Getting to Regression: The Workhorse of Quantitative Political Analysis

Department of Government
London School of Economics and Political Science

Correlation as Measure of Bivariate Relationship

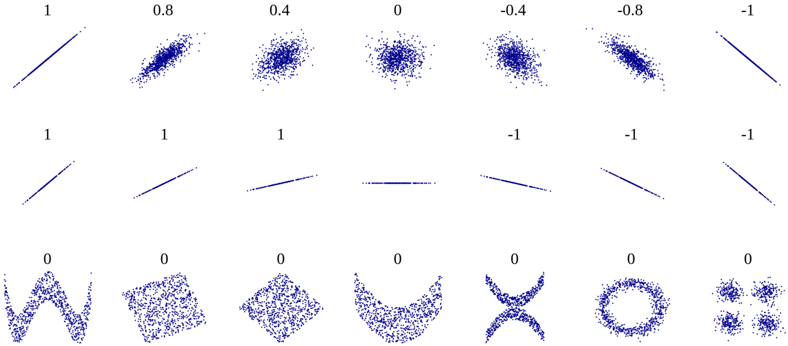
- Covariance:

$$Cov(X, Y) = \Sigma_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

Correlation as Measure of Bivariate Relationship

- Covariance:
  $$Cov(X,Y) = \Sigma_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$
- Correlation:
  $$Corr(X,Y) = r_{x,y} = \Sigma_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)s_x s_y}$$
  where $s_x = \sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}$

Correlation is linear!



Source: Wikimedia

Guess the Correlation!

1. Go to:
   http://guessthecorrelation.com/
2. Play a few rounds

Regression

- Definition: a statistical method for measuring the relationships between one variable and many other variables

Regression

- Definition: a statistical method for measuring the relationships between one variable and many other variables
- Uses of Regression
  1. Description
  2. Prediction
  3. Causal Inference

Regression

- Definition: a statistical method for measuring the relationships between one variable and many other variables
- Uses of Regression
  1. Description
  2. Prediction
  3. Causal Inference
- *Ordinary least squares* (OLS) regression

Interpretations of OLS

Interpretations of OLS

1. Line (or surface) of best fit
2. Ratio of $Cov(X, Y)$ and $Var(X)$
3. Minimizing residual sum of squares (SSR)

Interpretations of OLS

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

Bivariate Regression I

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

Bivariate Regression I

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

Bivariate Regression I

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

Three Equations

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

Three Equations

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

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

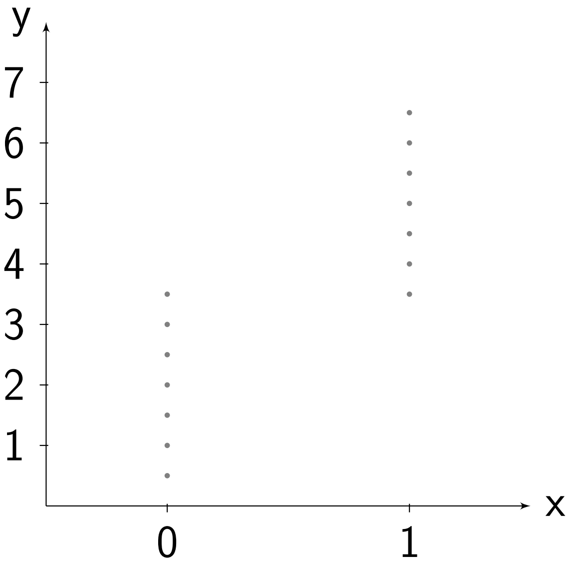
Three Equations

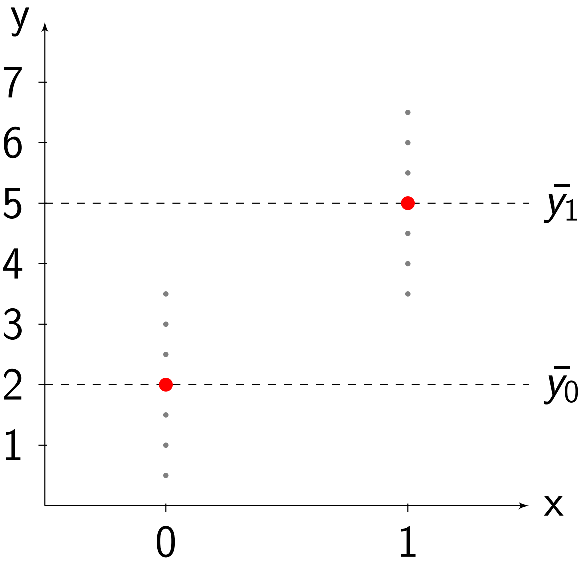
1. Population:
   $Y = \beta_0 + \beta_1 X \ \ (+\epsilon)$

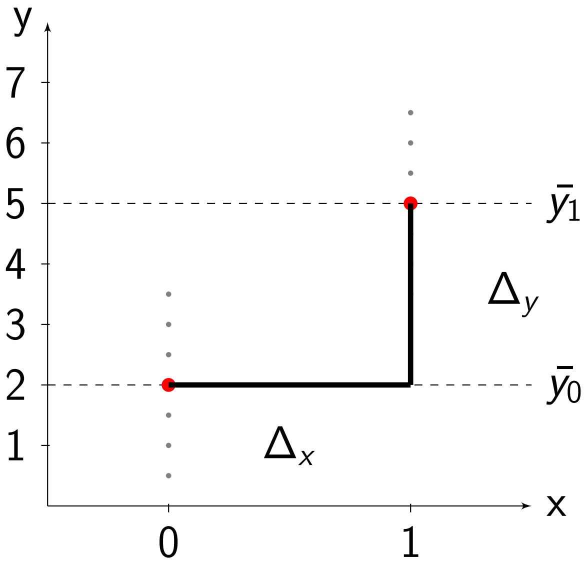
2. Sample estimate:
   $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + e$

3. Unit:

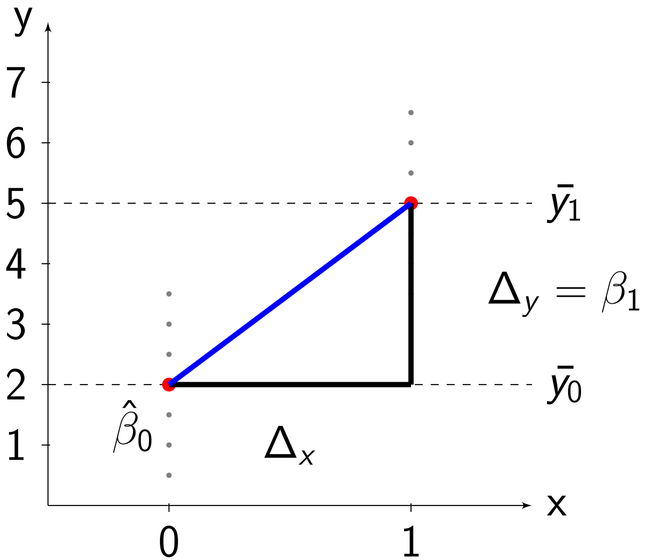
$$\begin{aligned} y_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \\ &= \bar{y}_{0i} + (y_{1i} - y_{0i}) x_i + (y_{0i} - \bar{y}_{0i}) \end{aligned}$$

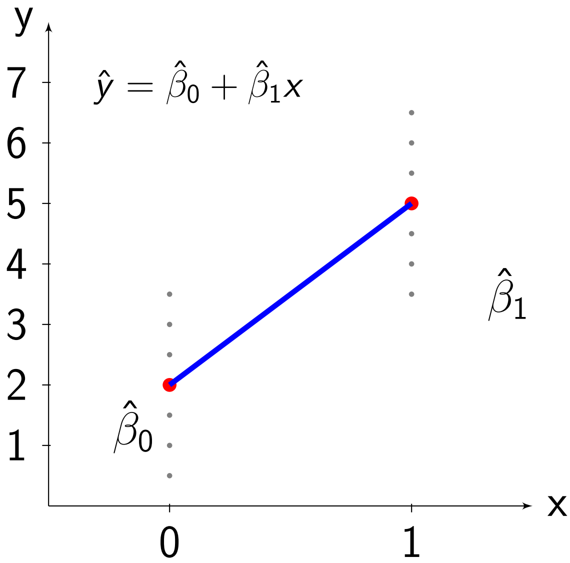

y
7
6
5
4
3
2
1
x
0
1

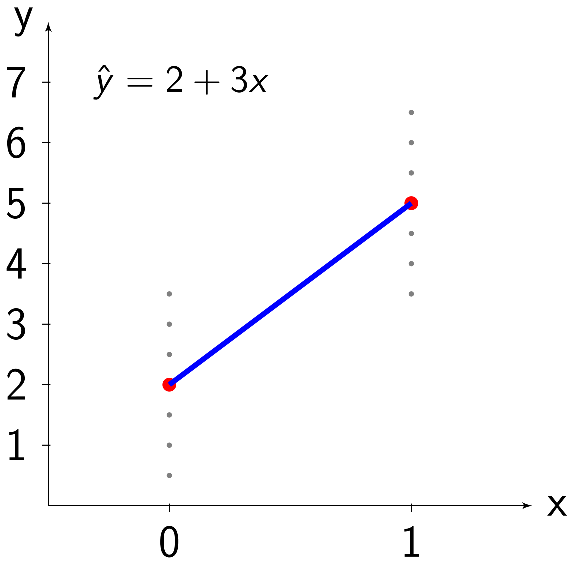

y
7
6
5
4
3
2
1
$\bar{y}_1$
$\bar{y}_0$
x
0
1

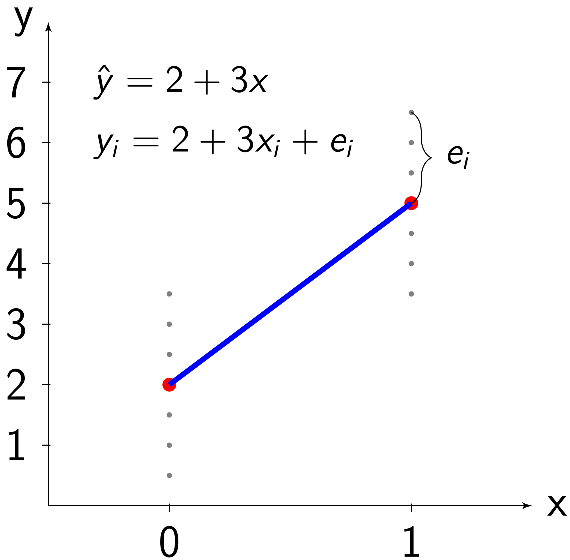


y
7
6
5
4
3
2
1
$\bar{y}_1$
Δ_y
$\bar{y}_0$
Δ_x
x
0
1


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
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
$\hat{\beta}_1$
$\hat{\beta}_0$
0
1
x


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



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

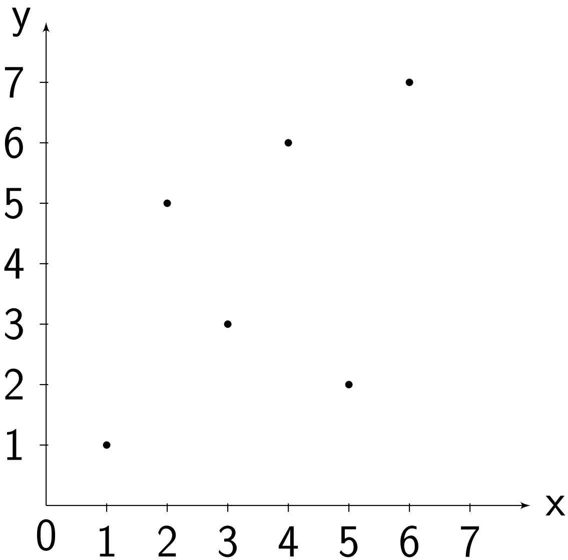
Questions?

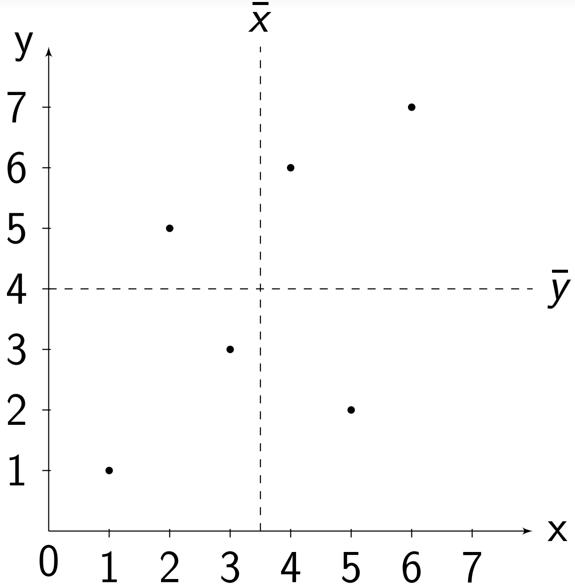
Continuous X

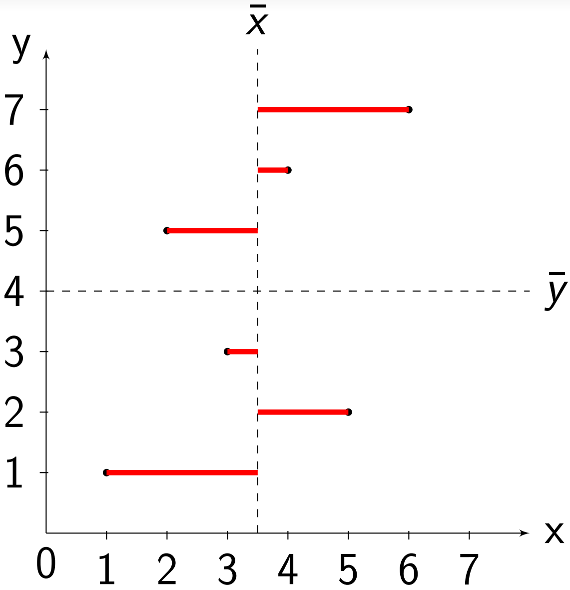
- If x is continuous, calculation is more complicated
- Rather than β_1 being the mean-difference in outcomes, it is the slope across *all* values of x
- $\hat{\beta}_1 = Cov(x, y)/Var(x)$

Calculations

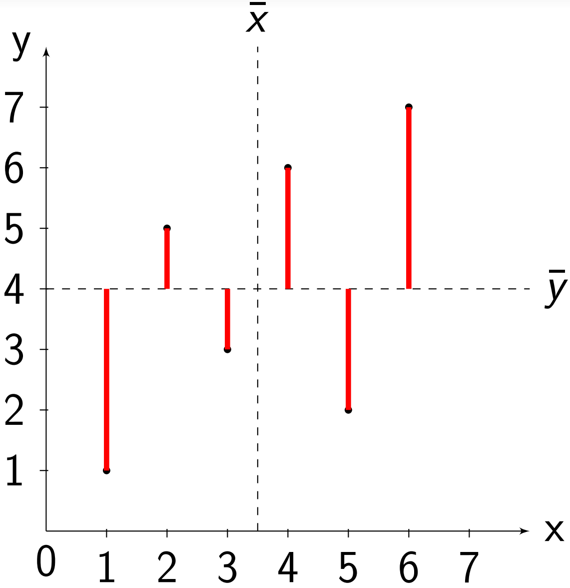
x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	1	?	?	?	?
2	5	?	?	?	?
3	3	?	?	?	?
4	6	?	?	?	?
5	2	?	?	?	?
6	7	?	?	?	?
$\bar{x}$	$\bar{y}$			$Cov(x, y)$	$Var(x)$


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



x̄
y
7
6
5
4
3
2
1
ȳ
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	?	?	?	?
$\bar{x}$	$\bar{y}$			$Cov(x, y)$	$Var(x)$

Calculations

If x is continuous, calculation is more complicated:
$\widehat{\beta_1} = Cov(x, y)/Var(x)$

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	$-2.\bar{6}$	-3	$-6.6\bar{6}$	6.25
2	5	$-1.\bar{3}$	$+1$	-2.00	2.25
3	3	$-0.\bar{6}$	-1	$-0.3\bar{3}$	0.25
4	6	$+0.\bar{3}$	$+2$	$-0.1\bar{6}$	0.25
5	2	$+1.\bar{6}$	-2	-2.50	2.25
6	7	$+2.\bar{3}$	$+3$	$-8.3\bar{3}$	6.25
3.5	$3.\bar{6}$			11	17.5

Calculations

If x is continuous, calculation is more complicated:
$\widehat{\beta_1} = Cov(x,y)/Var(x) = 11/17.5 = \mathbf{0.627}$

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	$-2.\bar{6}$	-3	$-6.6\bar{6}$	6.25
2	5	$-1.\bar{3}$	$+1$	-2.00	2.25
3	3	$-0.\bar{6}$	-1	$-0.3\bar{3}$	0.25
4	6	$+0.\bar{3}$	$+2$	$-0.1\bar{6}$	0.25
5	2	$+1.\bar{6}$	-2	-2.50	2.25
6	7	$+2.\bar{3}$	$+3$	$-8.3\bar{3}$	6.25
3.5	$3.\bar{6}$			11	17.5

Intercept $\hat{\beta}_0$

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

Intercept $\hat{\beta}_0$

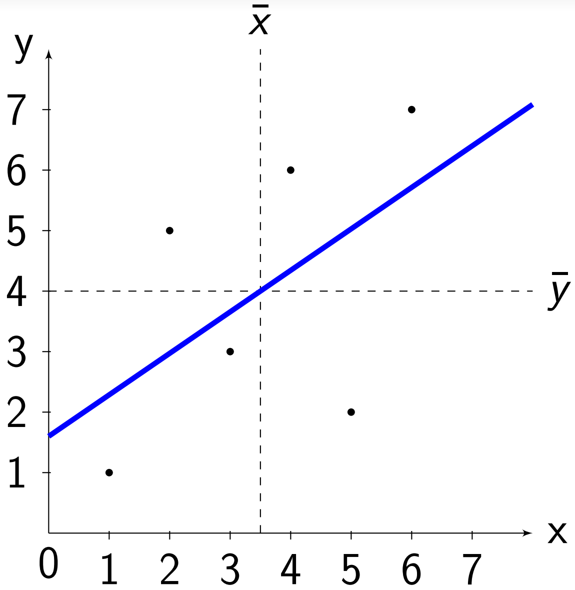
- Simple formula: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- Intuition: OLS fit always runs through point $(\bar{x}, \bar{y})$

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.: $\hat{\beta}_0 = 3.\bar{6} - 0.627 * 3.5 = 1.4\bar{6}$

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.: $\hat{\beta}_0 = 3.\bar{6} - 0.627 * 3.5 = 1.4\bar{6}$
- $\hat{y} = 1.4\bar{6} + 0.6857\hat{x}$



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

Systematic versus unsystematic components

Systematic versus unsystematic components

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

Systematic versus unsystematic components

- 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 β that minimizes the *residual sum of squares*

Why are there residuals?

Why are there residuals?

- Fundamental randomness

Why are there residuals?

- Fundamental randomness
- Measurement error

Why are there residuals?

- Fundamental randomness
- Measurement error
- Omitted variables

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 ($Cor(X, Y) = 0$)

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 ($Cor(X, Y) = 0$)

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 ($Cor(X, Y) = 0$)
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 ($Cor(X, Y) = 0$)
3. How many observations do we need?
   - $n \geq k$, where k is number of parameters to be estimated

Correlation/Regression Equivalence

- 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 $Cov(x, y)$

Correlation/Regression Equivalence

- 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 $Cov(x, y)$
- $R^2 = \hat{r}^2_{x,y} = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$

Questions about OLS?

Are Estimates Any Good?

Are Estimates Any Good?

1. Works mathematically
2. Linear relationship between X and Y
3. X is measured without error
4. No missing data (or MCAR)
5. No confounding (next week)

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

- 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?

Interpretation: Indicator

- $y = \hat{\beta}_0 + \hat{\beta}_1 x + e$
- β_0 is the estimate of $\bar{y}$ when $x = 0$
- β_1 is the difference: $\bar{y}_{x=1} - \bar{y}_{x=0}$

Interpretation: Categorical

- $y = \hat{\beta}_0 + \hat{\beta}_1 x_{x=1} + \hat{\beta}_2 x_{x=2} + \cdots + e$
- β_0 is the estimate of $\bar{y}$ when $x = 0$
- β_1 is the difference: $\bar{y}_{x=1} - \bar{y}_{x=0}$
- β_2 is the difference: $\bar{y}_{x=2} - \bar{y}_{x=0}$
- Need to select one category as the *reference category*!

Interpretation: Interval

- $y = \hat{\beta}_0 + \hat{\beta}_1 x + e$
- β_0 is the estimate of $\bar{y}$ when $x = 0$
- β_1 is the slope of the relationship between x and y
    - Slope is constant across full domain of x

Interpretation: Ordinal

- Two options:
  1. $y = \hat{\beta}_0 + \hat{\beta}_1 x + e$
  2. $y = \hat{\beta}_0 + \hat{\beta}_1 x_{x=1} + \hat{\beta}_2 x_{x=2} + \cdots + e$
- Have to choose whether to treat an ordinal variable as *categorical* or *interval*

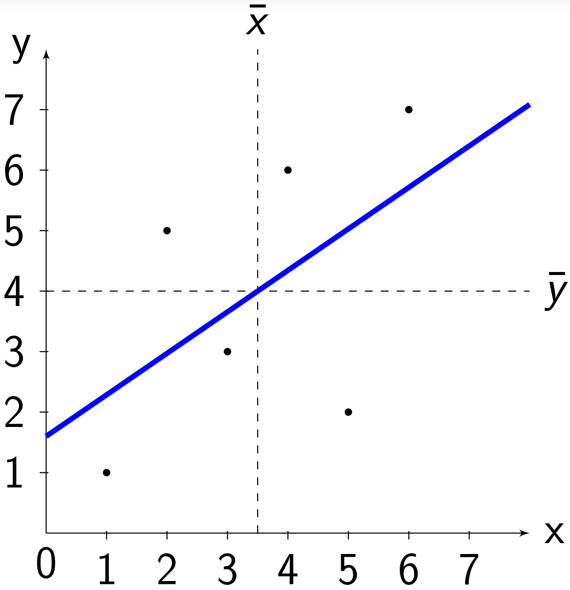
Questions?

What type of x variable is involved and how do we interpret the coefficient(s) on x for each of the following scenarios?

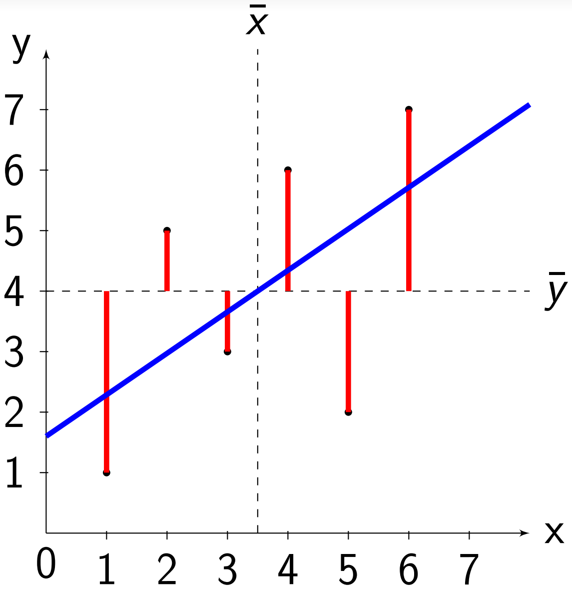
1. Body Mass Index (BMI) regressed on height
2. Monthly income ($) regressed on gender
3. Years of schooling regressed on birth region
4. Feeling thermometer toward Theresa May regressed on party affiliation
5. Weekly hours worked regressed on civil service pay grade

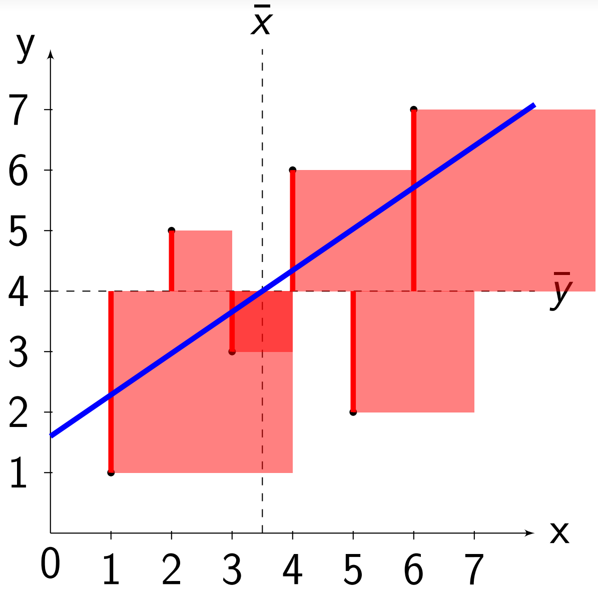
OLS Minimizes SSR

- Total Sum of Squares (SST): $\Sigma_{i=1}^{n}(y_i - \bar{y})^2$
- 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

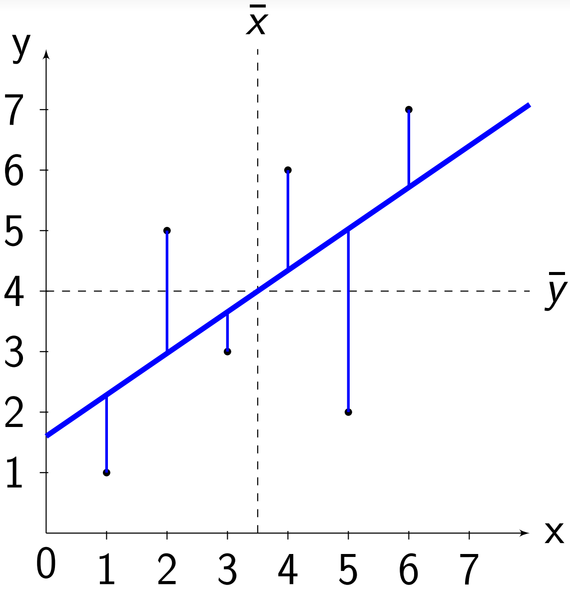


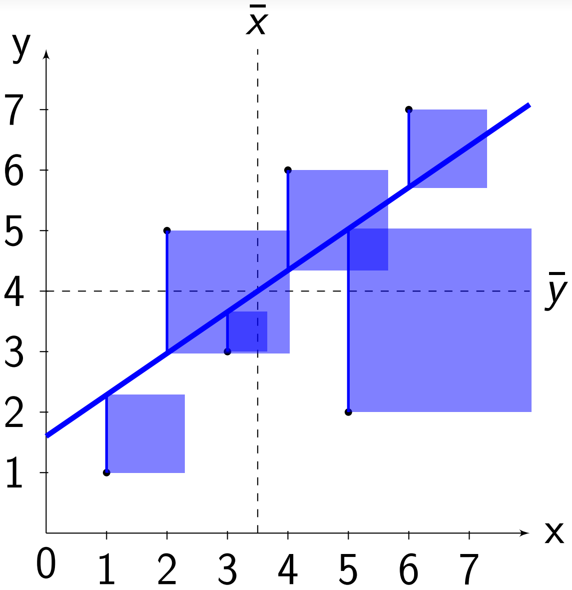
$\bar{x}$
y
7
6
5
4
3
2
1
$\bar{y}$
x
0
1
2
3
4
5
6
7


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

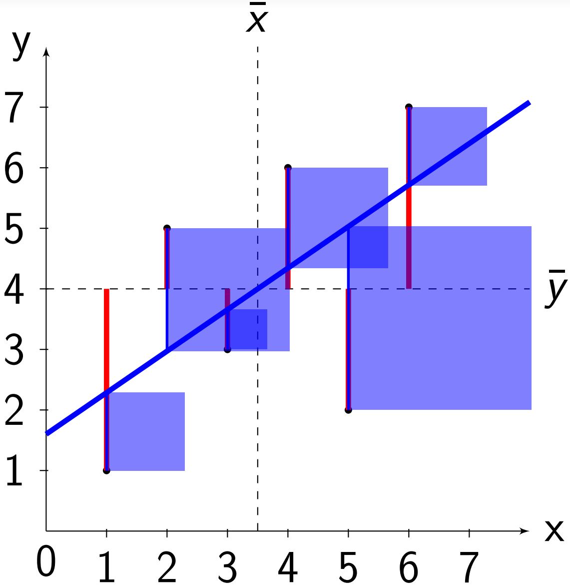


$\bar{x}$
y
7
6
5
4
3
2
1
$\bar{y}$
x
0
1
2
3
4
5
6
7

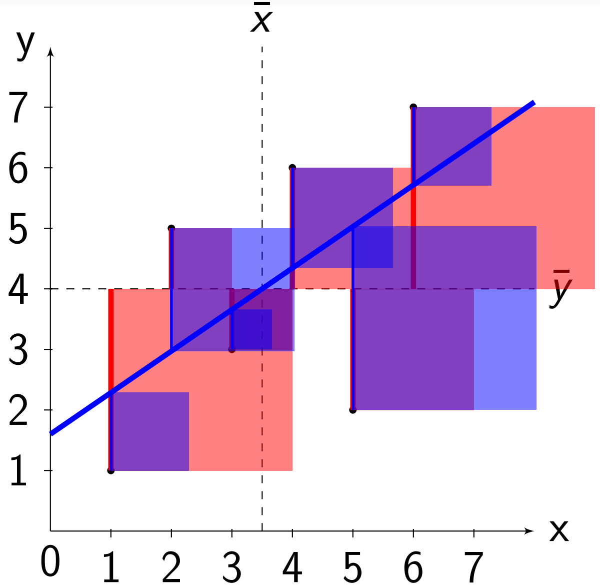

$\bar{x}$
y
7
6
5
4
3
2
1
$\bar{y}$
x
0 1 2 3 4 5 6 7



$\bar{x}$
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


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

RMSE (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 $sd(y)$)