Matching

A Tale of Two Colleges

Differences between a private and public college:

- **Tuition:** private colleges cost $20k/year more on average

- **School quality:** private colleges may have smaller classes, better teachers, smarter students etc.

- **Earnings potential:** wage premium for private college graduates

Does private college education increase future earnings?

Simple Public/Private Comparisons

Is a simple comparison of earnings between private and public college graduates *ceteris paribus*?

- Likely not *ceteris paribus.*
- Pre-treatment differences in earnings potential:
  - SAT scores, parental income, motivation etc.
  - These covariates may jointly determine school choice and future earnings
- Private and public school graduates are not comparable → selection bias

Matching

Suppose the only things that matter for future earnings are SAT score (as a proxy for ability) and school choice

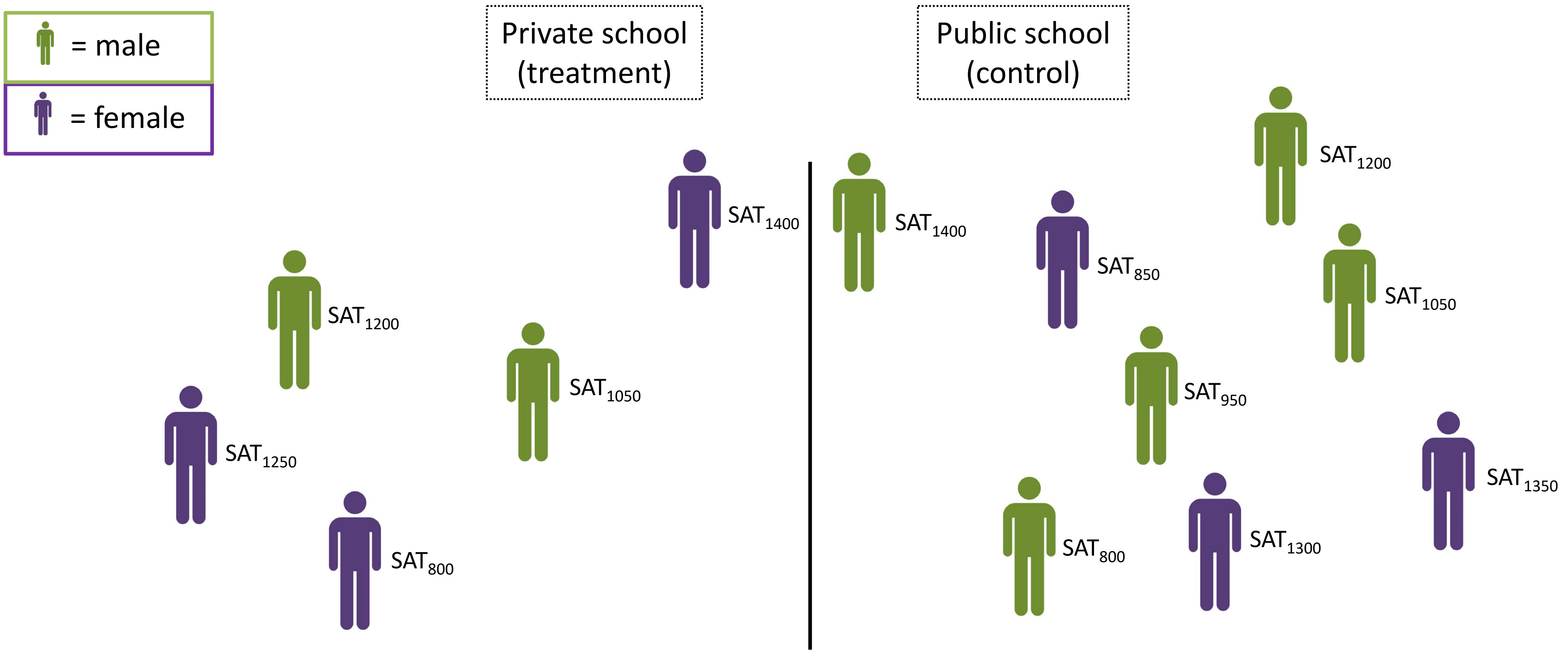
- For each private school alumni, identify a public school alumni with the same (or similar) SAT score → counterfactual for the missing potential outcome

- Compute average of the earnings differences within matched pairs → estimate of ATT

- Find matches for both private and public school alumni → estimate of ATE

Matching on a Single Covariate

Matching on Many Covariates



Matching

- Identify *observable* characteristics (i.e. covariates) that you think jointly determine treatment and outcome

- Take a treated unit and find a non-treated unit that has very similar covariate values → matched pairs

- Compute the average of the within matched-pair differences:

$$\hat{\tau}_{ATT} = \frac{1}{N_t} \sum_{i=1}^{N_t} (Y_i - Y_{j(i)})$$

  – where $Y_{j(i)}$ is the observed outcome of a control unit such that $X_{j(i)}$ is the closest value to X_i among all control units

Conditional Independence Assumption (CIA)

$$\left(Y_i^0, Y_i^1\right) \perp D_i | X_i$$

- Treatment is “as good as random” after controlling for covariates X_i
- No unmeasured confounders by assumption
- If CIA holds $\rightarrow$ selection on observables
- CIA is *not* testable; is the assumption plausible?

Common Support Assumption

$$0 < \Pr(D_i = 1|X_i) < 1$$

- For each value of X, there is a positive probability of being treated (and untreated)

- We need overlaps in covariates of treated and untreated units to find adequate matches

- Common support is testable

Exact Matching

- For each treated unit, find a control unit that has the exact same covariate values (i.e. $X_i = X_j$)
- If there are many control units with the same covariates, randomly select one of those control units to be the match
- Alternatively, take the average outcome of control units that are a perfect match
- Statistical programs will do the matches for you

Example – Exact Matching

Private

Student	SAT	Earnings
1	1400	105
2	1250	110
3	1200	100
4	1050	115
5	800	65
Average:		**99**
Matched avg:		**96.25**

Public

Student	SAT	Earnings
6	1400	100
7	1350	80
8	1300	85
9	1200	105
10	1050	95
11	950	70
12	850	65
13	800	60
		82.5

Matched Control

Student	SAT	Earnings
6	1400	100
---	---	---
9	1200	105
10	1050	95
13	800	60
		90

- ATT with exact matching: 96.25 – 90 = 6.25
- Simple difference in means: 99 – 82.5 = 16.5

Example – Approximate Matching

Student	SAT	Earnings
1	1400	105
2	1250	110
3	1200	100
4	1050	115
5	800	65

Student	SAT	Earnings
6	1400	100
7	1350	80
8	1300	85
9	1200	105
10	1050	95
11	950	70
12	850	65
13	800	60

Student	SAT	Earnings
6	1400	100
~~9~~	~~1200~~	~~105~~
9	1200	105
10	1050	95
13	800	60

Example – Approximate Matching

Student	SAT	Earnings
1	1400	105
2	1250	110
3	1200	100
4	1050	115
5	800	65
Average:		**99**

Student	SAT	Earnings
6	1400	100
7	1350	80
8	1300	85
9	1200	105
10	1050	95
11	950	70
12	850	65
13	800	60
		82.5

Student	SAT	Distance	Earnings
6	1400	0	100
9	1200	50	105
9	1200	0	105
10	1050	0	95
13	800	0	60
		10	**93**

- ATT with <u>approximate</u> matching: 99 – 93 = 6
- Simple difference in means: 99 – 82.5 = 16.5

Example – Approximate Matching with Averaging

Student	SAT	Earnings
1	1400	105
2	1250	110
3	1200	100
4	1050	115
5	800	65
Average:		**99**

Student	SAT	Earnings
6	1400	100
7	1350	80
8	1300	85
9	1200	105
10	1050	95
11	950	70
12	850	65
13	800	60
		82.5

Student	SAT	Distance	Earnings
6	1400	0	100
8/9	1250	0	95
9	1200	0	105
10	1050	0	95
13	800	0	60
		0	**91**

- ATT with approx. matching and averaging controls: 99 – 91 = 8
- Simple difference in means: 99 – 82.5 = 16.5

Nearest Neighbor Matching

For each treated unit, find a control unit that has the *closest* covariate values

To measure closeness, we need a *distance metric* which maps one or more covariate differences into a single number

- The *normalized Euclidean distance* scales each variable by the variable's variance → closeness is standardized across covariates

- The *Mahalanobis distance* additionally adjusts for the covariance in the data
  - If two covariates are highly correlated, their contribution to the distances should be smaller

Distance Metrics

Normalized Euclidean distance (for K covariates):

$$\left\|X_i - X_j\right\| = \sqrt{\sum_{k=1}^{K} \frac{\left(X_{ik} - X_{jk}\right)^2}{\hat{\sigma}_k^2}}$$

Mahalanobis distance:

$$\left\|X_i - X_j\right\| = \sqrt{\left(X_i - X_j\right)' \hat{\Sigma}_X^{-1} \left(X_i - X_j\right)}$$

- where $\hat{\Sigma}_X$ is the estimated variance-covariance matrix of X

Propensity Score Matching

Covariate matching may run into the “curse of dimensionality” problem

- It’s increasingly difficult to find adequate matches as the number of covariates on which we match increases

Propensity score matching compares units which, based on their observables, had similar probabilities (i.e. propensity scores) to get the treatment

- Criterion for a good match is based on a single variable
- Propensity score: $e_i(X_i) = Pr(D_i = 1|X_i)$

Propensity Score Matching

1. Estimate the propensity score using a probit or logit regression

2. Match each individual to the individual in the opposite treatment with the closest estimated value of the propensity score

3. Check that matched pairs have similar values of covariates X that you used to model the propensity score (checking for covariate balance)

4. Compute the differences between the treated and untreated individuals in each matched pair to get an estimate of the ATE

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Credits:
Alain Cohn
Assistant Professor of Information

Regression

Regression

Regression helps us to learn how two variables are related

Two ways to interpret regression:

1. Descriptive (correlation)
2. Causal

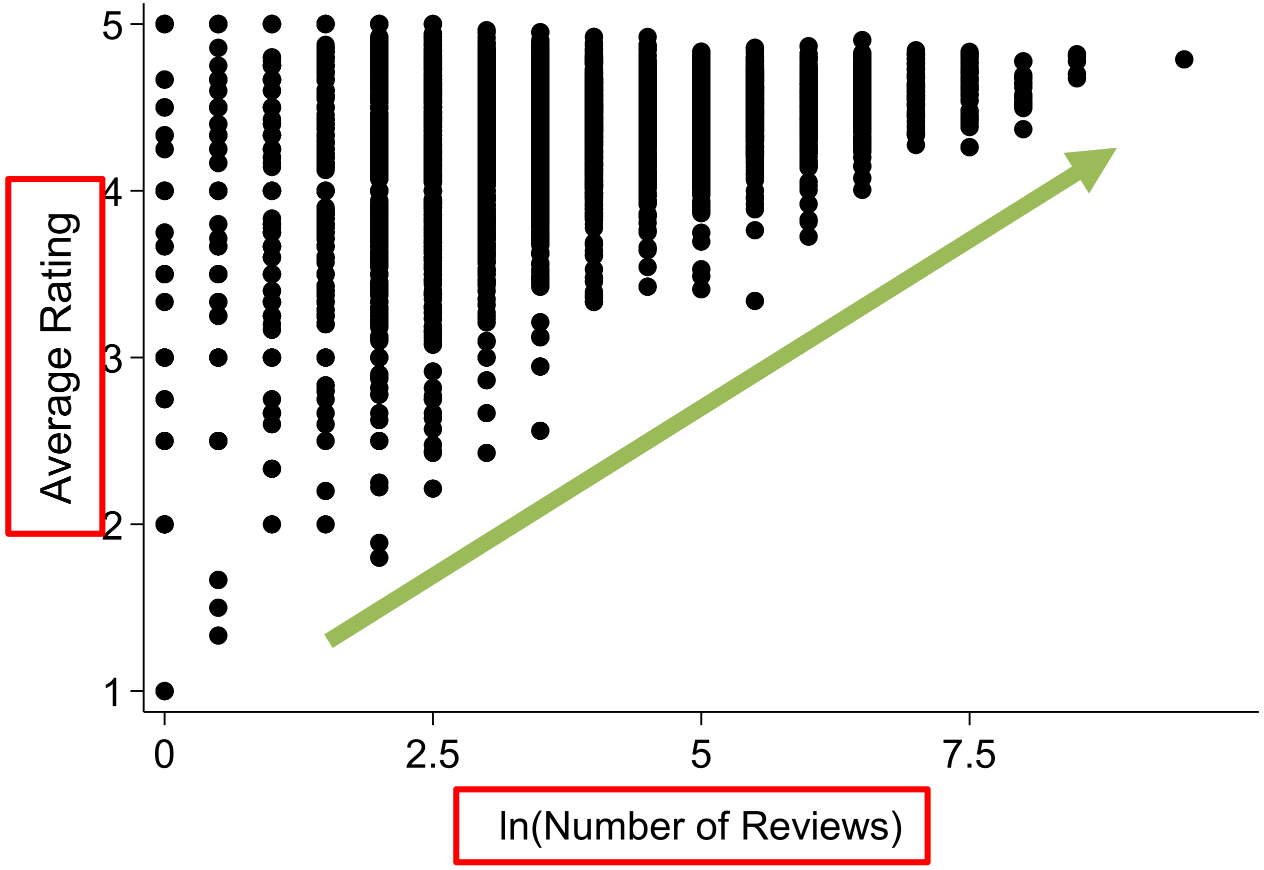
“Students with higher SAT scores earn higher wages later in life”

$$\neq$$

“SAT scores cause higher wages”

Population Relationship

- Population relationship between number of reviews and average rating of recipes on a cooking site
- Positive relationship between number of reviews and average rating
- We want to estimate the relationship between these two variables



Data from Timothy Lin (www.timlrx.com)

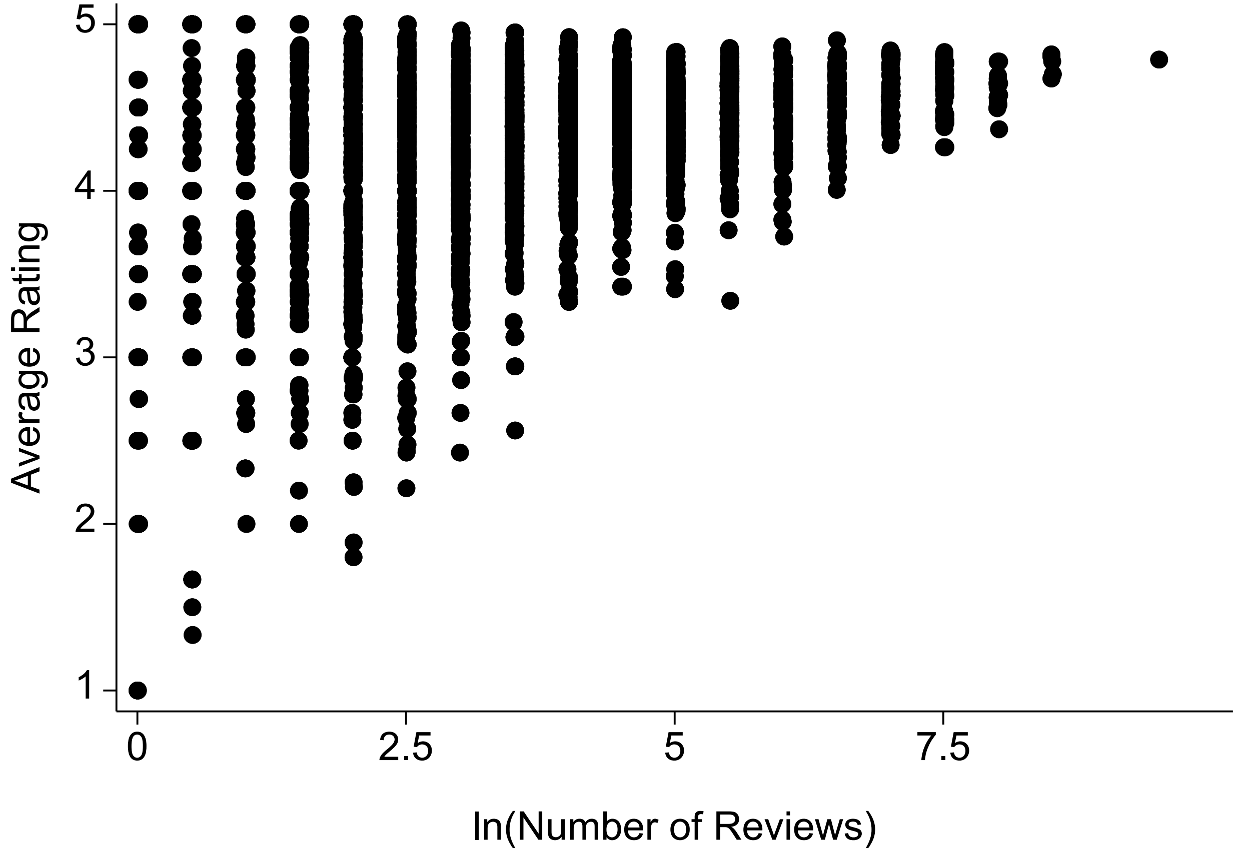
Conditional Expectation Function

Conditional expectation is the population average conditional on holding certain variables fixed:

$$E[Y_i|X_i = x]$$

Conditional expectation function (CEF) is the function that gives the mean of Y_i at various values of X_i:

$$\mu(X_i) = E[Y_i|X_i]$$



Properties of the CEF

The CEF is a useful tool to predict how Y_i changes as a function of X_i

We can split the outcome variable Y_i into two components, the CEF and an error u_i:

$$Y_i = E[Y_i|X_i] + u_i$$

- By definition: $E[u_i|X_i] = E[u_i] = 0$
- Y_i can be decomposed into the part "explained by X_i" and a part that is uncorrelated with X_i

The CEF is the function of X_i that best predicts (in a mean squared error sense) Y_i:

$$E[\left(Y_i - g(X_i)\right)^2] \geq E[\left(Y_i - \mu(X_i)\right)^2]$$

Regression and the CEF

Let's assume the CEF is linear:

$$\mu(X_i) = E[Y_i|X_i] = \alpha + \beta X_i$$

The intercept α is the conditional mean of Y_i if $X_i = 0$:

$$E[Y_i|X_i = 0] = \alpha$$

The slope β is the average change in the mean of Y_i for a one-unit increase in X_i:

$$E[Y_i|X_i = x + 1] - E[Y_i|X_i = x] = \beta$$

Regression = Best Linear Approximation of the CEF

The best linear function that approximates the CEF is given by:

$$(\alpha, \beta) = \underset{a,b}{\arg\min}\, E\left[\left(Y_i - (a + bX_i)\right)^2\right]$$

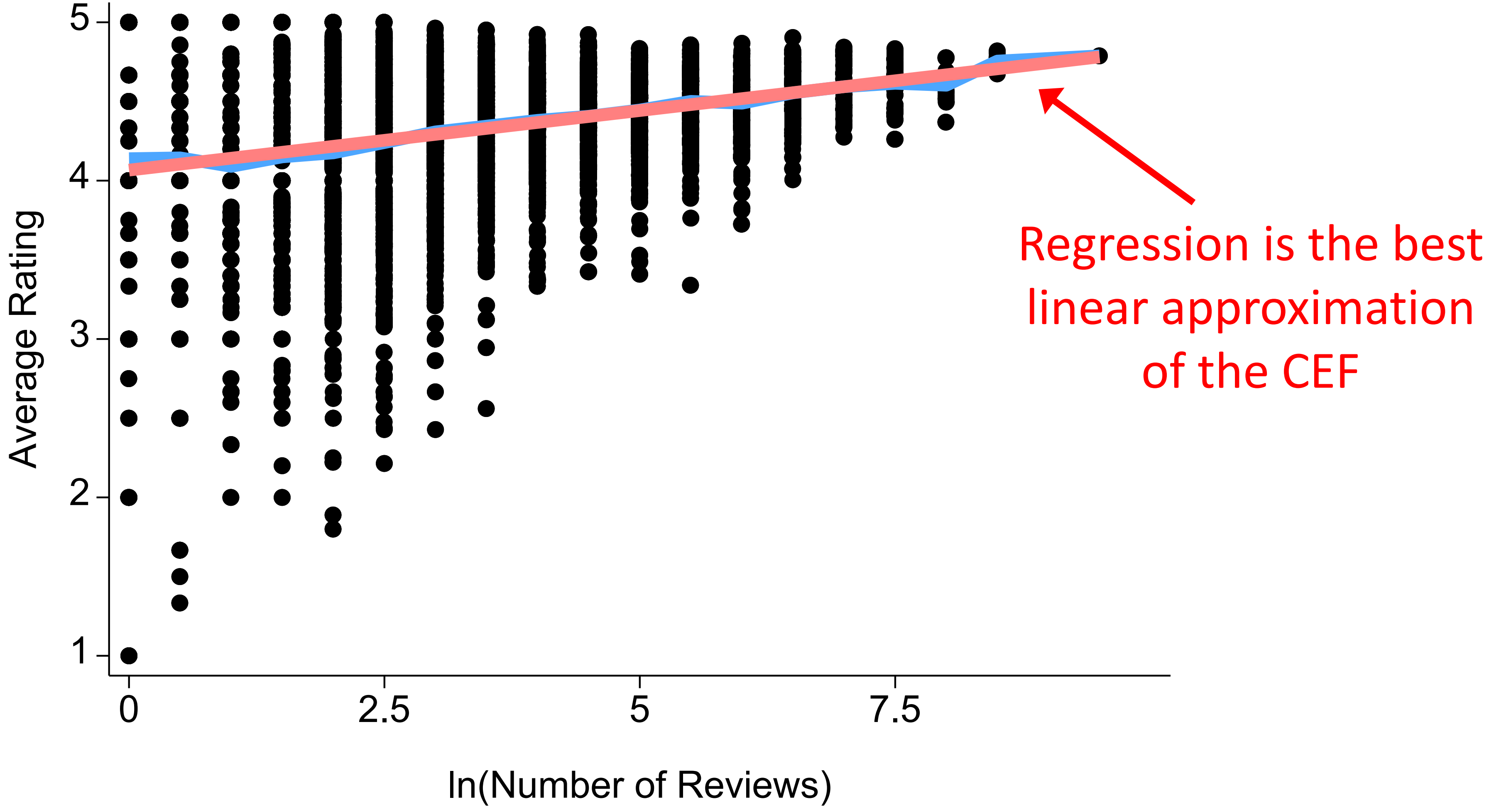
The solution is:

$$b = \beta = \frac{Cov[X_i, Y_i]}{V[X_i]} = \frac{E[(X_i - E[X_i])(Y_i - E[Y_i])]}{V[X_i]}$$

$$a = \alpha = E[Y_i] - \beta E[X_i]$$

The **population regression function** is the best linear approximation of the CEF even if the CEF is nonlinear

Population Regression Function



How To Estimate the Population Regression Line?

To get the sample line of best fit, we replace the population expectations with the sample versions:

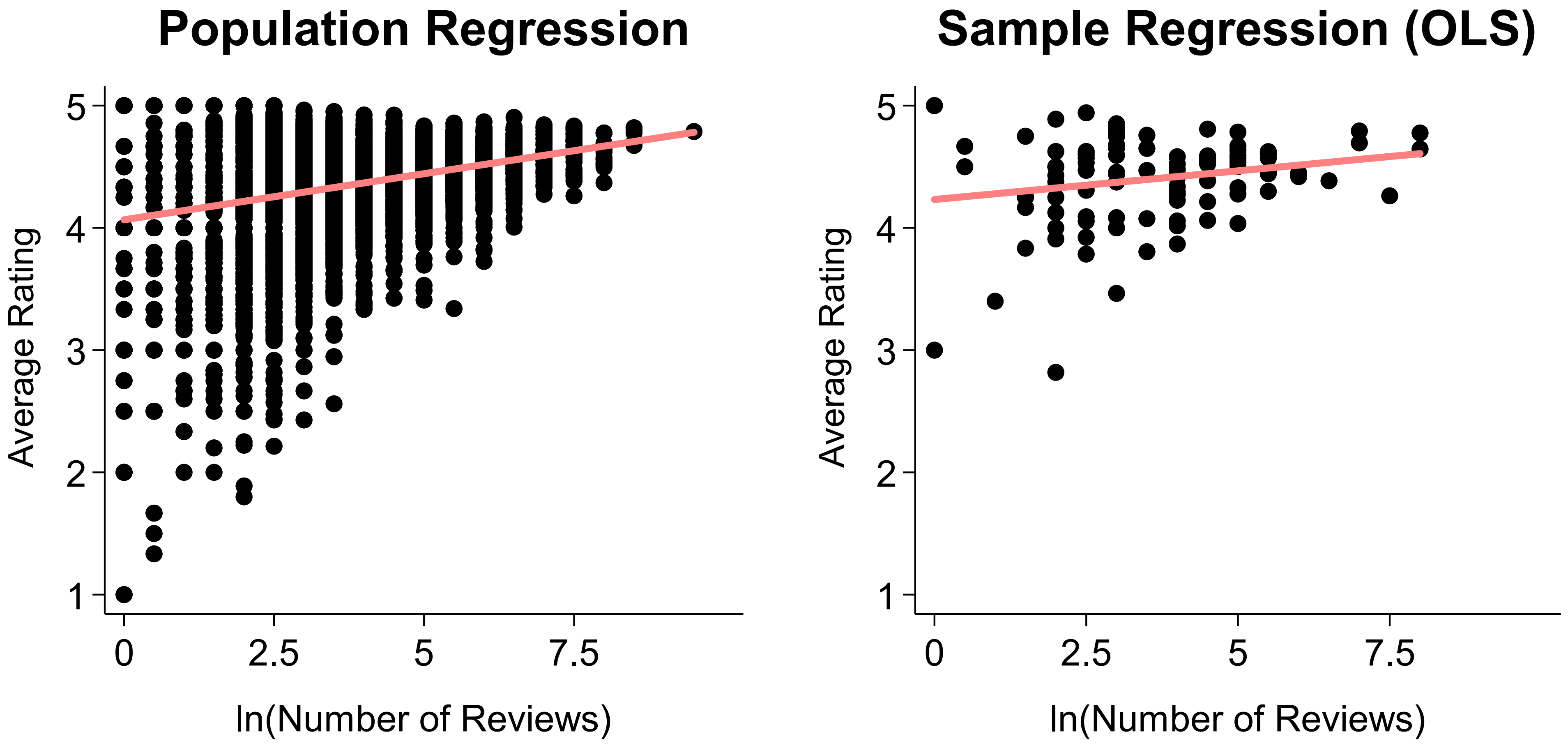
$$(\hat{\alpha}, \hat{\beta}) = \underset{a,b}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - (a + bX_i)\big)^2$$

$$b = \hat{\beta} = \frac{\text{Sample Covariance between } X_i \text{ and } Y_i}{\text{Sample Variance of } X_i} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$a = \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

This estimator is called **ordinary least squares (OLS)**

OLS: Sample Line of Best Fit



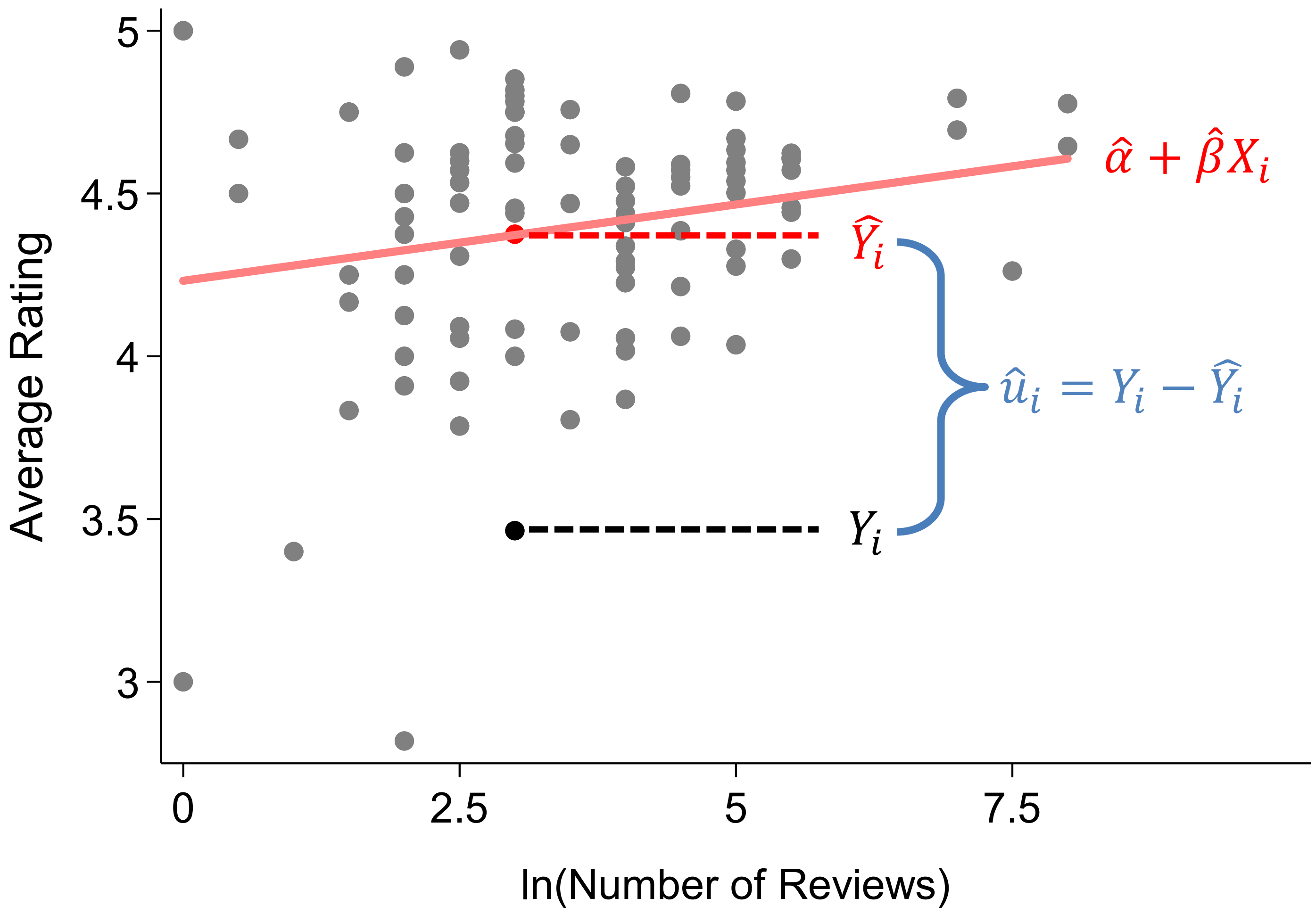
Intuition of the OLS Estimator

We define a **fitted value** of Y_i for a particular observation with explanatory variable X_i as:

$$\widehat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

The **residual** is the "mistake" we make:

$$\begin{aligned} \hat{u}_i &= Y_i - \widehat{Y}_i \\ &= Y_i - (\hat{\alpha} + \hat{\beta} X_i) \end{aligned}$$



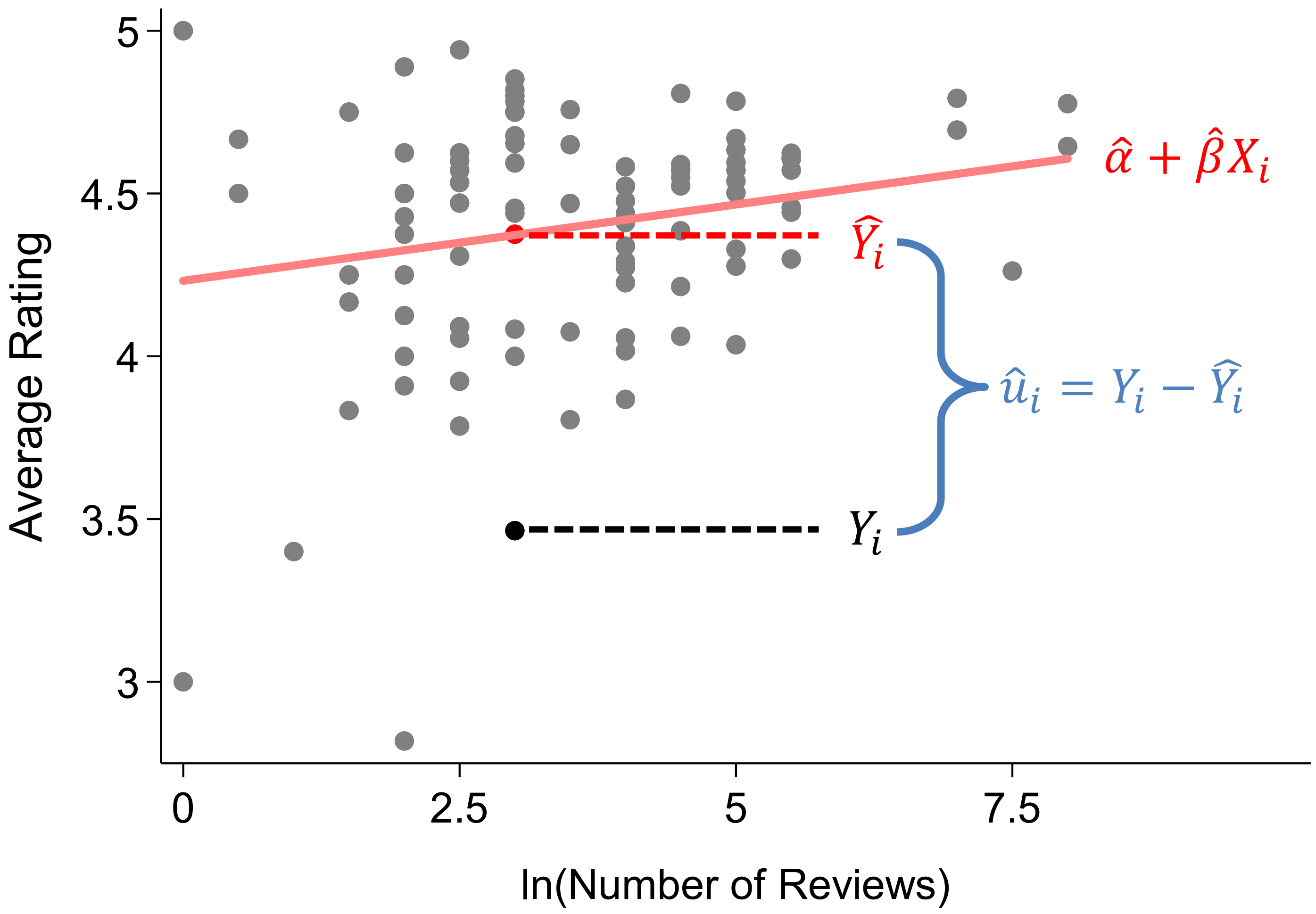
OLS Minimizes the Residuals

The residuals $\hat{u}_i$ tell us how well the line fits the data:

- Smaller residuals → better at predicting Y_i

OLS chooses $\hat{\alpha}$ and $\hat{\beta}$ so as to minimize the sum of squared residuals:

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\alpha} + \hat{\beta} X_i\right)^2$$



OLS Standard Errors

The constant variance assumption (*homoskedasticity*) helps to derive the sampling variance of $\hat{\beta}$:

$$V[u_i | X_i = x] = \sigma_u^2$$

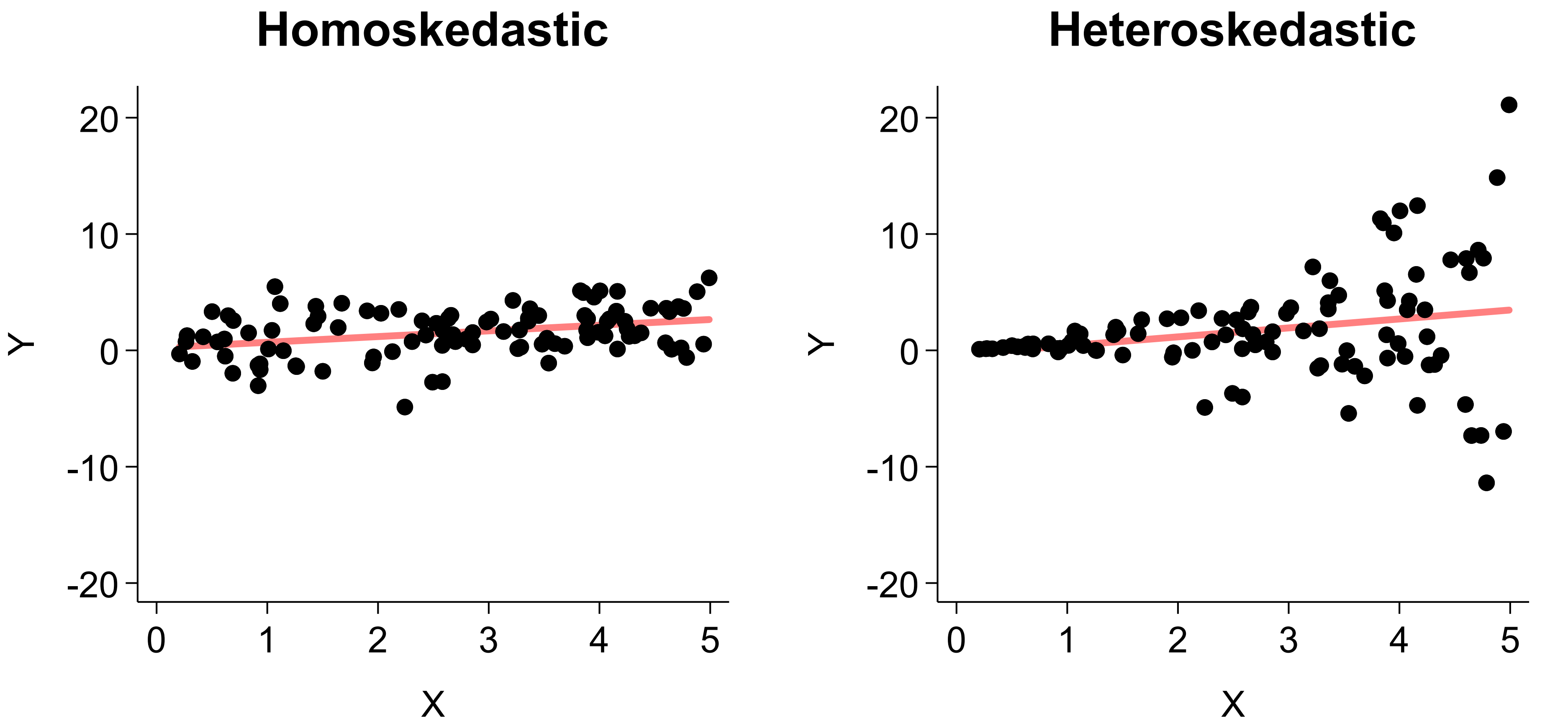
Use the residuals to estimate the unobserved variance of the errors:

$$\hat{\sigma}_u^2 = \frac{1}{n - 2} \sum_{i=1}^{n} \hat{u}_i^2$$

The standard error of the slope estimate in a bivariate regression is given by:

$$\widehat{SE}\left(\hat{\beta}\right) = \frac{\hat{\sigma}_u}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2}}$$

Homoskedasticity Assumption

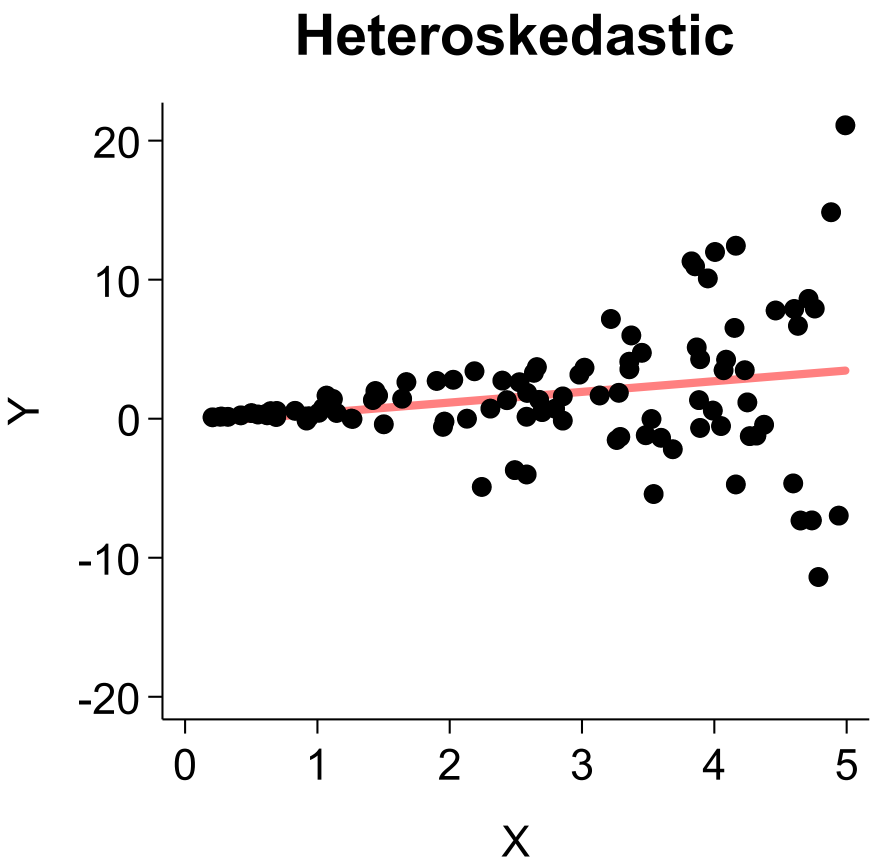


Robust Standard Errors

A common fix for heteroskedasticity is to estimate **"robust" standard errors**

- Most statistical software programs provide an option to calculate robust standard errors

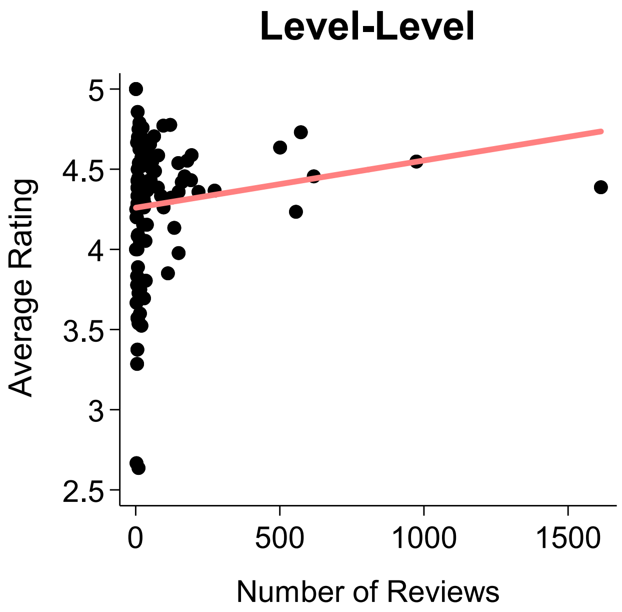
Robust standard errors are usually larger than conventional standard errors (but not always)

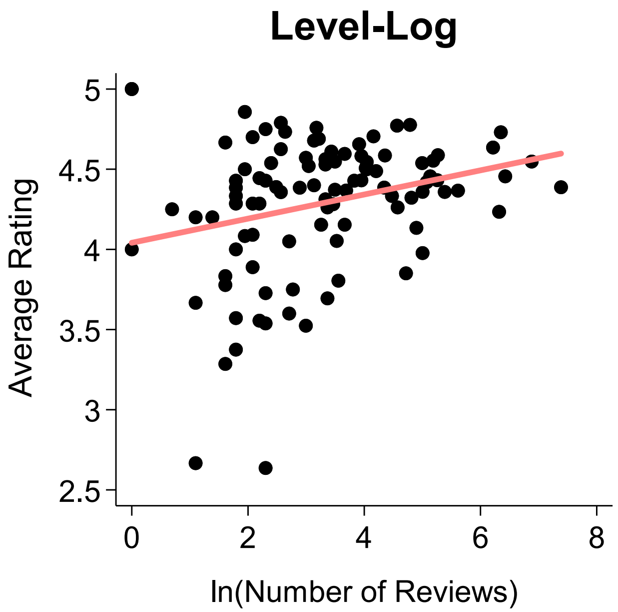


Models with Logs

What if we suspect that the relationship between X_i and Y_i is non-linear?

To account for the non-linearity, we can transform X_i or Y_i using the natural logarithm





Interpreting Logged Variables

Model	Equation	$\boldsymbol{\beta}$ **Interpretation**
Level-Level	$Y = \alpha + \beta X$	1-unit $\Delta X = \beta \Delta Y$
Log-Level	$\log(Y) = \alpha + \beta X$	1-unit $\Delta X \approx 100 \times \beta\%\Delta Y$
Level-Log	$Y = \alpha + \beta \log(X)$	1% $\Delta X \approx (\frac{\beta}{100})\ \Delta Y$
Log-Log	$\log(Y) = \alpha + \beta \log(X)$	1% $\Delta X \approx \beta\%\Delta Y$

Recipe Ratings and Number of Reviews

```
. regress recipe_stars ln_reviews, robust

Linear regression                               Number of obs     =        100
                                                F(1, 98)          =       6.41
                                                Prob > F          =     0.0130
                                                R-squared         =     0.0718
                                                Root MSE          =     .41814
```

A 1% increase in the number of reviews is associated with a (0.08/100) = 0.0008 increase in the average rating

recipe_stars	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
ln_reviews	.0753077	.029752	2.53	0.013	.0162659	.1343496
_cons	4.041276	.1280673	31.56	0.000	3.78713	4.295421

Recipe Ratings and Number of Reviews

```
. regress recipe_stars ln_reviews, robust

Line                                                             100
                                                                6.41
                                                              0.0130
                                                              0.0718
                                                              .41814
```

So do more reviews cause higher ratings? . . .

Probably not

recipe_stars	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
ln_reviews	$\boxed{.0753077}$	.029752	2.53	0.013	.0162659	.1343496
_cons	4.041276	.1280673	31.56	0.000	3.78713	4.295421

Recap of Regression

- Regression provides the best linear approximation of the CEF

- Regression is relatively simple (linear approach, easy interpretation, computational simplicity etc.)

- Statistical models are not meant to replicate the real world but to provide useful insights

- Regression is a flexible tool (e.g. $Y_i = \alpha \ + \beta X_i^2 + u_i$)

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Credits:
Alain Cohn
Assistant Professor of Information

Regression and Causality

Regression and Causality

When can we interpret a regression coefficient causally?

- Randomized experiments: coefficient on binary treatment is an estimate of the ATE
- Fancier techniques for observational data (regression discontinuity, differences-in-differences, instrumental variables etc.)
- **Controlled regression:** coefficients can be interpreted as causal effects but only if we control for all confounders

Zero Conditional Mean Assumption

The error u_i has expected value of 0 for any value of the explanatory variable X_i:

$$\boxed{E[u_i|X_i = x] = 0 \text{ for all values } x}$$

- Interpretation: all the other stuff that affects Y_i except X_i is the same at every level of X_i
- Plausible? Probably not
- Zero conditional mean is *not* testable (because the population regression function is unknown)

Example – Zero Conditional Mean Assumption

Suppose we want to estimate the effect of years of schooling on wages

- u is unobserved ability

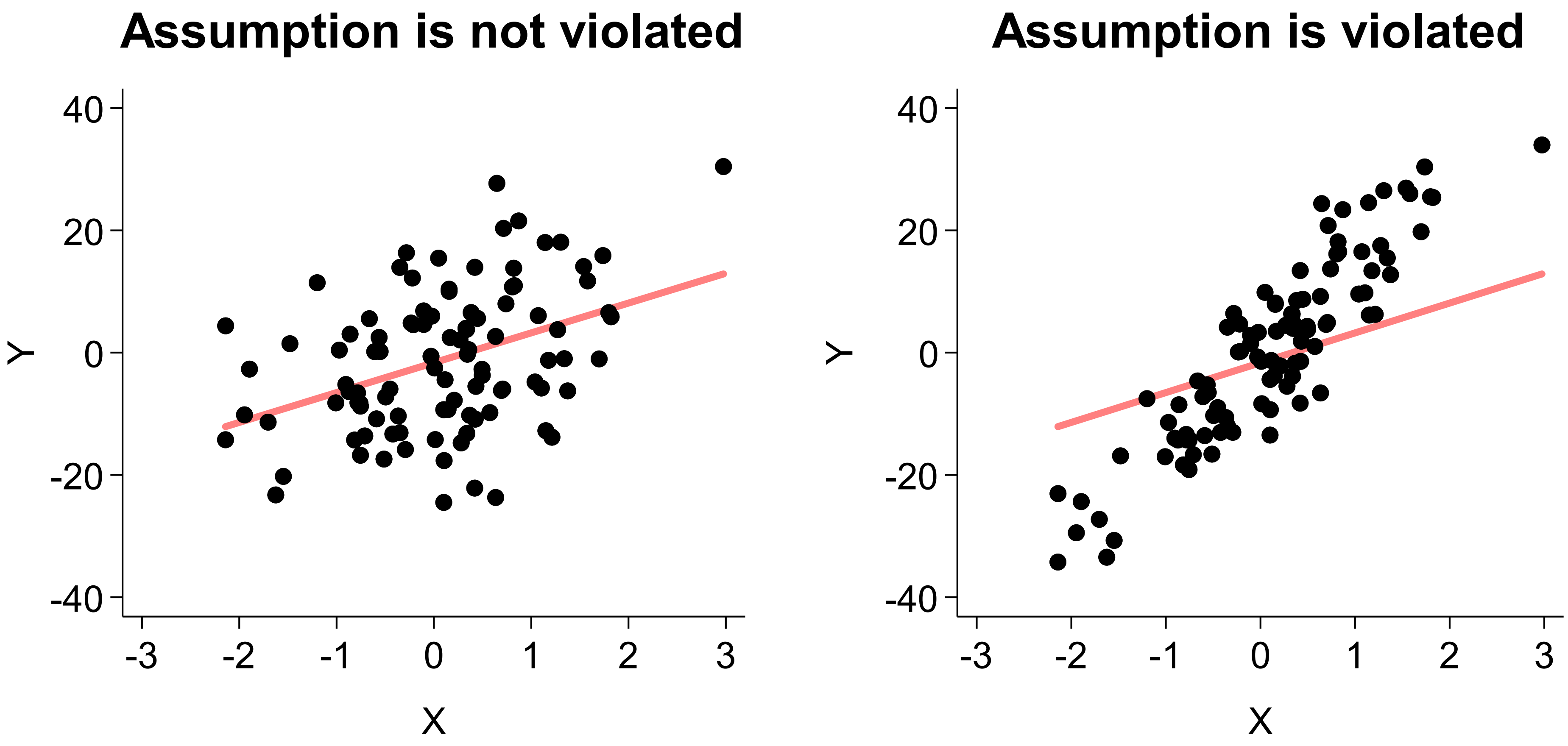
Mean independence requires that:

$$\mathrm{E}[\text{ability}|X_i = 8] = \mathrm{E}[\text{ability}|X_i = 12] = \mathrm{E}[\text{ability}|X_i = 16]$$

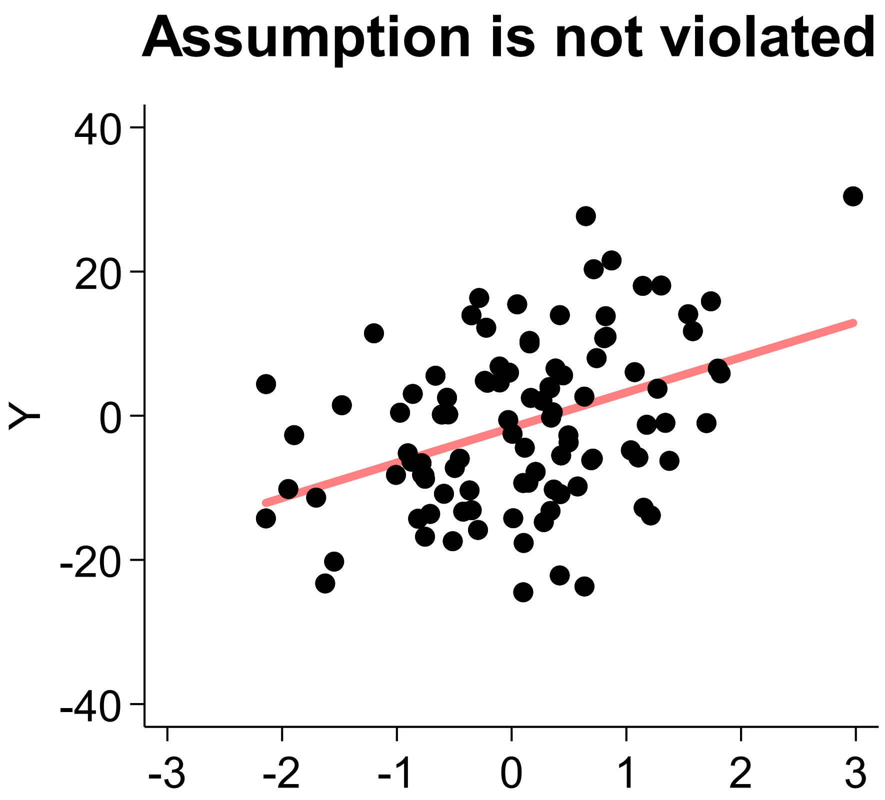
- As people choose education levels partly based on ability (selection bias), this assumption is almost certainly violated

- In addition, there may be other unobserved confounders than ability captured in u

Zero Conditional Mean Assumption



Zero Conditional Mean Assumption



When is this assumption most plausible?

- When X_i is randomly assigned in experimental data
- The X_i's are by design unrelated to the u_i's

A Tale of Two Colleges

Student	SAT	Earnings
1	1400	105
2	1250	110
3	1200	100
4	1050	115
5	800	65

Student	SAT	Earnings
6	1400	100
7	1350	80
8	1300	85
9	1200	105
10	1050	95
11	950	70
12	850	65
13	800	60

Regression Model

Let’s specify our population regression model:

$$\underline{Y_i} = \alpha + \beta \underline{P_i} + \gamma \underline{A_i} + \underline{u_i}$$

- Y_i is student i’s earnings later in life
- P_i is a dummy variable that equals to 1 if student i attended private college and 0 otherwise
- A_i is a discrete variable for student i’s SAT score (as a proxy for ability)
- u_i is the error term

Regression Model

Let's specify our population regression model:

$$Y_i = \alpha + \beta P_i + \gamma A_i + u_i$$

How can we interpret this model?

$$E[Y_i|P_i = 1] - E[Y_i|P_i = 0] = (\alpha + \beta + \gamma A_i) - (\alpha + \gamma A_i) = \beta$$

- β identifies the *causal* effect of private school on earnings

  … if we assume that private school attendance is "as good as random" conditional on SAT score (zero conditional mean assumption)

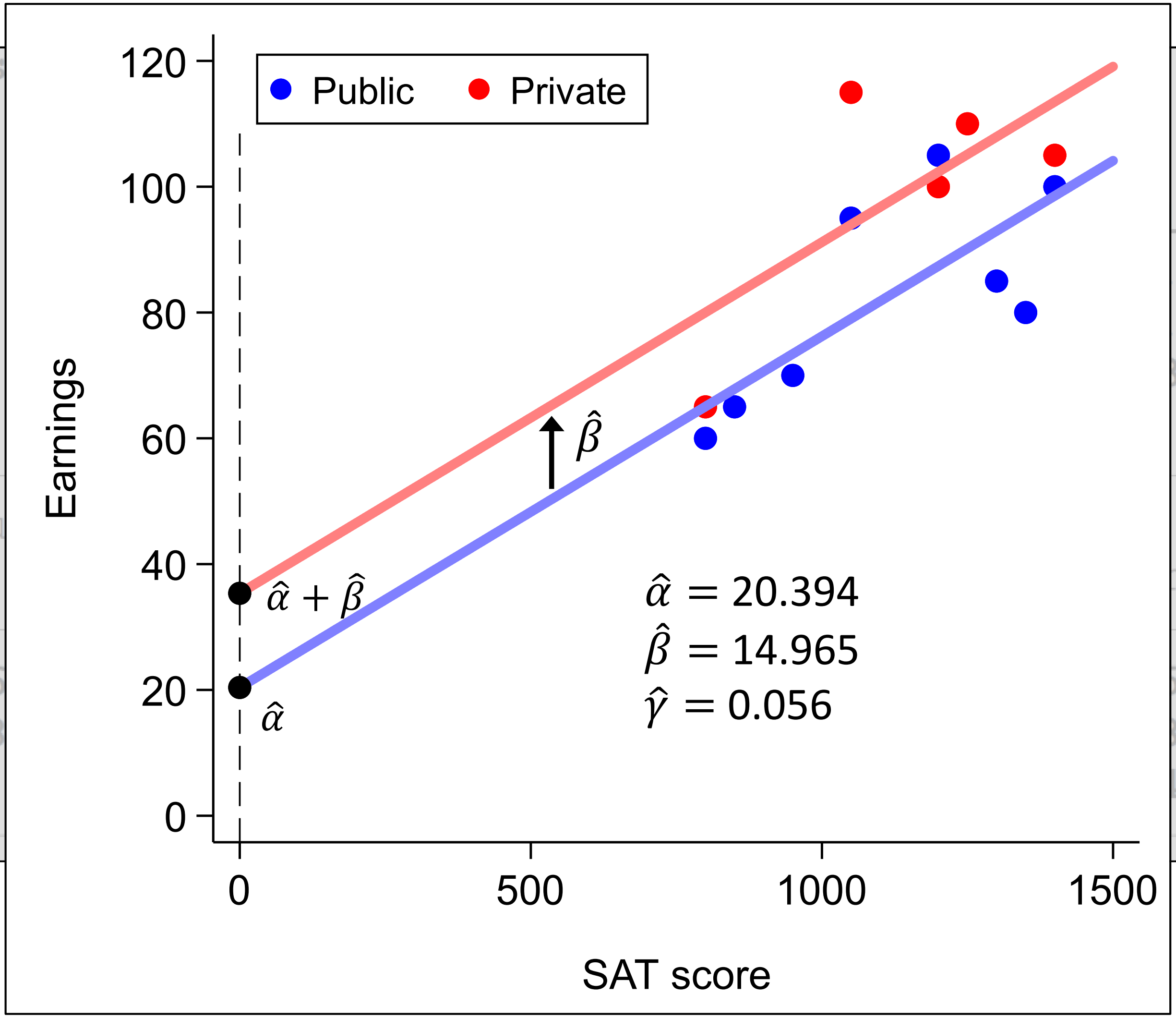
OLS Estimates

```
. reg earnings private sat, robust

Linear regression                               Number of obs     =         13
                                                F(2, 10)          =      19.54
                                                Prob > F          =     0.0004
                                                R-squared         =     0.6111
                                                Root MSE          =     13.017

------------------------------------------------------------------------------
             |               Robust
    earnings |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     private |   14.96479   7.896228     1.90   0.087    -2.629103    32.55868
         sat |   .0558259   .0148261     3.77   0.004     .0227913    .0888605
       _cons |   20.39373     15.724     1.30   0.224    -14.64153    55.42898
------------------------------------------------------------------------------
```

OLS Estimates

```
. reg earnings private sat, robus

Linear regression
                                              13
                                            .9.54
                                             0004
                                             6111
                                            3.017
```

earnings	Coef.	Robu Std.	val]
private	**14.96479**	7.896	5868
sat	**.0558259**	.0148	8605
_cons	**20.39373**	15.	2898



Omitted Variable Bias

$$Y_i = \alpha + \beta P_i + \cancel{\gamma A_i} + u_i$$

What if we run a regression without controlling for SAT scores?

```
. reg earnings private, robust
```

earnings	Coef.
private	16.5
_cons	82.5

```
. reg earnings private sat, robust
```

earnings	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
private	14.96479	7.896228	1.90	0.087	-2.629103	32.55868
sat	.0558259	.0148261	3.77	0.004	.0227913	.0888605
_cons	20.39373	15.724	1.30	0.224	-14.64153	55.42898

- Our estimate of the treatment effect gets larger
- But it's <u>biased</u> because we violate the zero conditional mean assumption:

$$E[u_i | P_i] \neq 0$$

Omitted Variable Bias

indicates "long"

Long regression: $$Y_i = \alpha^l + \beta^l P_i + \gamma^l A_i + u_i^l$$

Short regression: $$Y_i = \alpha^s + \beta^s P_i + \boxed{u_i^s}$$

$u_i^s = \gamma^l A_i + u_i^l$

How does $\widehat{\beta}^s$ relate to $\widehat{\beta}^l$?

Bias($\widehat{\beta}^s$) $= \widehat{\beta}^s - \widehat{\beta}^l$

Omitted Variable Bias Formula

Relationship between $\hat{\beta}^s$ and $\hat{\beta}^l$:

$$\hat{\beta}^s = \hat{\beta}^l + \boxed{\hat{\gamma}^l \, \hat{\pi}_1}$$

Omitted Variable Bias (OVB)

- $\hat{\pi}_1$ is the coefficient on P_i in a regression of A_i on P_i:

$$A_i = \pi_0 + \pi_1 P_i + v_i$$

OVB = ("effect" of A_i on Y_i) × ("effect" of P_i on A_i)
= (omitted → outcome) × (included → omitted)

Omitted Variable Bias

In practice we often have no choice but to omit A_i

Remember that by OLS: $$\hat{\pi}_1 = \frac{\widehat{Cov}(P_i, A_i)}{\hat{V}(P_i)}$$

We can sign the possible bias if we know the signs of the P_i and A_i relationship and the Y_i and A_i relationship

	$\widehat{Cov}(P_i, A_i) > 0$	$\widehat{Cov}(P_i, A_i) < 0$	$\widehat{Cov}(P_i, A_i) = 0$
$\hat{\gamma}^l > 0$	Positive bias	Negative bias	No bias
$\hat{\gamma}^l < 0$	Negative bias	Positive bias	No bias
$\hat{\gamma}^l = 0$	No bias	No bias	No bias

Example – Omitted Variable Bias

Suppose we don't observe students' SAT scores:

- We speculate that private school alumni have higher ability → $\widehat{Cov}(P_i, A_i) > 0$
- It's plausible that ability is positively related to earnings → $\hat{\gamma}^l > 0$

	$\widehat{Cov}(P_i, A_i) > 0$	$\widehat{Cov}(P_i, A_i) < 0$	$\widehat{Cov}(P_i, A_i) = 0$
$\hat{\gamma}^l > 0$	Positive bias	Negative bias	No bias
$\hat{\gamma}^l < 0$	Negative bias	Positive bias	No bias
$\hat{\gamma}^l = 0$	No bias	No bias	No bias

Example – Omitted Variable Bias

(omitted → outcome): $\hat{\gamma}^l = 0.05582$

```
. reg earnings private sat
```

earnings	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
private	14.96479	7.435714	2.01	0.072	-1.603013	31.53259
sat	.0558259	.0170084	3.28	0.008	.0179287	.093723
_cons	20.39373	19.47352	1.05	0.320	-22.99599	63.78344

(included → omitted): $\hat{\pi}_1 = 27.5$

```
. reg sat private
```

sat	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
private	27.5	131.5532	0.21	0.838	-262.0467	317.0467
_cons	1112.5	81.58584	13.64	0.000	932.9308	1292.069

Bias$(\hat{\beta}^s) = \hat{\gamma}^l \hat{\pi}_1 = 0.05582 \times 27.5 = 1.535$

Example – Omitted Variable Bias

Long regression: $\hat{\beta}^l = 14.96479$

```
. reg earnings private sat, robust
```

earnings	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
private	14.96479	7.896228	1.90	0.087	-2.629103	32.55868
sat	.0558259	.0148261	3.77	0.004	.0227913	.0888605
_cons	20.39373	15.724	1.30	0.224	-14.64153	55.42898

Short regression: $\hat{\beta}^s = 16.5$

```
. reg earnings private, robust
```

earnings	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
private	16.5	10.49889	1.57	0.144	-6.607902	39.6079
_cons	82.5	6.00071	13.75	0.000	69.29253	95.70747

$\text{Bias}(\hat{\beta}^s) = \hat{\beta}^s - \hat{\beta}^l = 16.5 - 14.96479 = 1.535$

Regression Sensitivity Analysis

We can never be sure whether our set of controls is enough to eliminate omitted variable bias

- Check sensitivity of regression estimates of treatment effects to the inclusion of controls
- If the coefficient on the treatment variable is stable after the inclusion of controls, we can take this as a sign that omitted variable bias is limited
- See Oster (2019) for a more advanced method to deal with selection on unobservables

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Credits:
Alain Cohn
Assistant Professor of Information