Ordinary Least Squares (Linear) Regression

Department of Political Science and Government
Aarhus University

February 17, 2015

Uses of Regression

1. Description
2. Prediction
3. Causal Inference

Descriptive Inference

1. We want to understand a *population* of cases
2. We cannot observe them all, so:
   1. Draw a *representative* sample
   2. Perform mathematical procedures on sample data
   3. Use assumptions to make inferences about population
   4. Express uncertainty about those inferences based on assumptions

Parameter Estimation

- We want to observe population *parameter* θ
- If we obtain a representative sample of population units:
  - Our sample statistic $\hat{\theta}$ is an unbiased estimate of θ
  - Our sampling procedure dictates how uncertain we are about the value of θ

An Example

- We want to know $\bar{Y}$ (population mean)
- Our *estimator* is the sample mean formula which produces the sample *estimate* $\bar{y}$:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{1}$$

- The *sampling variance* is our uncertainty:

$$Var(\bar{y}) = \frac{s^2}{n} \tag{2}$$

where s^2 = sample element variance

Uncertainty

- We never know θ
- Our $\hat{\theta}$ is an estimate that may not equal θ
  - Unbiased due to **Law of Large Numbers**
  - For $\bar{y}$: $N(Y, \sigma^2)$
- The size of sampling variance depends on:
  - Element variance
  - Sample size!
- Note: $SE(\bar{y}) = \sqrt{Var(\bar{y})}$
- We may want to know $\hat{\theta}$ per se, but we are mostly interested in it as an estimate of θ

Causal Inference

Causal Inference

1. Everything that goes into descriptive inference

Causal Inference

1. Everything that goes into descriptive inference
2. Plus, philosophical assumptions

Causal Inference

1. Everything that goes into descriptive inference
2. Plus, philosophical assumptions
3. Plus, randomization *or* perfectly specified model

Questions about philosophical assumptions?

Ways of Thinking About OLS

Ways of Thinking About OLS

1. Estimating Unit-level Causal Effect

Ways of Thinking About OLS

1. Estimating Unit-level Causal Effect
2. Ratio of $Cov(X, Y)$ and $Var(X)$

Ways of Thinking About OLS

1. Estimating Unit-level Causal Effect
2. Ratio of $Cov(X, Y)$ and $Var(X)$
3. Minimizing residual sum of squares (SSR)

Ways of Thinking About OLS

1. Estimating Unit-level Causal Effect
2. Ratio of $Cov(X, Y)$ and $Var(X)$
3. Minimizing residual sum of squares (SSR)
4. Line (or surface) of best fit

Bivariate Regression I

- Y is continuous
- X is a randomized treatment indicator/dummy $(0, 1)$
- How do we know if the treatment X had an effect on Y?

Bivariate Regression I

- Y is continuous
- X is a randomized treatment indicator/dummy $(0, 1)$
- How do we know if the treatment X had an effect on Y?
- Look at mean-difference: $E[Y_i|X_i = 1] - E[Y_i|X_i = 0]$

Three Equations

1. Population: $Y = \beta_0 + \beta_1 X \;\; (+\epsilon)$

2. Sample estimate: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

3. Unit:

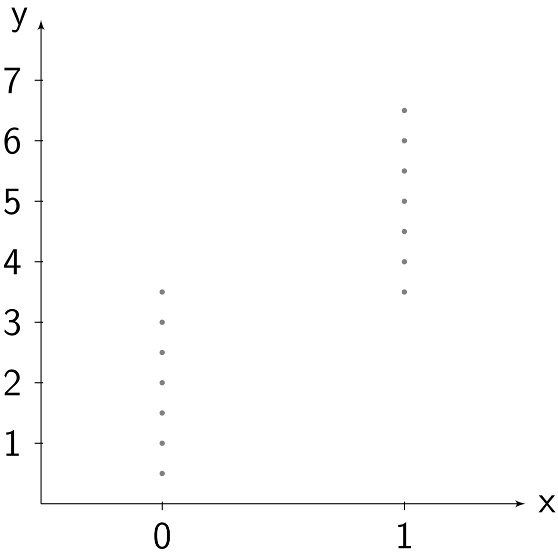
$$
\begin{aligned}
y_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \\
&= \bar{y}_{0i} + (y_{1i} - y_{0i}) x_i + (y_{0i} - \bar{y}_{0i})
\end{aligned}
$$

Bivariate Regression I

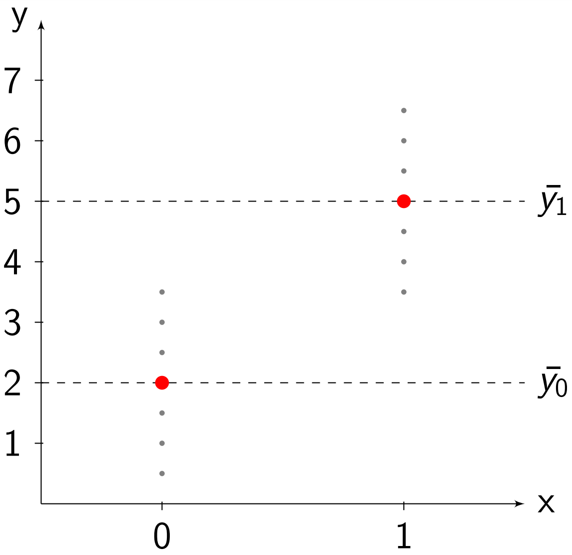
- Mean difference $(E[Y_i|X_i = 1] - E[Y_i|X_i = 0])$ is the regression line slope
- Slope (β) defined as $\frac{\Delta Y}{\Delta X}$

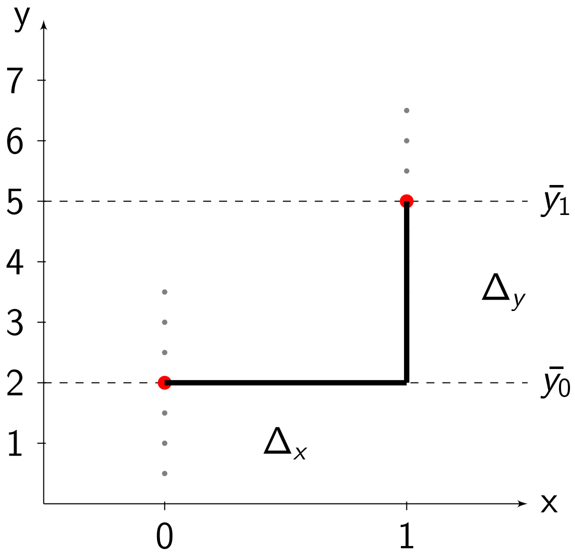
Bivariate Regression I

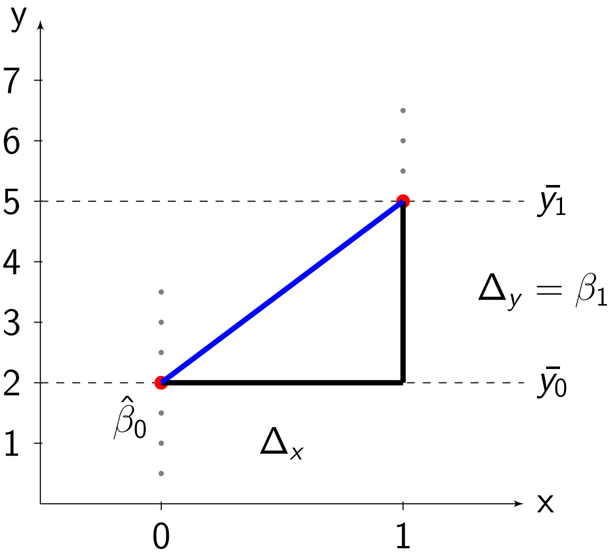
- Mean difference ($E[Y_i|X_i = 1] - E[Y_i|X_i = 0]$) is the regression line slope
- Slope (β) defined as $\frac{\Delta Y}{\Delta X}$
  - $\Delta Y = E[Y_i|X = 1] - E[Y_i|X = 0]$
  - $\Delta X = 1 - 0 = 1$

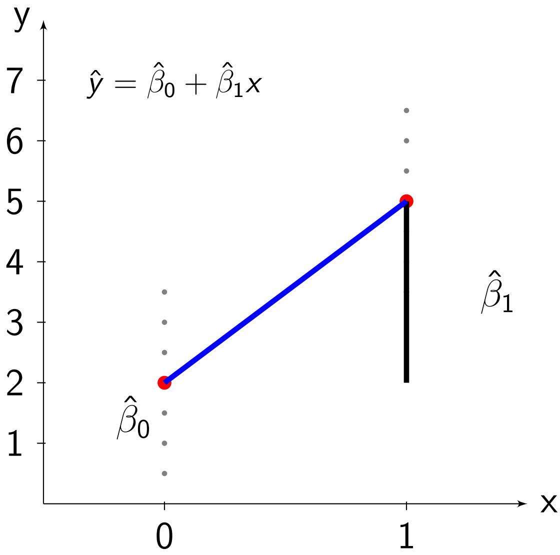


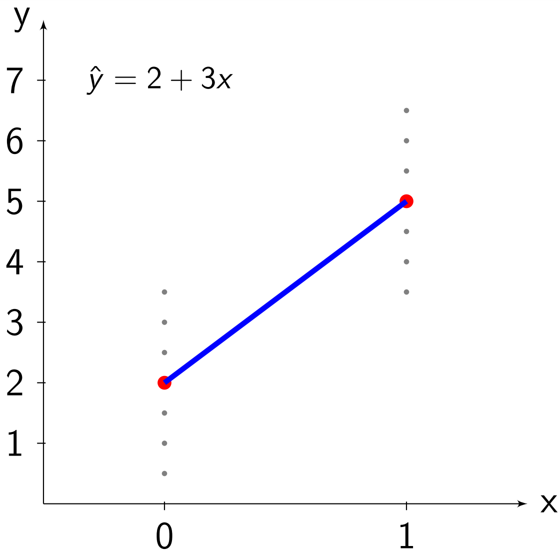
y
7
6
5
4
3
2
1
0
1
x

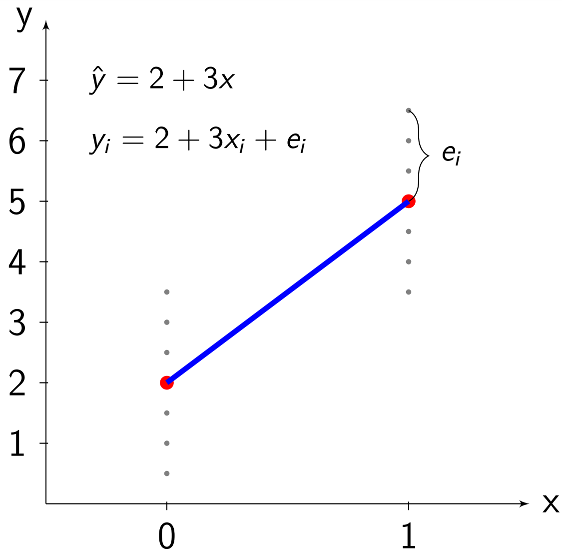

y
7
6
5
4
3
2
1
0
1
x
$\bar{y}_1$
$\bar{y}_0$


y
7
6
5
4
3
2
1
0
1
x
$\bar{y}_1$
$\bar{y}_0$
Δ_y
Δ_x


y
7
6
5
4
3
2
1
0
1
x
$\bar{y}_1$
$\bar{y}_0$
$\Delta_y = \beta_1$
$\hat{\beta}_0$
Δ_x


y
7
6
5
4
3
2
1
0
1
x
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
$\hat{\beta}_0$
$\hat{\beta}_1$


y
x
$\hat{y} = 2 + 3x$
1
2
3
4
5
6
7
0
1


y
7
6
5
4
3
2
1
0
1
x
$\hat{y} = 2 + 3x$
$y_i = 2 + 3x_i + e_i$
e_i

Systematic versus unsystematic component of the data

- Systematic: Regression line (slope)
  - Linear regression estimates the conditional means of the population data (i.e., $E[Y|X]$)

- Unsystematic: Error term is the deviation of observations from the line
  - The difference between each value y_i and $\hat{y}_i$ is the *residual*: e_i
  - OLS produces an estimate of the relationship between X and Y that minimizes the *residual sum of squares*

Why are there residuals?

Why are there residuals?

- Omitted variables
- Measurement error
- Fundamental randomness

Bivariate Regression I

- Mean difference $(E[Y_i|X_i = 1] - E[Y_i|X_i = 0])$ is the regression line slope
- Slope (β) defined as $\frac{\Delta Y}{\Delta X}$
  - $\Delta Y = E[Y_i|X = 1] - E[Y_i|X = 0]$
  - $\Delta X = 1 - 0 = 1$

Bivariate Regression I

- Mean difference $(E[Y_i|X_i = 1] - E[Y_i|X_i = 0])$ is the regression line slope
- Slope (β) defined as $\frac{\Delta Y}{\Delta X}$
  - $\Delta Y = E[Y_i|X = 1] - E[Y_i|X = 0]$
  - $\Delta X = 1 - 0 = 1$
- How do we know if this is a *significant* difference?
  - We'll come back to that

Ways of Thinking About OLS

1. Estimating Unit-level Causal Effect

Ways of Thinking About OLS

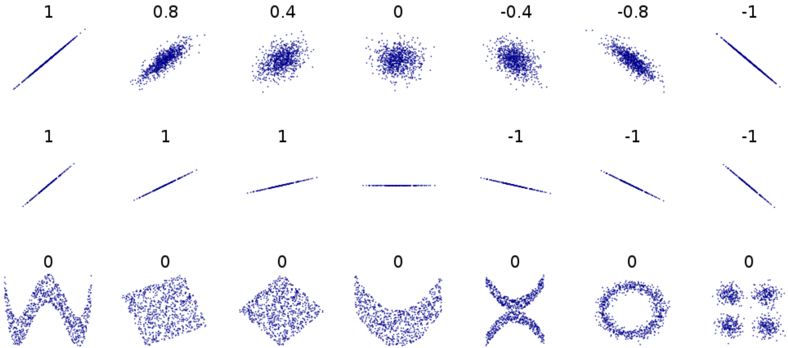
1. Estimating Unit-level Causal Effect
2. Ratio of $Cov(X, Y)$ and $Var(X)$

Bivariate Regression II

- Y is continuous
- X is continuous (and randomized)
- How do we know if the treatment X had an effect on Y?
  - Correlation coefficient (ρ)
  - Regression coefficient (slope; β_1)

Correlation Coefficient (ρ)

- Measures how well a scatterplot is represented by a straight (non-horizontal) line


1
0.8
0.4
0
-0.4
-0.8
-1
1
1
1
-1
-1
-1
0
0
0
0
0
0
0

Correlation Coefficient (ρ)

- Measures how well a scatterplot is represented by a straight (non-horizontal) line

Correlation Coefficient (ρ)

- Measures how well a scatterplot is represented by a straight (non-horizontal) line
- Formal definition: $\frac{Cov(X,Y)}{\sigma_X \sigma_y}$
- As a reminder:
  - $Cov(x, y) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
  - $s_x = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

OLS Coefficient (β_1)[1]

- Measures ΔY given ΔX

[1] Multivariate formula involves matrices; Week 20

OLS Coefficient (β_1)[1]

- Measures ΔY given ΔX
- Formal definition: $\frac{Cov(X,Y)}{Var(X)}$
- As a reminder:
  - $Cov(x, y) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
  - $Var(x) = \sum_{i=1}^{n}(x_i - \bar{x})^2$

[1] Multivariate formula involves matrices; Week 20

OLS Coefficient (β_1)[1]

- Measures ΔY given ΔX
- Formal definition: $\frac{Cov(X,Y)}{Var(X)}$
- As a reminder:
  - $Cov(x, y) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
  - $Var(x) = \sum_{i=1}^{n}(x_i - \bar{x})^2$
- $\hat{\rho}$ and $\hat{\beta}_1$ are just scaled versions of $\widehat{Cov}(x, y)$

[1] Multivariate formula involves matrices; Week 20

Minimum Mathematical Requirements

1. Do we need variation in X?

Minimum Mathematical Requirements

1. Do we need variation in X?
   - Yes, otherwise dividing by zero

Minimum Mathematical Requirements

1. Do we need variation in X?
   - Yes, otherwise dividing by zero
2. Do we need variation in Y?
   - No, $\hat{\beta}_1$ can equal zero

Minimum Mathematical Requirements

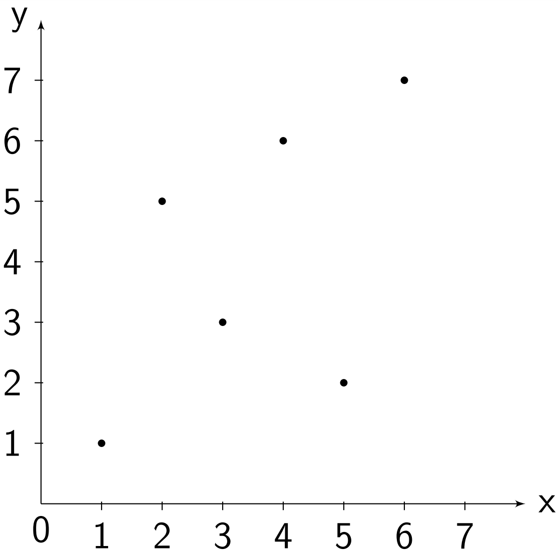
1. Do we need variation in X?
   - Yes, otherwise dividing by zero
2. Do we need variation in Y?
   - No, $\hat{\beta}_1$ can equal zero

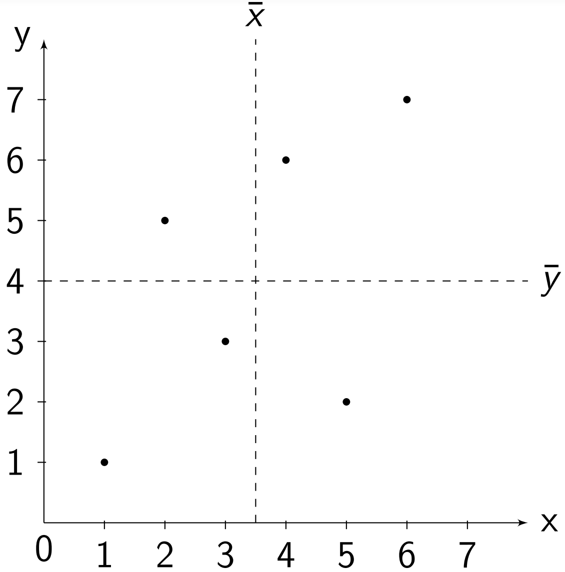
Minimum Mathematical Requirements

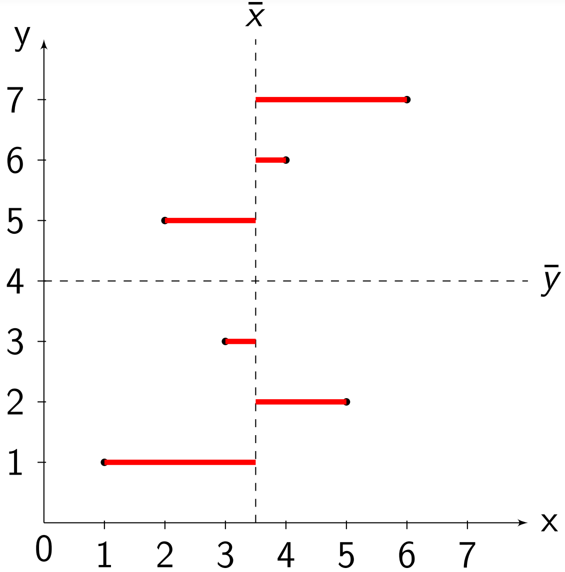
1. Do we need variation in X?
   - Yes, otherwise dividing by zero
2. Do we need variation in Y?
   - No, $\hat{\beta}_1$ can equal zero
3. How many observations do we need?

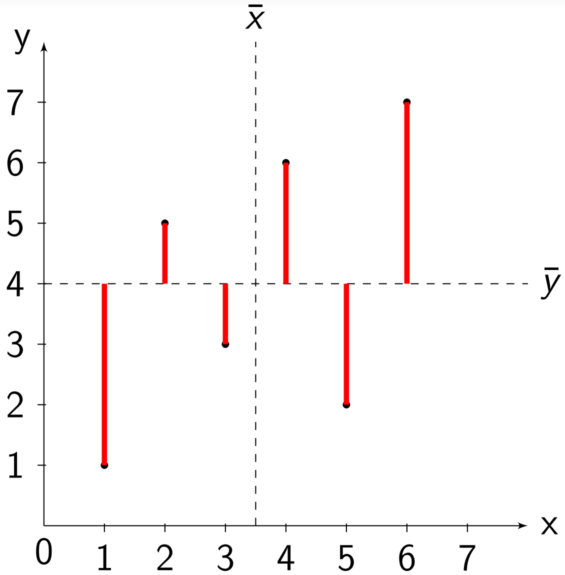
Minimum Mathematical Requirements

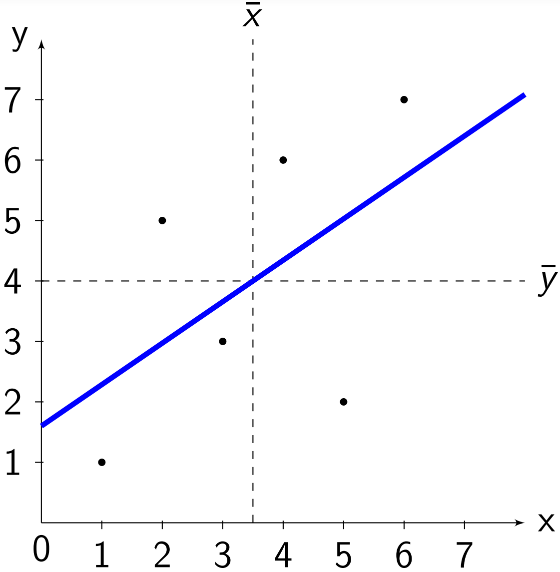
1. Do we need variation in X?
   - Yes, otherwise dividing by zero
2. Do we need variation in Y?
   - No, $\hat{\beta}_1$ can equal zero
3. How many observations do we need?
   - $n \geq k$, where k is number of parameters to be estimated


y
7
6
5
4
3
2
1
0
1 2 3 4 5 6 7
x


$\bar{x}$
y
7
6
5
4
3
2
1
$\bar{y}$
x
0
1 2 3 4 5 6 7


$\bar{x}$
y
7
6
5
4
3
2
1
$\bar{y}$
x
0
1
2
3
4
5
6
7


$\bar{x}$
y
7
6
5
4
3
2
1
$\bar{y}$
x
0
1
2
3
4
5
6
7


$\bar{x}$
y
7
6
5
4
3
2
1
$\bar{y}$
x
0
1
2
3
4
5
6
7

Calculations

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	1	?	?	?	?
2	5	?	?	?	?
3	3	?	?	?	?
4	6	?	?	?	?
5	2	?	?	?	?
6	7	?	?	?	?

Intercept $\hat{\beta}_0$

- Simple formula: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Intercept $\hat{\beta}_0$

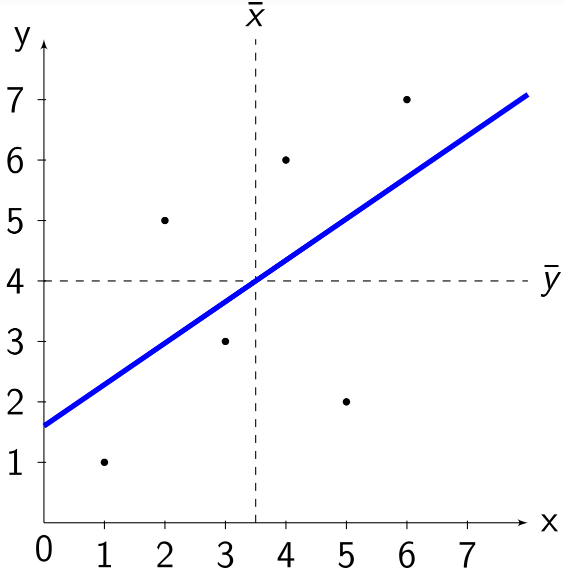
- Simple formula: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$
- Intuition: OLS fit always runs through point $(\bar{x}, \bar{y})$

Intercept $\hat{\beta}_0$

- Simple formula: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- Intuition: OLS fit always runs through point $(\bar{x}, \bar{y})$
- Ex.: $\hat{\beta}_0 = 4 - 0.6857 * 3.5 = 1.6$

Intercept $\hat{\beta}_0$

- Simple formula: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- Intuition: OLS fit always runs through point $(\bar{x}, \bar{y})$
- Ex.: $\hat{\beta}_0 = 4 - 0.6857 * 3.5 = 1.6$
- $\hat{y} = 1.6 + 0.6857\hat{x}$


$\bar{x}$
y
7
6
5
4
3
2
1
$\bar{y}$
x
0
1
2
3
4
5
6
7

Ways of Thinking About OLS

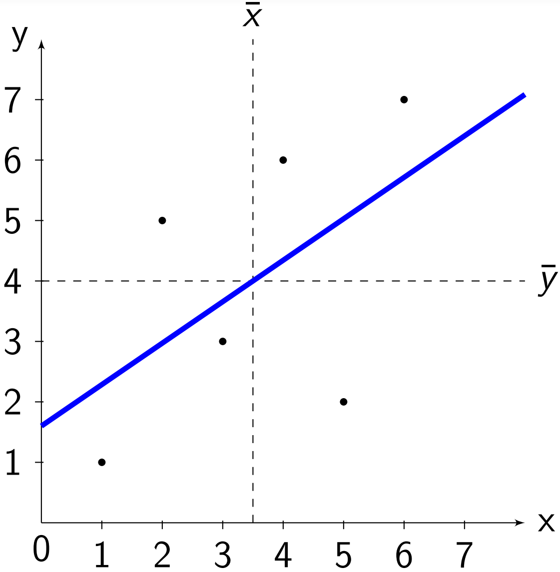
1. Estimating Unit-level Causal Effect
2. Ratio of $Cov(X, Y)$ and $Var(X)$

Ways of Thinking About OLS

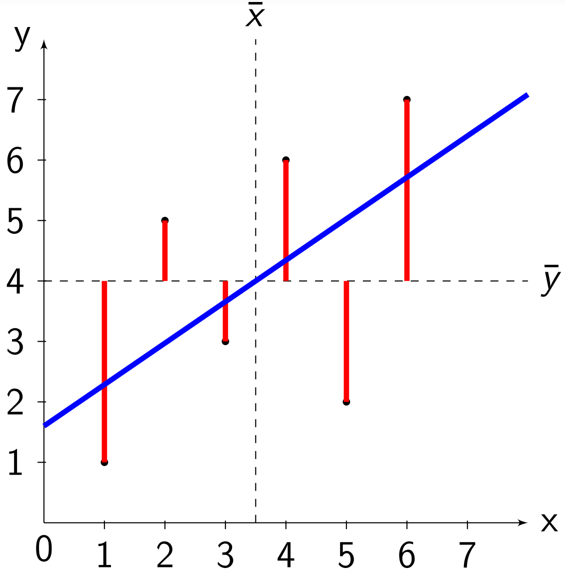
1. Estimating Unit-level Causal Effect
2. Ratio of $Cov(X, Y)$ and $Var(X)$
3. Minimizing residual sum of squares (SSR)

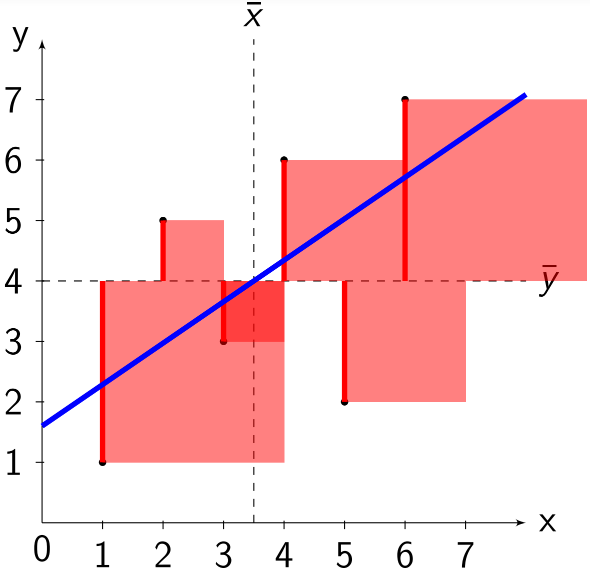
OLS Minimizes SSR

- Total Sum of Squares (SST): $\sum_{i=1}^{n}(y_i - \bar{y})^2$
- We can partition SST into two parts (ANOVA):
  - Explained Sum of Squares (SSE)
  - Residual Sum of Squares (SSR)
- $SST = SSE + SSR$
- OLS is the line with the lowest SSR

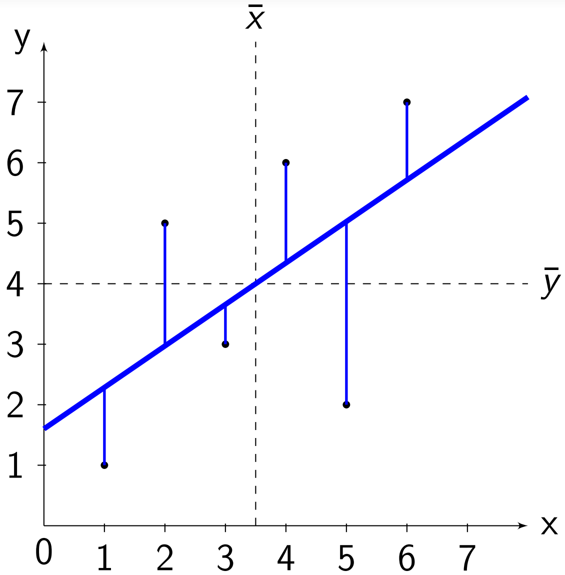


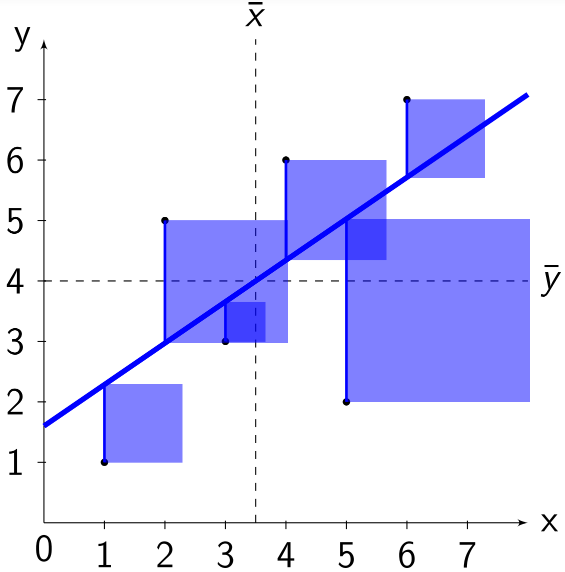
$\bar{x}$
y
7
6
5
4
3
2
1
$\bar{y}$
x
0
1
2
3
4
5
6
7

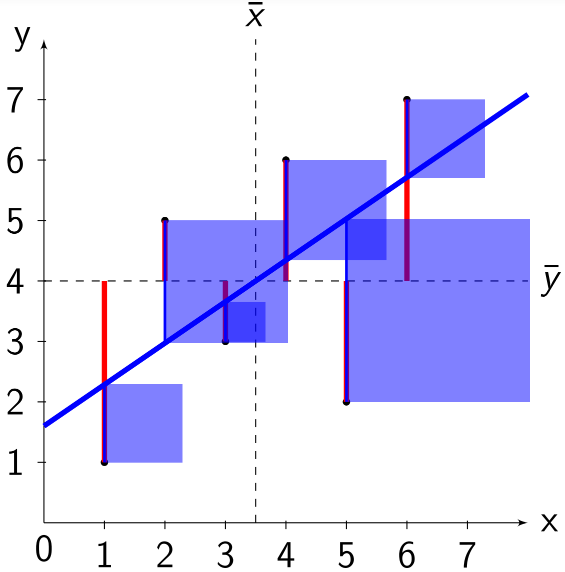

$\bar{x}$
y
7
6
5
4
3
2
1
$\bar{y}$
x
0
1
2
3
4
5
6
7

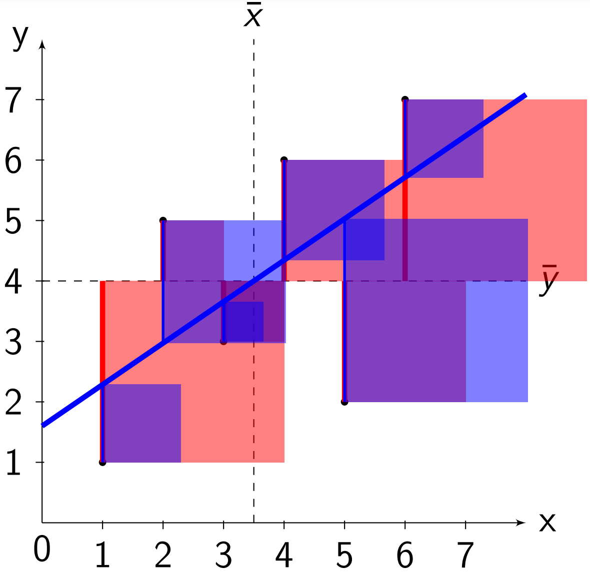


$\bar{x}$
y
x
$\bar{y}$
0
1
2
3
4
5
6
7


$\bar{x}$
y
7
6
5
4
3
2
1
$\bar{y}$
x
0
1
2
3
4
5
6
7


$\bar{x}$
y
7
6
5
4
3
2
1
0
$\bar{y}$
1 2 3 4 5 6 7
x


$\bar{x}$
y
7
6
5
4
3
2
1
0
1 2 3 4 5 6 7
x
$\bar{y}$


$\bar{x}$
y
7
6
5
4
3
2
1
0
1 2 3 4 5 6 7
x
$\bar{y}$

Questions about OLS calculations?

Are Our Estimates Any Good?

Yes, if:

1. Works mathematically
2. Causally valid theory
3. Linear relationship between X and Y
4. X is measured without error
5. No missing data (or MCAR; see Lecture 5)
6. No confounding

Linear Relationship

- If linear, no problems
- If non-linear, we need to transform
  - Power terms (e.g., x^2, x^3)
  - log (e.g., $log(x)$)
  - Other transformations
  - If categorical: convert to set of indicators
  - Multivariate interactions (next week)

Coefficient Interpretation Activity

- Four types of variables:
  1. Indicator (0,1)
  2. Categorical
  3. Ordinal
  4. Interval
- How do we interpret a coefficient on each of these types of variables?

Notes on Interpretation

- Effect β_1 is constant across values of x

Notes on Interpretation

- Effect β_1 is constant across values of x
- That is not true when there are:
  - Interaction terms (next week)
  - Nonlinear transformations (e.g., x^2)
  - Nonlinear regression models (e.g., logit/probit)

Notes on Interpretation

- Effect β_1 is constant across values of x
- That is not true when there are:
  - Interaction terms (next week)
  - Nonlinear transformations (e.g., x^2)
  - Nonlinear regression models (e.g., logit/probit)
- Interpretations are sample-level
  - Sample representativeness determines generalizability

Notes on Interpretation

- Effect β_1 is constant across values of x
- That is not true when there are:
  - Interaction terms (next week)
  - Nonlinear transformations (e.g., x^2)
  - Nonlinear regression models (e.g., logit/probit)
- Interpretations are sample-level
  - Sample representativeness determines generalizability
- Remember uncertainty
  - These are *estimates*, not population parameters

Measurement Error in Regressor(s)

- We want effect of x, but we observe x^*, where $x = x^* + w$:

$$\begin{aligned} y &= \beta_0 + \beta_1 x^* + \epsilon \\ &= \beta_0 + \beta_1 (x - w) + \epsilon \\ &= \beta_0 + \beta_1 x + (\epsilon - \beta_1 w) \\ &= \beta_0 + \beta_1 x + v \end{aligned}$$

Measurement Error in Regressor(s)

- Produces *attenuation*: as measurement error increases, $\beta_1 \to 0$
- Our coefficients fit the observed data
- But they are *biased* estimates of our population equation
  - This applies to all $\hat{\beta}$ in a multivariate regression
  - Direction of bias is unknown

Measurement Error in Y

- Not necessarily a problem
- If *random* (i.e., uncorrelated with x), it costs us precision
- If *systematic*, who knows?!
- If *censored*, see Lectures 11 and/or 12

Missing Data

- Missing data can be a big problem
- We will discuss it in Lecture 5

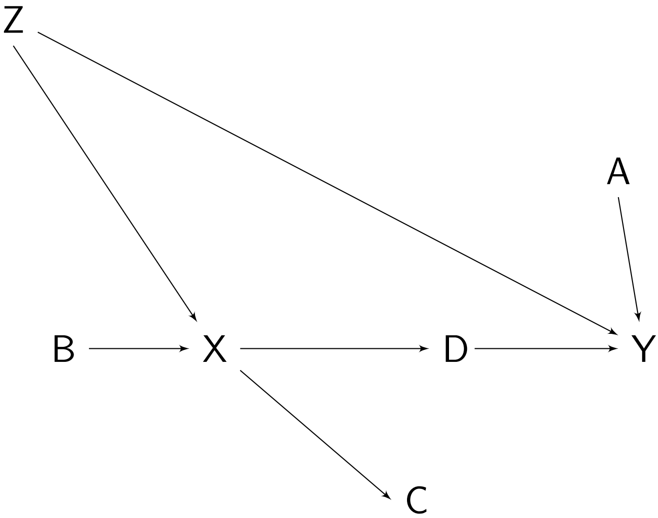
Confounding (Selection Bias)

- If x is not randomly assigned, potential outcomes are not independent of x
- Other factors explain why a unit i received their particular value x_i
- In matching, we obtain this *conditional independence* by comparing units that are identical on all confounding variables

Omitted Variables

$$\underbrace{E[Y_i|X_i = 1] - E[Y_i|X_i = 0] =}_{\text{Naive Effect}}$$

$$\underbrace{E[Y_{1i}|X_i = 1] - E[Y_{0i}|X_i = 1]}_{\text{Treatment Effect on Treated (ATT)}} + \underbrace{E[Y_{0i}|X_i = 1] - E[Y_{0i}|X_i = 0]}_{\text{Selection Bias}}$$


Z
A
B
X
D
Y
C

Omitted Variable Bias

- We want to estimate:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$$

- We actually estimate:

$$\begin{aligned}\tilde{y} &= \tilde{\beta}_0 + \tilde{\beta}_1 x + \epsilon \\ &= \tilde{\beta}_0 + \tilde{\beta}_1 x + (0 * z) + \epsilon \\ &= \tilde{\beta}_0 + \tilde{\beta}_1 x + \nu\end{aligned}$$

- Bias: $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$, where $\tilde{z} = \tilde{\delta}_0 + \tilde{\delta}_1 x$

Size and Direction of Bias

- Bias: $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$, where $\tilde{z} = \tilde{\delta}_0 + \tilde{\delta}_1 x$

	$Corr(x,z) < 0$	$Corr(x,z) > 0$
$\beta_2 < 0$	Positive	Negative
$\beta_2 > 0$	Negative	Positive

Aside: Three Meanings of "Endogeneity"

Formally endogeneity is when $Cov(X, \epsilon) \neq 0$

1. Measurement error in regressors
2. Omitted variables associated with included regressors
    - "Specification error"
    - Confounding
3. Lack of temporal precedence

Example: Englebert

- What is his research question?
- What is his theory? What does the graph look like?
- What is his analysis?

Common Conditioning Strategies

Common Conditioning Strategies

1. Condition on nothing (“naive effect”)

Common Conditioning Strategies

1. Condition on nothing ("naive effect")
2. Condition on some variables

Common Conditioning Strategies

1. Condition on nothing ("naive effect")
2. Condition on some variables
3. Condition on all observables

Common Conditioning Strategies

1. Condition on nothing (“naive effect”)
2. Condition on some variables
3. Condition on all observables

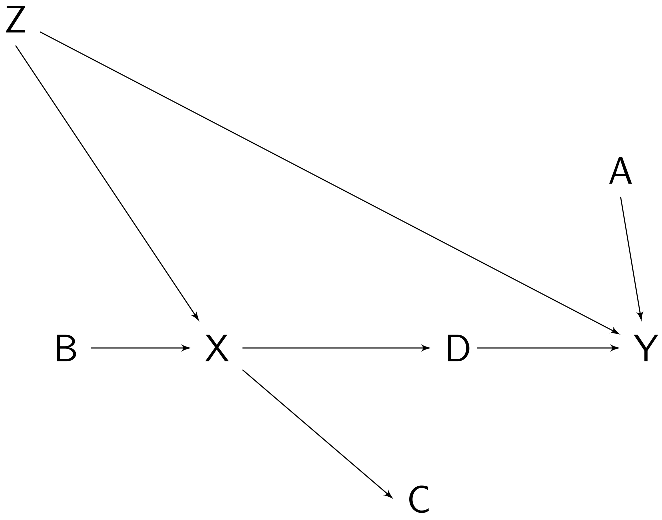
Which of these are good strategies?

What goes in our regression?

- Use theory to build causal models
  - Often, a causal graph helps
- Some guidance:

What goes in our regression?

- Use theory to build causal models
  - Often, a causal graph helps
- Some guidance:
  - Include confounding variables

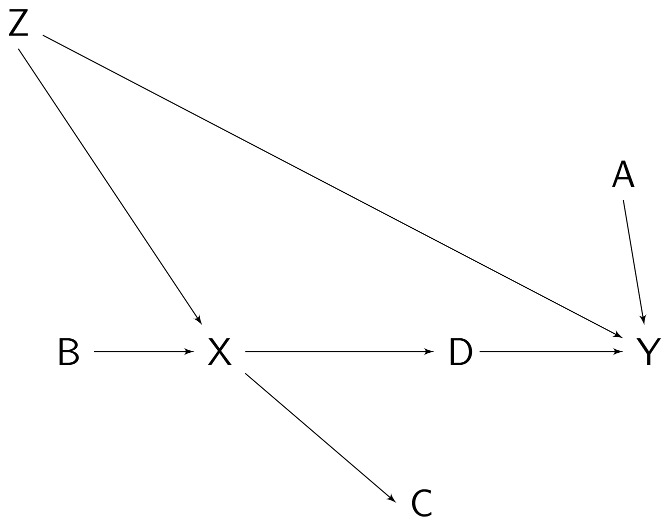

Z
A
B
X
D
Y
C

What goes in our regression?

- Use theory to build causal models
  - Often, a causal graph helps
- Some guidance:
  - Include confounding variables

What goes in our regression?

- Use theory to build causal models
  - Often, a causal graph helps
- Some guidance:
  - Include confounding variables
  - Do not include post-treatment variables


Z
A
B
X
D
Y
C

Post-treatment Bias

- We usually want to know the **total effect** of a cause
- If we include a mediator, D, of the $X \rightarrow Y$ relationship, the coefficient on X:
  - Only reflects the **direct** effect
  - Excludes the **indirect** effect of X through M
- So don't control for mediators!

What goes in our regression?

- Use theory to build causal models
  - Often, a causal graph helps
- Some guidance:
  - Include confounding variables
  - Do not include post-treatment variables

What goes in our regression?

- Use theory to build causal models
  - Often, a causal graph helps
- Some guidance:
  - Include confounding variables
  - Do not include post-treatment variables
  - Do not include *colinear* variables

Minimum Mathematical Requirements

1. Do we need variation in X?
   - Yes, otherwise dividing by zero
2. Do we need variation in Y?
   - No, $\hat{\beta}_1$ can equal zero
3. How many observations do we need?
   - $n \geq k$, where k is number of parameters to be estimated

Minimum Mathematical Requirements

1. Do we need variation in X?
   - Yes, otherwise dividing by zero
2. Do we need variation in Y?
   - No, $\hat{\beta}_1$ can equal zero
3. How many observations do we need?
   - $n \geq k$, where k is number of parameters to be estimated
4. Can we have highly correlated regressors?

Minimum Mathematical Requirements

1. Do we need variation in X?
   - Yes, otherwise dividing by zero
2. Do we need variation in Y?
   - No, $\hat{\beta}_1$ can equal zero
3. How many observations do we need?
   - $n \geq k$, where k is number of parameters to be estimated
4. Can we have highly correlated regressors?
   - Generally no (due to multicollinearity)

What goes in our regression?

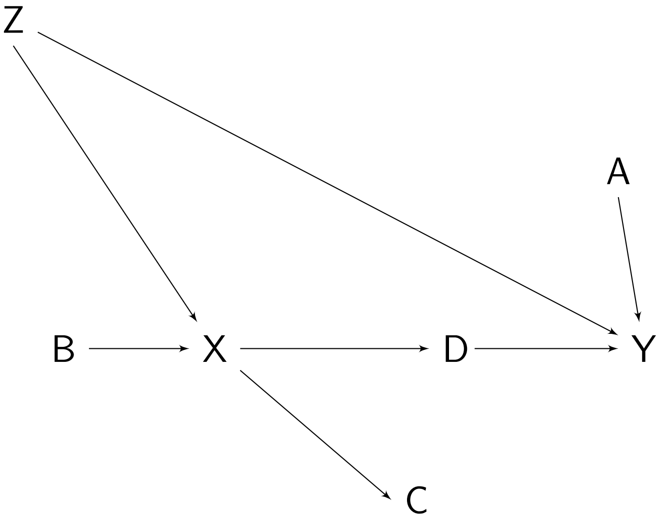
- Use theory to build causal models
  - Often, a causal graph helps
- Some guidance:
  - Include confounding variables
  - Do not include post-treatment variables
  - Do not include *colinear* variables

What goes in our regression?

- Use theory to build causal models
  - Often, a causal graph helps
- Some guidance:
  - Include confounding variables
  - Do not include post-treatment variables
  - Do not include *colinear* variables
  - Including irrelevant variables costs certainty

What goes in our regression?

- Use theory to build causal models
  - Often, a causal graph helps
- Some guidance:
  - Include confounding variables
  - Do not include post-treatment variables
  - Do not include *colinear* variables
  - Including irrelevant variables costs certainty
  - Including variables that affect Y alone increases certainty


Z
A
B
X
D
Y
C

Questions about specification?

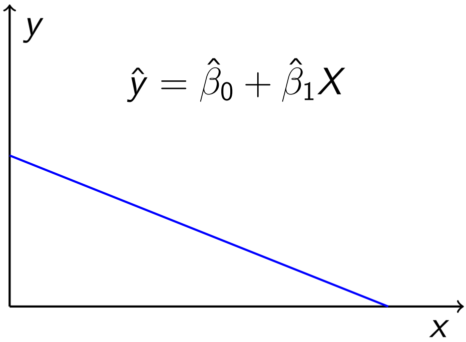
Multivariate Regression Interpretation

- All our interpretation rules from earlier still apply in a multivariate regression
- Now we interpret a coefficient as an effect "all else constant"
- Generally, not good to give all coefficients a causal interpretation
  - Think "forward causal inference"
  - We're interested in the $X \rightarrow Y$ effect
  - All other coefficients are there as "controls"

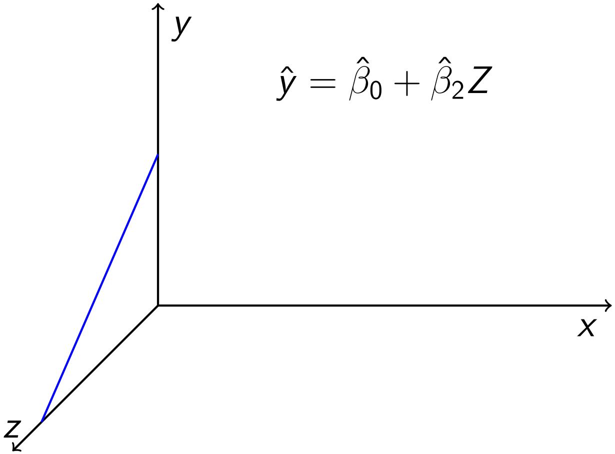
From Line to Surface I

- In simple regression, we estimate a **line**
- In multiple regression, we estimate a **surface**
- Each coefficient is the *marginal effect*, all else constant (at mean)
- This can be hard to picture in your mind

From Line to Surface II



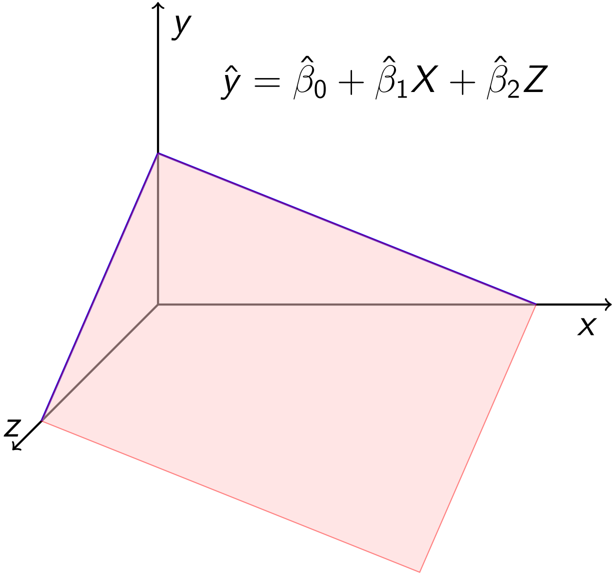
From Line to Surface II



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 Z$$

From Line to Surface II

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 Z$$



Are Our Estimates Any Good?

Yes, if:

1. Works mathematically
2. Causally valid theory
3. Linear relationship between X and Y
4. X is measured without error
5. No missing data (or MCAR; see Lecture 5)
6. No confounding

OLS is BLUE

- BLUE: Best Linear Unbiased Estimator
- Gauss Markov Assumptions:
  1. Linearity in parameters
  2. Random sampling
  3. No multicollinearity
  4. Exogeneity ($E[\epsilon|\mathbf{X}] = 0$)
  5. Homoskedasticity ($Var(\epsilon|\mathbf{X}) = \sigma^2$)
- Assumptions 1–4 prove OLS is unbiased
- Assumption 5 proves OLS is the *best* estimator

Squared vs. Absolute Errors

- Conventionally use Sum of Squared Errors
- Using absolute errors is also unbiased
- Sum of Squared Errors:
  - more heavily weights outliers
  - has a smaller variance
- Thus OLS is **B**est**LUE**

Goodness-of-Fit

- We want to know: "How good is our model?"

Goodness-of-Fit

- We want to know: "How good is our model?"
- We can answer:
  "How well does our model fit the observed data?"

Goodness-of-Fit

- We want to know: "How good is our model?"
- We can answer:
  "How well does our model fit the observed data?"
- Is this what we want to know?

Correlation

- Definition: $Corr(x, y) = \hat{r}_{x,y} = \frac{Cov(x,y)}{(n-1)s_x s_y}$
- Slope $\hat{\beta}_1$ and correlation $\hat{r}_{x,y}$ are simply different scalings of $Cov(x, y)$
- Interpretation: How well the bivariate relationship is summarized by a cloud of points?
- Units: none (range 0 to 1)

Coefficient of Determination (R^2)

- Definition: $R^2 = \hat{r}^2_{x,y} = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$
- Interpretation: How much of the total variation in y is explained by the model?
- But, R^2 increases simply by adding more variables
- So, Adjusted-$R^2 = R^2 - (1 - R^2)\frac{k}{n-k-1}$, where k is number of regressors
- Units: none (range 0 to 1)

Standard Error of the Regression (SER)

- "Root mean squared error" or just σ
- Definition: $\hat{\sigma} = \sqrt{\frac{SSR}{n-p}}$, where p is number of parameters estimated
- Interpretation: How far, on average, are the observed y values from their corresponding fitted values $\hat{y}$
  - $sd(y)$ is how far, on average, a given y_i is from $\bar{y}$
  - σ is how far, on average, a given y_i is from $\hat{y}_i$
- Units: same as y (range 0 to $sd(y)$)

The F-test

- Definition: Test of whether any of our coefficients differ from zero
  - In a bivariate regression, $F = t^2$
- Interpretation: Do any of the coefficients differ from zero?
  - Not a very interesting measure
- Units: none (range 0 to ∞)

```
. reg growth lcon

      Source |       SS       df       MS              Number of obs =      44
-------------+------------------------------           F(  1,    42) =    0.09
       Model |  .000038348     1  .000038348           Prob > F      =  0.7615
    Residual |  .017255198    42  .000410838           R-squared     =  0.0022
-------------+------------------------------           Adj R-squared = -0.0215
       Total |  .017293546    43  .000402175           Root MSE      =  .02027

------------------------------------------------------------------------------
      growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lcon |  -.0017819   .0058325    -0.31   0.761    -.0135524    .0099886
       _cons |   .0158988   .0390155     0.41   0.686    -.0628376    .0946353
------------------------------------------------------------------------------
```

The F-test for nested models

- Can use an F-test to compare fit of two nested models?
  - $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$
  - $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots$
- *Reduced* model is nested within *expanded* model
- Interpretation: Does adding additional variables significantly reduce SSR?

```
. nestreg: reg growth lcon (lconsq)

+-------------------------------------------------------------+
|       |          Block  Residual                     Change |
| Block |       F     df        df   Pr > F       R2    in R2 |
|-------+-----------------------------------------------------|
|     1 |    0.09      1        42   0.7615   0.0022          |
|     2 |    7.98      1        41   0.0073   0.1649   0.1626 |
+-------------------------------------------------------------+
```

Questions about model fit?

Inference from Sample to Population

- We want to know population parameter θ
- We only observe sample estimate $\hat{\theta}$
- We have a guess but are also uncertain

Inference from Sample to Population

- We want to know population parameter θ
- We only observe sample estimate $\hat{\theta}$
- We have a guess but are also uncertain

- What range of values for θ does our $\hat{\theta}$ imply?
- Are values in that range large or meaningful?

How Uncertain Are We?

- Our uncertainty depends on sampling procedures
- Most importantly, *sample size*
  - As $n \to \infty$, uncertainty $\to 0$
- We typically summarize our uncertainty as the *standard error*

Standard Errors (SEs)

- Definition: "The standard error of a sample estimate is the average distance that a sample estimate ($\hat{\theta}$) would be from the population parameter (θ) if we drew many separate random samples and applied our estimator to each."

- In bivariate regression: $Var(\hat{\beta}_1) = \frac{\frac{1}{n-2}SSR}{SST_x}$

- Thus, SE is a ratio of unexplained variance in y (weighted by sample size) and variance in x

- Units: same as coefficient ($\frac{y}{x}$)

What affects size of SEs?

- Larger variance in x means smaller SEs
- More unexplained variance in y means bigger SEs
- More observations reduces the numerator, thus smaller SEs
- Other factors:
  - Homoskedasticity
  - Clustering
- Interpretation:
  - Large SE: Uncertain about population effect size
  - Small SE: Certain about population effect size

Ways to Express Our Uncertainty

1. Standard Error
2. Confidence interval
3. t-statistic
4. p-value

```
. reg growth lcon

      Source |       SS       df       MS              Number of obs =      44
-------------+------------------------------           F(  1,    42) =    0.09
       Model |  .000038348     1  .000038348           Prob > F      =  0.7615
    Residual |  .017255198    42  .000410838           R-squared     =  0.0022
-------------+------------------------------           Adj R-squared = -0.0215
       Total |  .017293546    43  .000402175           Root MSE      =  .02027

------------------------------------------------------------------------------
      growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lcon |  -.0017819   .0058325    -0.31   0.761    -.0135524    .0099886
       _cons |   .0158988   .0390155     0.41   0.686    -.0628376    .0946353
------------------------------------------------------------------------------
```

Confidence Interval (CI)

- Definition: Were we to repeat our procedure of sampling, applying our estimator, and calculating a confidence interval *repeatedly* from the population, a fixed percentage of the resulting intervals would include the true population-level slope.
- Interpretation: If the confidence interval overlaps zero, we are uncertain if β differs from zero

Confidence Interval (CI)

- A CI is simply a range, centered on the slope
- Units: Same scale as the coefficient ($\frac{y}{x}$)
- We can calculate different CIs of varying *confidence*
  - Conventionally, $\alpha = 0.05$, so 95% of the CIs will include the β

t-statistic

- A measure of how large a coefficient is relative to our uncertainty about its size
- Typically used to test a formal null hypothesis:
  - No effect null: $t_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}}$
  - Any other null: $\frac{\hat{\beta}_1 - \alpha}{SE_{\hat{\beta}_1}}$, where α is our null hypothesis effect size

t-statistic

- A measure of how large a coefficient is relative to our uncertainty about its size
- Typically used to test a formal null hypothesis:
  - No effect null: $t_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}}$
  - Any other null: $\frac{\hat{\beta}_1 - \alpha}{SE_{\hat{\beta}_1}}$, where α is our null hypothesis effect size
- Note: The t-statistic from a t-test of mean-difference is the same as the t-statistic from a t-test on an OLS slope for a dummy covariate

p-value

- A summary measure in a hypothesis test
- General definition: “the probability of a statistic as extreme as the one we observed, if the null hypothesis was true, the statistic is distributed as we assume, and the data are as variable as observed”
- Definition in a regression context: “the probability of a slope as large as the one we observed . . . ”

The p-value is not:

- The probability that a hypothesis is true or false
- A reflection of our confidence or certainty about the result
- The probability that the true slope is in any particular range of values
- A statement about the importance or substantive size of the effect

Significance

1. Substantive significance

2. Statistical significance

Significance

1. Substantive significance
   - Is the effect size (or range of possible effect sizes) *important* in the real world?
2. Statistical significance

Significance

1. Substantive significance
   - Is the effect size (or range of possible effect sizes) *important* in the real world?
2. Statistical significance
   - Is the effect size (or range of possible effect sizes) larger than a predetermined threshold?
   - Conventionally, $p \leq 0.05$

Questions about inference?