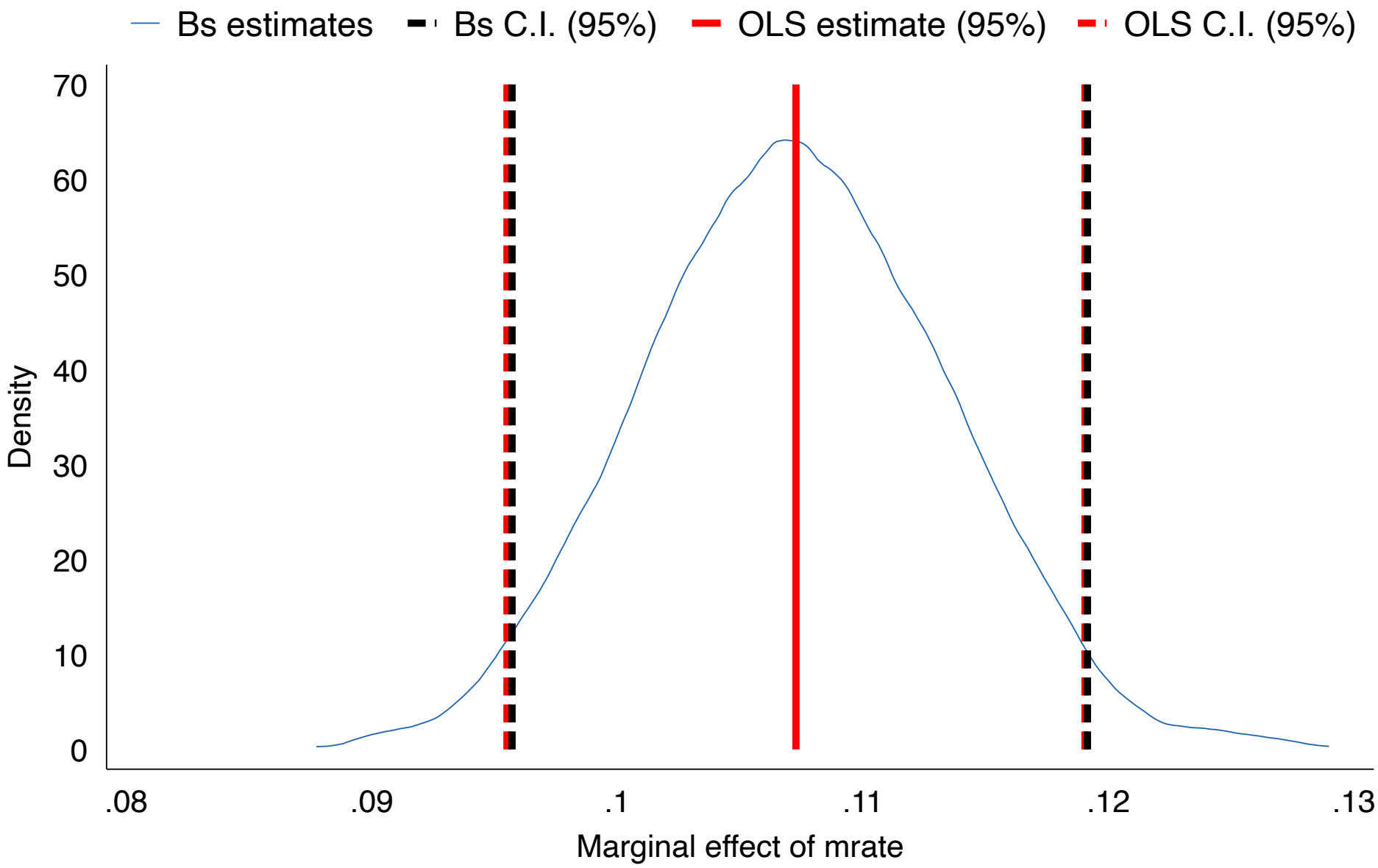
OLS with bootstrap 95% C.I.: 100 repetitions



OLS with bootstrap 95% C.I.: 500 repetitions



OLS with bootstrap 95% C.I.: 1000 repetitions



Topics

1. **Linear model with cross section data**
   - Identification and inference using the linear model
   - **Identification versus prediction**
2. **Violation of orthogonality**
   - IV Estimation of a General Equation
   - Two Stage Least Squares

What is OLS recovering?

Assume $E[y|x] = f(\mathrm{x})$, such that $y = f(x) + \epsilon$ where $E(\epsilon|\mathrm{x}) = 0$

- Look at **square deviations from** y

$$\begin{aligned}(y - \mathrm{xb})^2 &= [f(\mathrm{x}) + e - \mathrm{xb}]^2 \\ &= [f(\mathrm{x}) - \mathrm{xb}]^2 + 2[f(\mathrm{x}) - \mathrm{xb}] \cdot e + e^2 \\ E[(y - \mathrm{xb})^2] &= E\{[f(\mathrm{x}) - \mathrm{xb}]^2\} + \sigma_e^2\end{aligned}$$

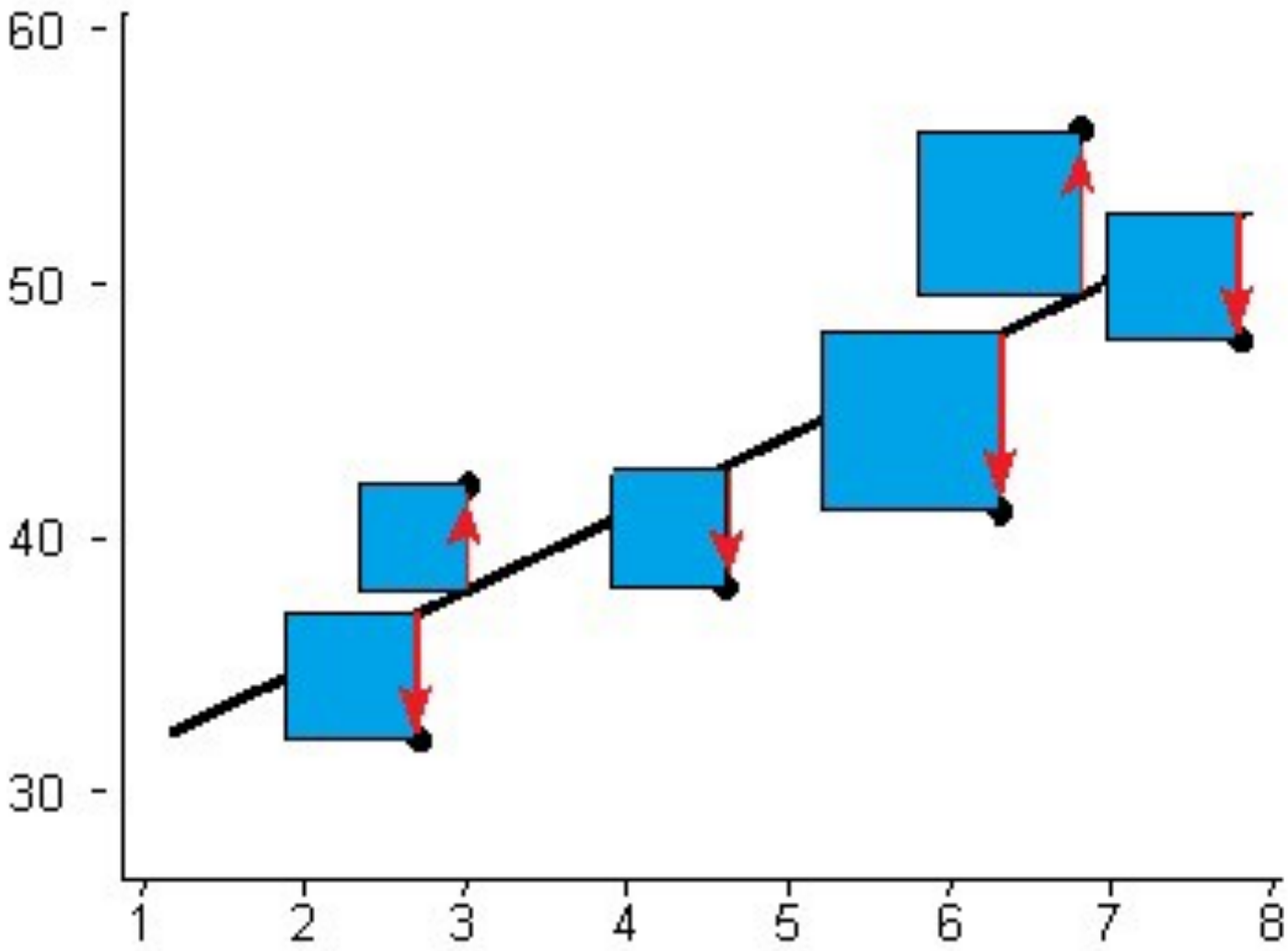
- Because β minimizes $E[(y - \mathrm{xb})^2]$, it also solves

$$\min_{\mathrm{b} \in \mathbb{R}^K} \; E\{[f(\mathrm{x}) - \mathrm{xb}]^2\}$$

(because σ_e^2 does not depend on b).

What is OLS recovering?

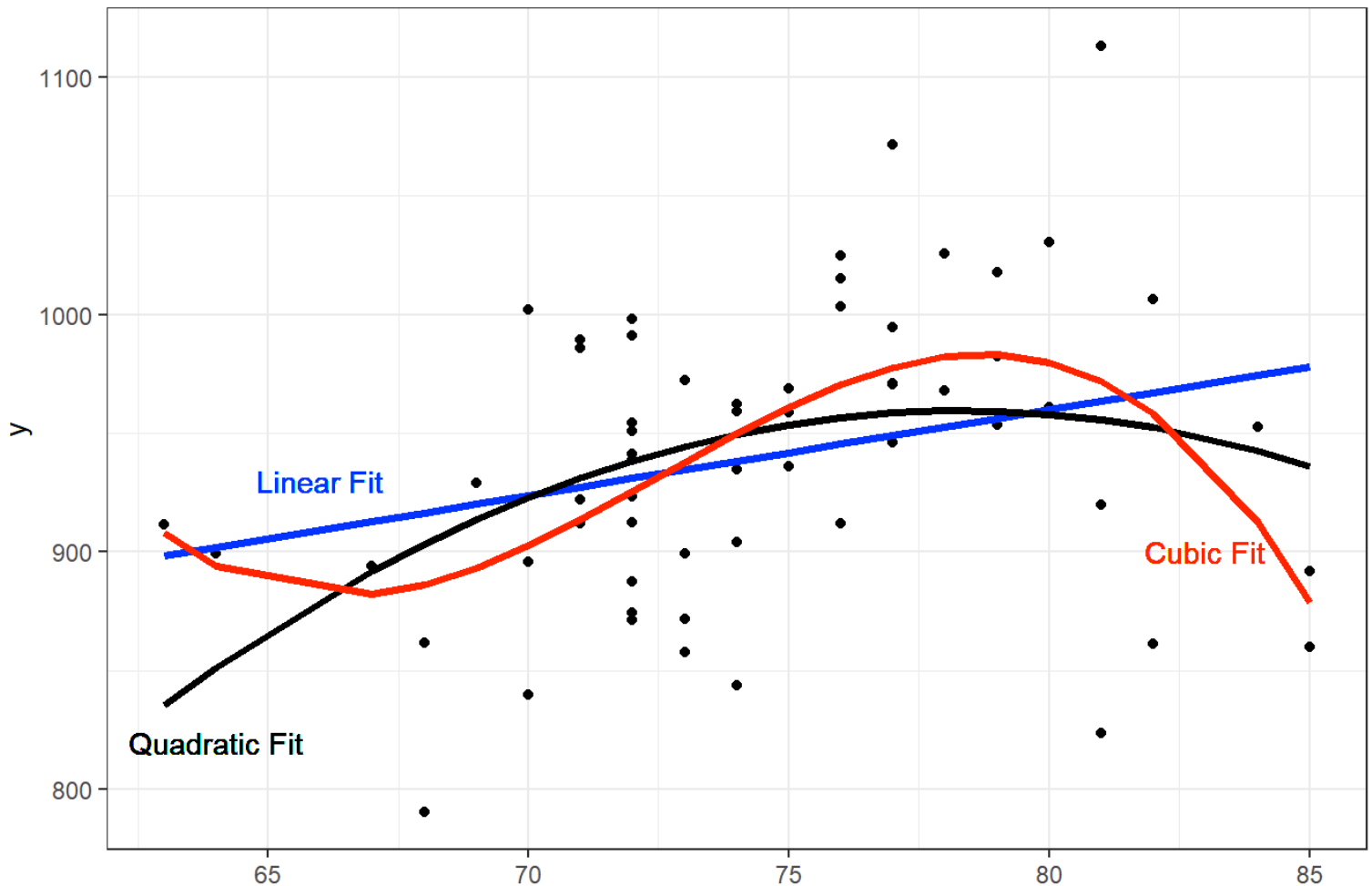
xβ is the **best mean squared error approximation to the true conditional mean function** $\mu(\mathrm{x}) = E(y|\mathrm{x})$.



A parallel with prediction (statistical learning)

Focus on **polynomial regression**:

$$E[y|x] = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x_3 + ...$$



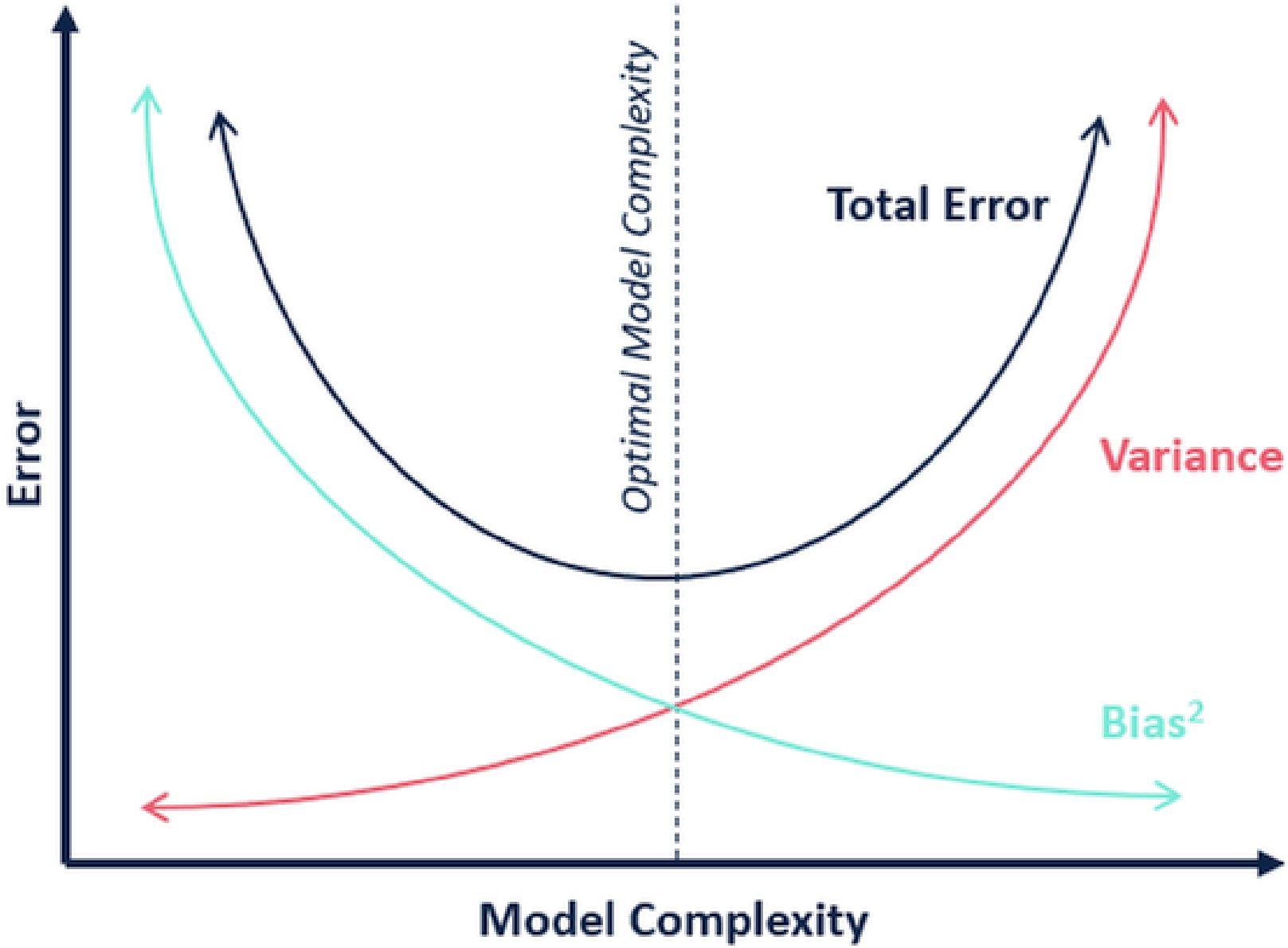
Choosing a model $E[Y|x]$

$$E[y|x] = f(x)$$

- Assume $f(x)$ minimizes $E[(Y - g(X))^2|X]$ over all functions g at all points $X = x$.
- $\epsilon = y - f(x)$ is the **irreducible error**
  - Even if we knew $f(x)$, we would still make errors in prediction, since at each $X = x$ there is typically a distribution of possible Y values.
- **Bias-variance trade-off**: for any estimate $\hat{f}(x)$ of $f(x)$:

$$E[(Y - \hat{f}(X))^2|X = x] = \underbrace{[f(x) - \hat{f}(x)]^2}_{\text{bias (reducible)}} + \underbrace{Var(\epsilon)}_{\text{variance (irreducible)}}$$

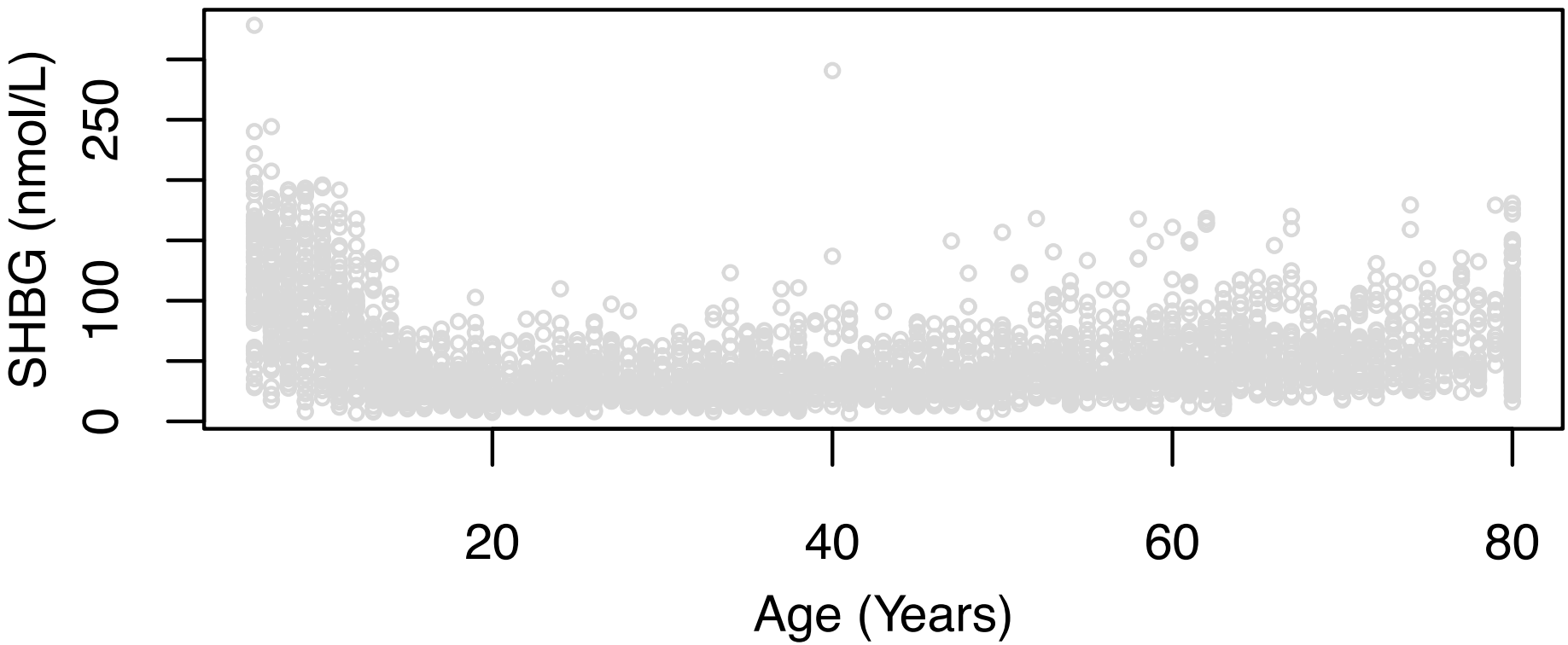
Choosing a model $Y = f(x)$



Global versus local fitting

Balance the **bias-variance trade-off**:

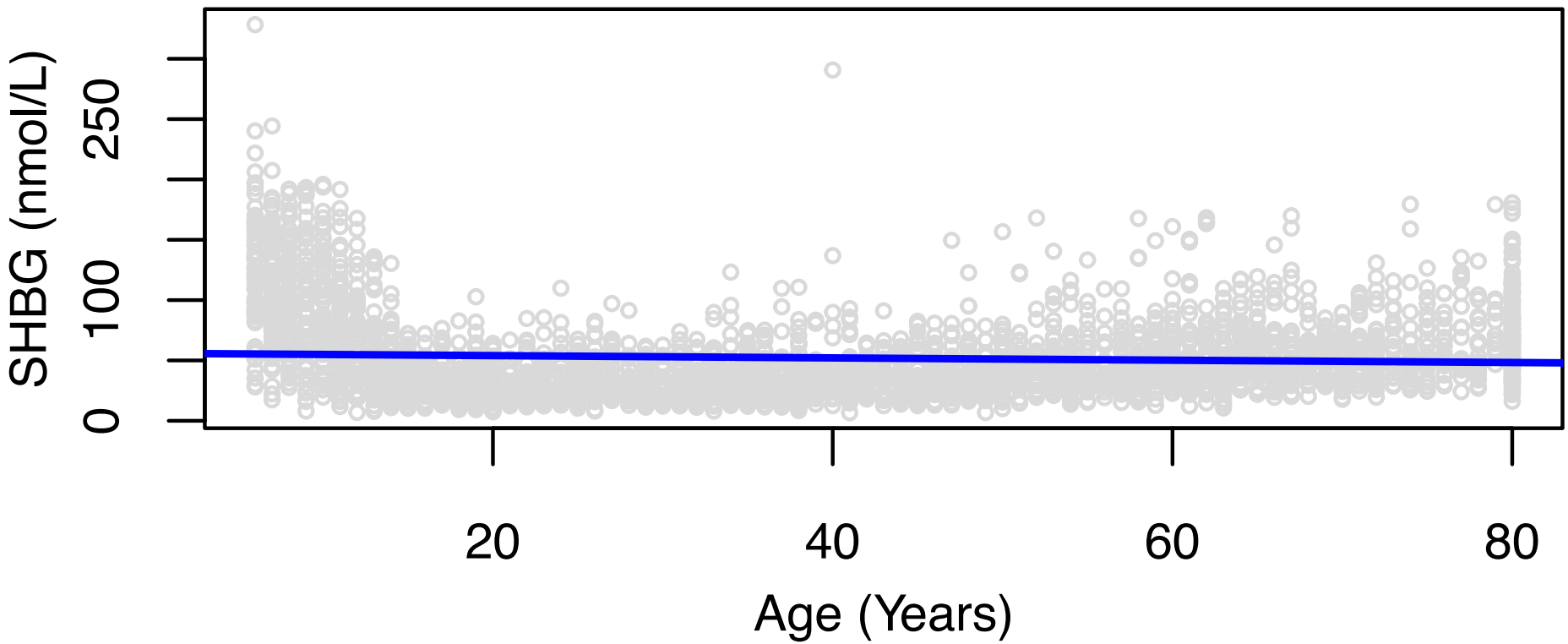
- $\uparrow$ flexibility of our model $\Rightarrow$ overfit the data and $\uparrow$ the variance.
- $\downarrow$ flexibility of our model $\Rightarrow$ poorly fit the data and $\uparrow$ our bias.



Global versus local fitting

Balance the **bias-variance trade-off**:

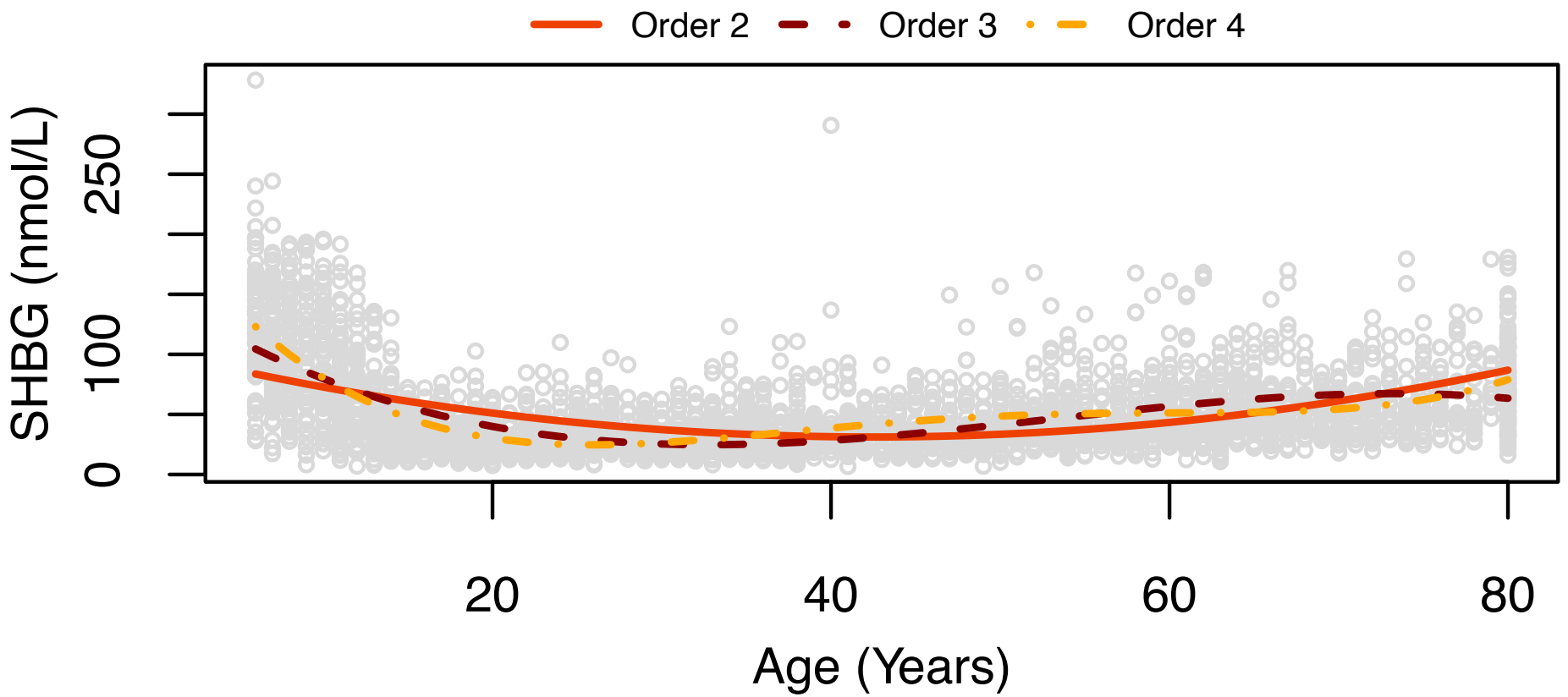
- $\uparrow$ flexibility of our model $\Rightarrow$ overfit the data and $\uparrow$ the variance.
- $\downarrow$ flexibility of our model $\Rightarrow$ poorly fit the data and $\uparrow$ our bias.



Global versus local fitting

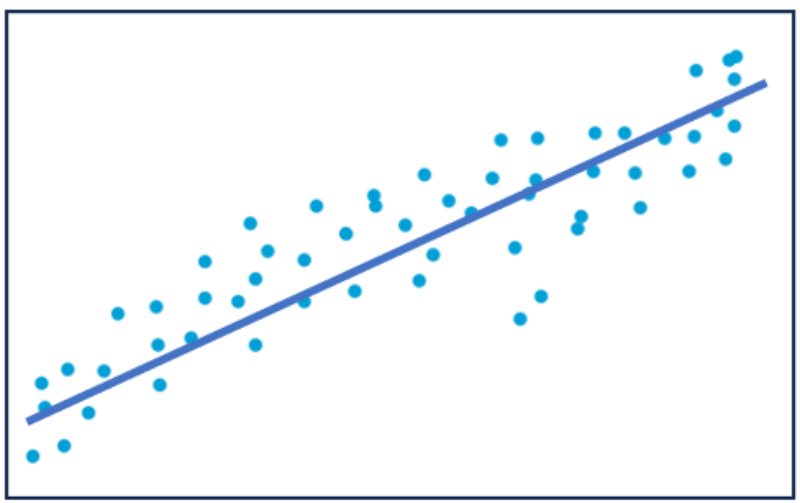
Balance the **bias-variance trade-off**:

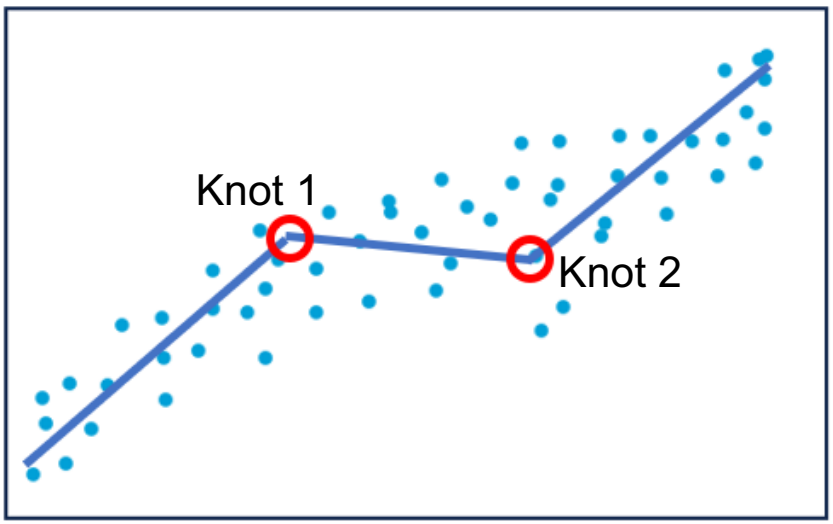
- ↑ flexibility of our model ⇒ overfit the data and ↑ the variance.
- ↓ flexibility of our model ⇒ poorly fit the data and ↑ our bias.



Polynomial vs spline regression

- **Limitation of polynomial regression:** non-locality
  - Fitted regression at any arbitrary point x depends on the data across the entire range
  - Changes to observed values near the boundary (e.g., min or max of x) can lead to changes in the fitted function far from that value.
- **Spline regression**: partition x into smaller intervals based on an arbitrary points and fit localized polynomials.





Spline regression

- **Knots (k)**
  - Points where the different piecewise polynomials are joined.
  - Divide the line into $(k+1)$ parts.
- **Degrees (d)**: degree of the polynomial in each part.
- **Spline**: OLS with polynomial expression for each segment
  - Example: k = 2, d = 2

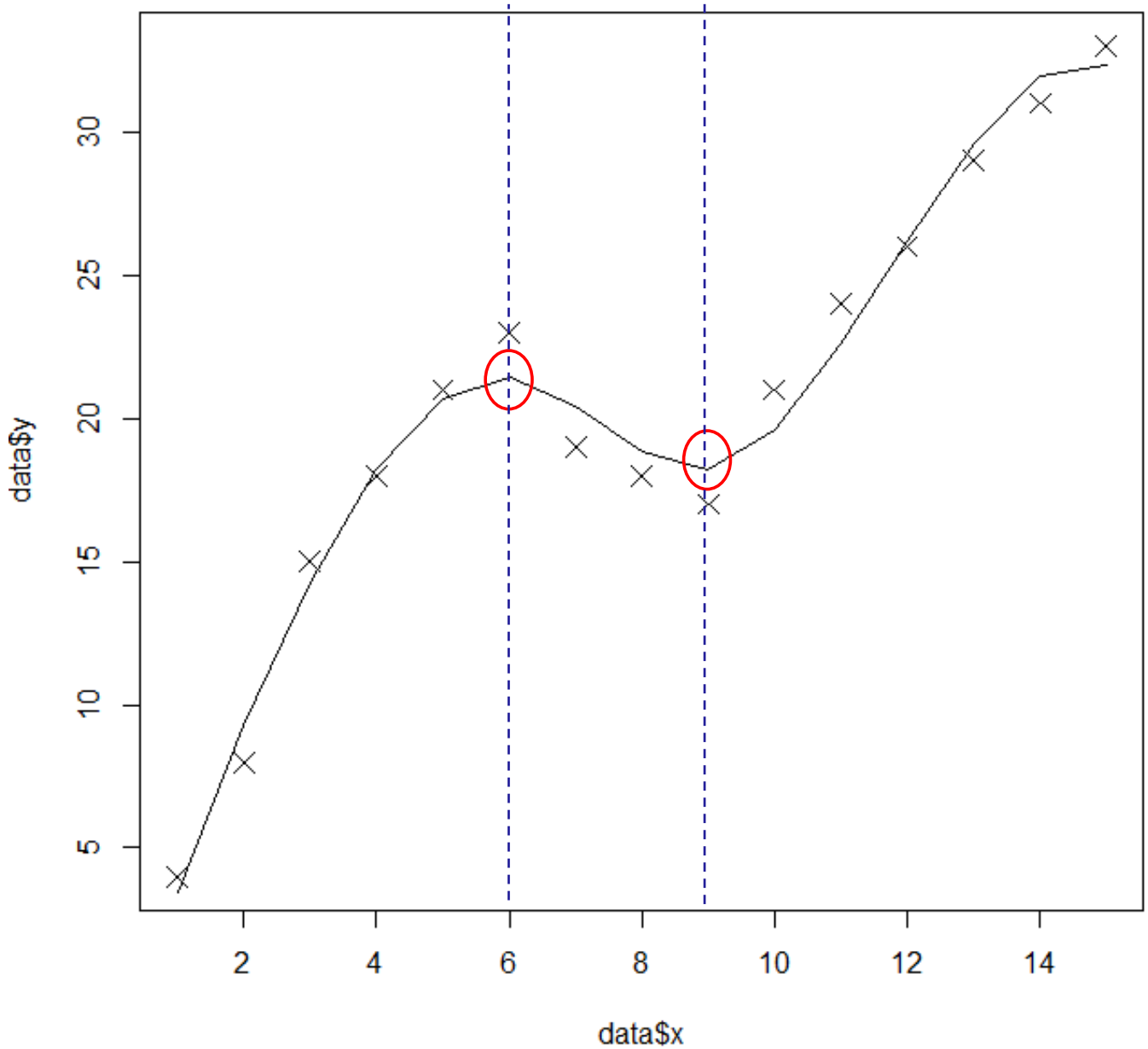
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 f(x, k_1) + \beta_4 f(x, k_2) + \epsilon$$

where for example

$$f(x, k_i) = max(0, x - k_i)^2$$

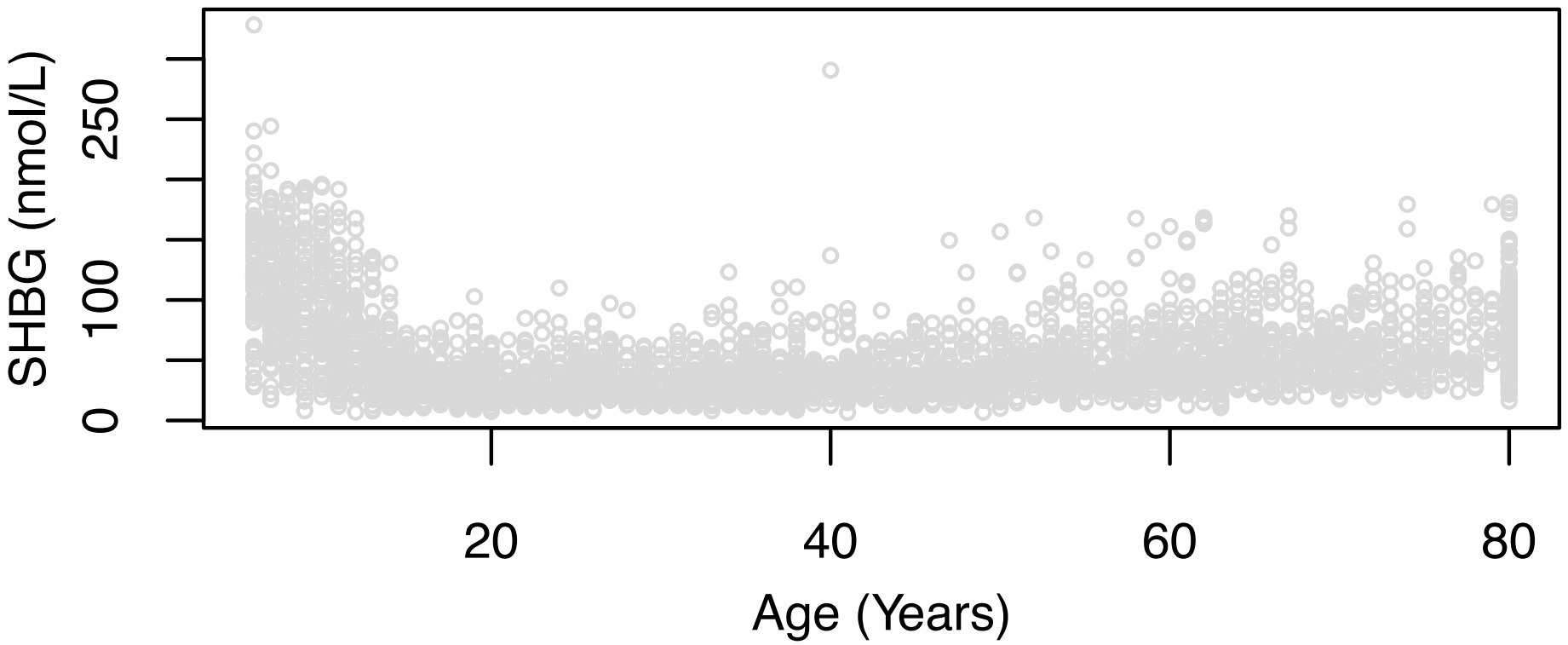
  - β_1 captures the overall linear trend.
  - β_2 captures the overall quadratic trend.
  - β_3 (β_4) captures additional curvature after k_1 (k_2).

Spline regression: quadratic spline



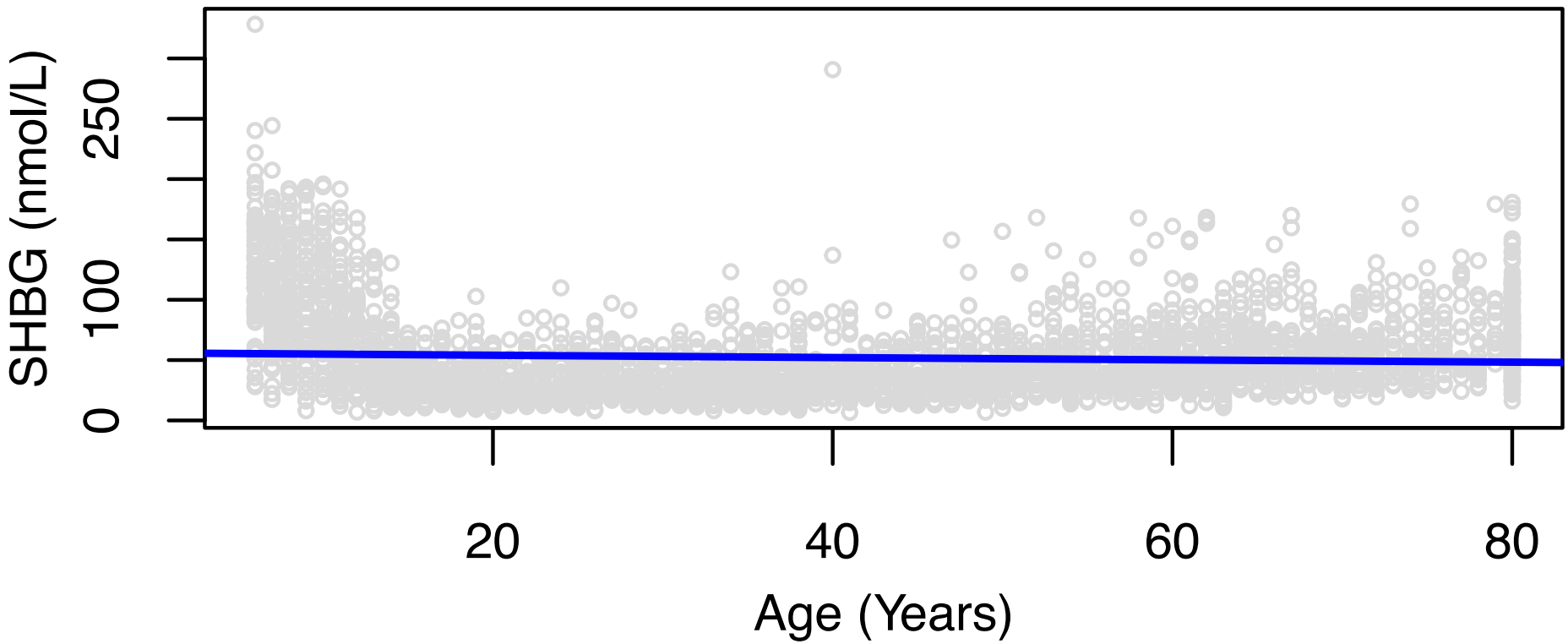
Polynomial regression versus spline

Data points Linear Higher degree polynomials Spline



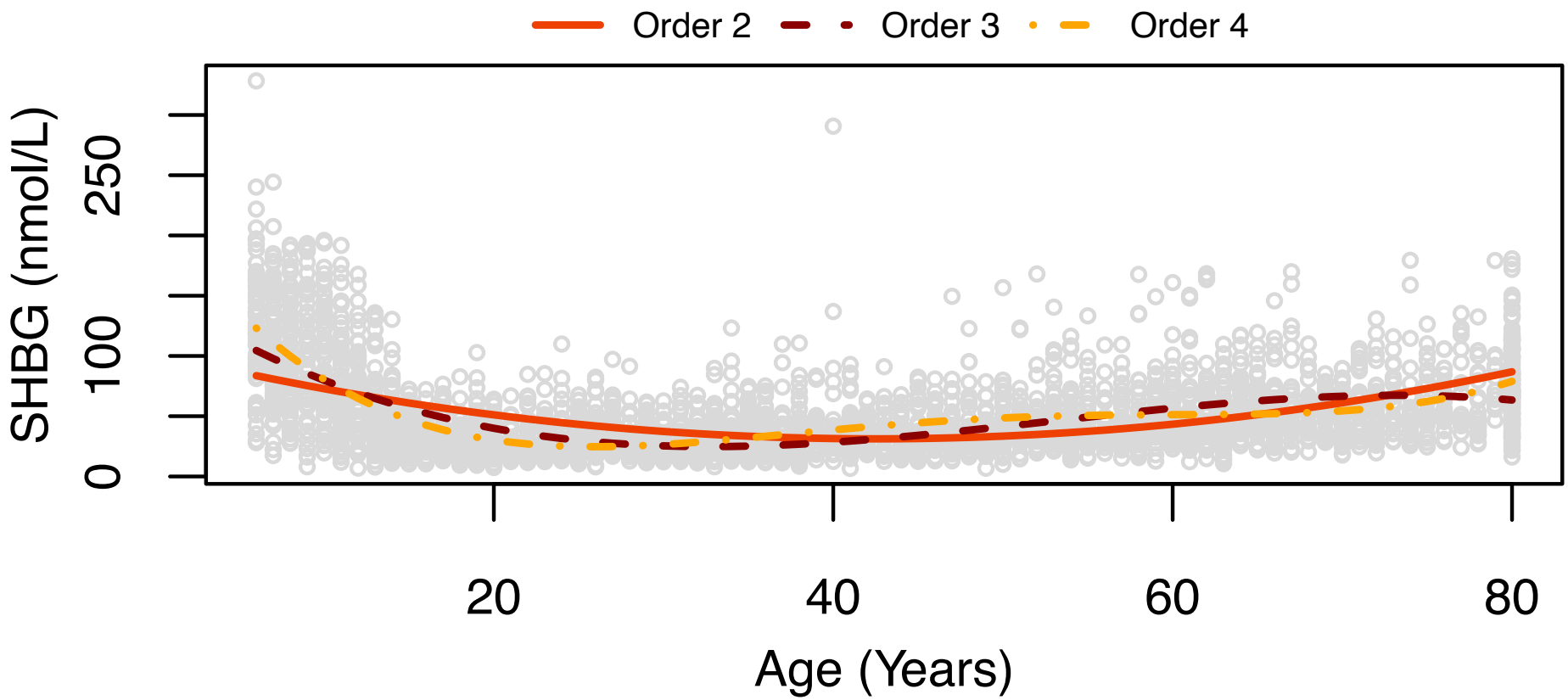
Polynomial regression versus spline

Data points Linear Higher degree polynomials Spline



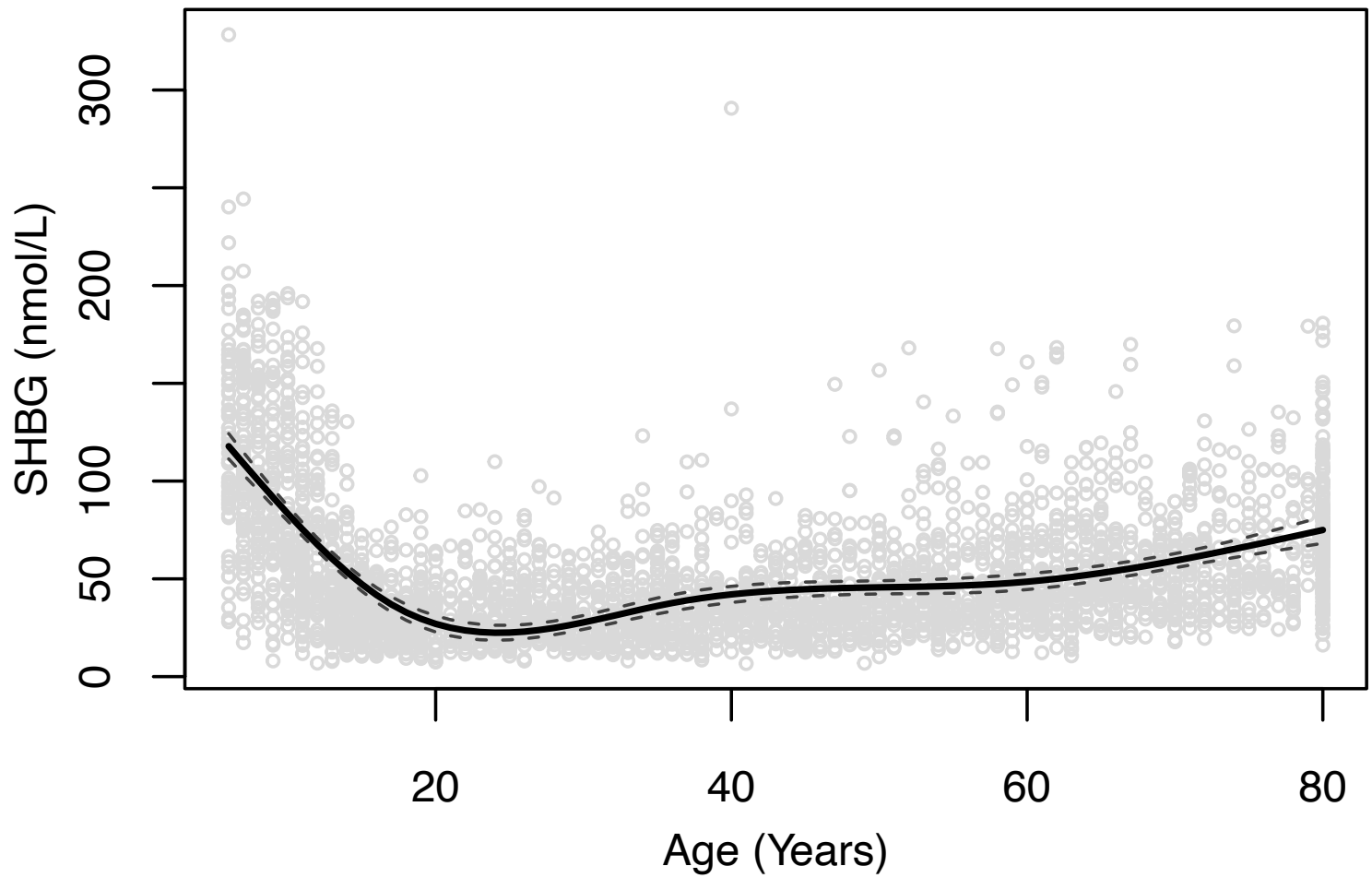
Polynomial regression versus spline

Data points Linear Higher degree polynomials Spline



Polynomial regression versus spline

Data points Linear Higher degree polynomials Spline



Topics

1. **Linear model with cross section data**
   - Identification and inference using the linear model
   - Identification versus prediction
2. **Violation of orthogonality**
   - **IV Estimation of a General Equation**
   - Two Stage Least Squares

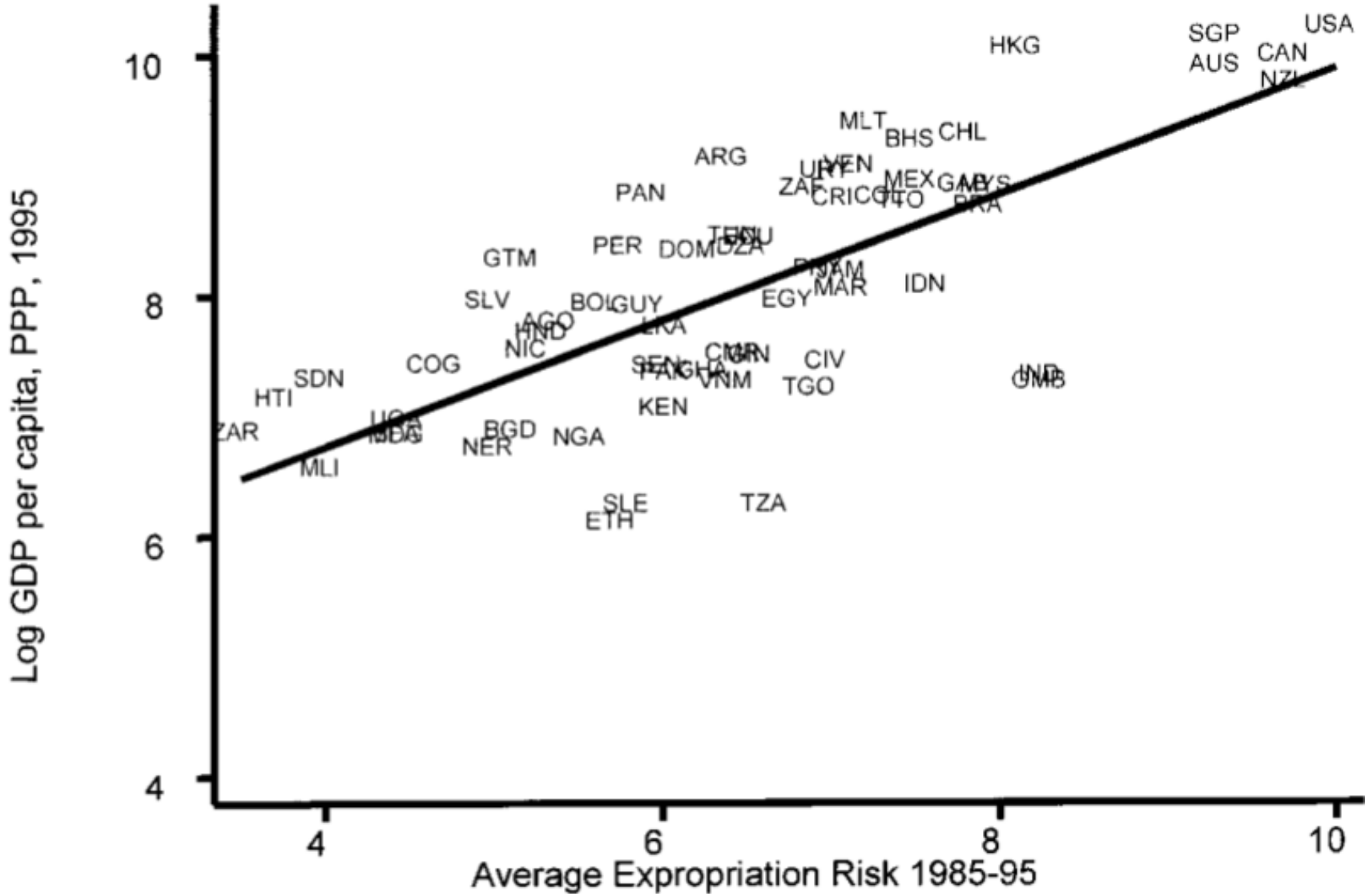
APPLICATION: Effect of institutions on economic performance (Acemoglu et al. AER 2001)

Interested in studying the causal effect of institutions on GDP:

$$GDP_i = \beta_0 + \beta_1 Institutions_i + X\gamma + \epsilon_i$$

- Institutions are the *humanly devised constraints* that structure political, economic and social interaction (North JEP 1991)
  - **informal constraints**: sanctions, taboos and codes of conduct, customs and traditions
  - **formal rules**: constitutions and laws, property rights
- Countries with better *institutions* will
  - invest more in physical and human capital + more efficient use
- **Correlation or causality?**

GDP per capita and current institutions



GDP per capita and current institutions

TABLE 2—OLS REGRESSIONS

	Whole world (1)	Base sample (2)	Whole world (3)	Whole world (4)	Base sample (5)	Base sample (6)	Whole world (7)	Base sample (8)
	Dependent variable is log GDP per capita in 1995						Dependent variable is log output per worker in 1988	
Average protection against expropriation risk, 1985–1995	0.54 (0.04)	0.52 (0.06)	0.47 (0.06)	0.43 (0.05)	0.47 (0.06)	0.41 (0.06)	0.45 (0.04)	0.46 (0.06)
Latitude			0.89 (0.49)	0.37 (0.51)	1.60 (0.70)	0.92 (0.63)		
Asia dummy				−0.62 (0.19)		−0.60 (0.23)		
Africa dummy				−1.00 (0.15)		−0.90 (0.17)		
"Other" continent dummy				−0.25 (0.20)		−0.04 (0.32)		
R^2	0.62	0.54	0.63	0.73	0.56	0.69	0.55	0.49
Number of observations	110	64	110	110	64	64	108	61

Notes: Dependent variable: columns (1)–(6), log GDP per capita (PPP basis) in 1995, current prices (from the World Bank's World Development Indicators 1999); columns (7)–(8), log output per worker in 1988 from Hall and Jones (1999). Average protection against expropriation risk is measured on a scale from 0 to 10, where a higher score means more protection against expropriation, averaged over 1985 to 1995, from Political Risk Services. Standard errors are in parentheses. In regressions with continent dummies, the dummy for America is omitted. See Appendix Table A1 for more detailed variable definitions and sources. Of the countries in our base sample, Hall and Jones do not report output per worker in the Bahamas, Ethiopia, and Vietnam.

- OLS estimates could be **biased upwards**
  1. rich countries can afford or prefer better institutions (*reverse causality*)
  2. unobserved determinants of both variables (*omitted variable bias*)

GDP per capita and current institutions

TABLE 2—OLS REGRESSIONS

	Whole world (1)	Base sample (2)	Whole world (3)	Whole world (4)	Base sample (5)	Base sample (6)	Whole world (7)	Base sample (8)
	Dependent variable is log GDP per capita in 1995						Dependent variable is log output per worker in 1988	
Average protection against expropriation risk, 1985–1995	0.54 (0.04)	0.52 (0.06)	0.47 (0.06)	0.43 (0.05)	0.47 (0.06)	0.41 (0.06)	0.45 (0.04)	0.46 (0.06)
Latitude			0.89 (0.49)	0.37 (0.51)	1.60 (0.70)	0.92 (0.63)		
Asia dummy				−0.62 (0.19)		−0.60 (0.23)		
Africa dummy				−1.00 (0.15)		−0.90 (0.17)		
"Other" continent dummy				−0.25 (0.20)		−0.04 (0.32)		
R^2	0.62	0.54	0.63	0.73	0.56	0.69	0.55	0.49
Number of observations	110	64	110	110	64	64	108	61

Notes: Dependent variable: columns (1)–(6), log GDP per capita (PPP basis) in 1995, current prices (from the World Bank's World Development Indicators 1999); columns (7)–(8), log output per worker in 1988 from Hall and Jones (1999). Average protection against expropriation risk is measured on a scale from 0 to 10, where a higher score means more protection against expropriation, averaged over 1985 to 1995, from Political Risk Services. Standard errors are in parentheses. In regressions with continent dummies, the dummy for America is omitted. See Appendix Table A1 for more detailed variable definitions and sources. Of the countries in our base sample, Hall and Jones do not report output per worker in the Bahamas, Ethiopia, and Vietnam.

- OLS estimates could be **biased upwards**
  1. rich countries can afford or prefer better institutions (*reverse causality*)
  2. unobserved determinants of both variables (*omitted variable bias*)

GDP per capita and current institutions

TABLE 2—OLS REGRESSIONS

	Whole world (1)	Base sample (2)	Whole world (3)	Whole world (4)	Base sample (5)	Base sample (6)	Whole world (7)	Base sample (8)
	Dependent variable is log GDP per capita in 1995						Dependent variable is log output per worker in 1988	
Average protection against expropriation risk, 1985–1995	0.54 (0.04)	0.52 (0.06)	0.47 (0.06)	0.43 (0.05)	0.47 (0.06)	0.41 (0.06)	0.45 (0.04)	0.46 (0.06)
Latitude			0.89 (0.49)	0.37 (0.51)	1.60 (0.70)	0.92 (0.63)		
Asia dummy				−0.62 (0.19)		−0.60 (0.23)		
Africa dummy				−1.00 (0.15)		−0.90 (0.17)		
"Other" continent dummy				−0.25 (0.20)		−0.04 (0.32)		
R^2	0.62	0.54	0.63	0.73	0.56	0.69	0.55	0.49
Number of observations	110	64	110	110	64	64	108	61

Notes: Dependent variable: columns (1)–(6), log GDP per capita (PPP basis) in 1995, current prices (from the World Bank's World Development Indicators 1999); columns (7)–(8), log output per worker in 1988 from Hall and Jones (1999). Average protection against expropriation risk is measured on a scale from 0 to 10, where a higher score means more protection against expropriation, averaged over 1985 to 1995, from Political Risk Services. Standard errors are in parentheses. In regressions with continent dummies, the dummy for America is omitted. See Appendix Table A1 for more detailed variable definitions and sources. Of the countries in our base sample, Hall and Jones do not report output per worker in the Bahamas, Ethiopia, and Vietnam.

- OLS estimates could be **biased upwards**
  1. rich countries can afford or prefer better institutions (*reverse causality*)
  2. unobserved determinants of both variables (*omitted variable bias*)

IV: identification when orthogonality fails

- **No such thing as an OLS or IV "model"**
  1. OLS and IV are different *estimation* methods that can be applied to the same model

- Population model is the **standard linear model**

$$y = \mathrm{x}\beta + u$$

  - x is $1 \times K$ and β is $K \times 1$

- **Instrumental variable**: $\mathrm{z} = (z_1, z_2, ..., z_L)$ be $1 \times L$ (call it the)
  - Suppose for the moment that $L = K$

Identification and exclusion restriction (IV.1)

- Follow the usual steps for identification in linear models

$$\begin{aligned} y &= \mathrm{x}\beta + u \\ \mathrm{z}'y &= \mathrm{z}'\mathrm{x}\beta + \mathrm{z}'u \\ E(\mathrm{z}'y) &= E(\mathrm{z}'\mathrm{x})\beta + E(\mathrm{z}'u) \end{aligned}$$

Exclusion restriction or **instrument exogeneity**

$$E(\mathrm{z}'u) = 0$$

- Same approach of orthogonality in OLS
- Used to cancel out the error term using expected values

Interpretation of exclusion restriction

- z contains **all exogenous elements** of x
  - Example: x_K is possibly endogenous and z_1 as an IV for x_K, then

$$\begin{aligned} \mathrm{x} &= (1, x_2, ..., x_{K-1}, x_K) \\ \mathrm{z} &= (1, x_2, ..., x_{K-1}, z_1) \end{aligned}$$

  - x_K is **excluded** from z
    - If one or more elements of x is correlated with u, z must contain some outside variables
    - The number of these outside variables (instruments) is at least equal to the number of endogenous variables

Rank condition (IV.2)

- Using the exclusion restriction we can write:

$$\begin{aligned} E(z'y) &= E(z'x)\beta \\ \beta &= [E(z'x)]^{-1}E(z'y) \end{aligned} \tag{7}$$

- β is identified if we assume a rank condition for $E(z'x)$

Rank condition or instrument relevance

$$rank\ E(z'x) = K$$

- **Note that IV.1 and IV.2 extends the conditions for OLS**
  - OLS is just a special IV case where $z = x$
  - All results on consistency and inference follows

Testing for the rank condition

Assume we have a single endogenous explanatory variable, x_K

1. Write the **reduced form** of x_K as

$$x_K = \delta_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + r_k \tag{8}$$

   - This is often called the **first-stage regression**
   - By definition, we require orthogonality to be valid

2. Rank condition holds if and only if

$$\theta_1 \neq 0 \tag{9}$$

   - Rule of thumb: F statistic > 10

Testing for the rank condition

Assume we have a single endogenous explanatory variable, x_K

1. Write the **reduced form** of x_K as

$$x_K = \delta_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + r_k \tag{8}$$

   - This is often called the **first-stage regression**
   - By definition, we require orthogonality to be valid

2. Rank condition holds if and only if

$$\theta_1 \neq 0 \tag{9}$$

   - Rule of thumb: F statistic > 10

APPLICATION: Effect of institutions on economic performance – IV strategy

$$GDP_i = \beta_0 + \beta_1 Institutions_i + X\gamma + \epsilon_i$$

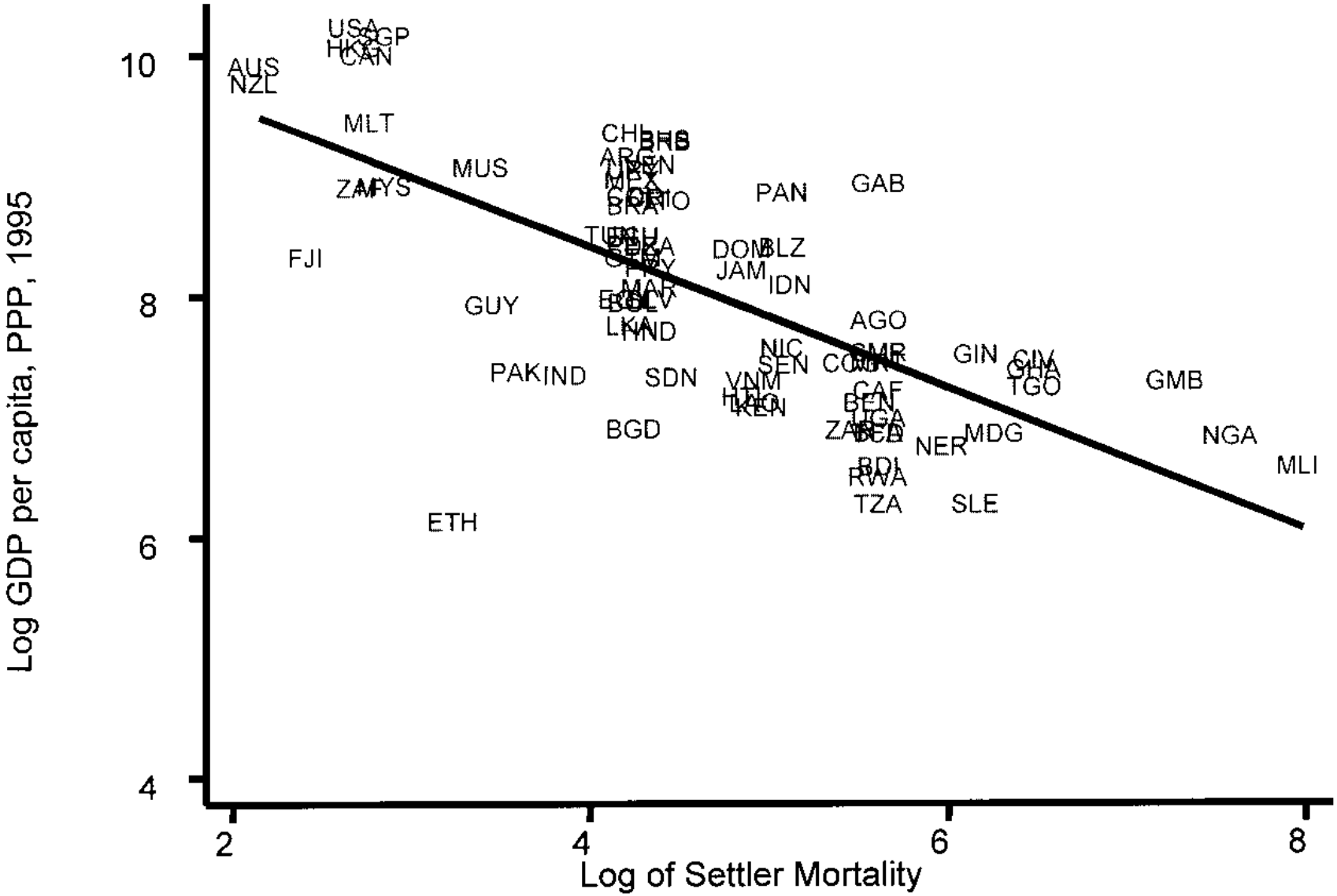
- **IV for** *Institutions*$_i$: mortality rates faced by the settlers during colonization as an instrument
  - Valid if mortality rates have no effect on income today other than through their influence on institutional development.

(potential) settler mortality ⇒ settlements

⇒ early institutions ⇒ current institutions

⇒ current performance.

Income and settler mortality (reduced-form)



The effect of institutions on GDP per capita

TABLE 4—IV REGRESSIONS OF LOG GDP PER CAPITA

	Base sample (1)	Base sample (2)	Base sample without Neo-Europes (3)	Base sample without Neo-Europes (4)	Base sample without Africa (5)	Base sample without Africa (6)	Base sample with continent dummies (7)	Base sample with continent dummies (8)	Base sample, dependent variable is log output per worker (9)
Average protection against expropriation risk 1985–1995	0.94 (0.16)	1.00 (0.22)	1.28 (0.36)	1.21 (0.35)	0.58 (0.10)	0.58 (0.12)	0.98 (0.30)	1.10 (0.46)	0.98 (0.17)
Latitude		−0.65 (1.34)		0.94 (1.46)		0.04 (0.84)		−1.20 (1.8)	
Asia dummy							−0.92 (0.40)	−1.10 (0.52)	
Africa dummy							−0.46 (0.36)	−0.44 (0.42)	
"Other" continent dummy							−0.94 (0.85)	−0.99 (1.0)	

- IV estimate is precisely estimated and large
  - Compare with OLS to understand bias (0.54 for base sample)

Topics

1. **Linear model with cross section data**
   - Identification and inference using the linear model
   - Identification versus prediction
2. **Violation of orthogonality**
   - IV Estimation of a General Equation
   - **Two Stage Least Squares**

Can we use multiple instruments?

In some cases, we have **more instruments than we need**

- Example: if we can use mother's education as an IV, why not father's education?

Identification versus **overidentification**:

1. When $L > K \Rightarrow$ model is potentially **overidentified**
2. When $L = K \Rightarrow$ model is **just identified**

When $L > K$, we can apply a flexible IV estimator, called **Two Stage Least Squares (2SLS)**

Two Stage Least Squares (2SLS)

$$\begin{aligned} y &= \mathrm{x}\beta + u \\ E(\mathrm{z}'u) &= 0 \\ L = \dim(\mathrm{z}) &\geq \dim(\mathrm{x}) = K \end{aligned}$$

- The best vector of instruments is the vector of linear projections of each element of x on z

$$\underset{1\times K}{\mathrm{x}} = \underset{1\times L}{\mathrm{z}} \cdot \underset{L\times K}{\Pi} + \underset{1\times K}{\mathrm{r}}$$

$$E(\mathrm{z}'\mathrm{r}) = 0$$

- **Intuition behind 2SLS**: if you have more than one instrument per endogenous variable, you can always recreate a single instrument using linear combinations of multiple instruments!

2SLS in practice: procedure

1. **First stage**
   - Run the regression X on Z to obtain $\hat{\Pi} = (Z'Z)^{-1}Z'X$.
   - Notice that exogenous variables act as their own instrument
   - Obtain the vector fitted values

$$\begin{aligned} \hat{X} &= Z\hat{\Pi} \\ &= Z(Z'Z)^{-1}Z'X \end{aligned}$$

2. **Second stage**: run the regression of y on fitted values $\hat{x}$

$$\hat{\beta}_{2SLS} = \left(N^{-1} \sum_{i=1}^{N} \hat{x}_i' \hat{x}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \hat{x}_i' y_i \right) = (\hat{X}'\hat{X})^{-1}\hat{X}'Y$$

Intepretation of 2SLS assumptions

- **IV.1 – Exogenous Instruments**

$$E(\hat{x}'u) = E(z'u) = 0$$

- **IV.2 – Rank Condition**:

  a) *rank* $E(z'z) = L$: rules out perfect collinearity among the exogenous variables (first stage)

  b) *rank* $E(\hat{x}'\hat{x}) = K$: it corresponds to *rank* $E(z'x) = K$ (second stage)

  c) $L \geq K$: I need <u>at least one</u> instrument for each endogenous variable

APPLICATION: Returns to schooling

Estimating the return to schooling via simple regression:

$$\log(wage_i) = \beta_0 + \beta_1 educ_i + u_i$$

- **OLS estimate biased**: $educ_i$ is clearly endogenous (orthogonality)
- Suggestions for z - are they valid as IV?
  - mother's education
  - number of siblings
  - distance to the nearest college at age 16
  - z is a randomly assigned education grant during high school

APPLICATION: Returns to schooling

Estimating the return to schooling via simple regression:

$$\log(wage_i) = \beta_0 + \beta_1 educ_i + u_i$$

- **OLS estimate biased**: $educ_i$ is clearly endogenous (orthogonality)
- **Suggestions for** z - are they valid as IV?
  - mother's education
  - number of siblings
  - distance to the nearest college at age 16
  - z is a randomly assigned education grant during high school

APPLICATION: Returns to schooling

Angrist and Krueger (QJE 1991) propose z binary, $z = 1$ if born in first quarter of year

- Individuals born in the beginning of the year start school at older age
- $\rightarrow$ can drop out after completing less years of schooling than individuals born near the end of the year

- What do I need for the **validity of this instrument**?
  1. **Relevance**: the quarter of birth is correlated with years of education - true?
  2. **Exclusion restriction**: Quarter of birth is not correlated with unobserved determinants of wage - true?

APPLICATION: Returns to schooling

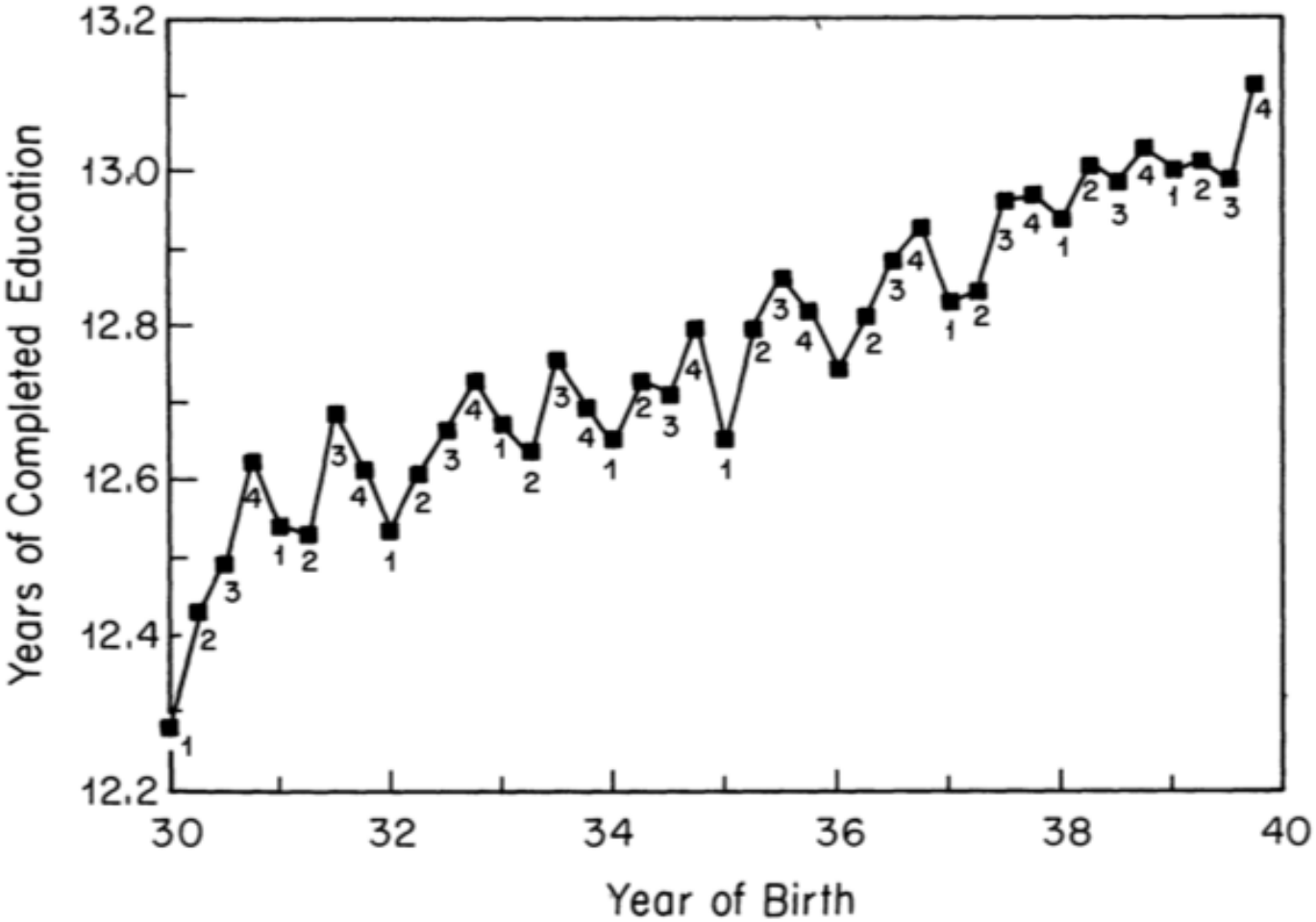


FIGURE I
Years of Education and Season of Birth
1980 Census
Note. Quarter of birth is listed below each observation.

APPLICATION: Returns to schooling

If identification strategy for IV is valid, compare OLS and 2SLS estimates to understand bias in OLS estimates

TABLE IV

OLS AND TSLS ESTIMATES OF THE RETURN TO EDUCATION FOR MEN BORN 1920–1929: 1970 CENSUS[a]

Independent variable	(1) OLS	(2) TSLS	(3) OLS	(4) TSLS	(5) OLS	(6) TSLS	(7) OLS	(8) TSLS
Years of education	0.0802	0.0769	0.0802	0.1310	0.0701	0.0669	0.0701	0.1007
	(0.0004)	(0.0150)	(0.0004)	(0.0334)	(0.0004)	(0.0151)	(0.0004)	(0.0334)
Race (1 = black)	—	—	—	—	0.2980	−0.3055	−0.2980	−0.2271
					(0.0043)	(0.0353)	(0.0043)	(0.0776)
SMSA (1 = center city)	—	—	—	—	0.1343	0.1362	0.1343	0.1163
					(0.0026)	(0.0092)	(0.0026)	(0.0198)
Married (1 = married)	—	—	—	—	0.2928	0.2941	0.2928	0.2804
					(0.0037)	(0.0072)	(0.0037)	(0.0141)
9 Year-of-birth dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
8 Region of residence dummies	No	No	No	No	Yes	Yes	Yes	Yes
Age	—	—	0.1446	0.1409	—	—	0.1162	0.1170
			(0.0676)	(0.0704)			(0.0652)	(0.0662)
Age-squared	—	—	−0.0015	−0.0014	—	—	−0.0013	−0.0012
			(0.0007)	(0.0008)			(0.0007)	(0.0007)
χ^2 [dof]	—	36.0 [29]	—	25.6 [27]	—	34.2 [29]	—	28.8 [27]

a. Standard errors are in parentheses. Sample size is 247,199. Instruments are a full set of quarter-of-birth times year-of-birth interactions. The sample consists of males born in the United States. The sample is drawn from the State, County, and Neighborhoods 1 percent samples of the 1970 Census (15 percent form). The dependent variable is the log of weekly earnings. Age and age-squared are measured in quarters of years. Each equation also includes an intercept.

Weak instrument

You find an instrumental variable, z, presenting these two properties:

1. **Exclusion restriction** is valid

$$Cov(z, \epsilon) = 0$$

2. **Relevance** is valid

$$Cov(z, x) \neq 0$$

but covariance is small (**z is a weak instrument**)

Intuition behind weak instruments

Identification of β_1 in a simple case: $y = \beta_0 + \beta_1 x + \epsilon$

- Apply the covariance operator

$$Cov(z, y) = \beta_1 Cov(z, x) + Cov(z, \epsilon)$$

- Identify β_1 using exogeneity assumption

$$\beta_1 = \frac{Cov(z, y)}{Cov(z, x)}$$

- β_1 **is identified!**

IV and OLS comparison: weak instruments

We can derive the estimators for OLS and IV:

$$\text{plim } \hat{\beta}_{1,IV} = \beta_1 + \frac{\sigma_\epsilon}{\sigma_x} \cdot \frac{Corr(z, \epsilon)}{Corr(z, x)}$$

$$\text{plim } \hat{\beta}_{1,OLS} = \beta_1 + \frac{\sigma_\epsilon}{\sigma_x} \cdot Corr(x, \epsilon)$$

- If z is a **weak instrument**, IV can produce a larger asymptotic bias than OLS
  - Common to see IV estimates larger in magnitude than OLS estimates
  - Weak instruments lead to large asymptotic standard errors

PROOF: IV and OLS comparison

- Replacing the population covariances with the sample covariances

$$\begin{aligned}\hat{\beta}_{1,IV} &= \frac{N^{-1}\sum_{i=1}^{N}(z_i-\bar{z})(y_i-\bar{y})}{N^{-1}\sum_{i=1}^{N}(z_i-\bar{z})(d_i-\bar{d})} \\ &= \beta_1 + \frac{N^{-1}\sum_{i=1}^{N}(z_i-\bar{z})\epsilon_i}{N^{-1}\sum_{i=1}^{N}(z_i-\bar{z})(d_i-\bar{d})}\end{aligned}$$

- Compare with OLS estimator

$$\hat{\beta}_{1,OLS} = \beta_1 + \frac{N^{-1}\sum_{i=1}^{N}(d_i-\bar{d})\epsilon_i}{N^{-1}\sum_{i=1}^{N}(d_i-\bar{d})^2}$$

- Then use the rule $Cov(d,\epsilon) = Corr(d,\epsilon)\sigma_d\sigma_\epsilon$

APPLICATION: Angrist and Evans (AER, 1998)

Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size

By JOSHUA D. ANGRIST AND WILLIAM N. EVANS*

Research on the labor-supply consequences of childbearing is complicated by the endogeneity of fertility. This study uses parental preferences for a mixed sibling-sex composition to construct instrumental variables (IV) estimates of the effect of childbearing on labor supply. IV estimates for women are significant but smaller than ordinary least-squares estimates. The IV are also smaller for more educated women and show no impact of family size on husbands' labor supply. A comparison of estimates using sibling-sex composition and twins instruments implies that the impact of a third child disappears when the child reaches age 13. (*JEL* J13, J22)

Data

Data are a subset from Angrist and Evans (AER, 1998), LABSUP.DTA.

```
. use labsup.dta

. * Women are black or Hispanic (possibly both).

. des hours nonmomi kids educ age black hispan samesex

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
hours           byte   %8.0g                  hours of work per week, mom
nonmomi         float  %9.0g                  'non-mom' income, $1000s
kids            byte   %8.0g                  number of kids
educ            byte   %8.0g                  mom's years of education
age             byte   %8.0g                  age of mom
black           byte   %8.0g                  =1 of black
hispan          byte   %8.0g                  =1 if hispanic
samesex         byte   %8.0g                  first two kids are of same sex
```

Summarize variables

```
. sum hours nonmomi kids educ age black hispan

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       hours |     31857    21.22011    19.49892           0         99
     nonmomi |     31857     31.7618    20.41241   -39.93675    157.438
        kids |     31857    2.752237    .9771916           2         12
        educ |     31857    11.00534    3.305196           0         20
         age |     31857    29.74175    3.613745          21         35
-------------+--------------------------------------------------------
       black |     31857    .4129705    .4923753           0          1
      hispan |     31857     .593182    .4912481           0          1

. count if hours == 0
13068

. count if hours == 40
11245
```

Tabulate number of kids

```
. tab kids

  number of |
       kids |      Freq.     Percent        Cum.
------------+-----------------------------------
          2 |     16,215       50.90       50.90
          3 |     10,014       31.43       82.33
          4 |      3,736       11.73       94.06
          5 |      1,374        4.31       98.37
          6 |        323        1.01       99.39
          7 |        134        0.42       99.81
          8 |         47        0.15       99.96
          9 |          6        0.02       99.97
         10 |          4        0.01       99.99
         11 |          2        0.01       99.99
         12 |          2        0.01      100.00
------------+-----------------------------------
      Total |     31,857      100.00
```

Gender of the first two kids

```
. tab samesex

  first two |
kids are of |
   same sex |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |     15,840       49.72       49.72
          1 |     16,017       50.28      100.00
------------+-----------------------------------
      Total |     31,857      100.00
```

- Not a surprise the distribution is around 50/50
  - Indicative of randomness
  - Not valid for all countries (i.e. sex-selective abortion)

OLS estimates

```
. reg hours kids nonmomi educ age agesq black hispan, robust

Linear regression                               Number of obs =   31857
                                                F(  7, 31849) =  377.87
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0727
                                                Root MSE      =  18.779

------------------------------------------------------------------------------
             |               Robust
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -2.325836   .1155164   -20.13   0.000    -2.552253   -2.099419
     nonmomi |  -.0578328   .0053515   -10.81   0.000     -.068322   -.0473436
        educ |   .5860083   .0374881    15.63   0.000     .5125302    .6594865
         age |   2.048793   .4483823     4.57   0.000     1.169946    2.927639
       agesq |  -.0277198   .0076957    -3.60   0.000    -.0428036    -.012636
       black |   1.058285    1.35088     0.78   0.433    -1.589492    3.706063
      hispan |  -5.114147    1.35152    -3.78   0.000    -7.763179   -2.465116
       _cons |  -10.44695   6.588891    -1.59   0.113    -23.36143    2.467528
------------------------------------------------------------------------------
```

- Each child beyond the first two reduces estimated hours by about 2.3 hours, other things fixed

Endogeneity: samesex as instrument

```
. reg kids samesex nonmomi educ age agesq black hispan, robust

Linear regression                               Number of obs =   31857
                                                F(  7, 31849) =  437.80
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1191
                                                Root MSE      =  .91724

------------------------------------------------------------------------------
             |               Robust
        kids |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     samesex |   .0703744   .0102783     6.85   0.000     .0502285    .0905202
     nonmomi |  -.0027871    .000257   -10.85   0.000    -.0032907   -.0022834
        educ |  -.0853676   .0020296   -42.06   0.000    -.0893457   -.0813895
         age |   .0589312   .0203278     2.90   0.004      .019088    .0987744
       agesq |   1.98e-06   .0003559     0.01   0.996    -.0006956    .0006995
       black |   .0128681   .0644422     0.20   0.842     -.113441    .1391772
      hispan |  -.0424722   .0644997    -0.66   0.510    -.1688941    .0839498
       _cons |   2.010258   .2930274     6.86   0.000     1.435913    2.584603
------------------------------------------------------------------------------
```

IV estimates

- Much bigger effect using IV, but only marginally statistically significant

```
. ivreg hours nonmomi educ age agesq black hispan (kids = samesex), robust

Instrumental variables (2SLS) regression              Number of obs =   31857
                                                      F(  7, 31849) =  304.81
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0583
                                                      Root MSE      =  18.924

------------------------------------------------------------------------------
             |               Robust
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -4.878903   3.013547    -1.62   0.105    -10.78557    1.027766
     nonmomi |  -.0649179   .0099359    -6.53   0.000    -.0843926   -.0454432
        educ |    .368042   .2595992     1.42   0.156    -.1407823    .8768664
         age |   2.200964   .4845126     4.54   0.000       1.2513    3.150627
       agesq |  -.0277443    .007744    -3.58   0.000     -.042923   -.0125657
       black |   1.094986   1.376742     0.80   0.426    -1.603482    3.793454
      hispan |  -5.217758   1.381364    -3.78   0.000    -7.925284   -2.510232
       _cons |  -5.253976   9.037541    -0.58   0.561     -22.9679    12.45995
------------------------------------------------------------------------------
Instrumented:  kids
Instruments:   nonmomi educ age agesq black hispan samesex
------------------------------------------------------------------------------
```

Weak instrument

- The partial correlation is even small

```
. corr kids samesex
(obs=31857)

             |     kids  samesex
-------------+------------------
        kids |   1.0000
     samesex |   0.0358   1.0000
```

- It's not surprising the IV estimate is much less precise than OLS
- A much larger sample size, as in Angrist and Evans, and another instrument – indicating a multiple second birth – help a lot with precision

Testing for endogeneity of x

Consider the standard linear model with IV

$$\begin{aligned} y &= \mathrm{x}\beta + u \\ E(\mathrm{z}'u) &= 0 \end{aligned}$$

- **How to test whether x is endogenous?**

$$H_0 : E(\mathrm{x}'u) = 0$$

- I can test for endogeneity only if I have an instrument!
  1. **Durbin-Wu-Hausman (DWH) test**
  2. **Regression-based Hausman test**

Testing for endogeneity of x

1. **Durbin-Wu-Hausman (DWH) test**
   - Takes the null $\hat{\beta}_{2SLS} - \hat{\beta}_{OLS} \xrightarrow{p} 0$ and provide limiting distribution
   - If all elements of x are exogenous then 2SLS and OLS should differ only due to sampling error.

2. **Regression-based Hausman test**
   1. Regress each endogenous variable in x_i on z_i to obtain a vector of residuals $\hat{v}_i$
   2. Run the regression y_i on $x_i, \hat{v}_i$ (**control function approach**)
   3. Test the coefficients in front of $\hat{v}_i$ are different from zero (if confirmed then x presents endogeneity)

Testing overidentifying restrictions

If more instruments than the number we need, we can test whether some of them are exogenous.

- From 2SLS: K exact moment conditions hold in the sample

$$N^{-1} \sum_{i=1}^{N} \hat{x}_i' \hat{u}_i = 0$$

- **Test for overidentifying restrictions** checks whether these conditions hold in the data
- Rejection indicates that one or more IVs fail the exogeneity requirement
  - We do not know which one though.

Labor supply application: OLS estimates

```
. reg hours kids nonmomi educ age agesq black hispan, robust

Linear regression                               Number of obs =   31857
                                                F(  7, 31849) =  377.87
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0727
                                                Root MSE      =  18.779

------------------------------------------------------------------------------
             |               Robust
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -2.325836   .1155164   -20.13   0.000    -2.552253   -2.099419
     nonmomi |  -.0578328   .0053515   -10.81   0.000     -.068322   -.0473436
        educ |   .5860083   .0374881    15.63   0.000     .5125302    .6594865
         age |   2.048793   .4483823     4.57   0.000     1.169946    2.927639
       agesq |  -.0277198   .0076957    -3.60   0.000    -.0428036    -.012636
       black |   1.058285    1.35088     0.78   0.433    -1.589492    3.706063
      hispan |  -5.114147    1.35152    -3.78   0.000    -7.763179   -2.465116
       _cons |  -10.44695   6.588891    -1.59   0.113    -23.36143    2.467528
------------------------------------------------------------------------------
```

Labor supply application: 2SLS estimates

```
. ivreg hours nonmomi educ age agesq black hispan (kids = samesex), robust

Instrumental variables (2SLS) regression          Number of obs =   31857
                                                  F(  7, 31849) =  304.81
                                                  Prob > F      =  0.0000
                                                  R-squared     =  0.0583
                                                  Root MSE      =  18.924

------------------------------------------------------------------------------
             |               Robust
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -4.878903   3.013547    -1.62   0.105    -10.78557    1.027766
     nonmomi |  -.0649179   .0099359    -6.53   0.000    -.0843926   -.0454432
        educ |    .368042   .2595992     1.42   0.156    -.1407823    .8768664
         age |   2.200964   .4845126     4.54   0.000       1.2513    3.150627
       agesq |  -.0277443    .007744    -3.58   0.000     -.042923   -.0125657
       black |   1.094986   1.376742     0.80   0.426    -1.603482    3.793454
      hispan |  -5.217758   1.381364    -3.78   0.000    -7.925284   -2.510232
       _cons |  -5.253976   9.037541    -0.58   0.561     -22.9679    12.45995
------------------------------------------------------------------------------
Instrumented:  kids
Instruments:   nonmomi educ age agesq black hispan samesex
------------------------------------------------------------------------------
```

Use two instruments in 2SLS

Use samesex and multi2nd (1 if second are twins) as IVs for kids.

- Estimate the **reduced form**

```
. reg kids samesex multi2nd nonmomi educ age agesq black hispan, robust

Linear regression                               Number of obs =   31857
                                                F(  8, 31848) =  410.77
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1244
                                                Root MSE      =  .91452

------------------------------------------------------------------------------
             |               Robust
        kids |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     samesex |     .07044   .0102481     6.87   0.000     .0503533    .0905267
    multi2nd |   .7632484   .0546856    13.96   0.000     .6560626    .8704342
     nonmomi |  -.0027879   .0002562   -10.88   0.000    -.0032901   -.0022858
        educ |  -.0853114   .0020267   -42.09   0.000    -.0892838   -.0813391
         age |   .0563395    .020282     2.78   0.005      .016586    .0960929
       agesq |   .0000436   .0003551     0.12   0.902    -.0006524    .0007396
       black |   .0105681   .0645589     0.16   0.870    -.1159698    .1371059
      hispan |  -.0420447   .0646128    -0.65   0.515    -.1686882    .0845988
       _cons |   2.043467   .2924263     6.99   0.000       1.4703    2.616634
------------------------------------------------------------------------------
```

Test relevance of instruments

```
. test samesex multi2nd

 ( 1)  samesex = 0
 ( 2)  multi2nd = 0

       F(  2, 31848) =  117.38
            Prob > F =    0.0000
```

- The two IV candidates are partially correlated with kids, both in the direction (positive) that we expect.
- Get the reduced form residuals: *predict v2h, resid*

Hausman regression-based test

Run regression of y on endogenous variables and fitted residuals (v2h)

```
. reg hours kids nonmomi educ age agesq black hispan v2h, robust

Linear regression                               Number of obs =   31857
                                                F(  8, 31848) =  330.79
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0727
                                                Root MSE      =  18.779

------------------------------------------------------------------------------
             |               Robust
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -2.986165   1.284302    -2.33   0.020    -5.503447   -.4688828
     nonmomi |  -.0596653   .0064263    -9.28   0.000     -.072261   -.0470696
        educ |   .5296332   .1154311     4.59   0.000     .3033839    .7558825
         age |    2.08815   .4545537     4.59   0.000     1.197208    2.979093
       agesq |  -.0277261   .0076958    -3.60   0.000    -.0428101   -.0126422
       black |   1.067778   1.350595     0.79   0.429     -1.57944    3.714995
      hispan |  -5.140945   1.352129    -3.80   0.000    -7.791169   -2.490721
         v2h |    .665256   1.290263     0.52   0.606     -1.86371    3.194222
       _cons |  -9.103833   7.093029    -1.28   0.199    -23.00644    4.798776
------------------------------------------------------------------------------
```

- Test statistic for v2h is about .52: little evidence of endogeity of kids.

Careful with s.e. $\Rightarrow$ 2SLS estimates

Compute the 2SLS estimates to have correct s.e.

```
. ivreg hours nonmomi educ age agesq black hispan (kids = samesex multi2nd),
        robust

Instrumental variables (2SLS) regression                Number of obs =   31857
                                                        F(  7, 31849) =  310.81
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.0717
                                                        Root MSE      =  18.789

------------------------------------------------------------------------------
             |               Robust
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -2.986165    1.28219    -2.33   0.020    -5.499307    -.473022
     nonmomi |  -.0596653   .0064235    -9.29   0.000    -.0722555   -.0470751
        educ |   .5296332   .1152961     4.59   0.000     .3036484     .755618
         age |    2.08815   .4545798     4.59   0.000     1.197156    2.979144
       agesq |  -.0277261   .0076979    -3.60   0.000    -.0428143    -.012638
       black |   1.067778   1.355563     0.79   0.431    -1.589178    3.724733
      hispan |  -5.140945   1.357096    -3.79   0.000    -7.800906   -2.480985
       _cons |  -9.103834   7.092956    -1.28   0.199     -23.0063    4.798632
------------------------------------------------------------------------------
Instrumented:  kids
Instruments:   nonmomi educ age agesq black hispan samesex multi2nd
------------------------------------------------------------------------------
```

Careful with s.e. $\Rightarrow$ 2SLS estimates

- OLS makes use of the following:

$$\sqrt{N}(\hat{\beta}_{OLS} - \beta) = \left(N^{-1} \sum_{i=1}^{N} \mathrm{x}_i' \mathrm{x}_i \right)^{-1} \left(N^{-1/2} \sum_{i=1}^{N} \mathrm{x}_i' u_i \right)$$

- 2SLS makes use of the following:

$$\sqrt{N}(\hat{\beta}_{2SLS} - \beta) = \left(N^{-1} \sum_{i=1}^{N} \hat{\mathrm{x}}_i' \hat{\mathrm{x}}_i \right)^{-1} \left(N^{-1/2} \sum_{i=1}^{N} \hat{\mathrm{x}}_i' u_i \right)$$