Recitation, Week 11

Ye Wang

New York University

POL-850

Spring 2018

Outline

1. Hypothesis testing and confidence interval
2. Correlation coefficient
3. Linear regression

Reveiw: Hypothesis testing

1. Data at hand

	Treatment	Control
Mean	2.1 years	2 years
σ	0.4 years	0.6 years
N	100	100

$$H_0 \text{ is } \mu_T = \mu_C$$
$$H_A \text{ is } \mu_T \neq \mu_C$$

2. The **significance level**: let's choose the 0.05 significance level

Review: Hypothesis testing

3. **Test-statistic**: then we can calculate the test statistic:

$$z = \frac{\overline{x}_T - \overline{x}_C}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_C^2}{n_C}}} = \frac{2.1 - 2}{\sqrt{\frac{0.4^2}{100} + \frac{0.6^2}{100}}} = 1.39$$

4. **p-value**: we get the corresponding p-value

$$Pr[Z < -1.39 \text{ or } Z > 1.39] = 0.1646$$

5. **Compare** our p-value with the 0.05 significance level

$$0.1646 > 0.05$$

We fail to reject the null hypothesis at the 0.05 significance level

Confidence interval for hypothesis testing

Actually, **we can test the alternative hypothesis using a confidence interval**

How can we construct the 95% confidence interval of $\mu_T - \mu_C$?

$$z = \frac{(\overline{x}_T - \overline{x}_C) - (\mu_T - \mu_C)}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_C^2}{n_C}}}$$

Confidence interval for hypothesis testing

Some hints from the confidence interval of μ:

1. The z score for the sample mean

$$z = \frac{\overline{x} - \mu}{SE} = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

2. The 95% confidence interval of the population mean μ:

$$\overline{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

Confidence interval for hypothesis testing

Some hints from the confidence interval of μ:

1. The z test statistic for the sample difference of means

$$z = \frac{(\overline{x}_T - \overline{x}_C) - (\mu_T - \mu_C)}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_C^2}{n_C}}}$$

2. The 95% confidence interval of the population difference of means $\mu_T - \mu_C$:

$$(\overline{x}_T - \overline{x}_C) \pm 1.96\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_C^2}{n_C}}$$

Confidence interval for hypothesis testing

1. Data

	Treatment	Control
Mean	2.1 years	2 years
σ	0.4 years	0.6 years
N	100	100

2. Confidence interval of $\mu_T - \mu_C$

$$(\overline{x}_T - \overline{x}_C) \pm 1.96\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_C^2}{n_C}}$$

Confidence interval for hypothesis testing

Computing the 95% confidence interval of $\mu_T - \mu_C$ yields,

$$0.1 \pm 0.14 \iff [-0.04, 0.15]$$

- Does this include the zero or not?

Confidence interval for hypothesis testing

Computing the 95% confidence interval of $\mu_T - \mu_C$ yields,

$$0.1 \pm 0.14 \iff [-0.04, 0.15]$$

- Does this include the zero or not? Yes
- **So what does this mean?**

Confidence interval for hypothesis testing

- If the 95% confidence interval **includes** zero, then the difference is **not statistically significant** at the 0.05 significance level

  $\Longleftrightarrow$ We **fail to reject** the H_0 at the 0.05 significance level

- If the 95% confidence interval does **not include** zero, then the difference is **statistically significant** at the 0.05 significance level

  $\Longleftrightarrow$ We **reject** the H_0 at the 0.05 significance level

Individual exercise 1

Is there a difference between life expectancy of two groups at the population level? Answer this question by constructing a 95% confidence interval

	Treatment	Control
Mean	4 years	2 years
σ	1 years	2 years
N	20	20

Hint:

$$(\overline{x}_T - \overline{x}_C) \pm 1.96\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_C^2}{n_C}}$$

Correlation coefficient

- Correlation coefficient is a measure of association
- We can use this coefficient when X and Y are interval variables
- It measures the **direction** and the **strength** of the relationship
- Its range is [-1,1]
- **-1** means "perfect negative association"
- **+1** means "perfect positive association"
- **0** means "no association"

Correlation coefficient

A quick quiz: True or False?

"the near-zero correlation coefficient means that there is no relationship between X and Y"

Correlation coefficient

A quick quiz: True or False?

"the near-zero correlation coefficient means that there is no relationship between X and Y"

This is **not true**! Why?

Correlation coefficient

A quick quiz: True or False?

"the near-zero correlation coefficient means that there is no relationship between X and Y"

This is **not true**! Why?

The correlation coefficient always measures **linear** association between X and Y

Correlation coefficient

The correlation coefficient is closed to **zero** but a non-linear relationship between X and Y!

Correlation coefficient

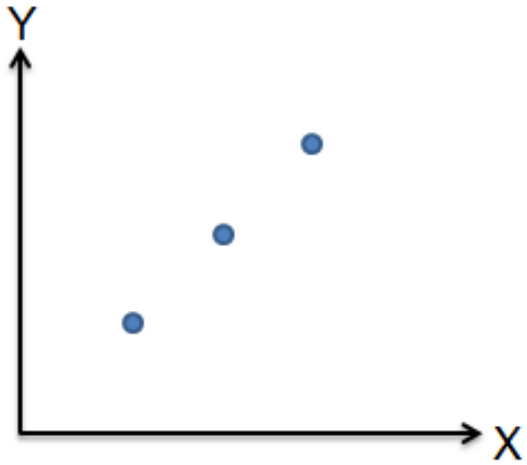
Therefore, the correlation coefficient means

- **-1**: "perfect **linear** negative association"
- **+1**: "perfect **linear** positive association"
- **0**: "no **linear** association"

Individual exercise 2

Suppose we have three points: (1, 1), (2, 2), and (3, 3). Verify that the correlation coefficient is one by hand calculation.

$$r = \frac{\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{s_X}\right)\left(\frac{Y_i-\overline{Y}}{s_Y}\right)}{n-1} \qquad s_X = \sqrt{\frac{\sum(X_i-\overline{X})^2}{n-1}}$$



Linear regression

What's linear regression?

- We assume a linear association between X and Y in population (note that ε are random errors)

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

- Here α, β, and ε are not observable. Why?

Linear regression

What's linear regression?

- We assume a linear association between X and Y in population (note that ε are random errors)

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

- Here α, β, and ε are not observable. Why?
- Because they are population parameters

Linear regression

What's linear regression?

- We assume a linear association between X and Y in population (note that ε are random errors)

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

- Here α, β, and ε are not observable. Why?
- Because they are population parameters
- Therefore, they need to be estimated (we should make a guess)

Linear regression

- We have a sample and we want to **estimate** α and β with our sample (e is called residuals)

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + e_i$$

- What is the difference between ϵ_i and e_i?
- Note that $\hat{\alpha}$ is constant, and $\hat{\beta}$ is coefficient

Linear regression

- We have a sample and we want to **estimate** α and β with our sample (e is called residuals)

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + e_i$$

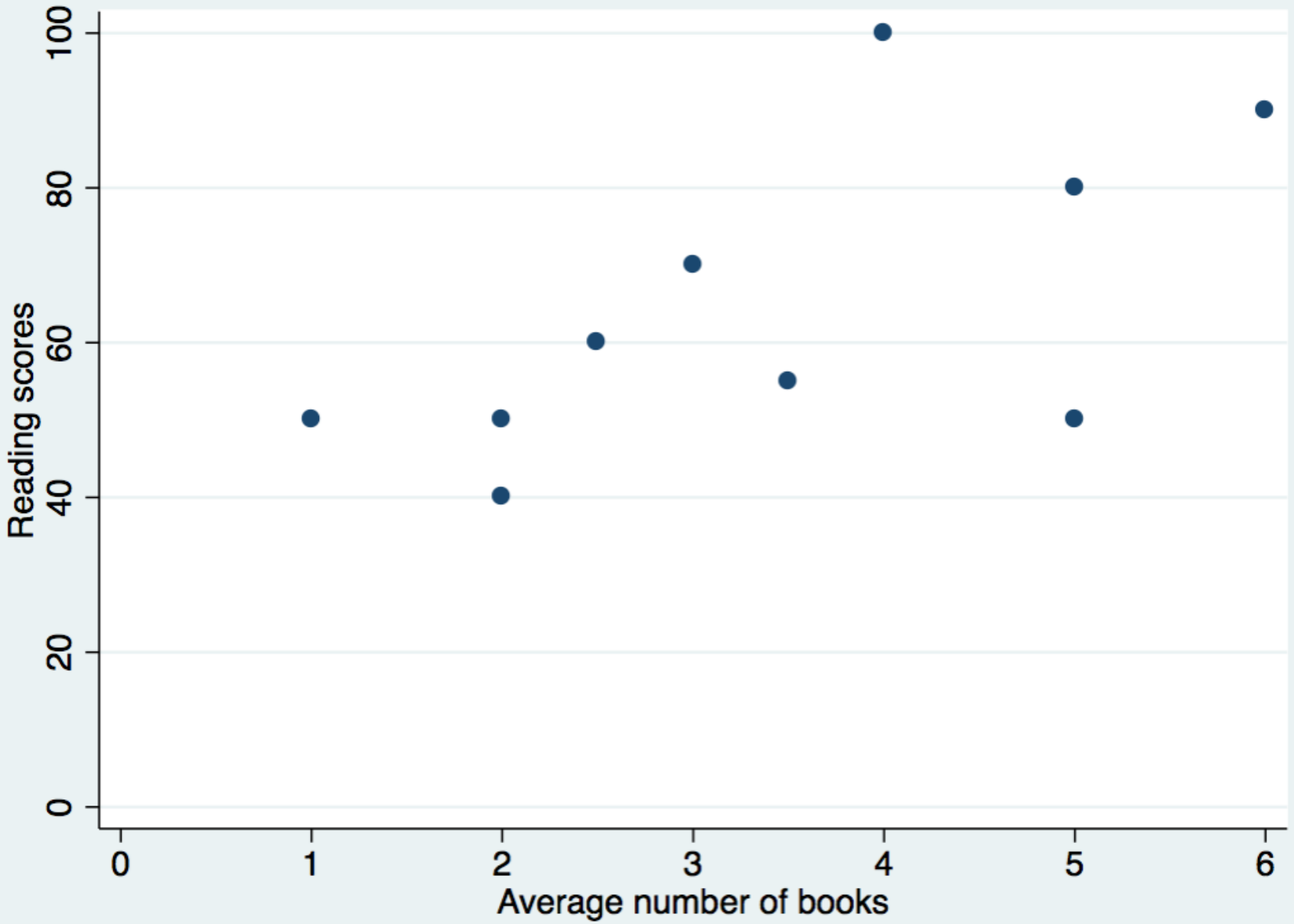
- What is the difference between ϵ_i and e_i?
- Note that $\hat{\alpha}$ is constant, and $\hat{\beta}$ is coefficient
- Our question becomes "how do we calculate $\hat{\alpha}$ and $\hat{\beta}$ based off of a sample?

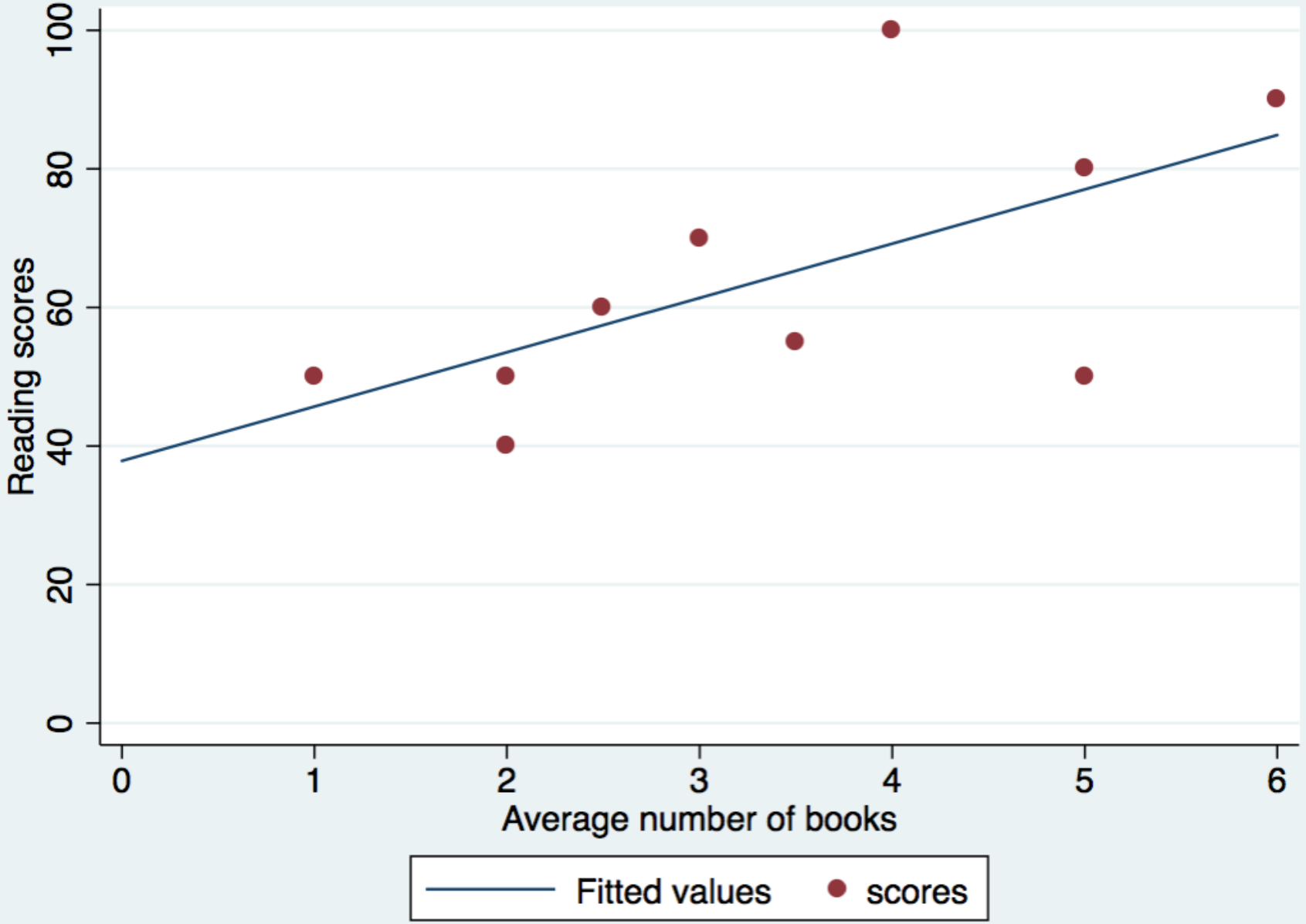
Linear regression

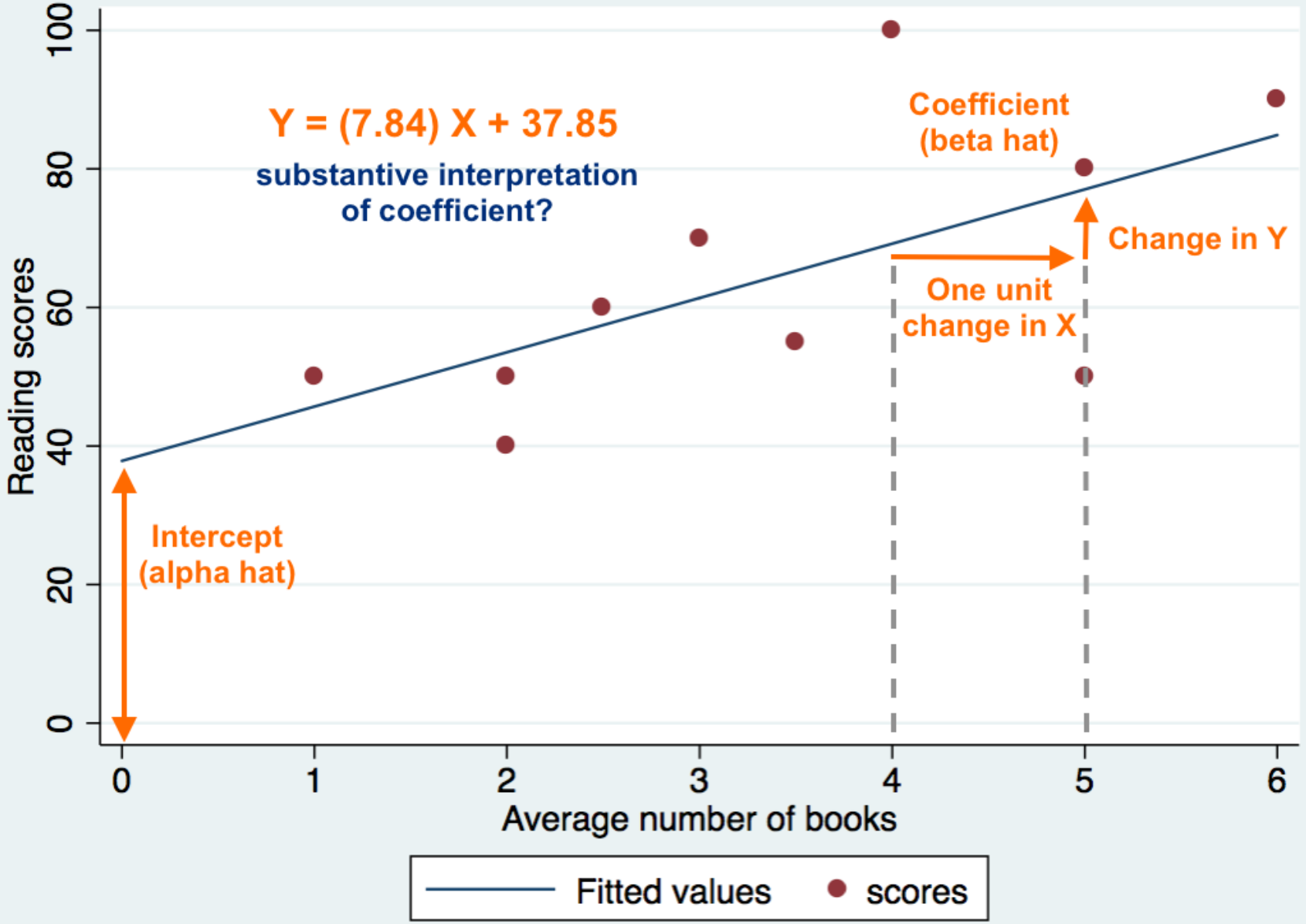
- We have a sample and we want to **estimate** α and β with our sample (e is called residuals)

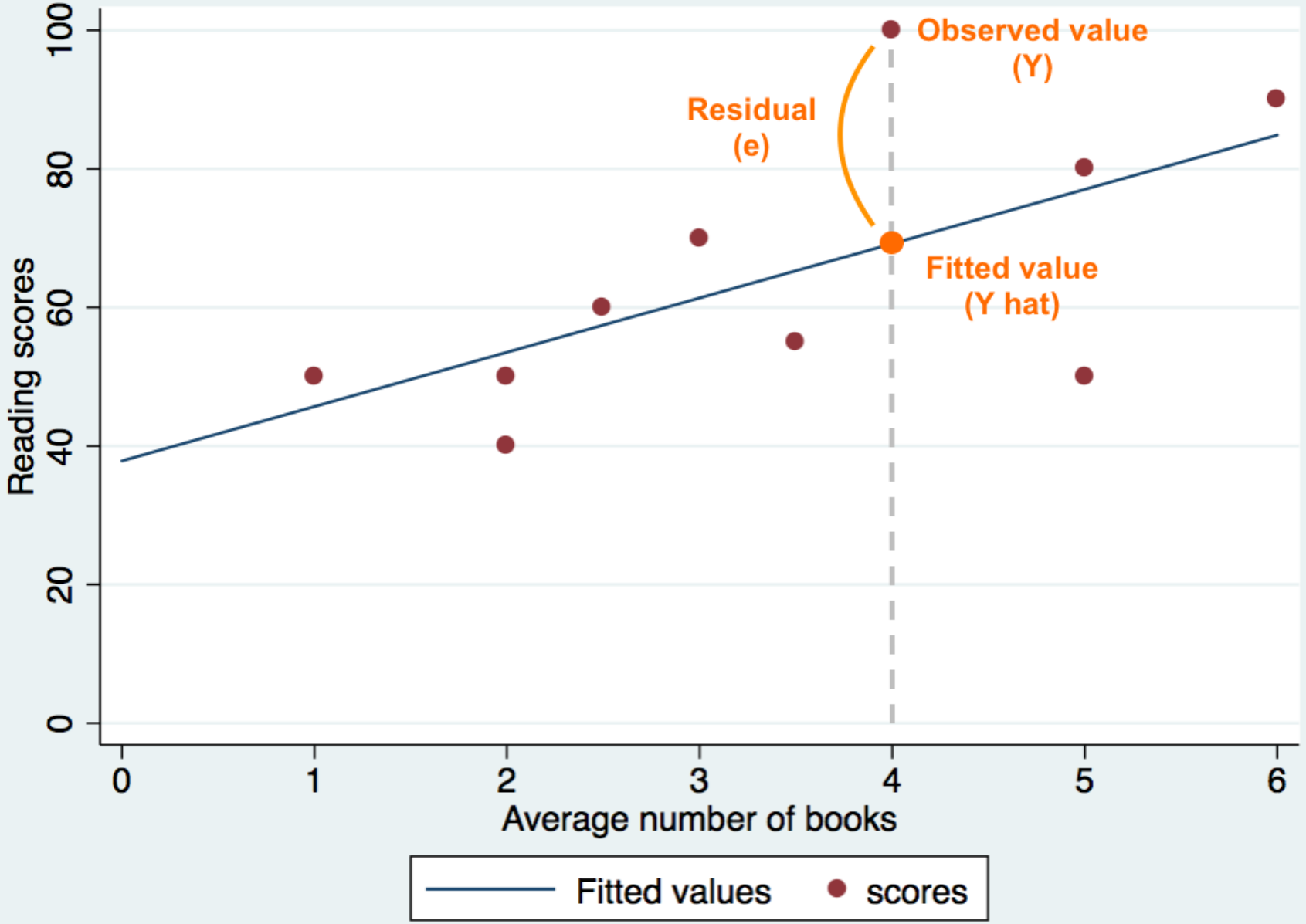
$$Y_i = \hat{\alpha} + \hat{\beta} X_i + e_i$$

- What is the difference between ϵ_i and e_i?
- Note that $\hat{\alpha}$ is constant, and $\hat{\beta}$ is coefficient
- Our question becomes "how do we calculate $\hat{\alpha}$ and $\hat{\beta}$ based off of a sample?
- Basically, we want to draw a line which fits "best" to data by choosing $\hat{\alpha}$ and $\hat{\beta}$









Linear regression

Goal: we want to draw a line which fits "best" to data by choosing $\hat{\alpha}$ and $\hat{\beta}$

- There could be many "best" ways of drawing a line
- One "best" way: **minimizes** the sum of squares of residuals

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

- where
  e_i: residuals ($e_i = Y_i - \hat{Y}_i$)
  Y_i: observed Y
  $\hat{Y}_i$: predicted Y or fitted value
  (we can calculate $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$)

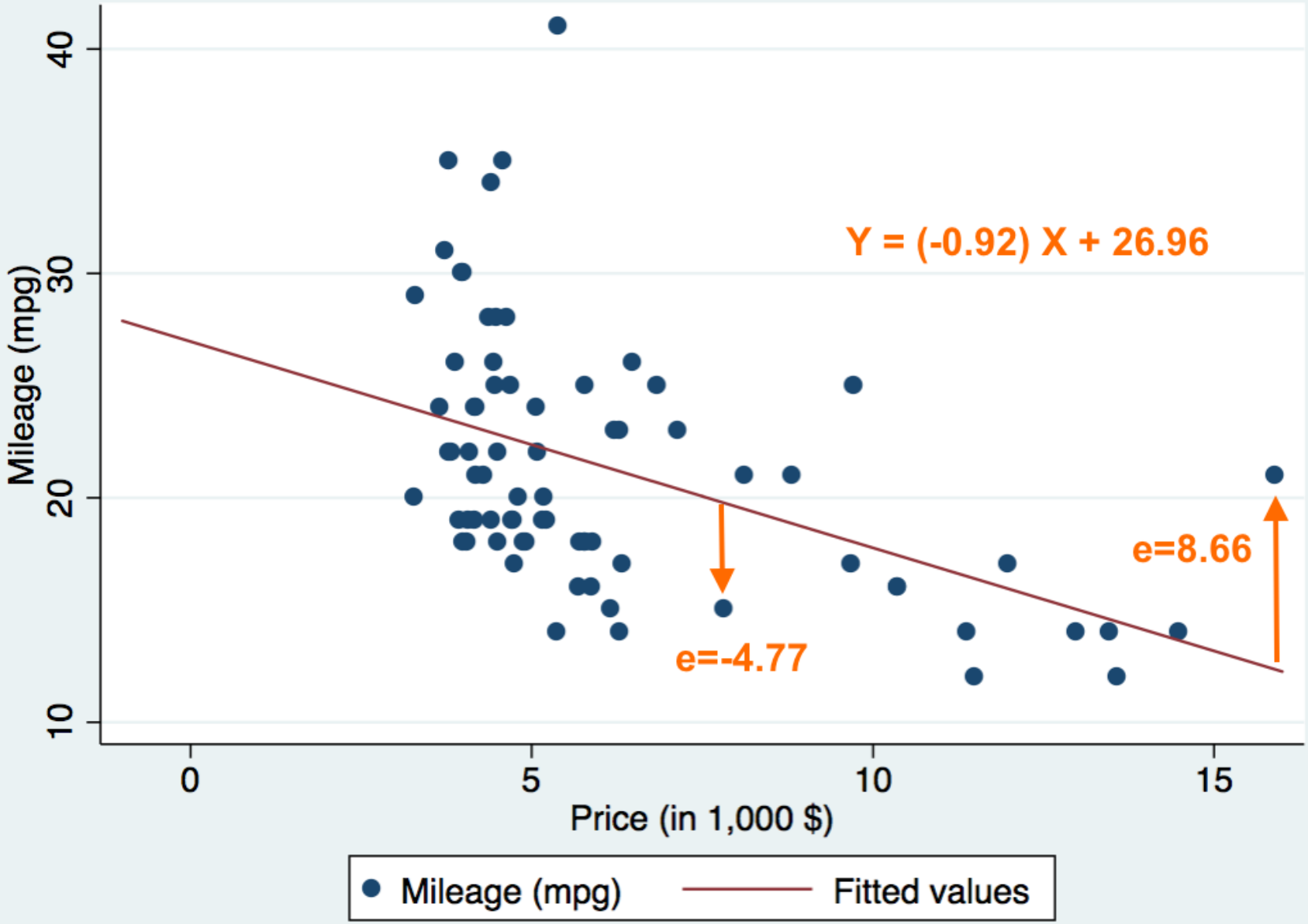
Linear regression

- No wonder why this is called ordinary **least squares** (OLS) estimation (also known as the OLS model)
- The OLS estimation draws a line which **minimizes** the sum of **squares** of residuals
- Intuitively, residuals mean variation of observed Y that the OLS model can't (or fails) to explain

Group exercise 3

A set of data about vintage 1978 automobiles sold in the US

Should you run linear regression, what would be $\hat{\alpha}$, $\hat{\beta}$ and e for selected points in this figure? Make your best guesses!



Linear regression

In the end, we want to know (at the population level)

1. if there is **any linear** relationship between X and Y
2. if any, the **direction** of the **strength** of association (i.e. the size and the sign of β)

Note that we interpret β as the change in Y we see for a one unit change in X
β and correlation coefficient contain the same information

Linear regression

We want to test the alternative hypothesis:

- H_0: there is no linear relationship between X and Y

$$\beta = 0$$

- H_A: there is linear relationship between X and Y

$$\beta \neq 0$$

Linear regression

Hypothesis testing is as follows:

1. Construct a hypothesis
2. Collect data
3. Choose a significance level (e.g. 0.05)
4. Calculate t-statistic (done by STATA)

$$t = \frac{\hat{\beta}}{\text{SE}(\hat{\beta})}$$

5. Calculate p-value (done by STATA)
6. Compare our p-value with the significance level
7. Reject or fail to reject the null hypothesis

Linear regression

Alternatively we can construct a 95% confidence interval of β:

$$\hat{\beta} \pm t_{0.025,n-k-1} \times \text{SE}(\hat{\beta})$$

where

n: the total number of observations
k: the number of independent variables

Easy to calculate since STATA provides $\hat{\beta}$ and $\text{SE}(\hat{\beta})$

- If the CI contains zero, then we fail to reject H_0: $\beta = 0$
- If the CI does not contain zero, then we reject H_0: $\beta = 0$

Group exercise 4

This is the result of an OLS estimation between average homework score (X) and exam score (Y) of some students:

scores	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hw_average	.0007281	.1357555	0.01	0.996	-.2678875	.2693436
_cons	82.22634	10.94083	7.52	0.000	60.57804	103.8746

1. What's the alternative hypothesis?
2. What is the coefficient ($\hat{\beta}$)?
3. What's the substantive interpretation of the coefficient?
4. What is the constant ($\hat{\alpha}$)?
5. What's the p-value for hypothesis testing?
6. Would you reject the null hypothesis?

Why linear?

The real relationship between X and Y could be: $Y_i = f(X_i) + \epsilon_i$

1. From calculus, we know that $f(X_i)$ can be approximated by a polynomial of X_i (Taylor expansion)
2. When $f(X_i)$ is not linear, we could approximate it using a linear function
3. The question becomes: How good is this approximation?
4. Usually it is not bad (that's why linear regression is so popular)

Key takeaways

You should be able to answer these questions:

- When comparing group means, the confidence interval includes zero means what?

Key takeaways

You should be able to answer these questions:

- When comparing group means, the confidence interval includes zero means what?
  We fail to reject the null hypothesis
- Correlation coefficient?

Key takeaways

You should be able to answer these questions:

- When comparing group means, the confidence interval includes zero means what?
  We fail to reject the null hypothesis
- Correlation coefficient?
  A measure of association like λ, Cramer's V, and γ
  (it tells us about direction and strength of a relationship)
- Linear regression?

Key takeaways

You should be able to answer these questions:

- When comparing group means, the confidence interval includes zero means what?
  We fail to reject the null hypothesis
- Correlation coefficient?
  A measure of association like λ, Cramer's V, and γ
  (it tells us about direction and strength of a relationship)
- Linear regression?
  Assuming a linear association between X and Y in population, estimate that relationship (choose $\hat{\alpha}$ and $\hat{\beta}$) with a sample
- What's the OLS estimation?

Key takeaways

You should be able to answer these questions:

- When comparing group means, the confidence interval includes zero means what?
  We fail to reject the null hypothesis
- Correlation coefficient?
  A measure of association like λ, Cramer's V, and γ
  (it tells us about direction and strength of a relationship)
- Linear regression?
  Assuming a linear association between X and Y in population, estimate that relationship (choose $\hat{\alpha}$ and $\hat{\beta}$) with a sample
- What's the OLS estimation?
  Ordinary least squares estimation (minimizing the sum of squares of residuals)