Conditional Means and Least Squares Regression

Econ 6105, Fall 2024

Prof. Josh Abel

(Greene chapters 4, 5.1-3, 6.3, 6.5, 7.5, 8.3-4)

Economic model

Consider the following "data-generating process" / "structural model":

$$y = y(x_1, x_2, ..., x_K) + \epsilon_i \tag{1}$$

- A bunch of x variables causally determine y via some relationship, $f : R^K \rightarrow R^1$
- There is also a disturbance, ϵ, that impacts y, separately from any of the xs

This model generates some joint distribution:

$$(y, x_1, ..., x_K) \sim f(y, x_1, ..., x_K) \tag{2}$$

We may may be interested in the conditional distribution:

$$y|(x_1, ..., x_K) \sim f_{y|x}(y|x_1, ..., x_K) \tag{3}$$

Dataset

Consider the following dataset of N observations:

$$\begin{bmatrix} y_1 \\ \dots \\ yi \\ \dots \\ y_N \end{bmatrix}, \begin{bmatrix} x_{11} & x_{21} & \dots & x_{K1} \\ \dots & \dots & \dots & \dots \\ x_{1i} & x_{2i} & \dots & x_{Ki} \\ \dots & \dots & \dots & \dots \\ x_{1N} & x_{2N} & \dots & x_{KN} \end{bmatrix}$$

Moonshot: Estimating the conditional distribution

In principle, can estimate the full conditional distribution.

1. Define $I_i^{a,b_1,\ldots b_K} = \begin{cases} 1 \text{ if } y_i = a, x_{1i} = b_1, \ \ldots, \ x_{Ki} = b_K \\ 0 \text{ otherwise} \end{cases}$
2. Define $J_i^{b_1,\ldots b_K} = \begin{cases} 1 \text{ if } x_{1i} = b_1, \ \ldots, \ x_{Ki} = b_K \\ 0 \text{ otherwise} \end{cases}$
3. $\hat{P}(y = a, x_1 = b_1, \ldots, x_K = b_K) = \frac{\sum_{i=1}^{N} I_i^{a,b_1,\ldots b_K}}{\sum_{i=1}^{N} J_i^{b_1,\ldots b_K}}$

I.e. estimate the population conditional distribution with the sample conditional distribution. If you have a random sample, this:

- Is unbiased
- Will converge to the population distribution as $N \to \infty$ ("consistent")

We almost never do this. Why not?

Moonshot: Estimating the conditional distribution

In principle, can estimate the full conditional distribution.

1. Define $I_i^{a,b_1,\ldots b_K} = \begin{cases} 1 \text{ if } y_i = a, x_{1i} = b_1, \ \ldots, \ x_{Ki} = b_K \\ 0 \text{ otherwise} \end{cases}$
2. Define $J_i^{b_1,\ldots b_K} = \begin{cases} 1 \text{ if } x_{1i} = b_1, \ \ldots, \ x_{Ki} = b_K \\ 0 \text{ otherwise} \end{cases}$
3. $\hat{P}(y = a, x_1 = b_1, \ldots, x_K = b_K) = \frac{\sum_{i=1}^{N} I_i^{a,b_1,\ldots b_K}}{\sum_{i=1}^{N} J_i^{b_1,\ldots b_K}}$

I.e. estimate the population conditional distribution with the sample conditional distribution. If you have a random sample, this:

- Is unbiased
- Will converge to the population distribution as $N \to \infty$ ("consistent")

We almost never do this. Why not?

1. When $N << \infty$, this might be very erratic and unreliable
   - Bias-Variance Tradeoff!
2. Even if we get the true distribution, it's hard to interpret!

Scaling back: Estimating the conditional mean

Let's be less greedy. Instead of the whole distribution, let's just try to get the conditional *mean*, sometimes called **the regression function**:

$$E[y|x_1, ..., x_K]. \tag{4}$$

Building off our previous work, we can estimate population means with sample means:

$$\hat{E}[y|x_1 = b_1, ..., x_K = b_K] = \frac{\sum_{i=1}^{N} y_i \cdot J_i^{b_1,...b_K}}{\sum_{i=1}^{N} J_i^{b_1,...b_K}} \tag{5}$$

If you have a random sample, this is:

- Is unbiased
- Consistent

Nonparametric Regression

The approach on the previous slide is known as **non-parametric regression**

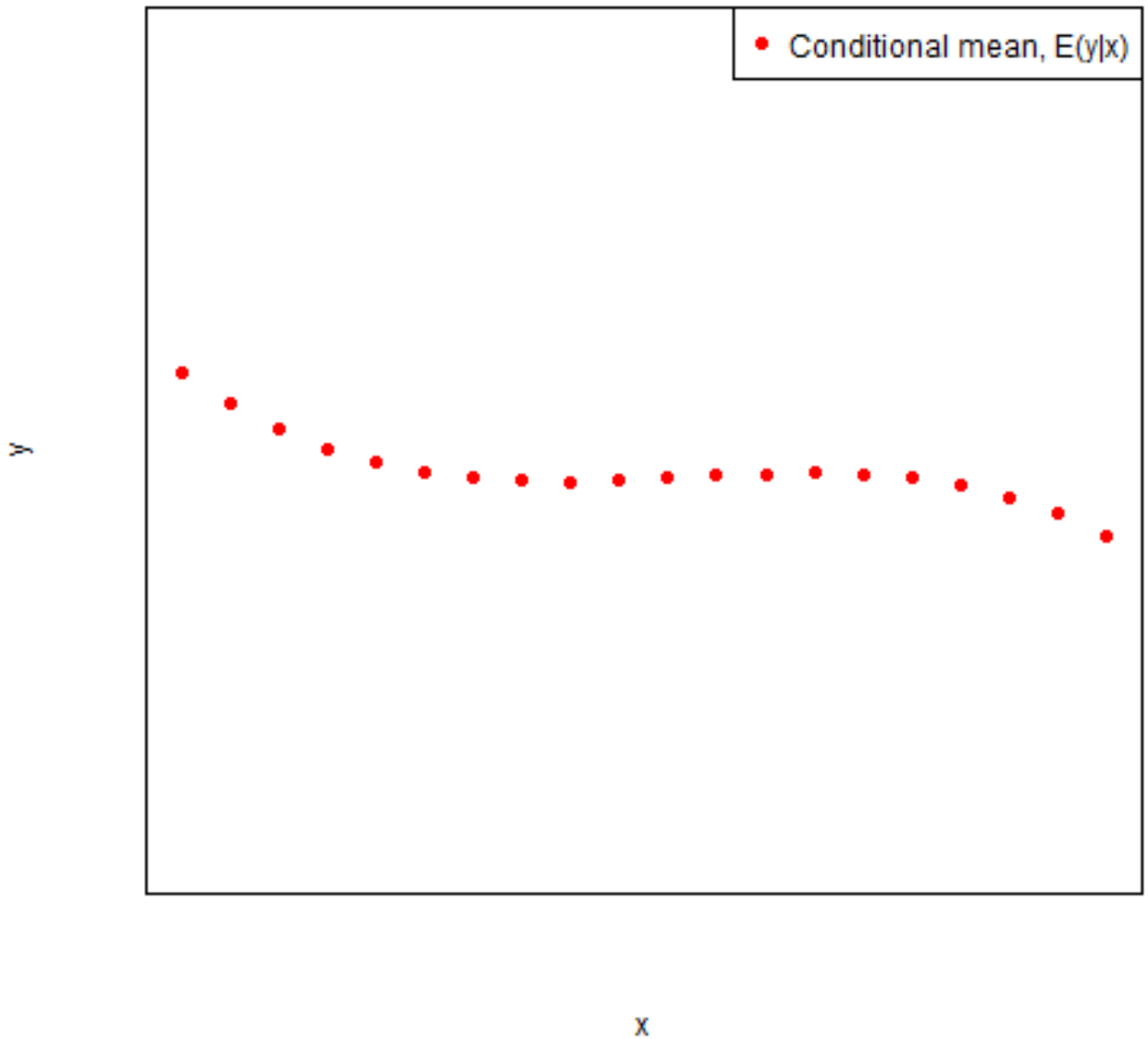
1. Partition domain of **x** into cells
2. Within each cell, calculate the mean of y

Very powerful, practical approach:

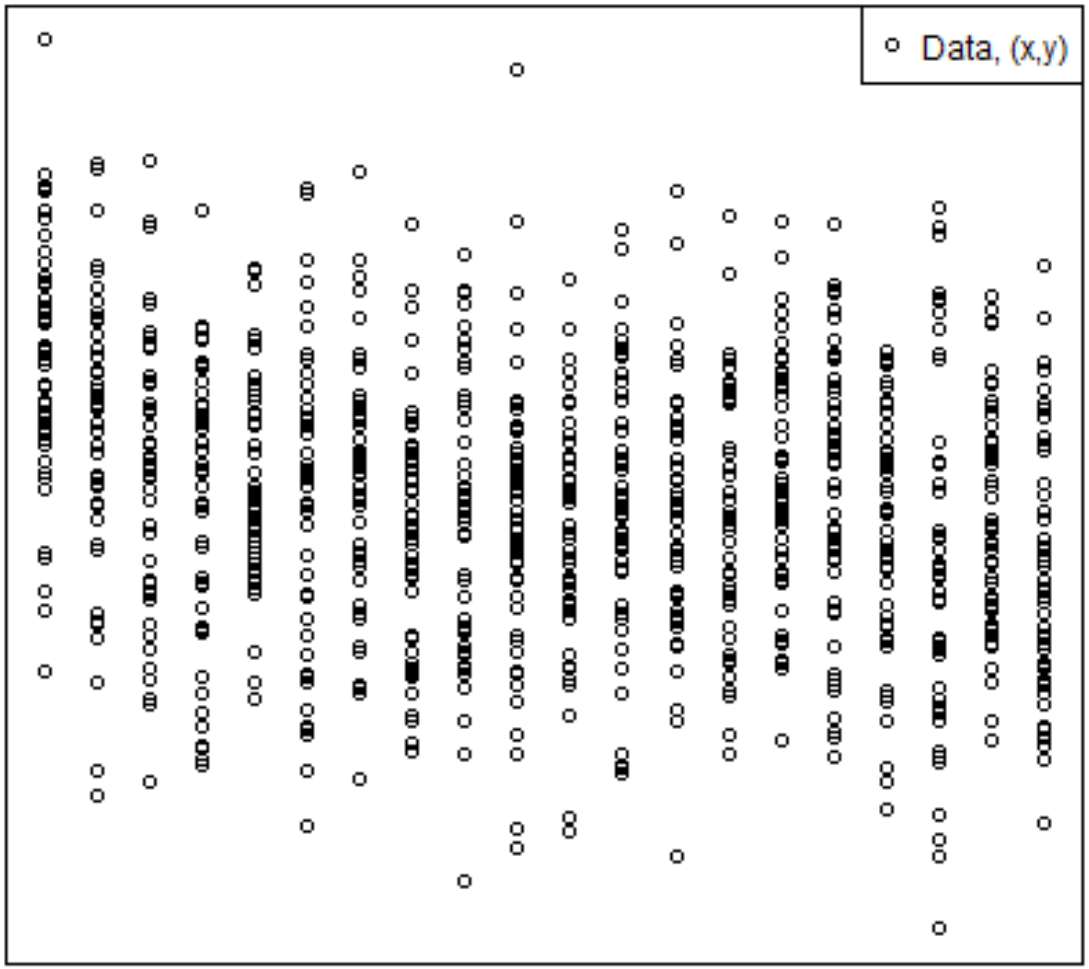
- No assumptions!
- Nothing fancier than calculating averages!

(Note: if any x is continuous, you may need to something slightly fancy...)

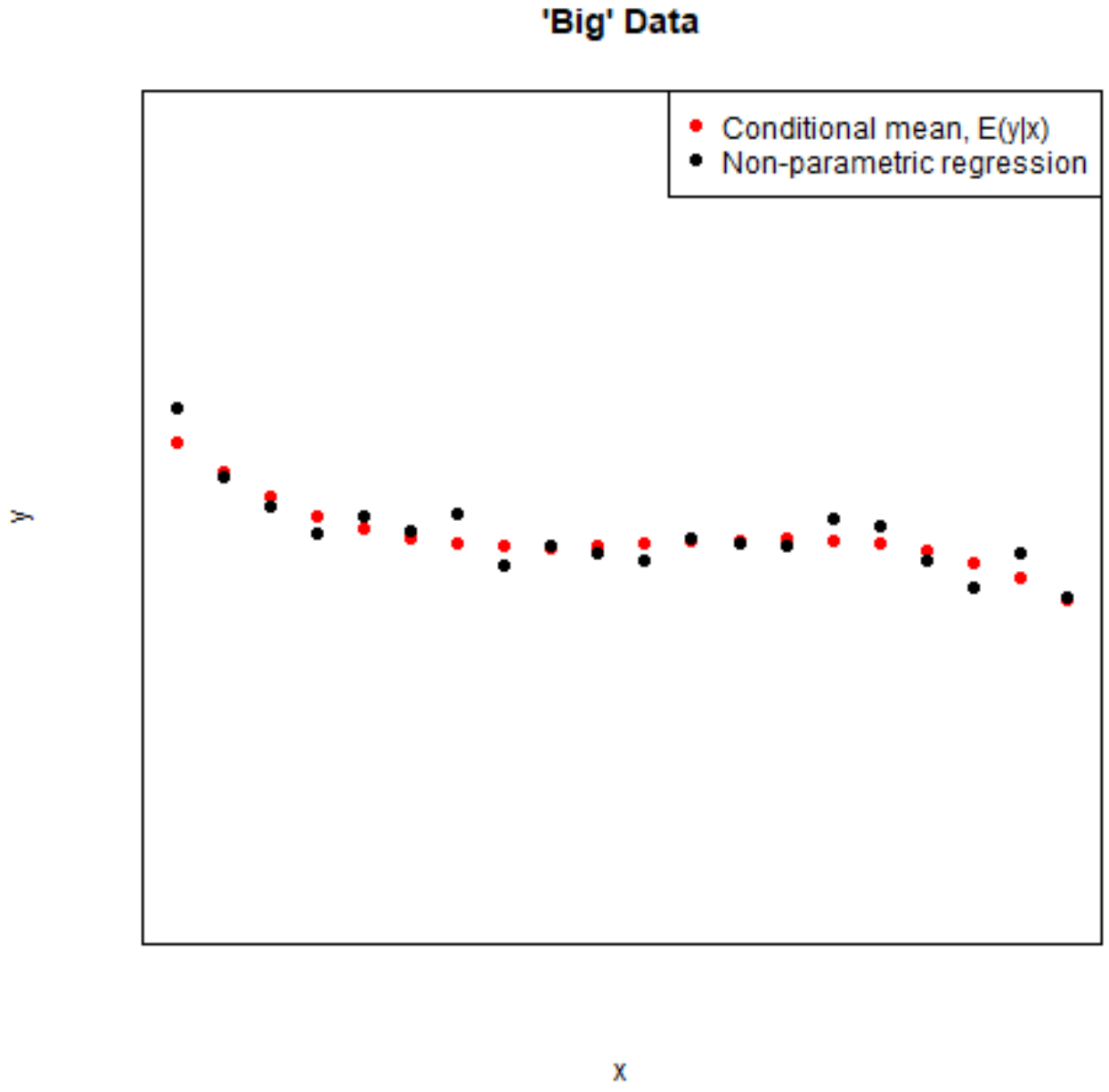
Key (made-up) example



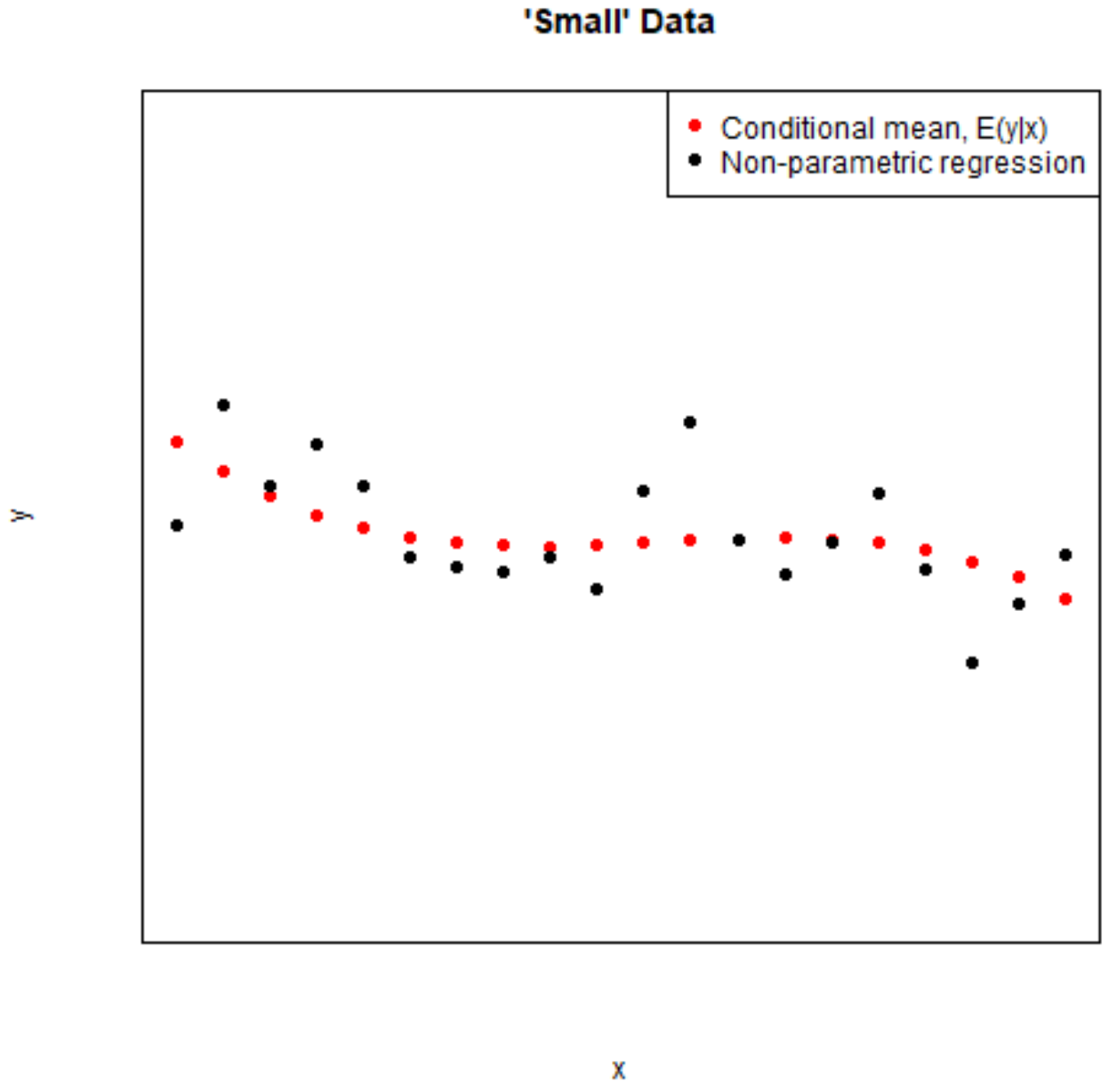
Key (made-up) example



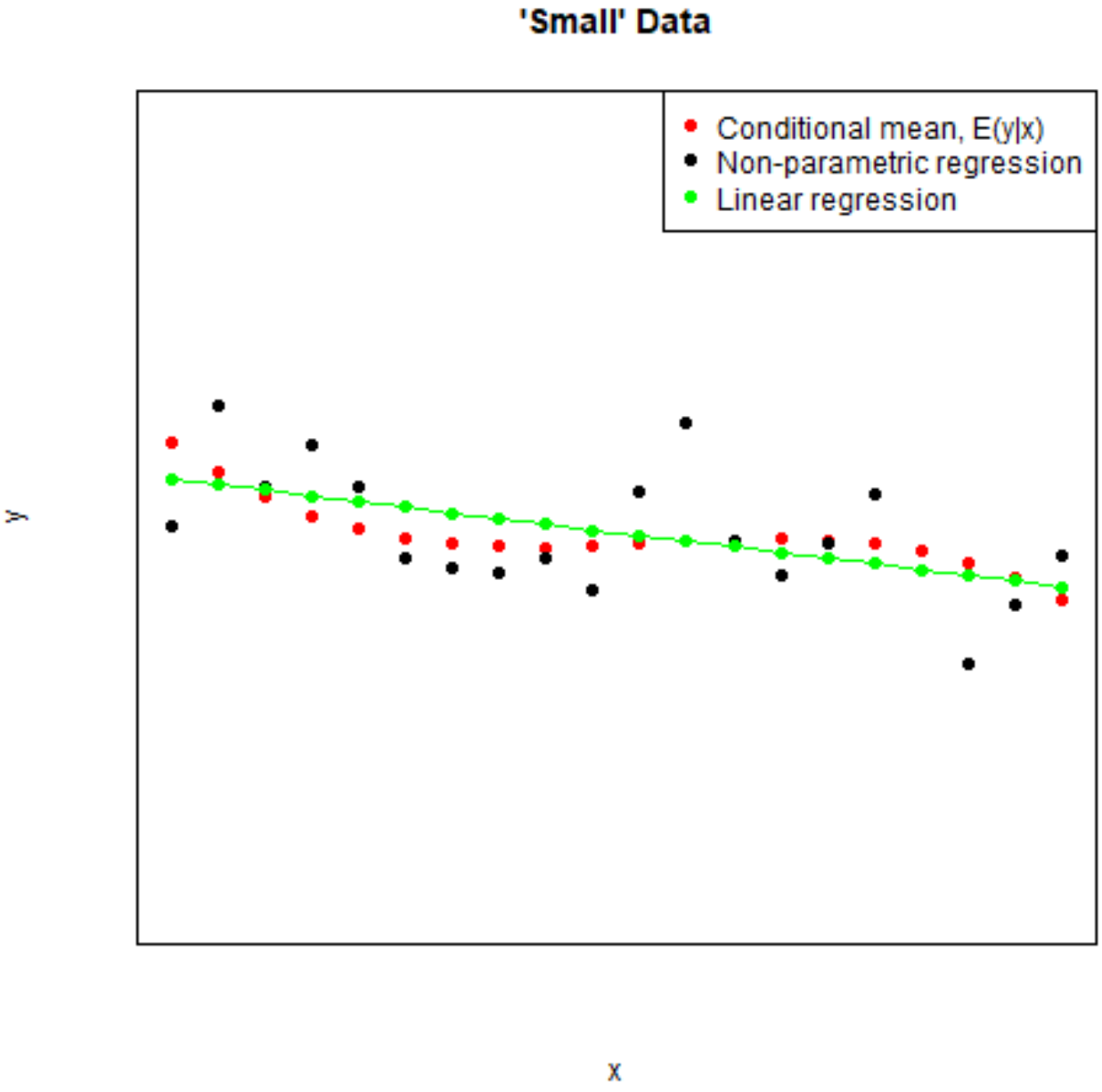
Key (made-up) example



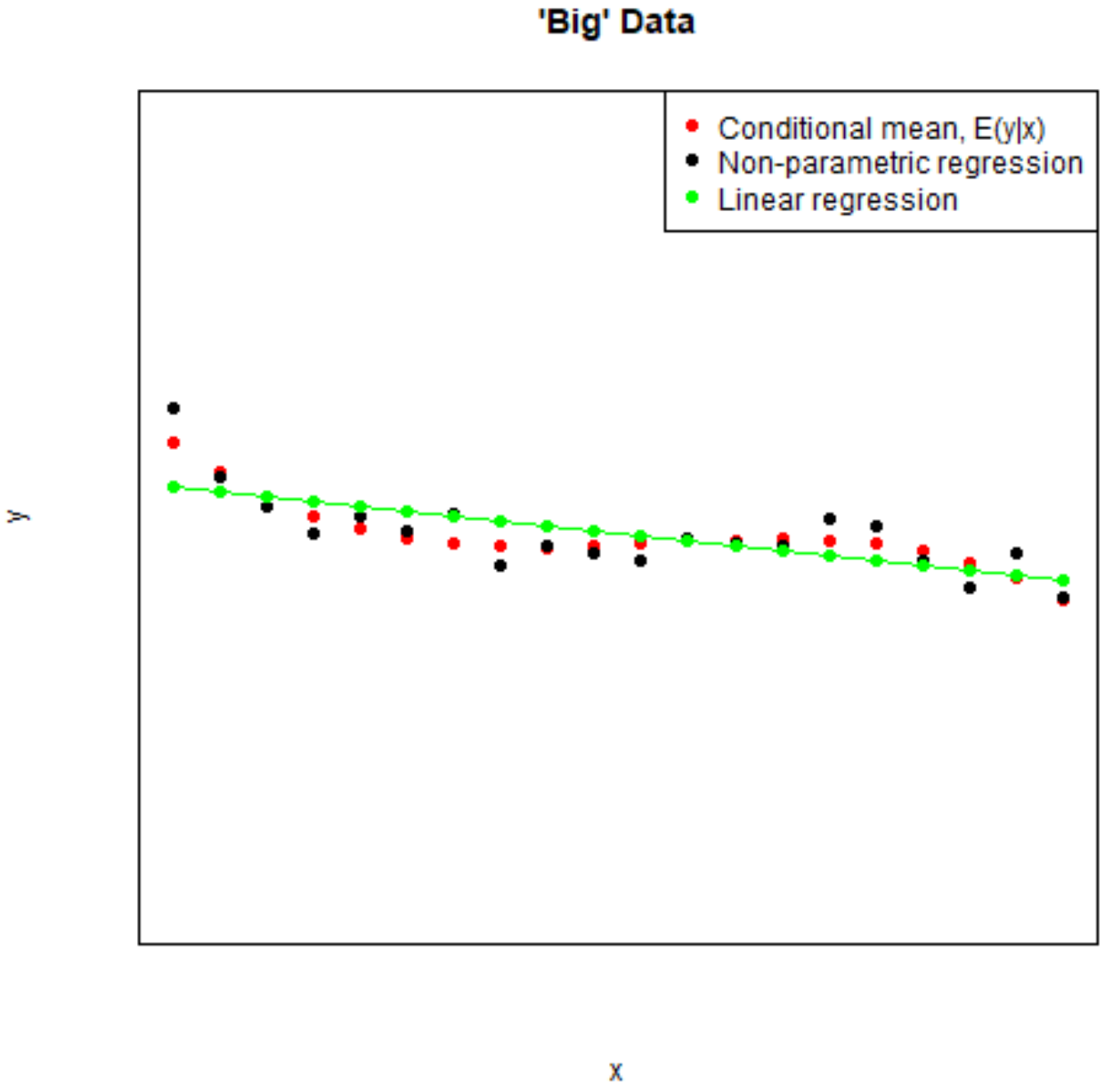
Key (made-up) example



Key (made-up) example



Key (made-up) example



Takeaways

1. Estimating the conditional mean is already a concession to the BVT
   - Justifies estimating the (conditional) mean rather than *distribution*
2. We can vary the flexibility of our conditional mean estimator
   - Most flexible: nonparametric regression
     - Pro: No assumptions/bias. Will nail any pattern with large N
     - Con: Unreliable/erratic when N is small
   - Least flexible: linear model
     - Pro: Relatively stable, even for small N
     - Con: Imposes untested – likely false – assumptions on analysis
   - Compromise: quadratic, cubic, etc.
     - The machinery of the linear model easily translates to these extensions
     - (In fact, the linear model nests the nonparametric approach as a special case, too.)

Linear regression model

Based on that motivation, we will typically assume a regression function of the form:

$$y_i = \beta_0 + \beta_1 \cdot x_{1i} + ... + \beta_K \cdot x_{Ki} + \epsilon_i. \tag{6}$$

The dataset can be represented in matrix notation:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon. \tag{7}$$

The model is "linear in the parameters:"

- No β_1^2 or β_1/β_0, etc.

Can still estimate non-linear relationships between y and some x, for instance by including x^2, x^3, etc in the regression

Ordinary Least Squares (OLS)

To estimate β, **Ordinary Least Squares (OLS)** solves the following minimization problem:

$$\hat{\beta} = \text{argmin}_{\tilde{\beta}} \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i \tilde{\beta} \right)^2 = \text{argmin}_{\tilde{\beta}} \sum_{i=1}^{N} \tilde{\epsilon}_i^2, \tag{8}$$

where $\tilde{\epsilon}_i \equiv y_i - \mathbf{x}_i \tilde{\beta}$.
Because we are using MSE criterion, resulting function is interpreted as a conditional mean:

$$\hat{E}[y_i | \mathbf{x}_i] = \mathbf{x}_i \hat{\beta}. \tag{9}$$

Sometimes useful to represent the solution as:

$$y_i = \mathbf{x}_i \hat{\beta} + \hat{\epsilon}_i, \tag{10}$$

or

$$y_i = \hat{y}_i + \hat{\epsilon}_i, \tag{11}$$

where $\hat{y}_i \equiv x_i \hat{\beta}$.

Univariate OLS solution

In univariate case ($y = \beta_0 + \beta_1 \cdot x + \epsilon$), the OLS estimator...

- ...is explicitly solved as:

$$\hat{\beta}_1^{OLS} = \frac{cov(y, x)}{var(x)} \tag{12}$$

$$\hat{\beta}_0^{OLS} = \bar{y} - \hat{\beta}_1^{OLS} \cdot \bar{x} \tag{13}$$

- ...is implicitly characterized by "orthogonality conditions:"

$$E[\hat{\epsilon}^{OLS}] = 0 \tag{14}$$

$$E[\hat{\epsilon}^{OLS} \cdot x] = 0 \tag{15}$$

Understanding the univariate OLS estimator

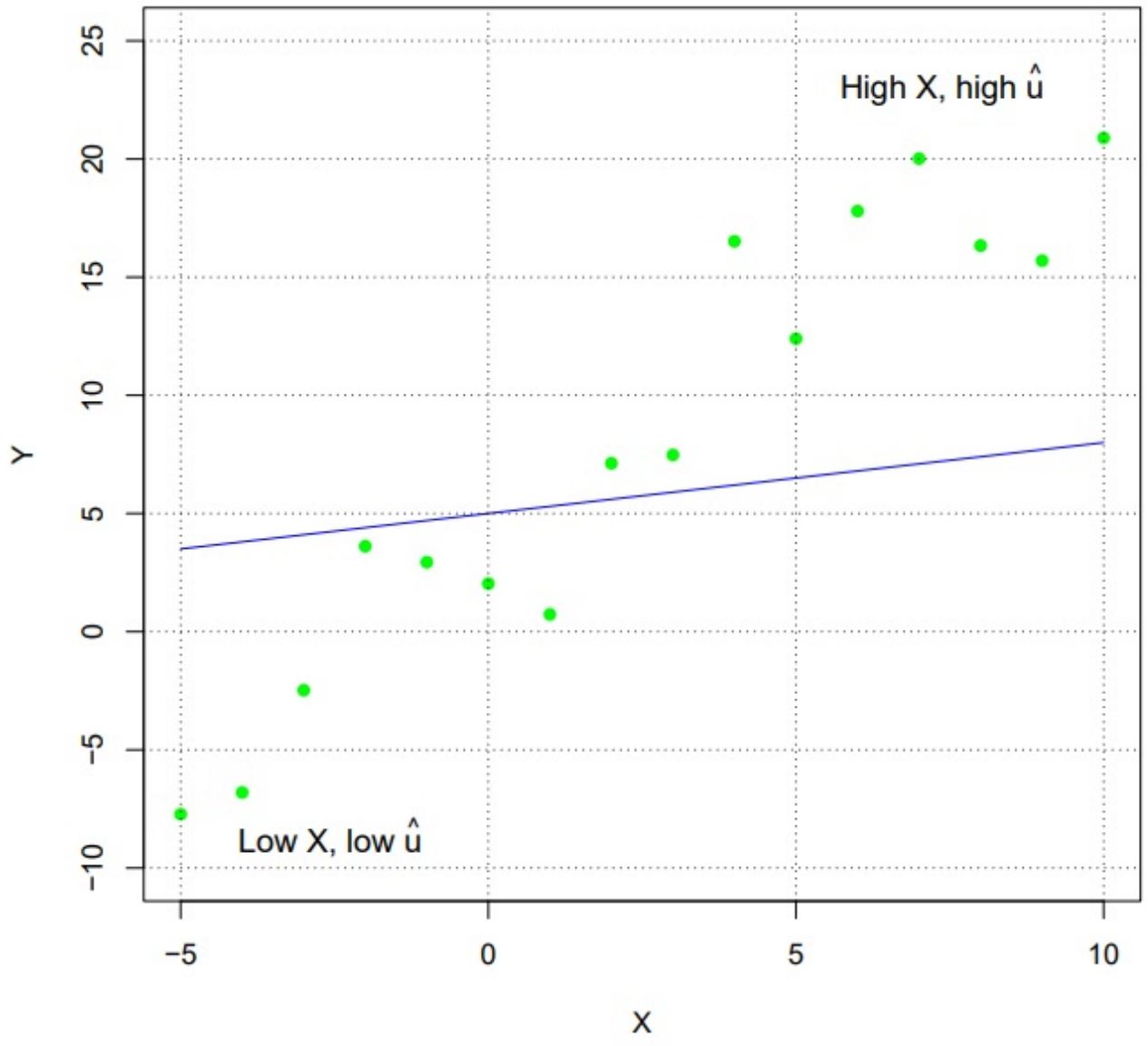
Explicit solution:

- "Slope parameter" $\hat{\beta}_1$ is the covariance of x with y, normalized to be in terms of a 1-unit change of x
- "Intercept parameter" $\hat{\beta}_0$ makes sure the model is right on average

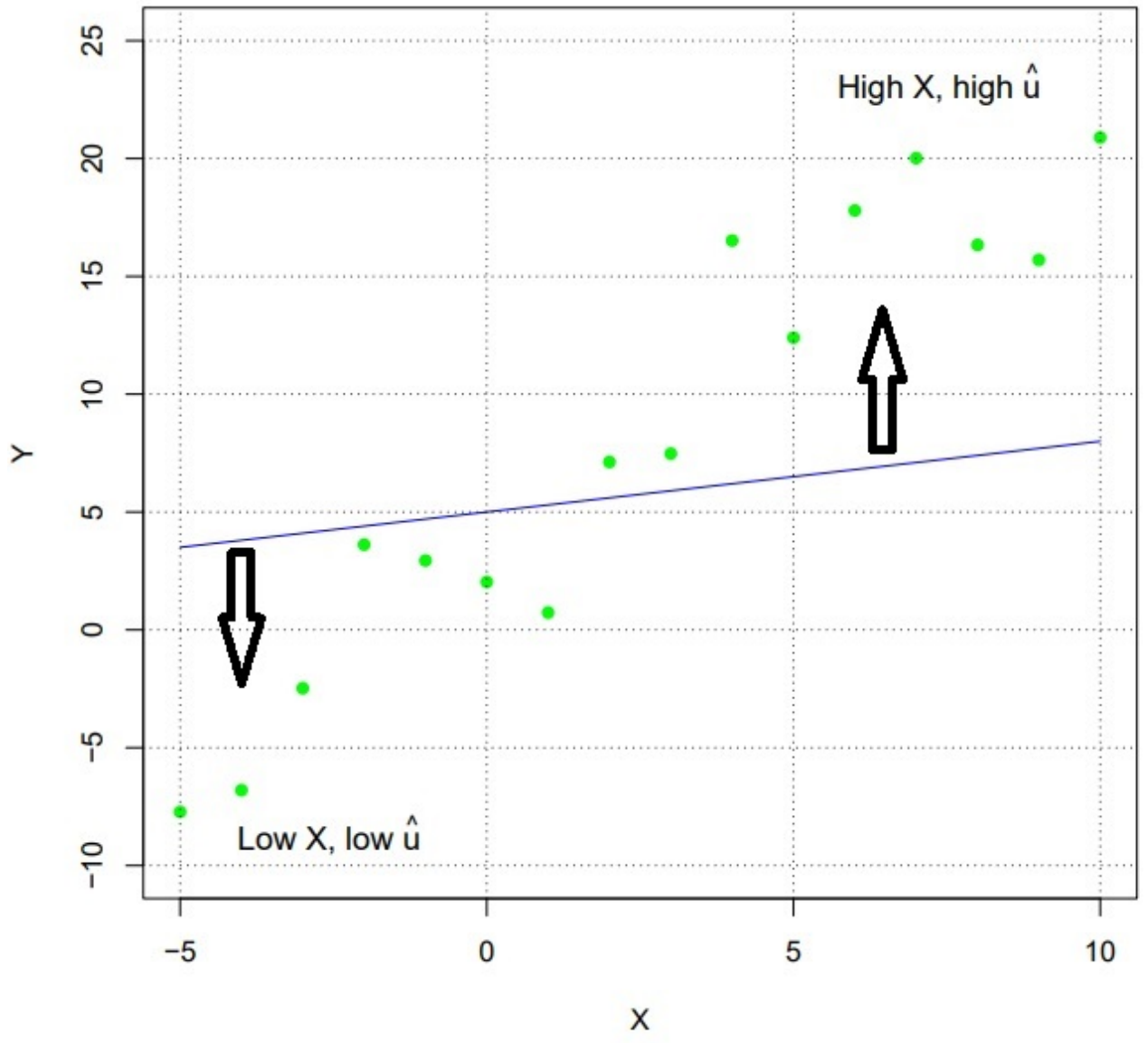
Implicit solution:

- Residuals $\hat{\epsilon}$ are uncorrelated with regressor x
- Residuals are zero on average

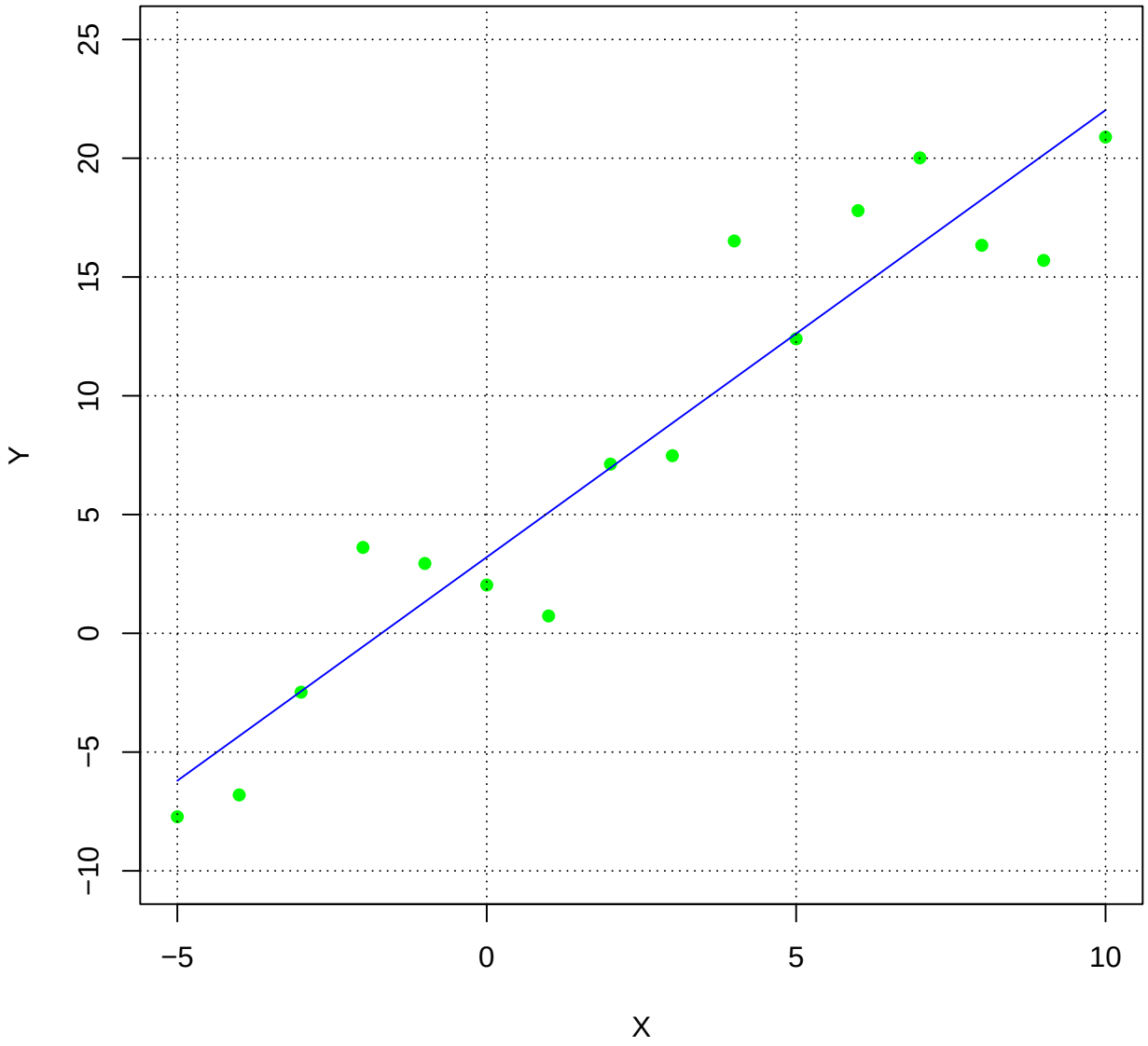
Orthogonality condition, visualized



Orthogonality condition, visualized



Orthogonality condition, visualized



Multivariate OLS

We can add more x variables:

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + ... + \beta_K \cdot x_{iK} + \epsilon_i \tag{16}$$

The explicit solution approach gets gnarly, but the implicit characterization via orthogonality conditions scales up easily:

$$\begin{aligned} E[\hat{\epsilon}] &= 0 \\ E[\hat{\epsilon} \cdot \mathbf{x}_1] &= 0 \\ &... \\ E[\hat{\epsilon} \cdot \mathbf{x}_K] &= 0, \end{aligned}$$

or more compactly:

$$E[\mathbf{X}'\hat{\epsilon}] = \mathbf{0}. \tag{17}$$

Equation counting

OLS is such a workhorse because it always generates the same number of equations as unknowns

- Each β_k generates an orthogonality condition
- (This assumes none of the xs are linear combinations of each other, but if they are, you have not through your model well anyway.)

Better yet, the equations are all linear in the $\hat{\beta}$s

- E.g. in univariate case, we have:
  1. $E[y] - \hat{\beta}_0 - \hat{\beta}_1 \cdot E[x] = 0$
  2. $E[y \cdot x] - \hat{\beta}_0 \cdot E[x] - \hat{\beta}_1 \cdot E[x^2] = 0$

So our linear algebra training tells us we are guaranteed a unique solution

- "Full rank"

General explicit form of OLS solution

Recall, OLS is characterized by orthogonality conditions:

$$\begin{aligned} 0 &= \frac{1}{n} \cdot \mathbf{X}'\hat{\epsilon} \\ &= \mathbf{X}'(y - \mathbf{X}\hat{\beta}) \\ \mathbf{X}'\mathbf{X}\hat{\beta} &= \mathbf{X}'y \\ \hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y \end{aligned}$$

Inference with $\hat{\beta}$

Define the variance (or covariance, or variance-covariance) of a vector as:

$$VCV(\hat{\beta}) = \begin{bmatrix} \sigma^2_{\hat{\beta}_1} & \sigma_{\hat{\beta}_1,\hat{\beta}_2} & \dots & \sigma_{\hat{\beta}_1,\hat{\beta}_K} \\ \sigma_{\hat{\beta}_1,\hat{\beta}_2} & \sigma^2_{\hat{\beta}_2} & \dots & \sigma_{\hat{\beta}_2,\hat{\beta}_K} \\ \dots & \dots & \dots & \dots \\ \sigma_{\hat{\beta}_1,\hat{\beta}_K} & \sigma_{\hat{\beta}_2,\beta_K} & \dots & \sigma^2_{\hat{\beta}_K} \end{bmatrix} \tag{18}$$

- Diagonal contains the square of each coefficient's SE
- Off-diagonal is each coefficient's covariance with each other one

Note that CLT gives that distribution of $\hat{\beta}_k$ approaches $N(0, \sigma^2_{\hat{\beta}_k})$

- So we conduct hypothesis tests and create confidence intervals just as before!
- Only wrinkle is computing VCV

Sandwich Estimator of VCV

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$$

We can think of the uncertainty as being driven by the ϵs hidden inside the ys. Heuristically:

$$VCV\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y\right) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'VCV(y)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

Because of random sampling, $VCV(y)$ is a diagonal matrix:

$$VCV(y) = \begin{bmatrix} \sigma^2_{\epsilon_1} & 0 & 0 & \dots & 0 \\ 0 & \sigma^2_{\epsilon_2} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \sigma^2_{\epsilon_n} \end{bmatrix}$$

Sandwich Estimator of VCV (2)

$VCV(y)$ can be estimated as:

$$\hat{VCV}(y) = \begin{bmatrix} \hat{\epsilon}_1^2 & 0 & 0 & \dots & 0 \\ 0 & \hat{\epsilon}_2^2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \hat{\epsilon}_n^2 \end{bmatrix} \tag{19}$$

So:

$$\hat{VCV}(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{VCV}(y)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \tag{20}$$

with $\hat{VCV}(y)$ defined in Equation 19.
This is known as **heteroskedasticity-robust** because it allows the different observations to have different variances.
Assuming **homoskedasticity**, i.e. $\sigma_{\epsilon_i}^2 = \sigma^2 \; \forall i$, we have:

$$VCV(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \tag{21}$$

Intuition from the homoskedastic univariate case

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \cdot \frac{\frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2}{\frac{1}{n}\sum_{i=1}^{n}\left[(X_i - \bar{X})^2\right]}$$

- Your standard error is driven by 3 things

Intuition from the homoskedastic univariate case

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \cdot \frac{\frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2}{\frac{1}{n}\sum_{i=1}^{n}\left[(X_i - \bar{X})^2\right]}$$

- Your standard error is driven by 3 things
  - Sample size (n) – just like sample mean!
    - Larger samples will have smaller standard errors – more information reduces uncertainty

Intuition from the homoskedastic univariate case

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \cdot \frac{\frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2}{\frac{1}{n}\sum_{i=1}^{n}\left[(X_i - \bar{X})^2\right]}$$

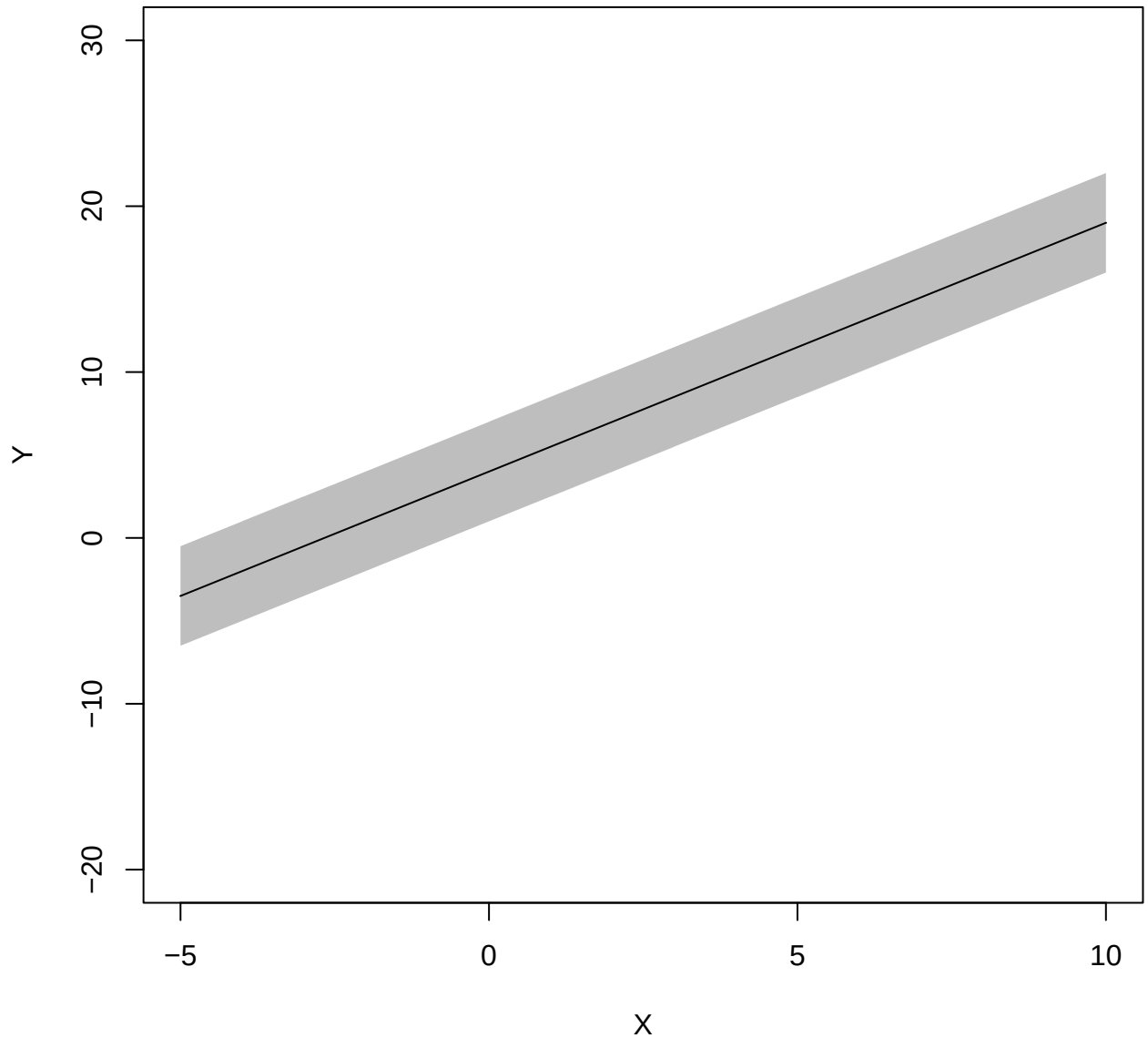
- Your standard error is driven by 3 things
  - Sample size (n) – just like sample mean!
    - Larger samples will have smaller standard errors – more information reduces uncertainty
  - Variance of residuals (numerator of expression) – very similar to sample mean!
    - The noisier the data (higher variance), the greater the uncertainty

Intuition from the homoskedastic univariate case

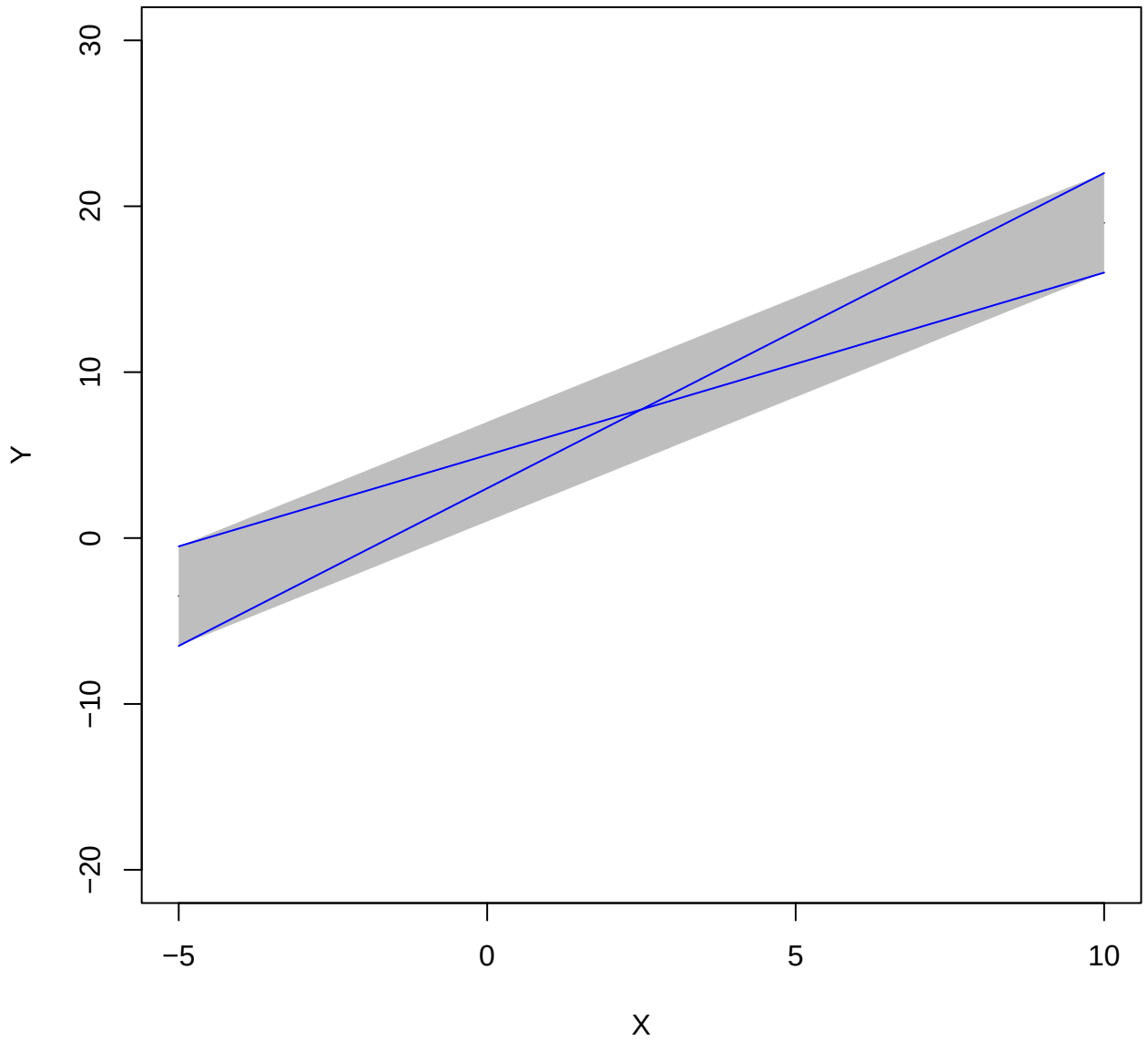
$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \cdot \frac{\frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2}{\frac{1}{n}\sum_{i=1}^{n}\left[(X_i - \bar{X})^2\right]}$$

- Your standard error is driven by 3 things
  - Sample size (n) – just like sample mean!
    - Larger samples will have smaller standard errors – more information reduces uncertainty
  - Variance of residuals (numerator of expression) – very similar to sample mean!
    - The noisier the data (higher variance), the greater the uncertainty
  - Variance of explanatory variable (denominator of expression)
    - To see how Y varies when X varies, we need X to vary! The more it does, the less uncertainty we have about the relationship.
    - Imagine if all observations had the same X. You literally could not estimate a relationship between X and Y (variance of ∞).

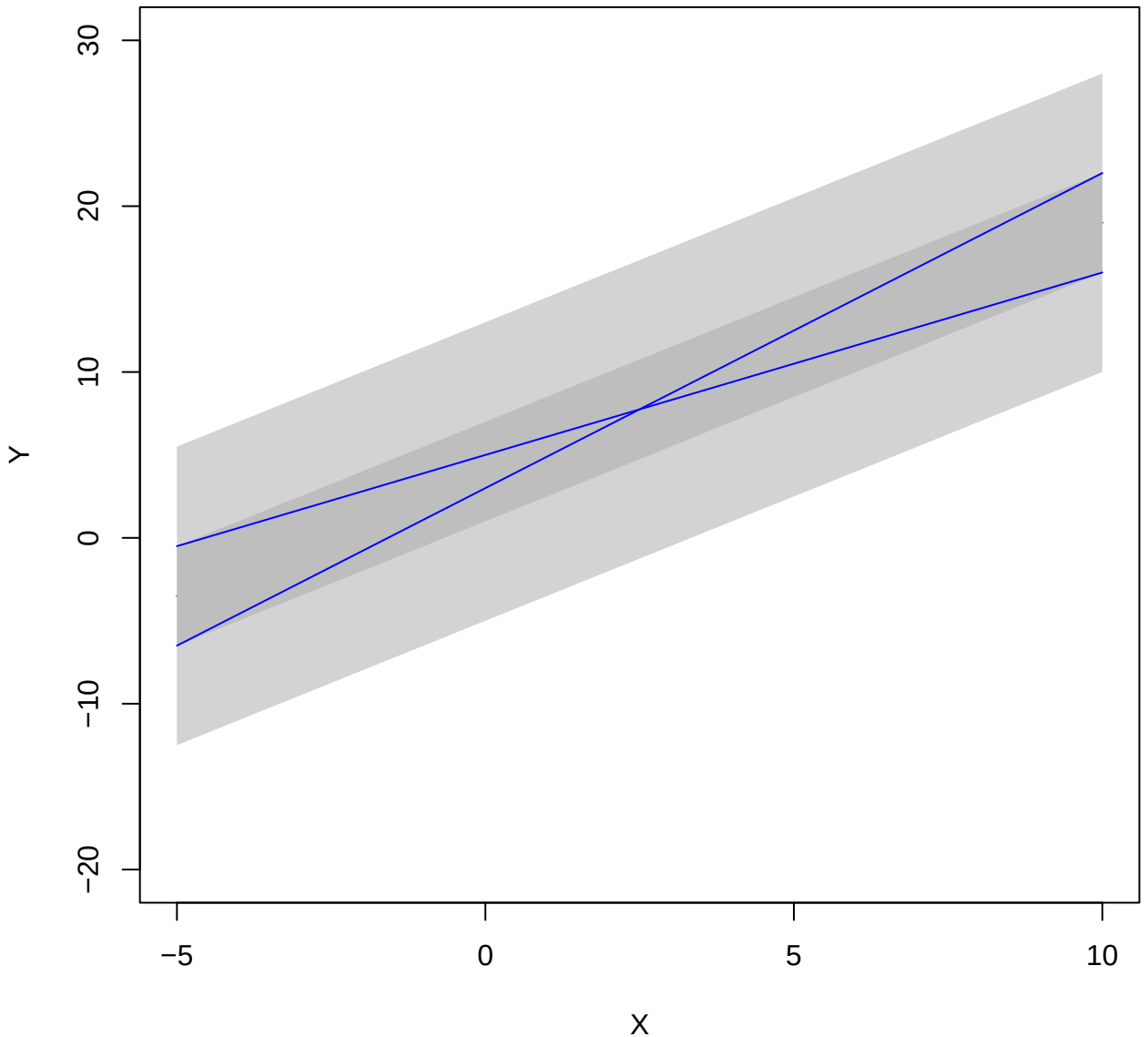
Effect of larger error variance



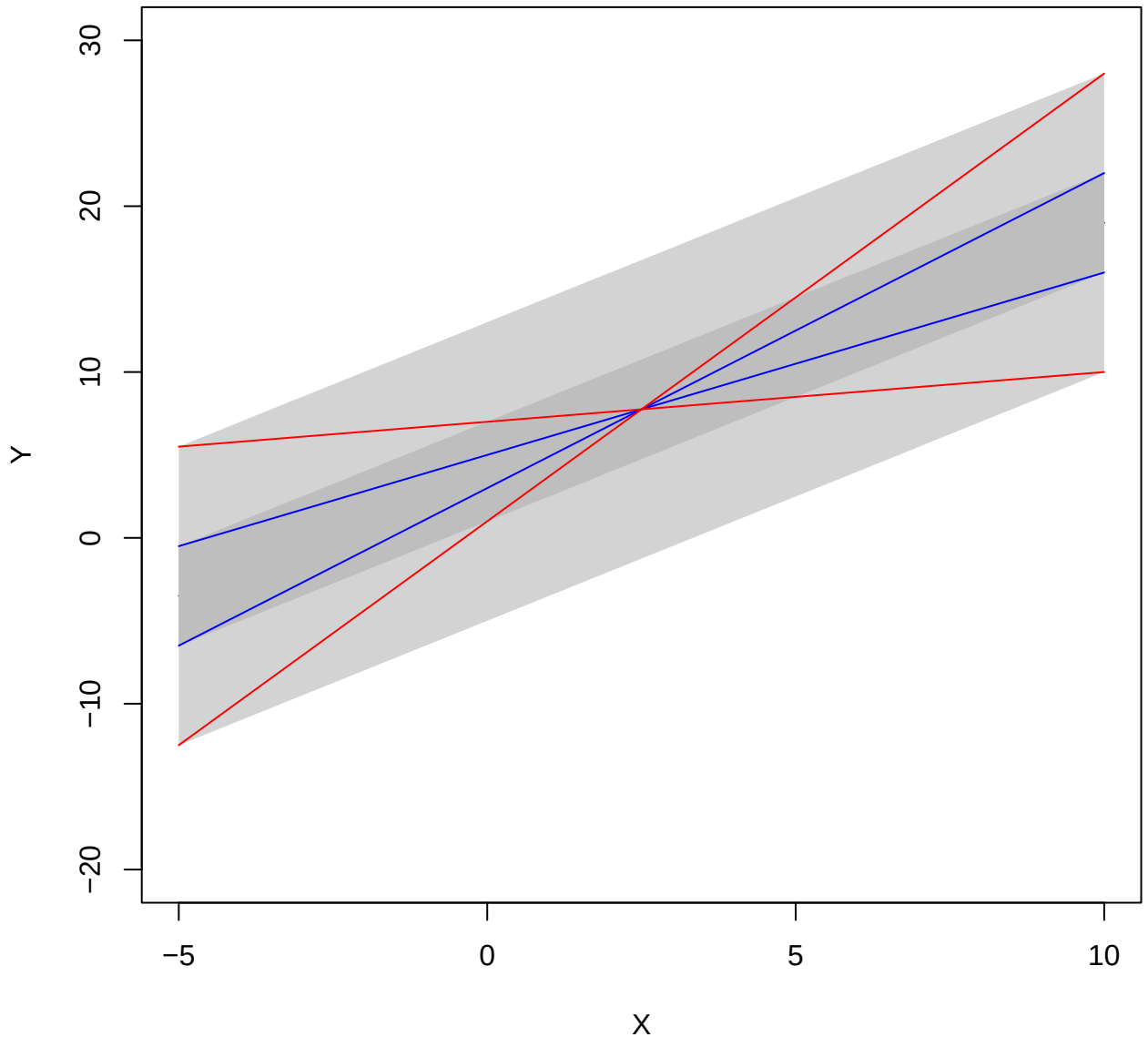
Effect of larger error variance



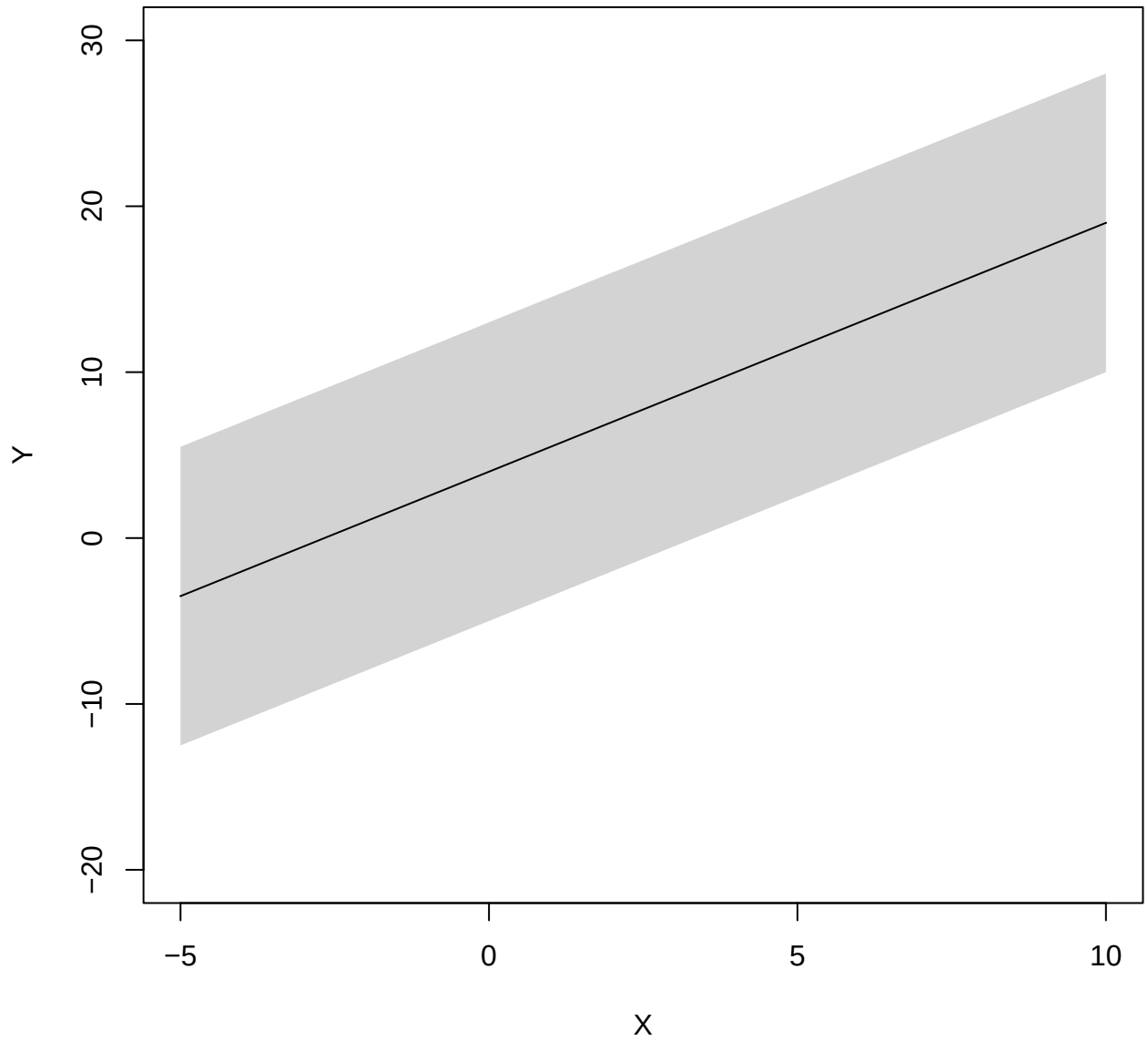
Effect of larger error variance



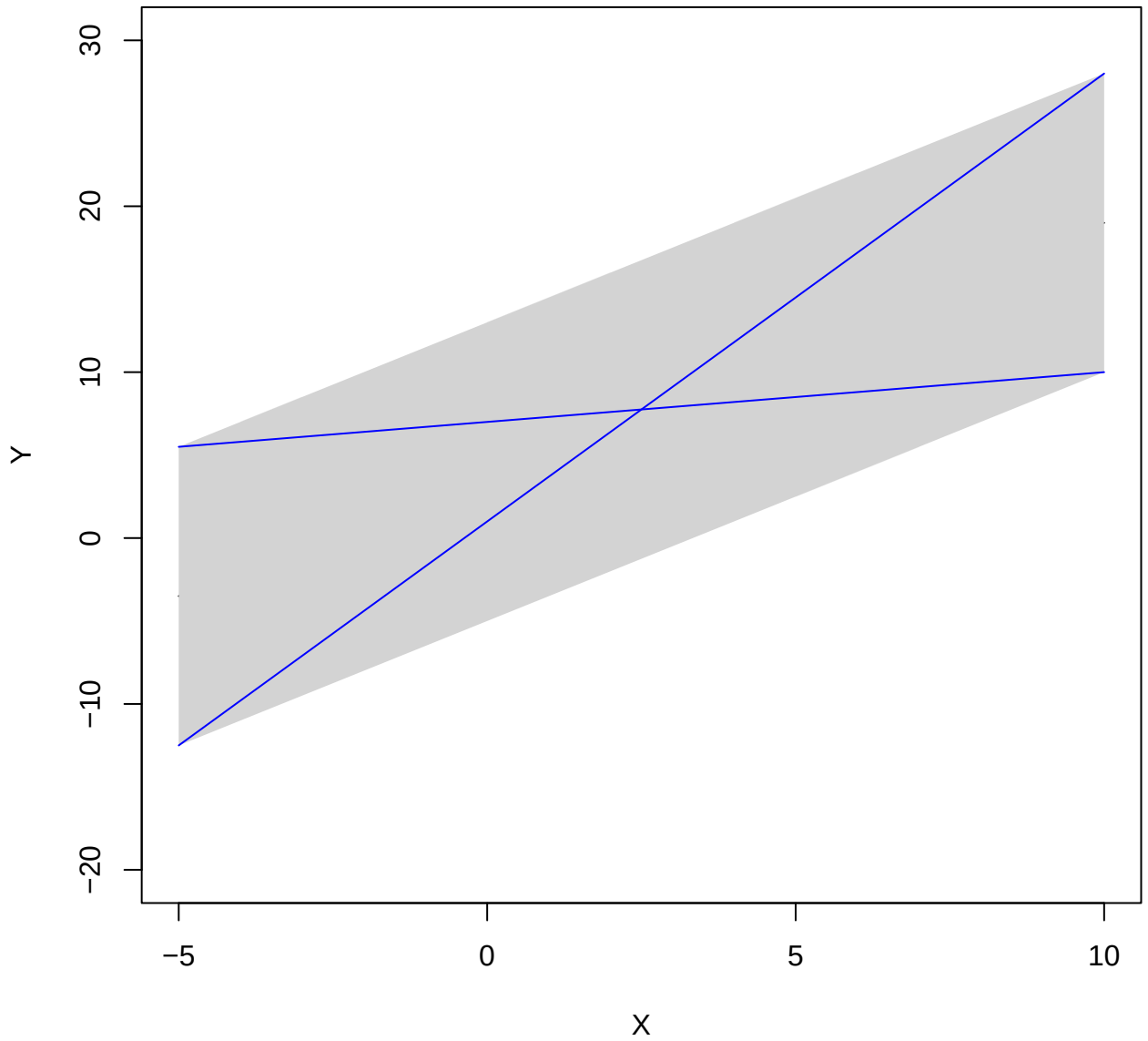
Effect of larger error variance



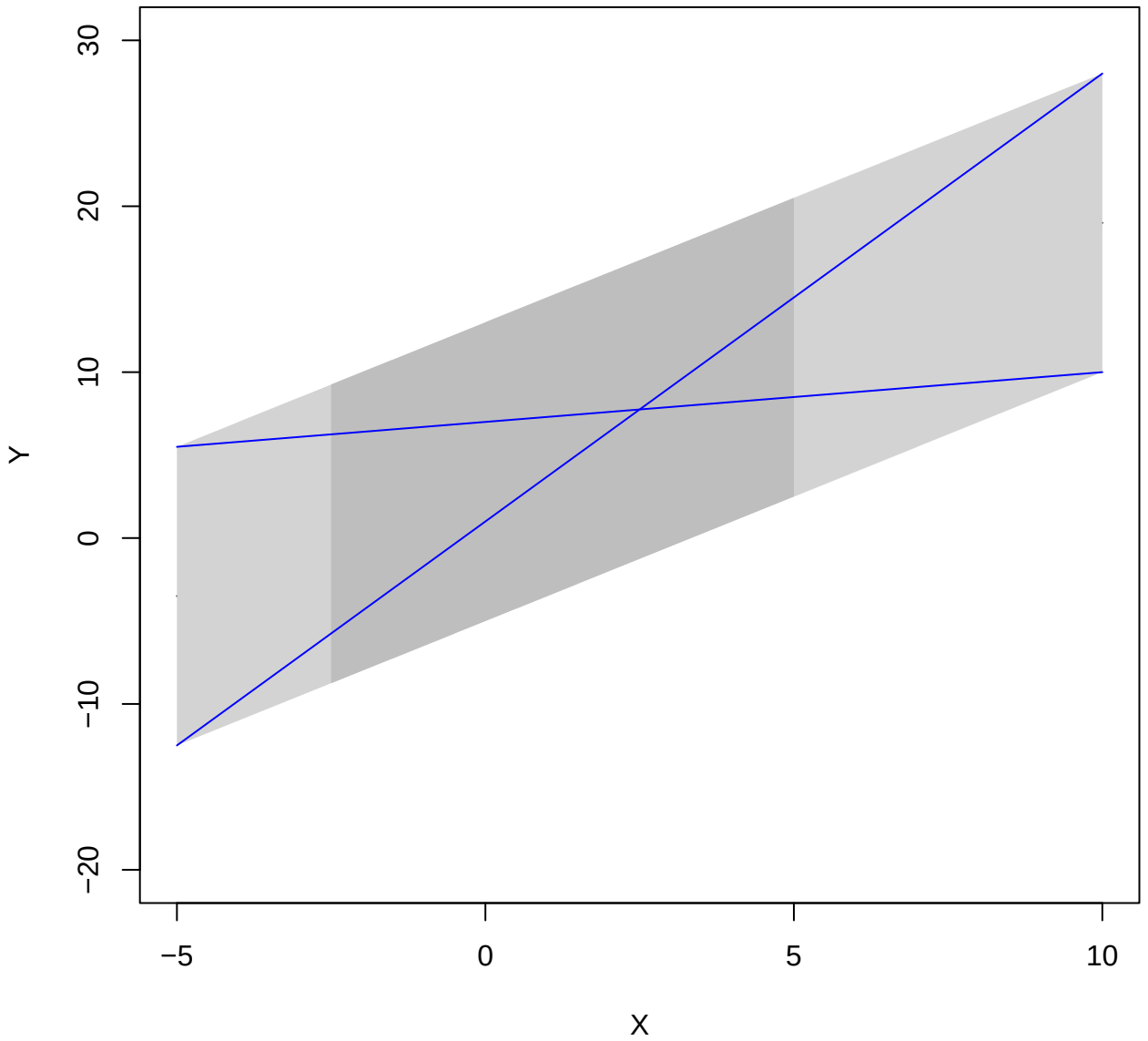
Effect of larger variance in X



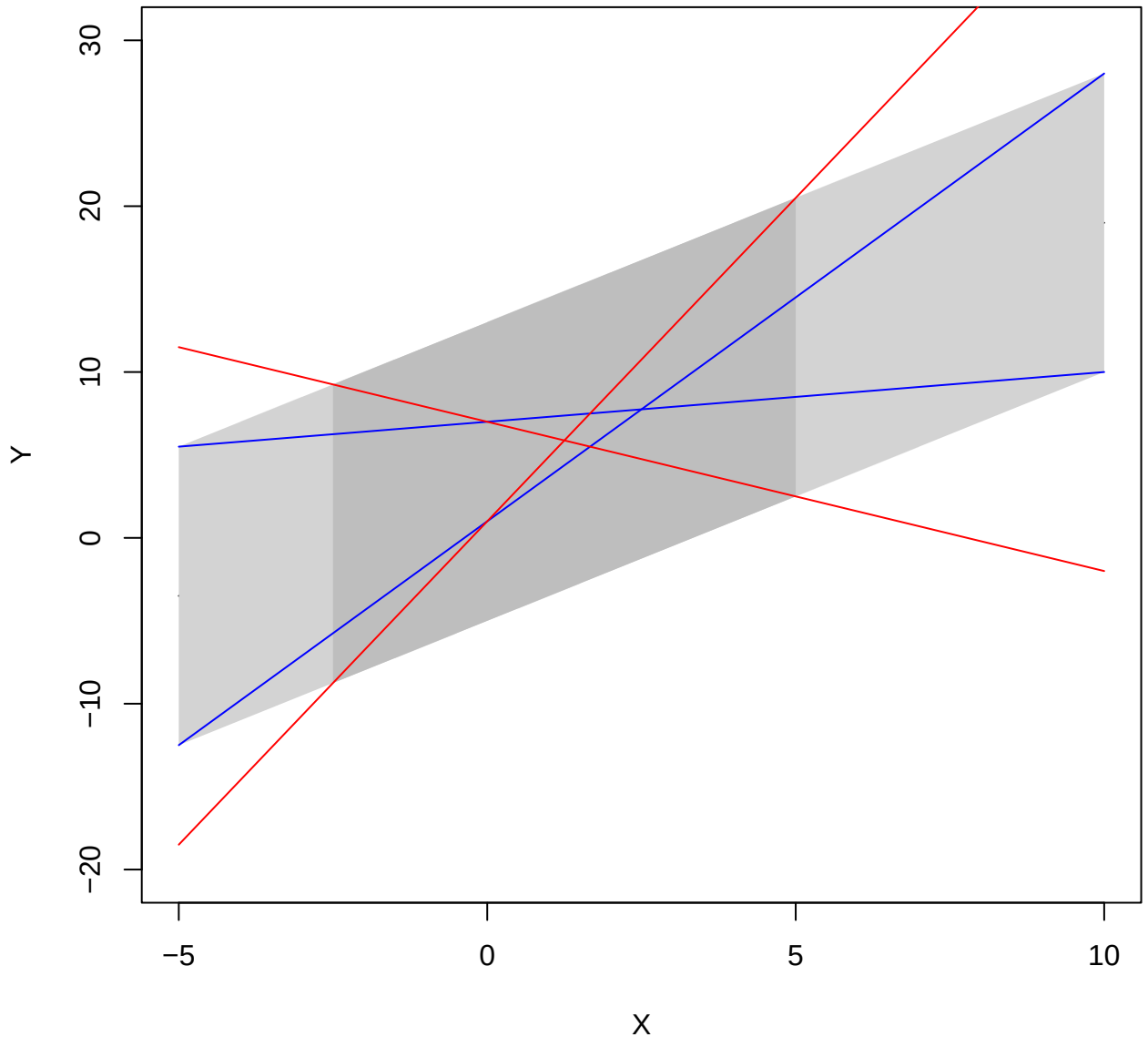
Effect of larger variance in X



Effect of larger variance in X



Effect of larger variance in X



Residual regression

Consider the linear regression function:

$$E[y] = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_K \cdot x_K \tag{22}$$

Consider an auxilary regression of one x on all others (WLOG, just use x_1):

$$x_1 = \alpha_0 + \alpha_2 \cdot x_2 + ... + \alpha_K \cdot x_K + \epsilon_{x_1} \tag{23}$$

Finally, consider this "residual regression:"

$$E[y] = \eta_0 + \eta_1 \cdot \epsilon_{x_1} \tag{24}$$

You will show the following on your homework:

1. $\eta_0 = \bar{y}$
2. $\eta_1 = \beta_1$

Residual regression, explained

Consider a variable included in a multivariate regression (e.g. x_1)...

- ...it has some coefficient in that regression (e.g. β_1).

That coefficient is the same as the *univariate* coefficient from a regression of y on:

- the-part-of-x_1-that-the-other-xs-can't-explain

In other words, β_1 is "identified off of" the variation in x_1 that is uncorrelated with the other xs.

- People with higher education may earn more...
- But to the extent that they also have higher IQs and that explains some of their high income...
- We're not going to let that affect $\beta_{\text{Education}}$

The most important idea in applied empirical analysis

Applied econometrics is all about seeing how variation in x explains variation in y.
Choosing an empirical approach is all about answering the question:

- "What part of x's variation should I use to 'identify' its effect on y?"

Potential answers:

1. All of it: Univariate Regression
2. The part that is not related to other observables: Multivariate Regression
3. The part that occurs within some categorical partition: Fixed Effects, Diff-in-Diff
4. A part that is good-as-random: Instrumental Variables
5. A part that is actually random: Experiments

All of these "techniques" are just formalizations of deeper ideas about how you want to identify x's "impact" on y.

Fixed Effects, example

$$E[\ln(\text{Earn})_{ij}] = \beta_0 + \beta_1 \cdot \text{Female}_i + \beta_2 \cdot \text{Education}_i + \alpha_j \quad (25)$$

Modeling (log) earnings of person i who works in occupation j as a function of their gender, education, and an "occupation Fixed Effect."

- Technically, the regression is using a series of indicator variables, $I_{j=1}$,...$I_{j=J}$, and so there is actually a large sum:

$$\sum_{k=1}^{J} \alpha_k \cdot I_{j=k}$$

- Since the indicators are all 0 except for occupation j, people often use the shorthand in Equation 25

If $\hat{\beta}_1 = -0.2$, we estimate that women earn 20% less than men *within occupations*, controlling for education

Goldin (2014)

TABLE 1—RESIDUAL GENDER DIFFERENCES IN EARNINGS AND THE ROLE OF OCCUPATION

Sample	Variables included	Coefficient on female	Standard error	R^2
Full-time	Basic	−0.248	0.00101	0.112
Full-time	Basic, time	−0.193	0.00100	0.163
Full-time	Basic, time, education	−0.247	0.000905	0.339
Full-time	Basic, time, education, occupation	−0.192	0.00104	0.453
All	Basic	−0.320	0.00105	0.102
All	Basic, time	−0.196	0.000925	0.353
All	Basic, time, education	−0.245	0.000847	0.475
All	Basic, time, education, occupation	−0.191	0.000963	0.563
Full-time, BA	Basic	−0.285	0.00159	0.131
Full-time, BA	Basic, time	−0.230	0.00158	0.177
Full-time, BA	Basic, time, education	−0.233	0.00155	0.216
Full-time, BA	Basic, time, education, occupation	−0.163	0.00158	0.374
All, BA	Basic	−0.384	0.00173	0.119
All, BA	Basic, time	−0.227	0.00151	0.380
All, BA	Basic, time, education	−0.229	0.00148	0.407
All, BA	Basic, time, education, occupation	−0.163	0.00151	0.525

Notes: "Basic" regression is the log of annual earnings regressed on the female dummy, age as a quartic, race, and year. "Time" adds log hours per week and log weeks. "Education" adds dummies for education categories (and those above a BA for the college graduate sample). "Occupation" adds three-digit occupation dummies. "Full-time" is 35 and above hours per week and 40 and above weeks per year. "All" includes workers 25 to 64 years old with positive earnings and positive hours worked during the past year. The "full-time" sample consists of full-time, full-year individuals 25 to 64 years old excluding those in the military using trimmed annual earnings data (exceeding 1,400 hours × 0.5 × 2009 minimum wage). The "BA" sample includes workers with at least a college or university bachelor's degree. The number of observations is 2,603,968 for full-time, 3,291,168 for all, 964,705 for full-time BA or more, and 1,162,638 for all BA or more.

Diff-in-Diff

Suppose your data spans before and after some treatment in $t = T$.

- Some subjects never receive treatment ($\text{Treat}_i = 0$)
- Others ($\text{Treat}_i = 1$) are untreated for $t < T$ and treated for $t \geq T$

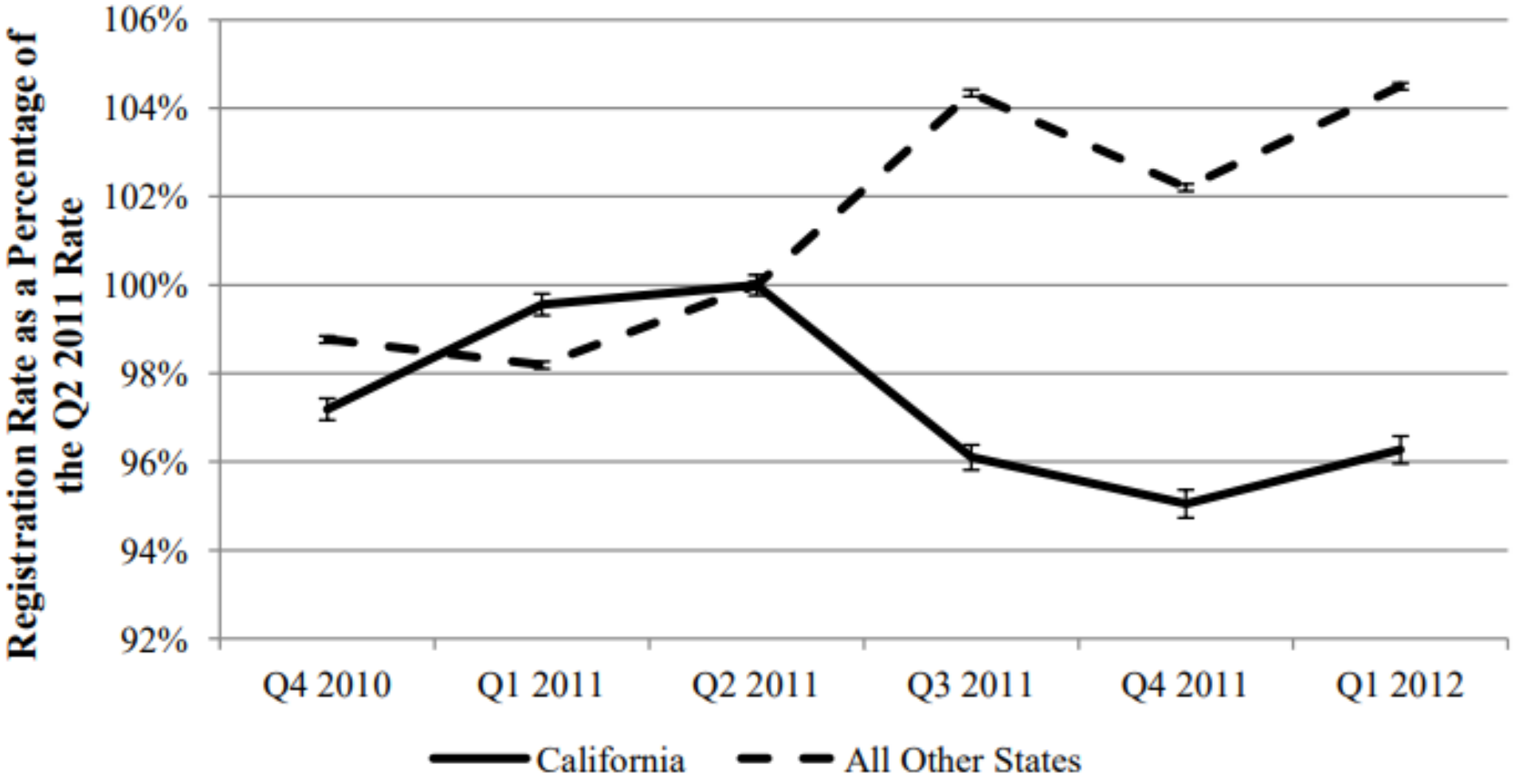
Defining $\text{Post}_t = 1$ if $t \geq T$ and 0 otherwise, run:

$$E[Y_{it}] = \beta_0 + \beta_1 \cdot \text{Treat}_i + \beta_2 \cdot \text{Post}_t + \beta_3 \cdot (\text{Treat}_i \cdot \text{Post}_t) \tag{26}$$

β_3 represents:

- The average change in Y for the treated group from pre- to post-treatment, *minus the change experienced by the control group.*
- This is known as a **difference-in-differences estimator**.
  Also sometimes called **two-way fixed effects (TWFE) estimator** because it's evaluating the impact:
  1. Within the treatment group, across time periods
  2. Within the treatment period, across groups

Kessler and Roth (2014)



Kessler and Roth (2014)

Table 2: Change from Opt-in to Active Choice on CA Registration Decisions

	Organ Donor Registration Rates			
	Quarterly Rates by State		Registration Decisions	
	(1)	(2)	(3)	(4)
Post*California	-0.022	-0.022	-0.027	-0.024
	(0.006)***	(0.007)***	(0.007)***	(0.007)***
Post	0.014	0.014	0.019	0.015
	(0.006)**	(0.007)**	(0.007)**	(0.007)***
California	-0.174		-0.123	
	(0.031)***		(0.042)***	
Constant	0.445	0.439	0.394	0.380
	(0.031)***	(0.003)***	(0.042)***	(0.003)***
State FE	No	Yes	No	Yes
Observations	162	162	65,856,108	65,856,108

Endogeneity problem

Structural model:

$$Y_i = \beta_0 + \beta_1 \cdot T_i + \epsilon_i, \tag{27}$$

with $corr(T_i, \epsilon_i) \neq 0$
$\hat{\beta}_1^{OLS} \neq \beta_1$. Orthogonality condition yields:

$$E\left[T_i \cdot \hat{\epsilon}_i^{OLS}\right] = 0 \tag{28}$$

- T_i is correlated with ϵ_i but not with $\hat{\epsilon_i}^{OLS}$
- We recovered bad estimates of ϵ
- ϵ and β are jointly estimated, so estimate of β is bad, too

Instrumental variables

Consider a variables Z such that:

1. $corr(Z_i, T_i) \neq 0$ ("Relevance Condition")
   - Z_i is predictive of T_i
2. $corr(Z_i, \epsilon_i) = 0$ ("Exclusion Condition")
   - Z_i has to direct impact on Y_i (not hidden in ϵ_i)
     - This is **not** testable: we don't observe ϵ_i

Using an Instrumental Variable

Structural model:

$$Y_i = \beta_0 + \beta_1 \cdot T_i + \epsilon_i,$$

Consider 3 regressions:

$$E[T_i] = \beta_0^{FS} + \beta_1^{FS} \cdot Z_i \tag{29}$$

$$E[Y_i] = \beta_0^{RF} + \beta_1^{RF} \cdot Z_i \tag{30}$$

$$E[Y_i] = \beta_0^{SS} + \beta_1^{SS} \cdot \hat{T}_i, \tag{31}$$

where $\hat{T}_i$ are fitted values from Regression 29.
You will show on HW that there are two ways to accurately estimate β_1:

1. $\hat{\beta}_1 = \hat{\beta}_1^{RF} / \hat{\beta}_1^{FS}$
2. $\hat{\beta}_1 = \hat{\beta}_1^{SS}$

An intuition for why the IV helps

$$Y_i = \beta_0 + \beta_1 \cdot T_i + \epsilon_i,$$

$$E[T_i] = \beta_0^{FS} + \beta_1^{FS} \cdot Z_i$$

$$E[Y_i] = \beta_0^{SS} + \beta_1^{SS} \cdot \hat{T}_i,$$

$\hat{T}$ retains only the variation in T that can be explained by Z

- It is the part of T being driven by forces that have no independent impact on Y

So when we use only that part of T – the as-good-as-random part – it is like we are randomly assigning T

We used the IV to dump a bunch of the variation in T that we thought could be tainted (i.e. correlated with other determinants of Y hidden inside ϵ)

Abel and Fuster (2021)

TABLE 3—REGRESSION ESTIMATES OF THE TREATMENT EFFECT O
MORTGAGE DEFAULT AND NON-MORTGAG

	Panel A. Mortgage default			
	(1)	(2)	(3)	(4)
IV				
Basis points	−2.68	−2.46	−2.17	−3.94
(SE)	(0.77)	(0.77)	(0.77)	(1.03)
OLS				
Basis points	−4.30	−4.11	−4.10	−4.27
(SE)	(0.25)	(0.27)	(0.26)	(0.29)
Quarter FEs	✓	✓	✓	✓
Zip-code FEs	✓	✓	✓	✓
Observables	✓	✓	✓	✓
Q-by-zip FEs		✓	✓	✓
Guar. lag FEs			✓	✓
Cohort FEs				✓

The Great Tensions of applied empirical work in economics

To estimate the causal effect of x on y, economists do some combination of:

1. Strip out variation in x that is correlated with other factors (i.e. controls)
2. Focus on variation in x that is driven by random factors (i.e. instruments)
3. Focus on settings in which techniques 1 and 2 will be most effective (e.g. policy change)

All of these amount to ignoring some of x's variation. This has 2 major negative consequences:

1. Statistical power: losing variation in x can drive up the SE on its coefficient
   - Many clever empirical ideas have been thwarted by large SEs
2. External relevance: by focusing on a very specific setting, you may get the right causal effect in that setting, but you have to think about whether it applies elsewhere