Indicator and Interaction Variables

Econ 2560, Fall 2023

Prof. Josh Abel

(Chapters 5.3, 8.3)

Introduction

- We have learned the core idea underlying regression
  - "How does (residual) variation in X explain variation in Y?"
- The next 2 slide decks explore very useful common and useful wrinkles that can help you tailor your regression to your research question
- This slide deck looks at 2 that tie very naturally to multivariate regression
  1. Dummy/indicator variables
  2. Interactions
- Everything we've learned still holds, we're just applying it in new settings!

Indicator variables

- Gender is an example of an **indicator variable**
- An indicator is a binary variable equal to 1 if a condition is met and 0 otherwise
- E.g.

$$\text{Gender} = \begin{cases} 1 \text{ if Female} \\ 0 \text{ otherwise} \end{cases}$$

- An indicator is a special case of a **categorical variable**, which can have more than 2 values – note they are "unordered"

$$\text{Race} = \begin{cases} 1 \text{ if White, Hispanic} \\ 2 \text{ if White, non-Hispanic} \\ 3 \text{ if Non-White Hispanic} \\ 4 \text{ if Black} \\ 5 \text{ if Asian} \\ 6 \text{ otherwise} \end{cases}$$

A univariate regression function with an indicator variable

$$E[\text{Income}_i|\text{Gender}_i] = \beta_0 + \beta_1 \cdot \text{Gender}_i$$

- What is the expected income of a male?

A univariate regression function with an indicator variable

$$E[\text{Income}_i|\text{Gender}_i] = \beta_0 + \beta_1 \cdot \text{Gender}_i$$

- What is the expected income of a male?
  - β_0

A univariate regression function with an indicator variable

$$E[\text{Income}_i|\text{Gender}_i] = \beta_0 + \beta_1 \cdot \text{Gender}_i$$

- What is the expected income of a male?
  - β_0
- What is the expected income of a female?

A univariate regression function with an indicator variable

$$E[\text{Income}_i|\text{Gender}_i] = \beta_0 + \beta_1 \cdot \text{Gender}_i$$

- What is the expected income of a male?
  - β_0
- What is the expected income of a female?
  - $\beta_0 + \beta_1$
- So β_1 is the difference in average incomes between men and women
  - So regression gives us an easy way to test for the difference in means
  - Just do inference on β_1!

Results

Outcome: Earnings	
Constant	80.95
(SE)	(1.84)
Gender	-32.99
	(2.07)

- Do men and women earn different amounts on average?

Results

Outcome: Earnings	
Constant	80.95
(SE)	(1.84)
Gender	-32.99
	(2.07)

- Do men and women earn different amounts on average?
  - Yes! The t-stat for the null hypothesis that $\beta_1 = 0$ is $\frac{-32.99}{2.07} = -15.94 << -1.96$.

A regression function with multiple indicator variables

- How do you interpret $\hat{\beta}_1$ in this regression:

$$\text{Income}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{White}_i$$

A regression function with multiple indicator variables

- How do you interpret $\hat{\beta}_1$ in this regression:

$$\text{Income}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{White}_i$$

  - Estimated difference in expected earnings between White and Non-White people

A regression function with multiple indicator variables

- How do you interpret $\hat{\beta}_1$ in this regression:

$$\text{Income}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{White}_i$$

  - Estimated difference in expected earnings between White and Non-White people

- How do you interpret $\hat{\beta}_1$ in this regression:

$$\text{Income}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{White}_i + \hat{\beta}_2 \cdot \text{Black}_i$$

A regression function with multiple indicator variables

- How do you interpret $\hat{\beta}_1$ in this regression:

$$\text{Income}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{White}_i$$

  - Estimated difference in expected earnings between White and Non-White people

- How do you interpret $\hat{\beta}_1$ in this regression:

$$\text{Income}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{White}_i + \hat{\beta}_2 \cdot \text{Black}_i$$

  - ...White people and people who are neither White nor Black

- Why can't OLS estimate this regression:

$$\text{Income}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{White}_i + \hat{\beta}_2 \cdot \text{Black}_i + \hat{\beta}_3 \cdot \text{Other}_i$$

A regression function with multiple indicator variables

- How do you interpret $\hat{\beta}_1$ in this regression:

$$\text{Income}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{White}_i$$

  - Estimated difference in expected earnings between White and Non-White people

- How do you interpret $\hat{\beta}_1$ in this regression:

$$\text{Income}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{White}_i + \hat{\beta}_2 \cdot \text{Black}_i$$

  - ...White people and people who are neither White nor Black

- Why can't OLS estimate this regression:

$$\text{Income}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{White}_i + \hat{\beta}_2 \cdot \text{Black}_i + \hat{\beta}_3 \cdot \text{Other}_i$$

  - Multicollinearity! $\text{White}_i + \text{Black}_i + \text{Other}_i = 1$.
  - Need to omit one group: $\hat{\beta}_0$ is their mean

Aside: Non-parametric regression is a form of OLS

- Define the following indicator variables:

$$E_2 = \begin{cases} 1 & \text{if Educ } = 2 \\ 0 & \text{otherwise} \end{cases}, \; E_3 = \begin{cases} 1 & \text{if Educ } = 3 \\ 0 & \text{otherwise} \end{cases} \; ,\ldots, \; E_{20} = \begin{cases} 1 & \text{if Educ } = 20 \\ 0 & \text{otherwise} \end{cases}$$

- Now consider the following regression:

$$E[\text{Income}_i | \{E_{ki}\}] = \beta_0 + \beta_2 \cdot E_{2i} + \beta_3 \cdot E_{3i} + \ldots \beta_{20} \cdot E_{20i}$$

- This shows the mean very every level of education – a non-parametric regression!
  - $E[\text{Income}_i | \text{Educ}_i = 1] = \beta_0$
  - $E[\text{Income}_i | \text{Educ}_i = 2] = \beta_0 + \beta_2$
  - $E[\text{Income}_i | \text{Educ}_i = 8] = \beta_0 + \beta_8$
  - Etc., and we can use the SEs to test for differences
- Takeaway: OLS is as flexible as you want it to be.

A multivariate regression function with an indicator variable

- We can also test for a "gender pay gap" while controlling for education, using:

$$E[\text{Income}_i|\text{Education}_i, \text{Gender}_i] = \beta_0 + \beta_1 \cdot \text{Gender}_i + \beta_2 \cdot \text{Education}_i$$

- Is it a good idea to "control for" education?

A multivariate regression function with an indicator variable

- We can also test for a "gender pay gap" while controlling for education, using:

$$E[\text{Income}_i|\text{Education}_i, \text{Gender}_i] = \beta_0 + \beta_1 \cdot \text{Gender}_i + \beta_2 \cdot \text{Education}_i$$

- Is it a good idea to "control for" education?
  - This one does a better job testing for discrimination in the labor market, as it ensures that gender pay gap is not driven by differential education
  - However, if part of the discrimination is in the education system, this "over-controls" and actually neutralizes some of the effect we should be capturing
- What is the difference in expected income between a male and female with the same education?

A multivariate regression function with an indicator variable

- We can also test for a "gender pay gap" while controlling for education, using:

$$E[\text{Income}_i|\text{Education}_i, \text{Gender}_i] = \beta_0 + \beta_1 \cdot \text{Gender}_i + \beta_2 \cdot \text{Education}_i$$

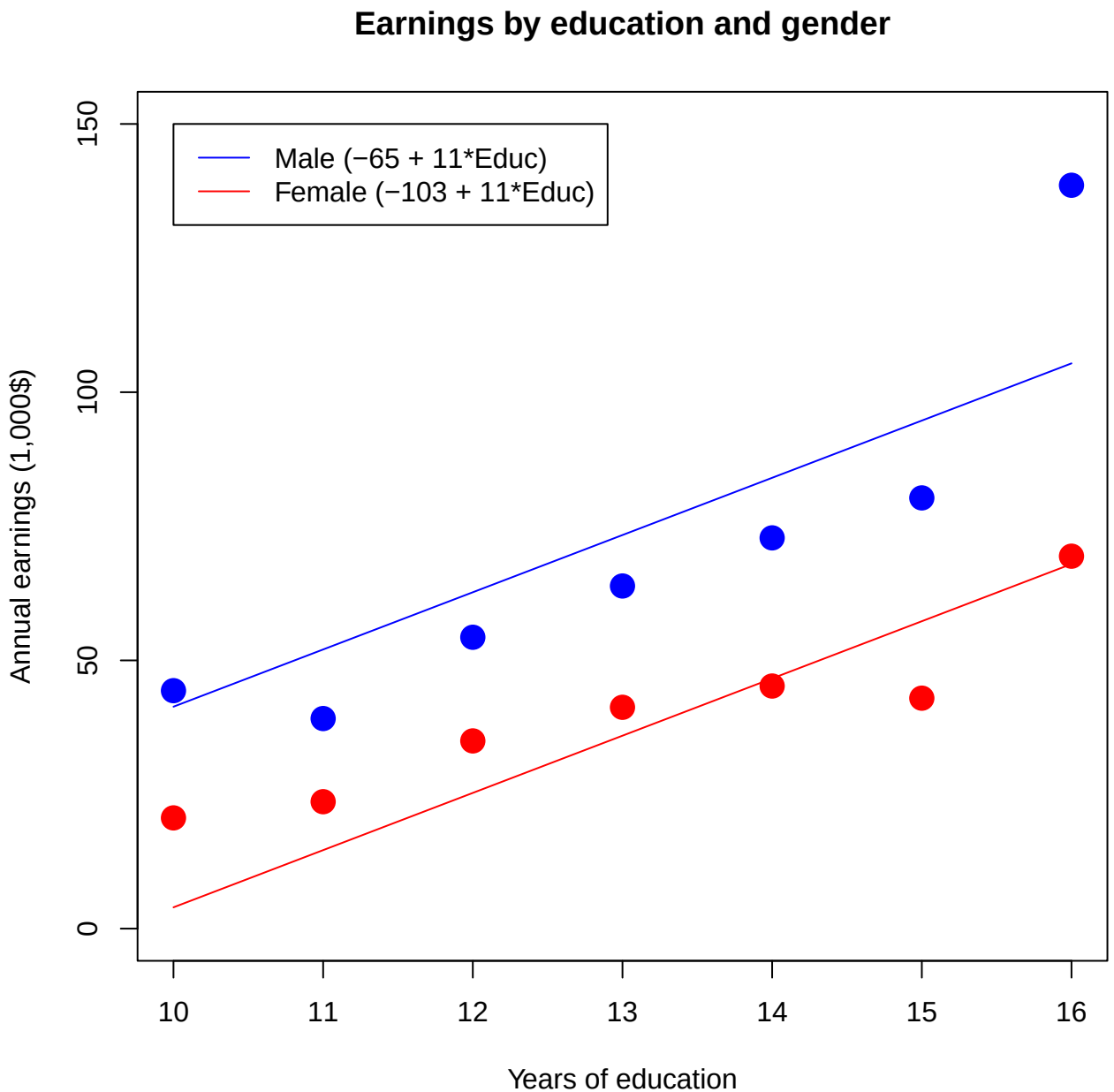
- Is it a good idea to "control for" education?
  - This one does a better job testing for discrimination in the labor market, as it ensures that gender pay gap is not driven by differential education
  - However, if part of the discrimination is in the education system, this "over-controls" and actually neutralizes some of the effect we should be capturing
- What is the difference in expected income between a male and female with the same education?
  - β_1
- The model assumes the same effect of education across Gender, but different intercepts
  - Male intercept is β_0, female intercept is $\beta_0 + \beta_1$

Results

Outcome: Earnings		
Constant	80.95	-65.30
(SE)	(1.84)	(6.33)
Gender	-32.99	-37.40
	(2.07)	(1.99)
Education		10.67
		(0.52)

- Gender pay gap rose.
  - We can understand this through OVB:
    - Women are actually more educated
    - Education raises income
    - So omitting Education means that some of Education's effect was captured by Gender

A regression function with an indicator variable, results



Interpreting indicator variables

- An indicator generates a level shift (moves the constant) between categories
- In some cases, we might care about the shift itself
  - "Men earn more than women, and this is not driven by differences in education between men and women."
- In other cases, it is useful as a control
  - "A year of education increases income by \$10,000/year, and this is not driven by differences in education between men and women."

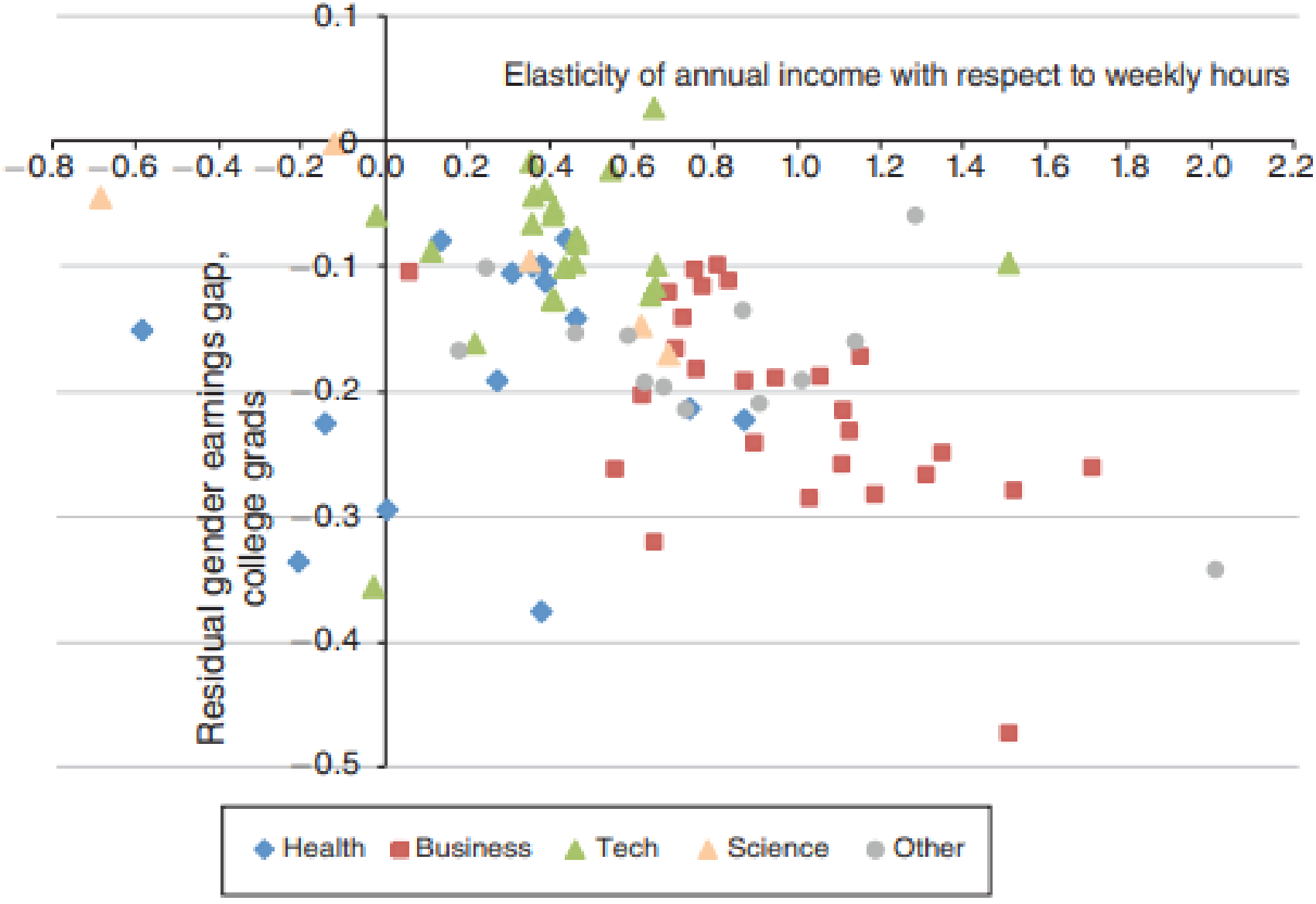
Goldin (2014)

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

Sample	Variables included	Coefficient on female	Standard error
Full-time	Basic	−0.248	0.00101
Full-time	Basic, time	−0.193	0.00100
Full-time	Basic, time, education	−0.247	0.000905
Full-time	Basic, time, education, occupation	−0.192	0.00104
All	Basic	−0.320	0.00105
All	Basic, time	−0.196	0.000925
All	Basic, time, education	−0.245	0.000847
All	Basic, time, education, occupation	−0.191	0.000963
Full-time, BA	Basic	−0.285	0.00159
Full-time, BA	Basic, time	−0.230	0.00158
Full-time, BA	Basic, time, education	−0.233	0.00155
Full-time, BA	Basic, time, education, occupation	−0.163	0.00158
All, BA	Basic	−0.384	0.00173
All, BA	Basic, time	−0.227	0.00151
All, BA	Basic, time, education	−0.229	0.00148
All, BA	Basic, time, education, occupation	−0.163	0.00151

Notes: "Basic" regression is the log of annual earnings regressed on the female dummy, age as a quarti

Goldin (2014)


Elasticity of annual income with respect to weekly hours
Residual gender earnings gap, college grads
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.8 −0.6 −0.4 −0.2 0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2
Health
Business
Tech
Science
Other

Bolotnyy and Emanuel (2021)

Table 4: Gender Differences in Weekly Earnings

	(1)	(2)	(3)	(4)
Female	-160.10***	-158.70***	-145.60***	-138.20***
	(10.17)	(9.92)	(1.41)	(1.58)
Seniority Decile		2.71***	3.06***	3.02***
		(0.15)	(0.16)	(0.16)
Dependents=1			2.76	26.82
			(16.61)	(22.95)
Married=1				52.48***
				(13.94)
Female × Dependents			-33.23	-53.57
			(25.43)	(30.76)
Female × Married				-6.97
				(28.36)
Dependents × Married				-71.67*
				(33.50)
Female × Dependents × Married				85.65
				(64.84)
Constant	1447.30***	1296.30***	1316.00***	1302.70***
	(5.86)	(9.48)	(9.71)	(10.50)
Male Mean	1447.30	1447.30	1447.30	1447.30
Adjusted R^2	0.025	0.053	0.064	0.066
Observations	682,583	682,583	571,344	571,344

Allowing slopes to differ

- Gender indicator captured that male income is typically above female income
- But it looked male slope was also higher
- How would we interpret that?

Allowing slopes to differ

- Gender indicator captured that male income is typically above female income
- But it looked male slope was also higher
- How would we interpret that?
  - Men get higher returns from the marginal year of education

A regression with an interaction

- A regression with an **interaction** can allow for different slopes

$$
\begin{aligned}
&E[\text{Income}_i|\text{Education}_i, \text{Gender}_i] \\
&\quad = \beta_0 + \beta_1 \cdot \text{Gender}_i + \beta_2 \cdot \text{Education}_i + \beta_3 \cdot \text{Gender}_i \cdot \text{Education}_i
\end{aligned}
$$

- Now men and women have different slopes

A regression with an interaction

- A regression with an **interaction** can allow for different slopes

$$E[\text{Income}_i|\text{Education}_i, \text{Gender}_i]$$
$$= \beta_0 + \beta_1 \cdot \text{Gender}_i + \beta_2 \cdot \text{Education}_i + \beta_3 \cdot \text{Gender}_i \cdot \text{Education}_i$$

- Now men and women have different slopes
  - Men: β_2
  - Women: $\beta_2 + \beta_3$

A regression with an interaction

- A regression with an **interaction** can allow for different slopes

$$E[\text{Income}_i|\text{Education}_i, \text{Gender}_i]$$
$$= \beta_0 + \beta_1 \cdot \text{Gender}_i + \beta_2 \cdot \text{Education}_i + \beta_3 \cdot \text{Gender}_i \cdot \text{Education}_i$$

- Now men and women have different slopes
  - Men: β_2
  - Women: $\beta_2 + \beta_3$
- β_3 is the additional impact of a year of education for a woman relative to a man
- Note: if including interaction, (almost) always need to include the two variables separately as well!

A regression with an interaction, results

Outcome: Earnings			
Constant	80.95	-65.30	-122.27
(SE)	(1.84)	(6.33)	(23.78)
Gender	-32.99	-37.40	80.79
	(2.07)	(1.99)	(12.75)
Education		10.67	14.82
		(0.52)	(0.91)
Gender·Educ			-8.49
			(1.00)

A regression with an interaction, results

Outcome: Earnings			
Constant	80.95	-65.30	-122.27
(SE)	(1.84)	(6.33)	(23.78)
Gender	-32.99	-37.40	80.79
	(2.07)	(1.99)	(12.75)
Education		10.67	14.82
		(0.52)	(0.91)
Gender·Educ			-8.49
			(1.00)

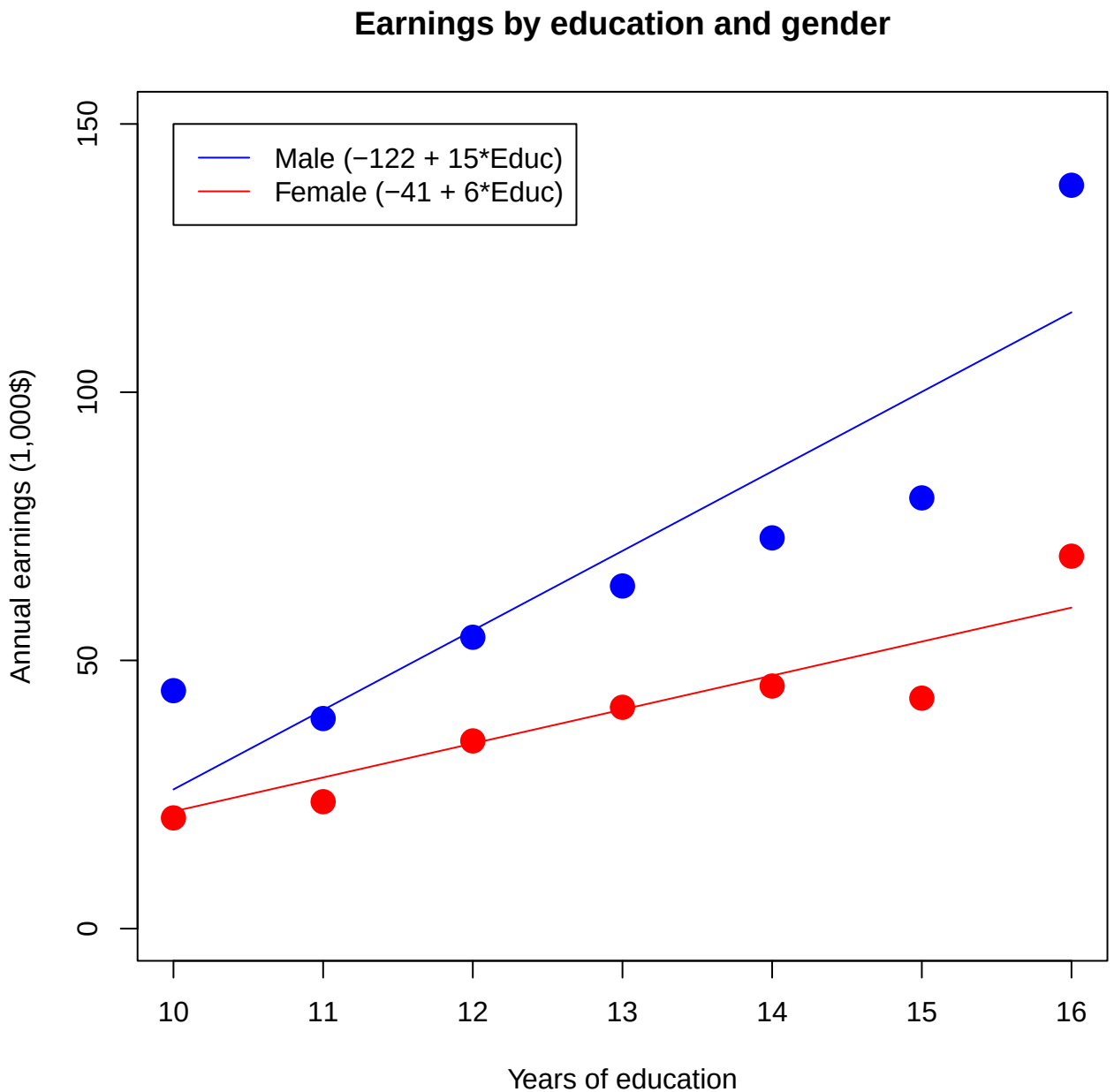
- Gender coefficient now positive. Do women earn more than men?

A regression with an interaction, results

Outcome: Earnings			
Constant	80.95	-65.30	-122.27
(SE)	(1.84)	(6.33)	(23.78)
Gender	-32.99	-37.40	80.79
	(2.07)	(1.99)	(12.75)
Education		10.67	14.82
		(0.52)	(0.91)
Gender·Educ			-8.49
			(1.00)

- Gender coefficient now positive. Do women earn more than men?
  - No – that only strictly applies to people with 0 years of education!
  - Women and men with 10 years of education earn roughly the same amount
    - 80.79 vs. 10·-8.49
  - Above that, men earn more
- As always, slopes and intercepts have to be looked at together

A regression with an interaction, results



Interaction of continuous variables

- We can also interact two continuous variables together

$$
\begin{aligned}
&E[\text{Income}_i|\text{Education}_i, \text{IQ}_i] \\
&\quad = \beta_0 + \beta_1 \cdot \text{Education}_i + \beta_2 \cdot \text{IQ}_i + \beta_3 \cdot \text{IQ}_i \cdot \text{Education}_i
\end{aligned}
$$

Interaction of continuous variables

- We can also interact two continuous variables together

$$E[\text{Income}_i|\text{Education}_i, \text{IQ}_i] = \beta_0 + \beta_1 \cdot \text{Education}_i + \beta_2 \cdot \text{IQ}_i + \beta_3 \cdot \text{IQ}_i \cdot \text{Education}_i$$

- β_3 measures "complementarities" between IQ and Education

$$\frac{\partial^2 E[\text{Income}|\text{Education}_i, \text{IQ}_i]}{\partial \text{Education}_i \partial \text{IQ}_i} = \beta_3$$

Interaction of continuous variables

- We can also interact two continuous variables together

$$E[\text{Income}_i|\text{Education}_i, \text{IQ}_i] = \beta_0 + \beta_1 \cdot \text{Education}_i + \beta_2 \cdot \text{IQ}_i + \beta_3 \cdot \text{IQ}_i \cdot \text{Education}_i$$

- β_3 measures "complementarities" between IQ and Education

$$\frac{\partial^2 E[\text{Income}|\text{Education}_i, \text{IQ}_i]}{\partial \text{Education}_i \partial \text{IQ}_i} = \beta_3$$

  - $\beta_3 > 0$ means that an additional year of education leads to a bigger increase in income for people with higher IQs

A regression with a continuous interaction, results

Outcome: Earnings					
Constant	80.95	-65.30	-122.27	-53.17	32.13
(SE)	(1.84)	(6.33)	(11.59)	(6.58)	(10.46)
Gender	-32.99	-37.40	80.79		
	(2.07)	(1.99)	(12.75)		
Education		10.67	14.82	6.73	0.33
		(0.52)	(0.91)	(0.52)	(0.84)
Gender·Educ			-8.49		
			(1.00)		
AFQT				0.54	-1.26
				(0.04)	(0.25)
AFQT·Educ					0.13
					(0.02)

Interpreting a continuous interaction

$$E[\text{Income}_i|\text{Education}_i, \text{IQ}_i]$$
$$= 32.13 + 0.33 \cdot \text{Education}_i - 1.26 \cdot \text{IQ}_i + 0.13 \cdot \text{IQ}_i \cdot \text{Education}_i$$

- In the 90^{th} percentile of AFQT, the marginal year of education:

Interpreting a continuous interaction

$$E[\text{Income}_i|\text{Education}_i, \text{IQ}_i]$$
$$= 32.13 + 0.33 \cdot \text{Education}_i - 1.26 \cdot \text{IQ}_i + 0.13 \cdot \text{IQ}_i \cdot \text{Education}_i$$

- In the 90^{th} percentile of AFQT, the marginal year of education:
  - increases income by $0.33 + 0.13 \cdot 90 = 12.03$ (thousand \$/yr)

Interpreting a continuous interaction

$$E[\text{Income}_i|\text{Education}_i, \text{IQ}_i]$$
$$= 32.13 + 0.33 \cdot \text{Education}_i - 1.26 \cdot \text{IQ}_i + 0.13 \cdot \text{IQ}_i \cdot \text{Education}_i$$

- In the 90^{th} percentile of AFQT, the marginal year of education:
  - increases income by $0.33 + 0.13 \cdot 90 = 12.03$ (thousand \$/yr)
- In the 10^{th} percentile of AFQT, the marginal year of education:

Interpreting a continuous interaction

$$E[\text{Income}_i|\text{Education}_i, \text{IQ}_i]$$
$$= 32.13 + 0.33 \cdot \text{Education}_i - 1.26 \cdot \text{IQ}_i + 0.13 \cdot \text{IQ}_i \cdot \text{Education}_i$$

- In the 90^{th} percentile of AFQT, the marginal year of education:
  - increases income by $0.33 + 0.13 \cdot 90 = 12.03$ (thousand \$/yr)
- In the 10^{th} percentile of AFQT, the marginal year of education:
  - increases income by $0.33 + 0.13 \cdot 10 = 1.63$ (thousand \$/yr)

Interpreting a continuous interaction

$$E[\text{Income}_i|\text{Education}_i, \text{IQ}_i]$$
$$= 32.13 + 0.33 \cdot \text{Education}_i - 1.26 \cdot \text{IQ}_i + 0.13 \cdot \text{IQ}_i \cdot \text{Education}_i$$

- In the 90^{th} percentile of AFQT, the marginal year of education:
  - increases income by $0.33 + 0.13 \cdot 90 = 12.03$ (thousand \$/yr)
- In the 10^{th} percentile of AFQT, the marginal year of education:
  - increases income by $0.33 + 0.13 \cdot 10 = 1.63$ (thousand \$/yr)
- With an 8^{th} grade education, being 1 AFQT centile higher:

Interpreting a continuous interaction

$$E[\text{Income}_i|\text{Education}_i, \text{IQ}_i]$$
$$= 32.13 + 0.33 \cdot \text{Education}_i - 1.26 \cdot \text{IQ}_i + 0.13 \cdot \text{IQ}_i \cdot \text{Education}_i$$

- In the 90^{th} percentile of AFQT, the marginal year of education:
  - increases income by $0.33 + 0.13 \cdot 90 = 12.03$ (thousand \$/yr)
- In the 10^{th} percentile of AFQT, the marginal year of education:
  - increases income by $0.33 + 0.13 \cdot 10 = 1.63$ (thousand \$/yr)
- With an 8^{th} grade education, being 1 AFQT centile higher:
  - increases income by -1.26 + 0.13·8 = -0.22 (i.e. decreases income!)

Interpreting a continuous interaction

$$E[\text{Income}_i|\text{Education}_i, \text{IQ}_i]$$
$$= 32.13 + 0.33 \cdot \text{Education}_i - 1.26 \cdot \text{IQ}_i + 0.13 \cdot \text{IQ}_i \cdot \text{Education}_i$$

- In the 90^{th} percentile of AFQT, the marginal year of education:
  - increases income by $0.33 + 0.13 \cdot 90 = 12.03$ (thousand \$/yr)
- In the 10^{th} percentile of AFQT, the marginal year of education:
  - increases income by $0.33 + 0.13 \cdot 10 = 1.63$ (thousand \$/yr)
- With an 8^{th} grade education, being 1 AFQT centile higher:
  - increases income by $-1.26 + 0.13 \cdot 8 = -0.22$ (i.e. decreases income!)
- With 16 years of education, being 1 AFQT centile higher:

Interpreting a continuous interaction

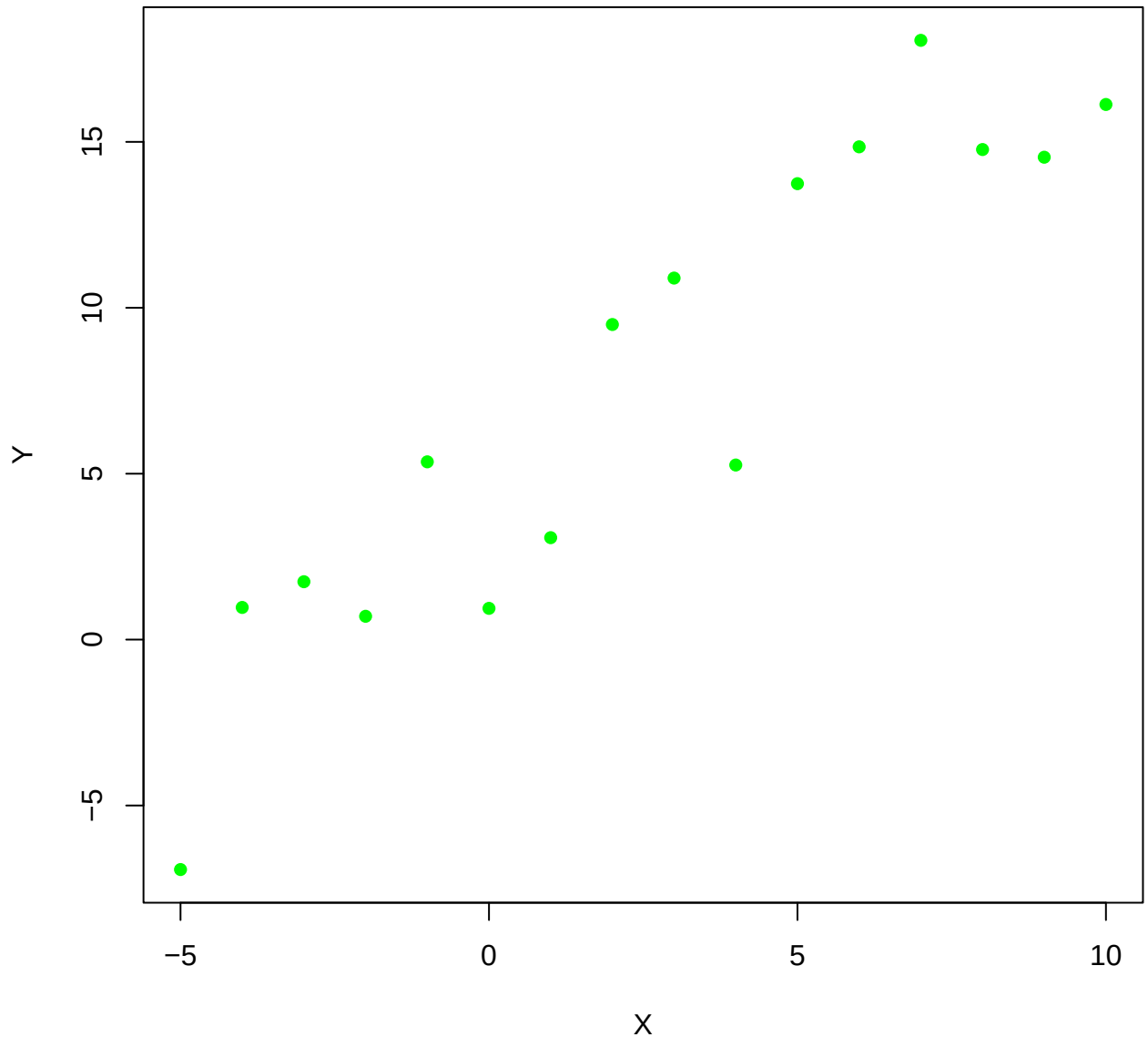
$$E[\text{Income}_i|\text{Education}_i, \text{IQ}_i]$$
$$= 32.13 + 0.33 \cdot \text{Education}_i - 1.26 \cdot \text{IQ}_i + 0.13 \cdot \text{IQ}_i \cdot \text{Education}_i$$

- In the 90^{th} percentile of AFQT, the marginal year of education:
  - increases income by $0.33 + 0.13 \cdot 90 = 12.03$ (thousand \$/yr)
- In the 10^{th} percentile of AFQT, the marginal year of education:
  - increases income by $0.33 + 0.13 \cdot 10 = 1.63$ (thousand \$/yr)
- With an 8^{th} grade education, being 1 AFQT centile higher:
  - increases income by $-1.26 + 0.13 \cdot 8 = -0.22$ (i.e. decreases income!)
- With 16 years of education, being 1 AFQT centile higher:
  - increases income by $-1.26 + 0.13 \cdot 16 = 0.82$

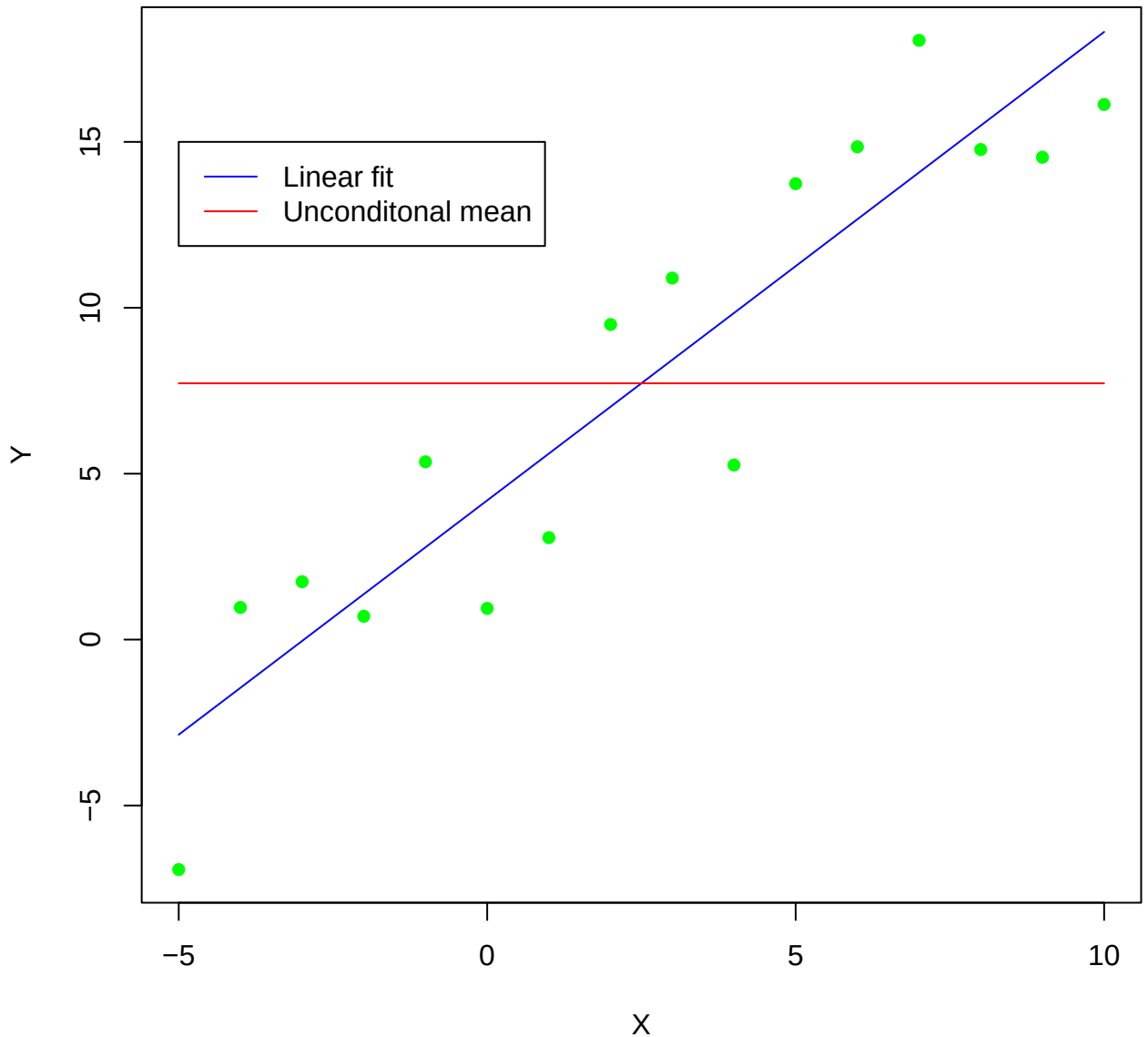
Sidebar: R^2

- R^2 is a measure of how well a model fits the data
- It answers the following question:
  - "How much closer to the actual data (Y_i) do I get by using my fitted values ($\hat{Y}_i$) than if I had just guessed the sample mean ($\bar{Y}$)?"

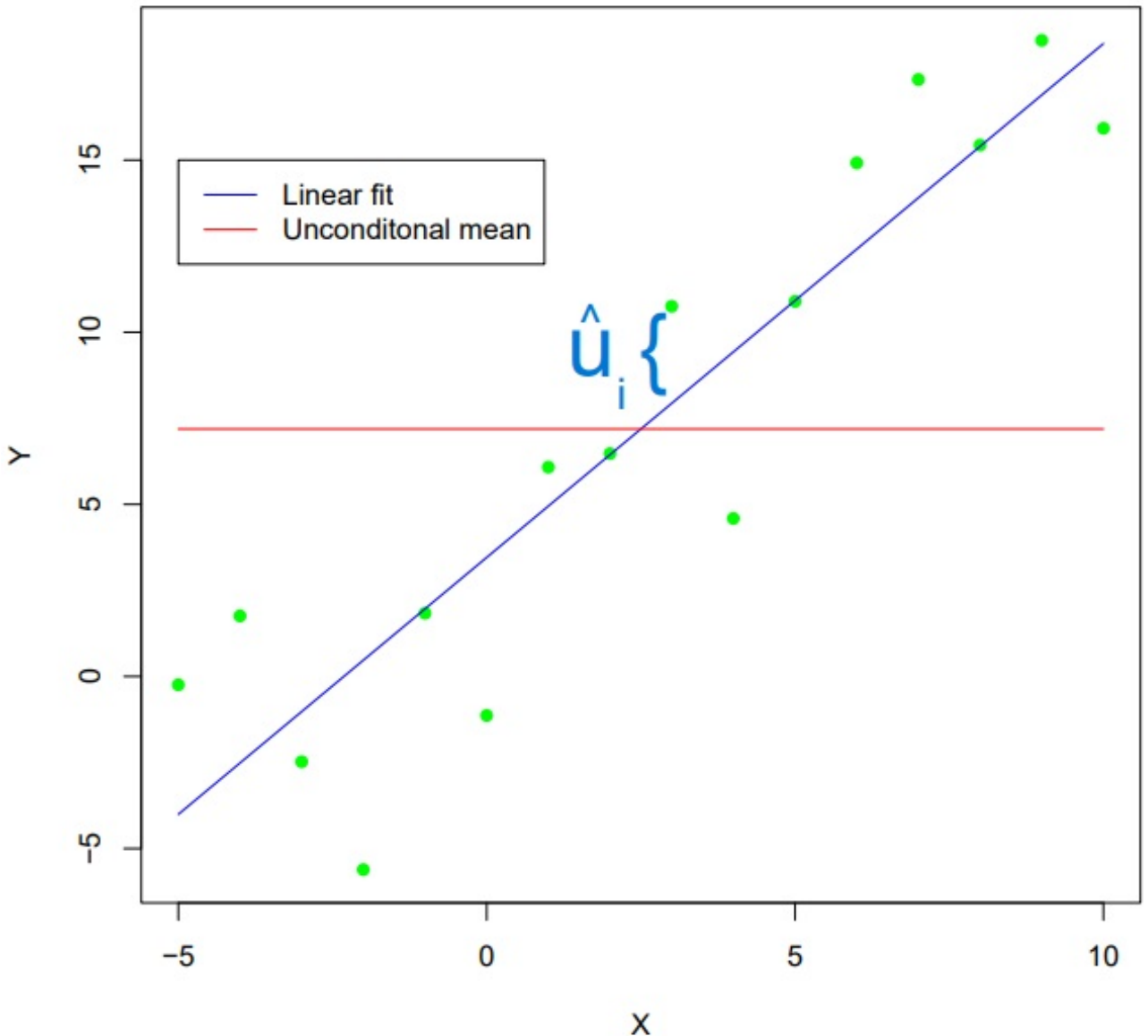
R^2 visual



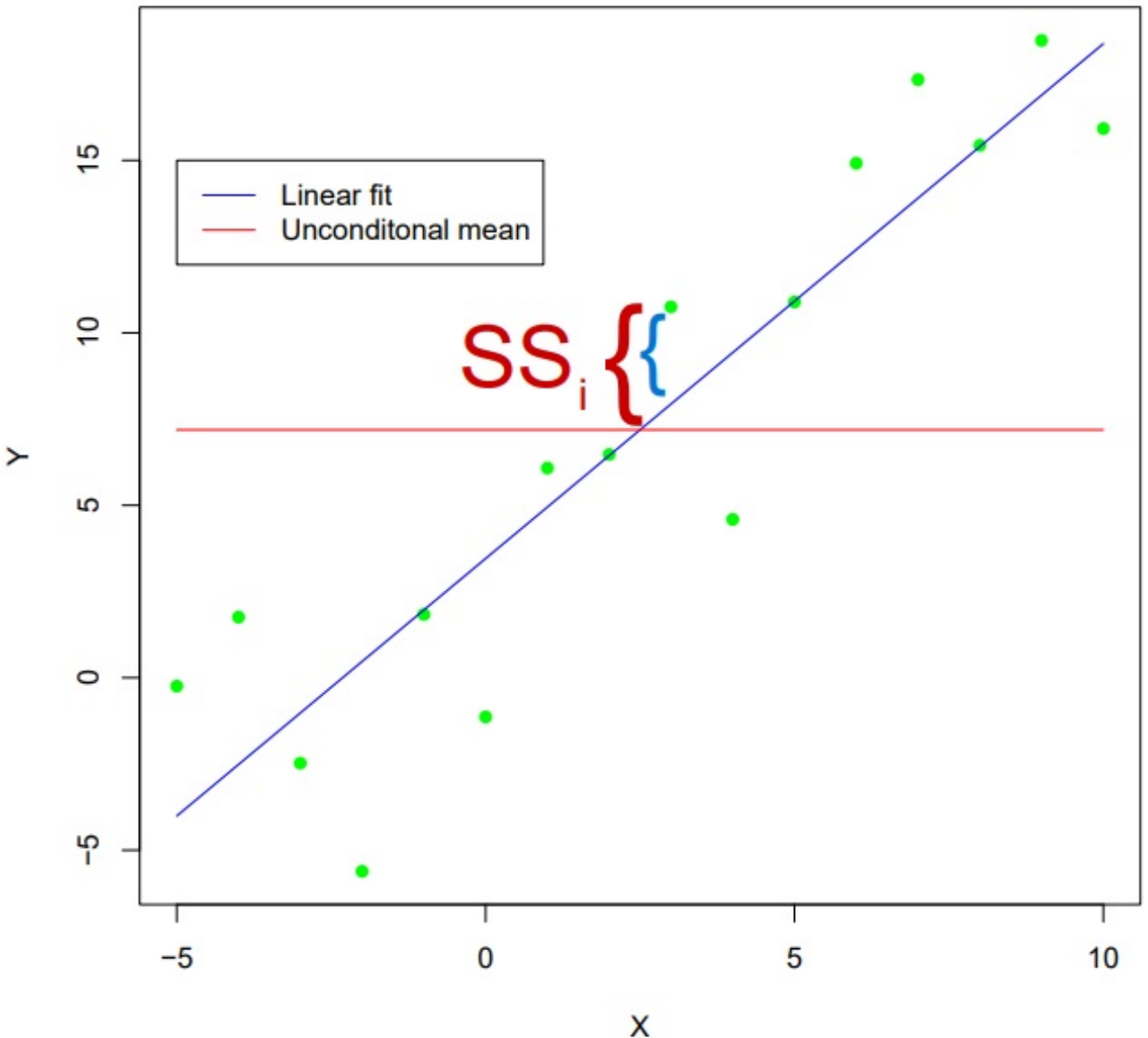
R^2 visual



R^2 visual



R^2 visual



R^2 formula

$$TSS = \sum_{i=1}^{N} (Y_i - \bar{Y})^2$$

$$SSR = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{N} \hat{u}_i^2$$

$$R^2 = 1 - \frac{SSR}{TSS}$$

- Share of squared residuals we eliminate by using fitted values rather than mean

Comments on R^2

- R^2 is a measure of in-sample predictive accuracy
  - OLS maximizes R^2
- $R^2 \in [0, 1]$; $R^2 = \rho^2$
  - Measures tightness of relationship, not slope
- It is not a summary measure of how "good" the model is
  - Can easily come up with inane models with high R^2s
- Can only compare R^2 across models if they have the same Y_i
- **Adding variables to the model will always increase $\mathbf{R^2}$**
  - If a variable were truly useless, the model would give it a coefficient of 0, so R^2 wouldn't change
  - If there is any predictive content, R^2 will increase

Adjusted R^2

- Because of the problems with R^2, economists sometimes refer to "adjusted R^2"

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS},$$

  where k is the number of regressors.
- Recall $R^2 = 1 - \frac{SSR}{TSS}$, so if $n >> k$, $\bar{R}^2 \approx R^2$
- $\bar{R}^2$ is like R^2, but it punishes models with many regressors
  - If you add a useless variable, R^2 will increase but $\bar{R}^2$ will probably fall