Today's Agenda

Multiple OLS Regression Modeling

1. Fitting the models
2. Interpreting the models
3. Evaluating the models

Justin Leinaweaver (Summer 2023)

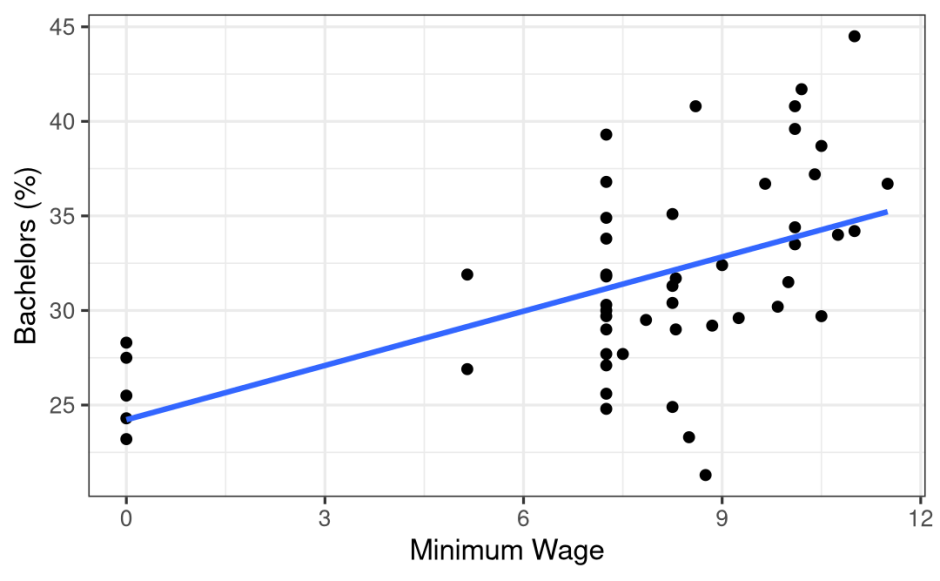
Session 2: Simple OLS Regression

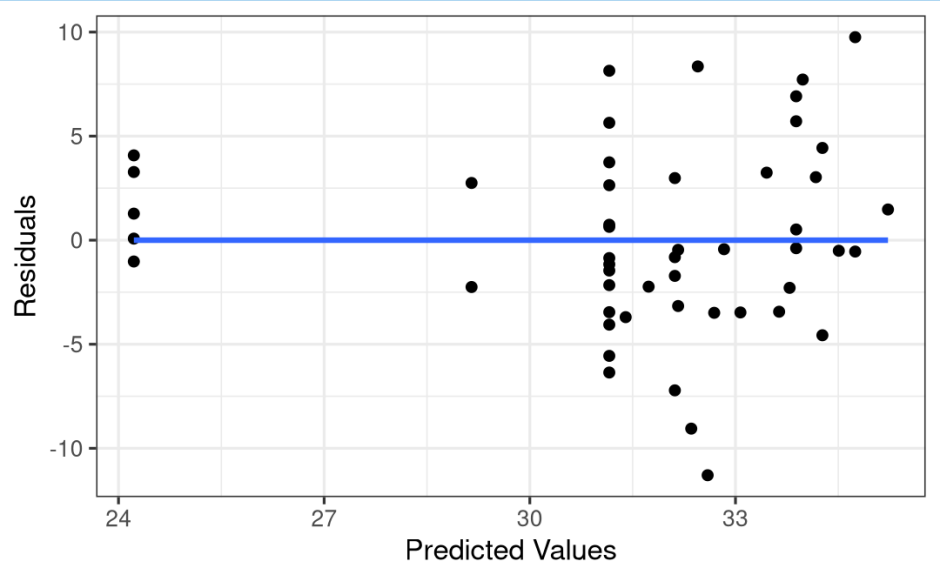
Estimating a linear relationship between two variables

1. Fitting in Excel
2. Interpreting the results
3. Evaluating the fit (4 steps)
4. Making point estimates

Do states that pay a higher minimum wage see more people finish college?

1. Visualize the relationship (scatter plot)
2. Regress bachelors on min_wage
3. Evaluate the fit (four steps)
4. Make three predictions
    - Minimum wage at the minimum, median, and the maximum





	Bachelors (%)
Minimum Wage	0.96*
	(0.22)
Constant	24.22*
	(1.79)
Observations	50
Adjusted R^2	0.27
Residual Std. Error	4.53 (df = 48)
F Statistic	19.36* (df = 1; 48)
Note:	*p<0.05

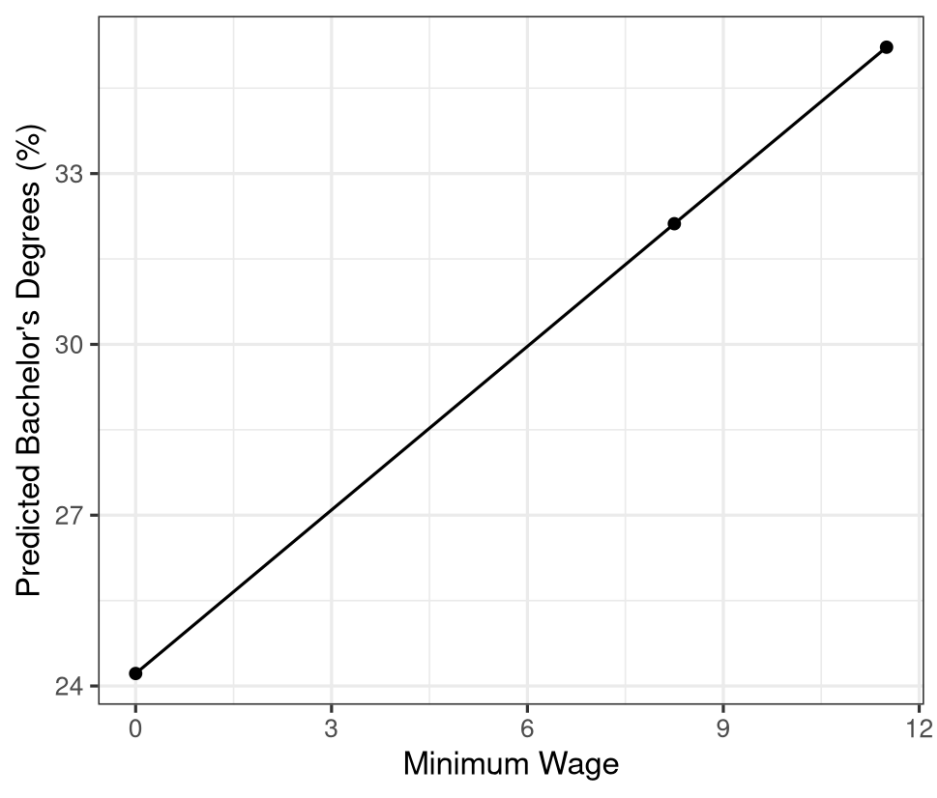
	Bachelors (%)
Minimum Wage	0.96*
	(0.22)
Constant	24.22*
	(1.79)
Observations	50
Adjusted R^2	0.27
Residual Std. Error	4.53 (df = 48)
F Statistic	19.36* (df = 1; 48)
Note:	*p<0.05

Minimum Wage	**Predicted Bachelors**
Minimum = 0	24.22
Median = 8.25	32.12
Maximum = 11.50	35.22

A Marginal Effects Plot

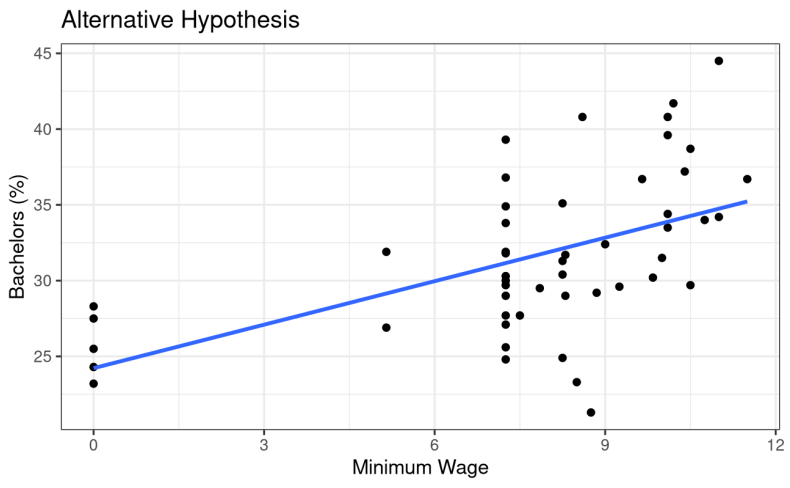
Minimum Wage	Predicted Bachelors
0.00	24.22
8.25	32.12
11.50	35.22

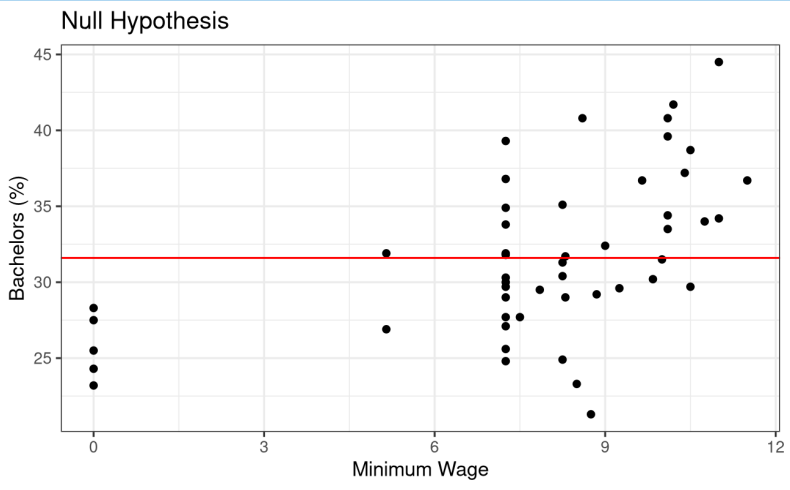
Presenting OLS Regression Results



	Bachelors (%)
Minimum Wage	0.96*
	(0.22)
Constant	24.22*
	(1.79)
Observations	50
Adjusted R^2	0.27
Residual Std. Error	4.53 (df = 48)
F Statistic	19.36* (df = 1; 48)
Note:	*p<0.05

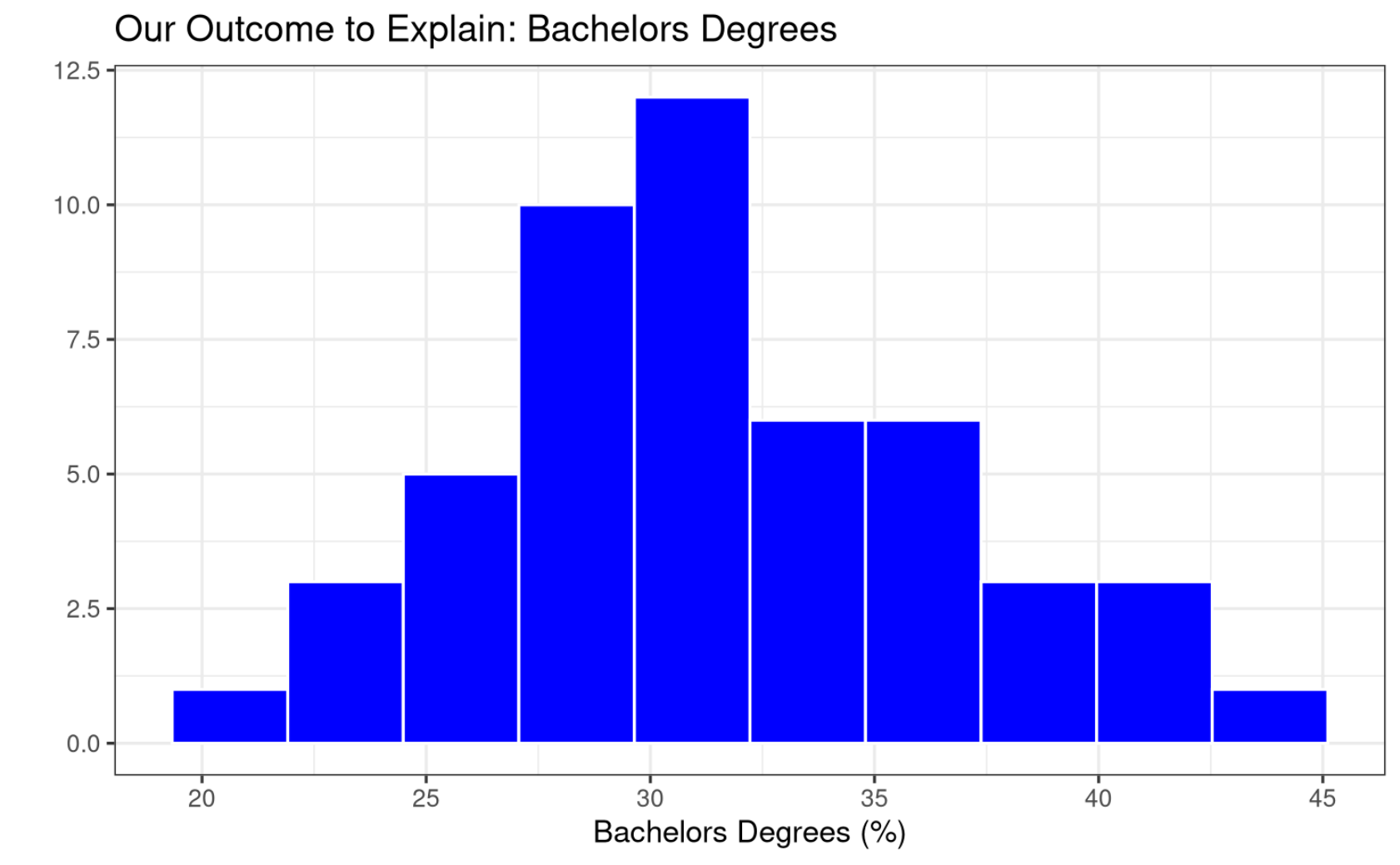
Statistical Significance: Take Two!



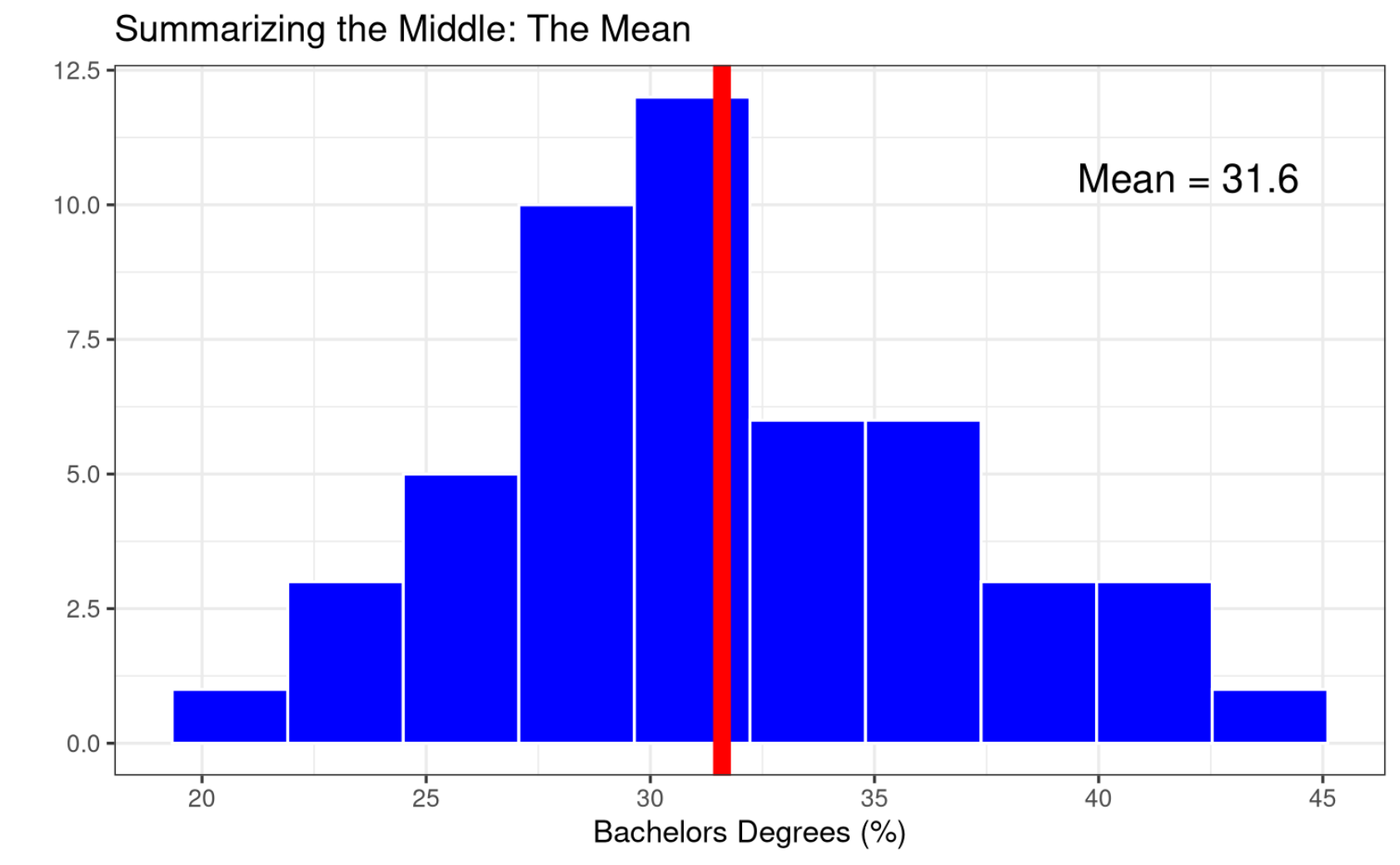


	Bachelors (%)
Minimum Wage	0.96*
	(0.22)
Constant	24.22*
	(1.79)
Observations	50
Adjusted R^2	0.27
Residual Std. Error	4.53 (df = 48)
F Statistic	19.36* (df = 1; 48)
Note:	*p<0.05

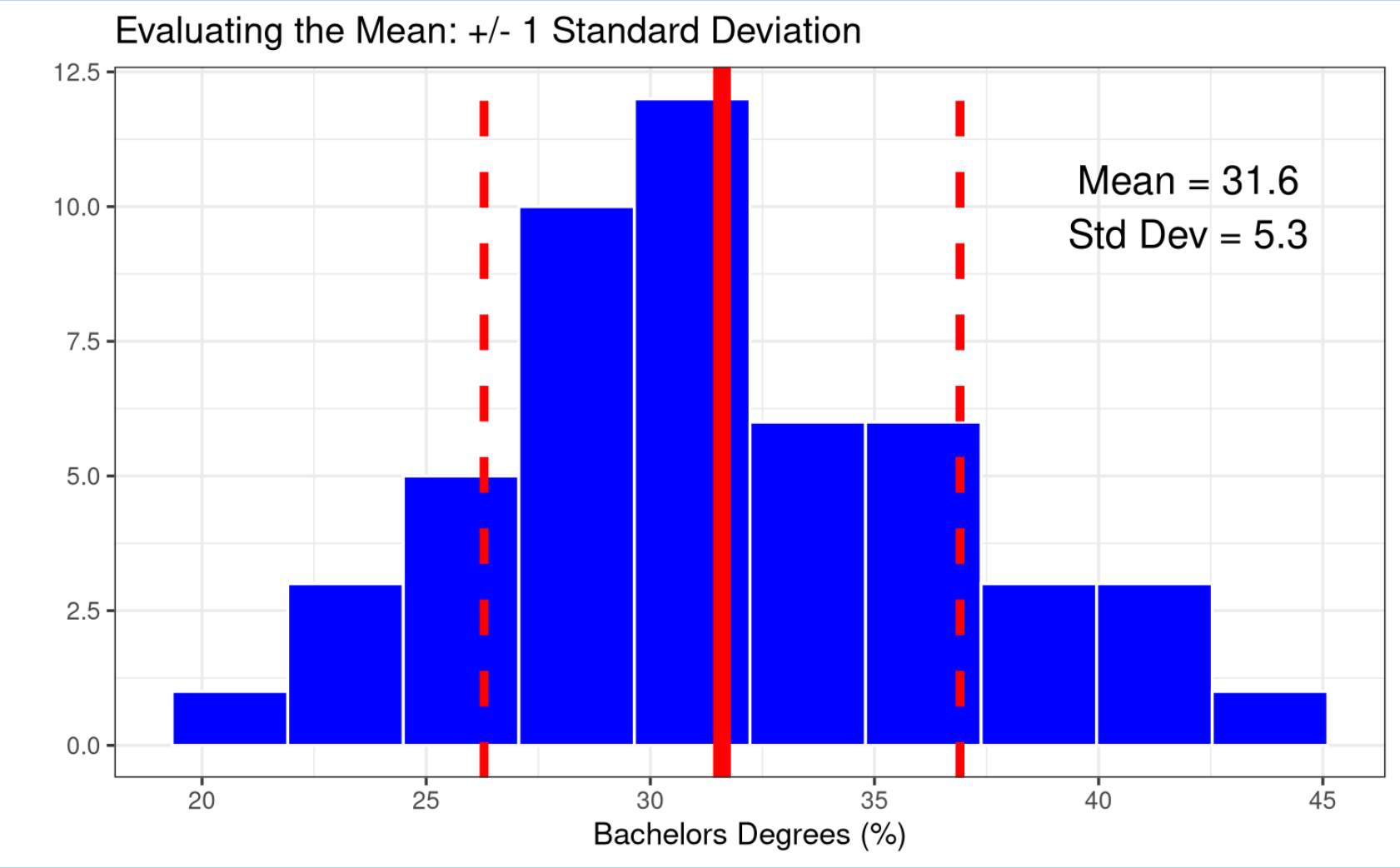
Statistical Significance: What is the Standard Error?



Statistical Significance: What is the Standard Error?



Statistical Significance: What is the Standard Error?

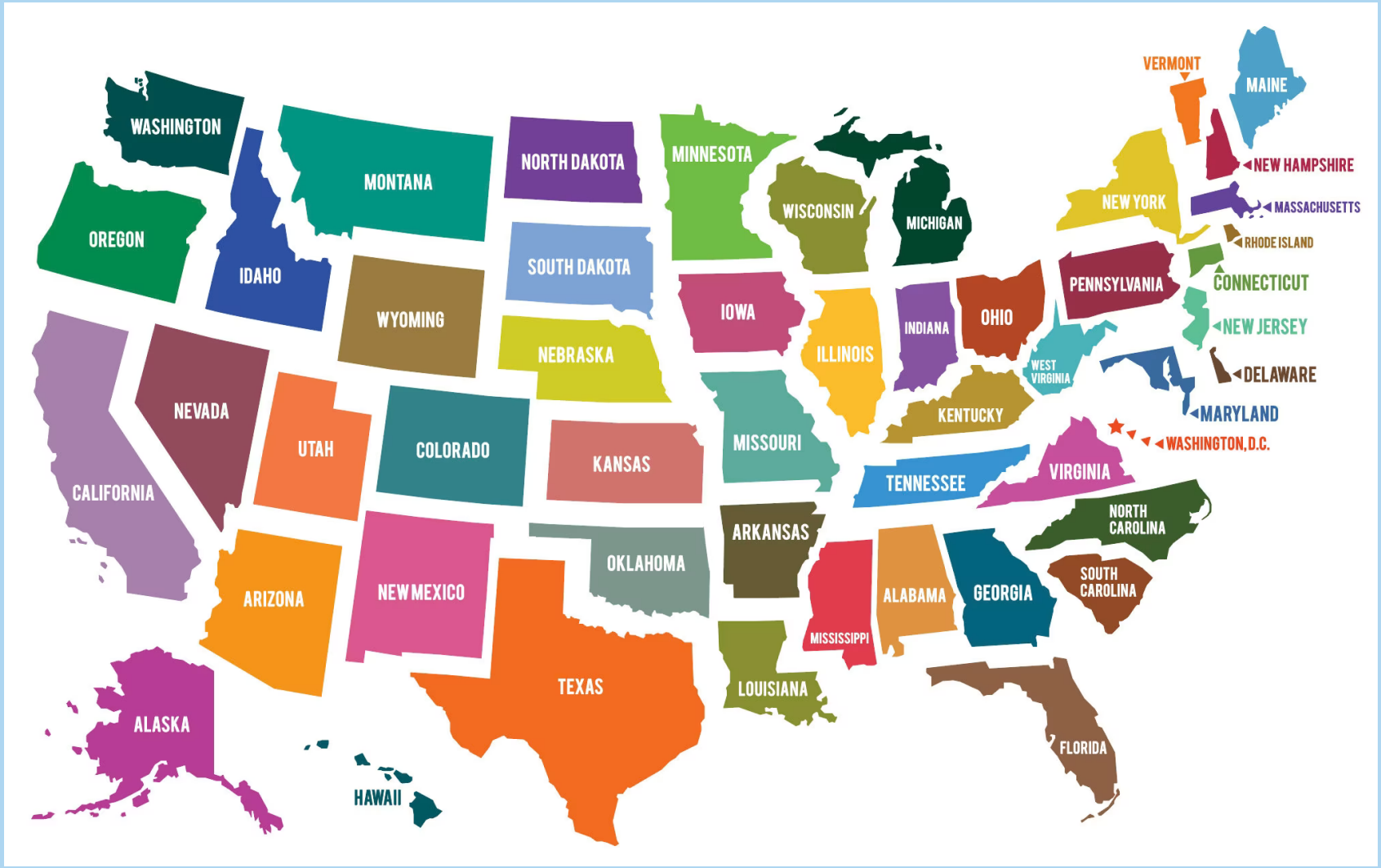


Statistical Significance: What is the Standard Error?

state	year	min_wage	bachelors
New Mexico	2018	7.50	27.7
Virginia	2018	7.25	39.3
Alabama	2018	0.00	25.5
North Dakota	2018	7.25	29.7
Tennessee	2018	0.00	27.5
South Carolina	2018	0.00	28.3
Nebraska	2018	9.00	32.4
Montana	2018	8.30	31.7
Oklahoma	2018	7.25	25.6
Maine	2018	10.00	31.5

	Bachelors (%)
Minimum Wage	0.96*
	(0.22)
Constant	24.22*
	(1.79)
Observations	50
Adjusted R^2	0.27
Residual Std. Error	4.53 (df = 48)
F Statistic	19.36* (df = 1; 48)
Note:	*p<0.05

Statistical Significance: What is the Standard Error?



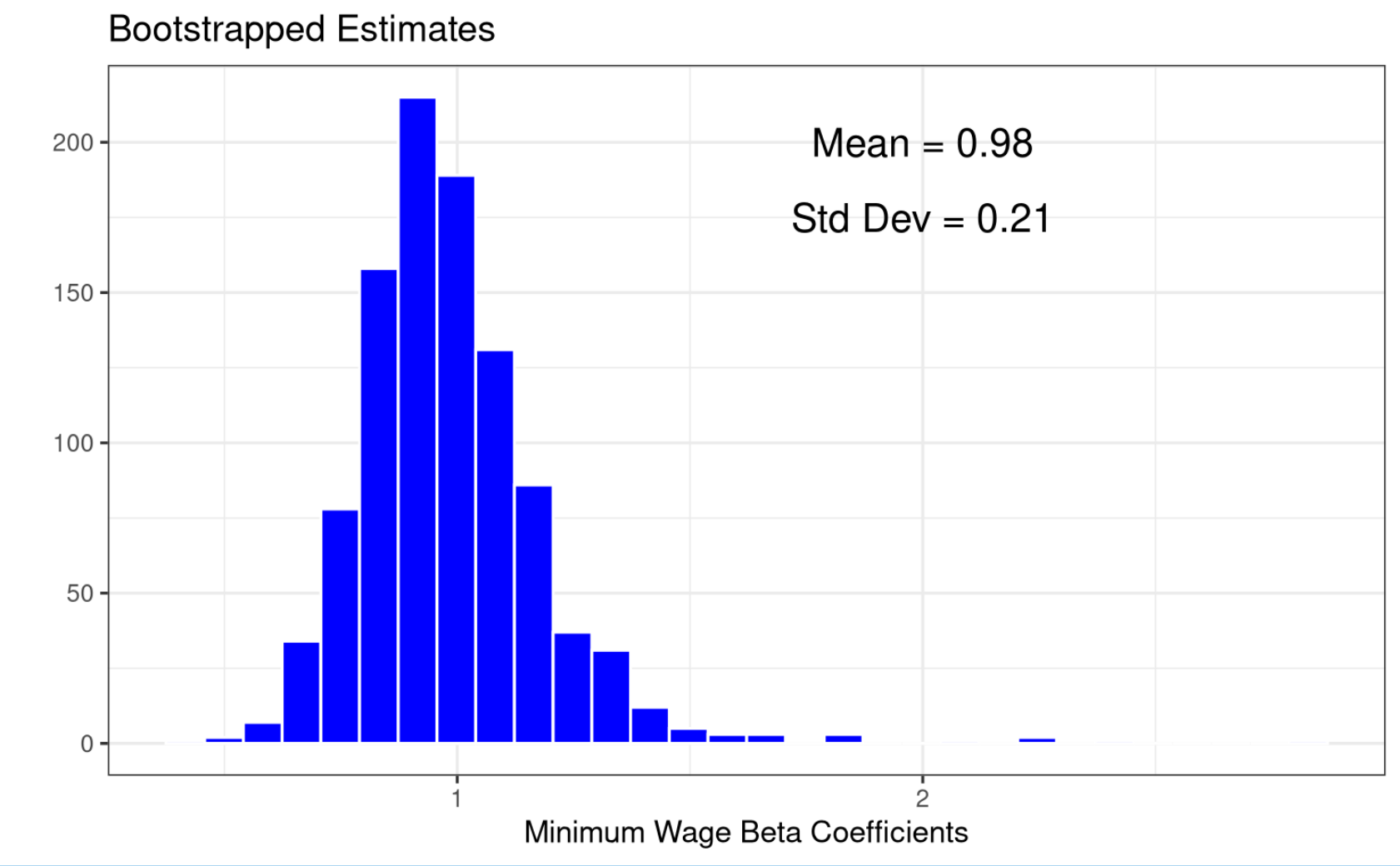
Statistical Significance: What is the Standard Error?

	Original	New Sample
(Intercept)	24.22*	24.76*
	(1.79)	(1.42)
min_wage	0.96*	1.00*
	(0.22)	(0.18)
Num.Obs.	50	50
R2	0.287	0.400
R2 Adj.	0.273	0.387
RMSE	4.44	4.09
* p < 0.05		

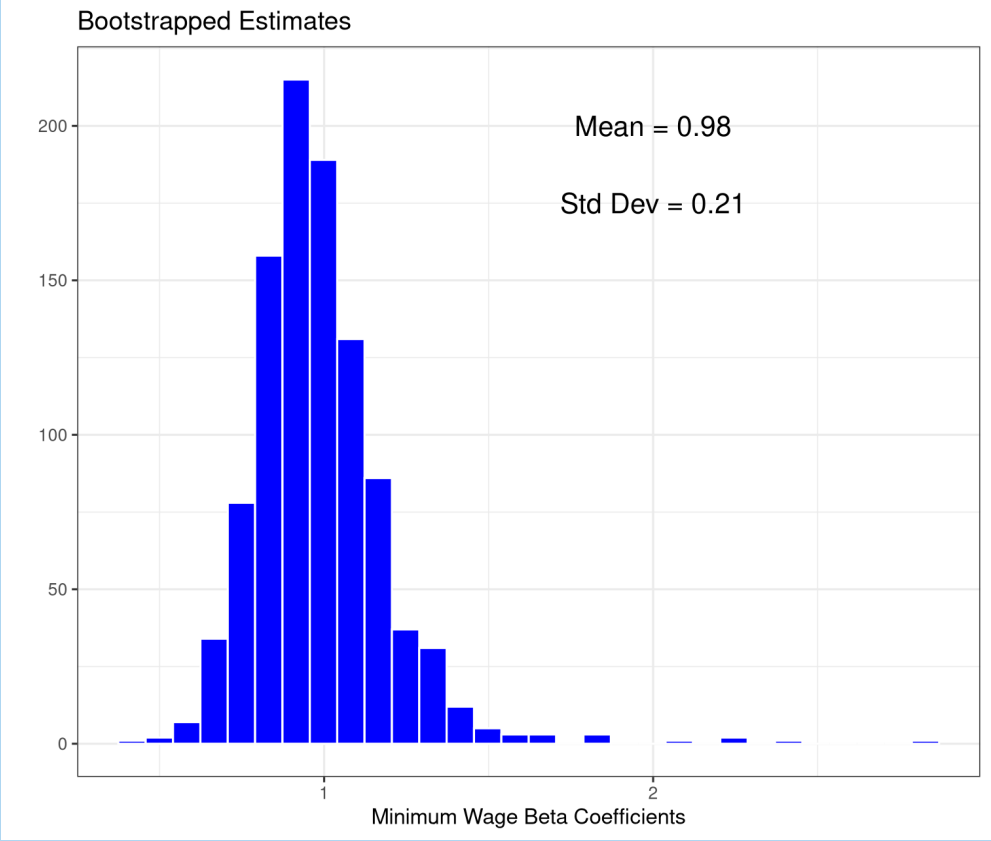
Statistical Significance: What is the Standard Error?

	Original	1	2	3	4	5
(Intercept)	24.22*	25.09*	25.27*	25.47*	23.45*	23.86*
	(1.79)	(1.70)	(1.16)	(1.54)	(1.84)	(1.45)
min_wage	0.96*	0.87*	0.98*	0.81*	1.11*	1.03*
	(0.22)	(0.21)	(0.14)	(0.20)	(0.21)	(0.17)
Num.Obs.	50	50	50	50	50	50
R2	0.287	0.273	0.491	0.263	0.362	0.440
R2 Adj.	0.273	0.258	0.481	0.248	0.349	0.429
RMSE	4.44	3.69	3.60	3.86	4.01	4.06
* p < 0.05						

Statistical Significance: What is the Standard Error?



Statistical Significance: What is the Standard Error?

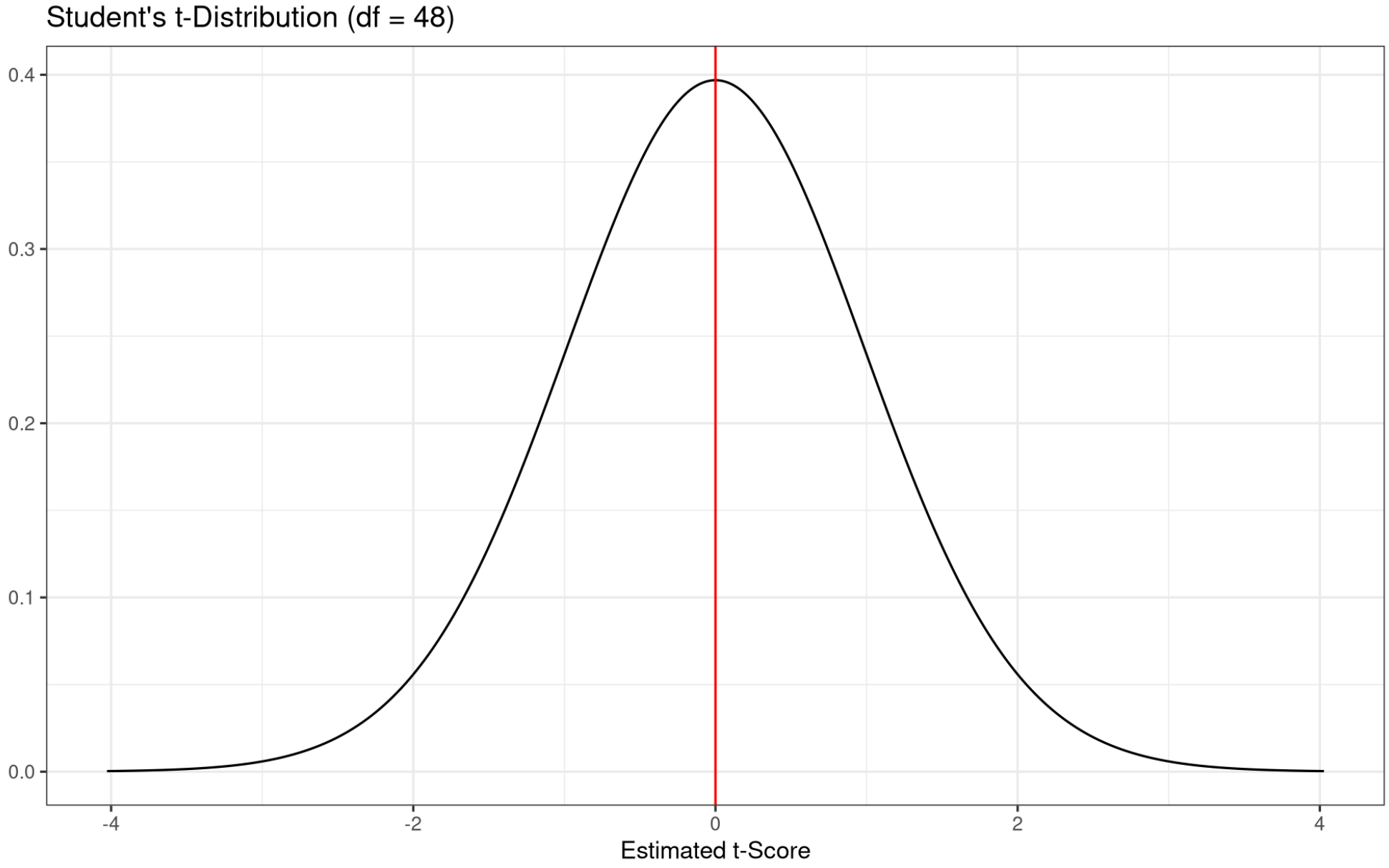


	Bachelors (%)
Minimum Wage	0.96*
	(0.22)
Constant	24.22*
	(1.79)
Observations	50
Adjusted R^2	0.27
Residual Std. Error	4.53 (df = 48)
F Statistic	19.36* (df = 1; 48)
Note:	*p<0.05

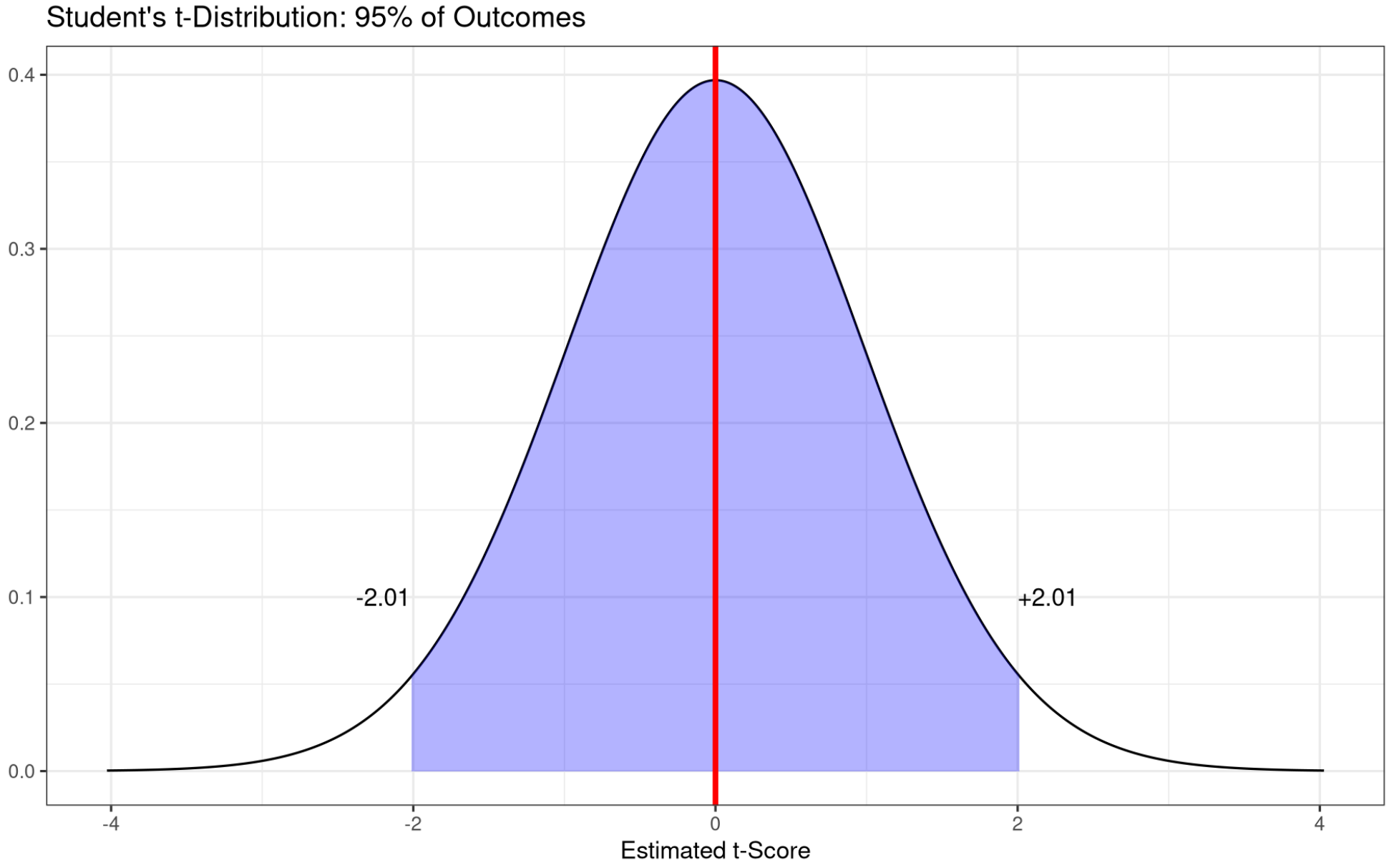
Statistical Significance: What is the t Value?

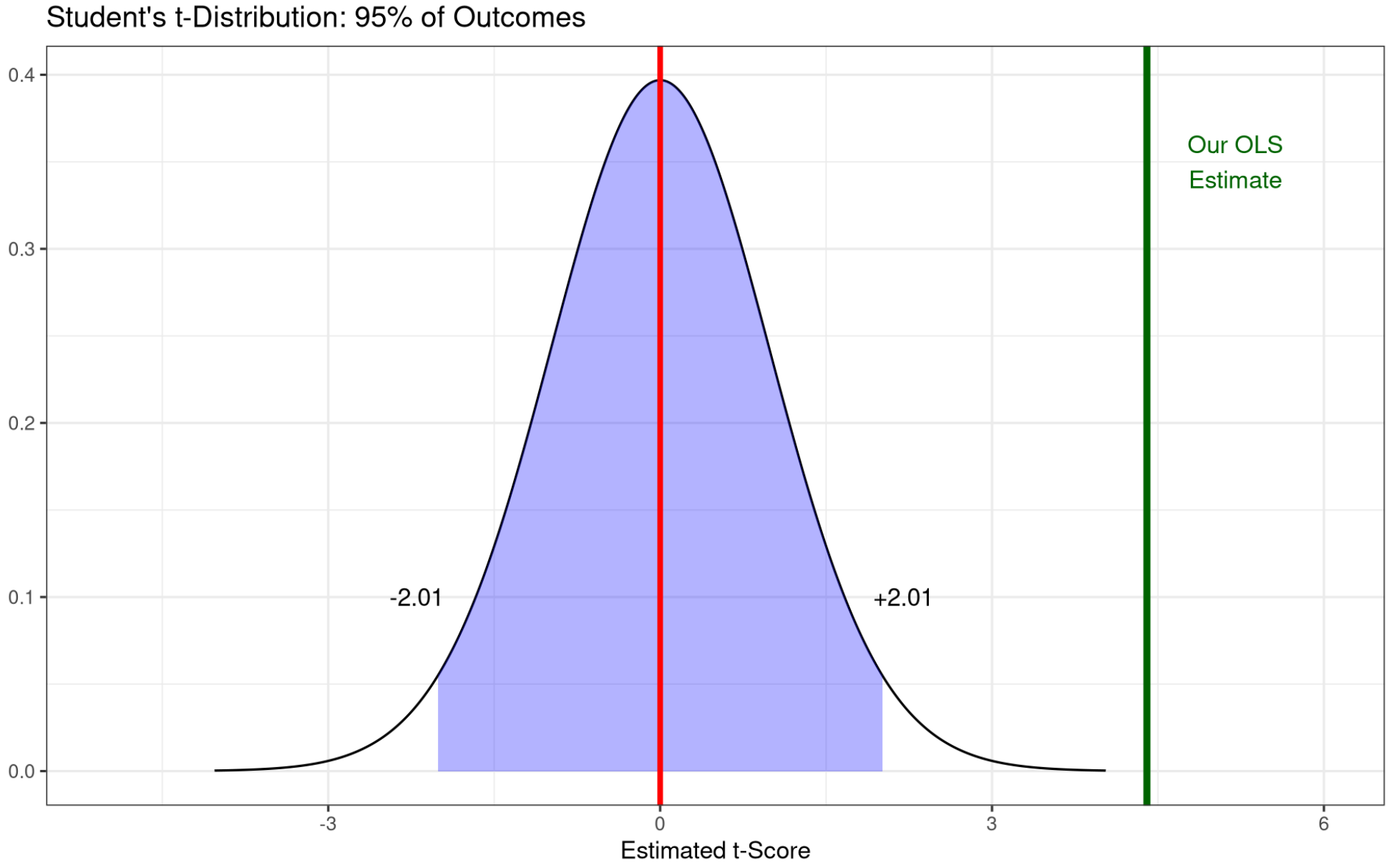
Rescaling our coefficient into SE-units

$$\text{T-statistic} = \frac{\text{coefficient}}{\text{standard error}}$$


Student's t-Distribution (df = 48)
0.4
0.3
0.2
0.1
0.0
-4
-2
0
2
4
Estimated t-Score


Student's t-Distribution: 99% of Outcomes
0.4
0.3
0.2
0.1
0.0
-2.68
+2.68
-4
-2
0
2
4
Estimated t-Score


Student's t-Distribution: 95% of Outcomes
0.4
0.3
0.2
0.1
0.0
-2.01
+2.01
-4
-2
0
2
4
Estimated t-Score


Student's t-Distribution: 95% of Outcomes
0.4
0.3
0.2
0.1
0.0
-2.01
+2.01
Our OLS
Estimate
-3
0
3
6
Estimated t-Score

Simple linear regression formula

$$Y = \alpha + \beta X$$

Multiple linear regression formula

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + + \beta_k X_k$$

Is there a significant gender difference in yearly earnings?

Three Approaches

1) Group means,

2) Box plots, and

3) OLS regression

Is there a significant gender difference in income?

Calculate the mean earnk_adj for each gender using a pivot table

Is there a significant gender difference in income?

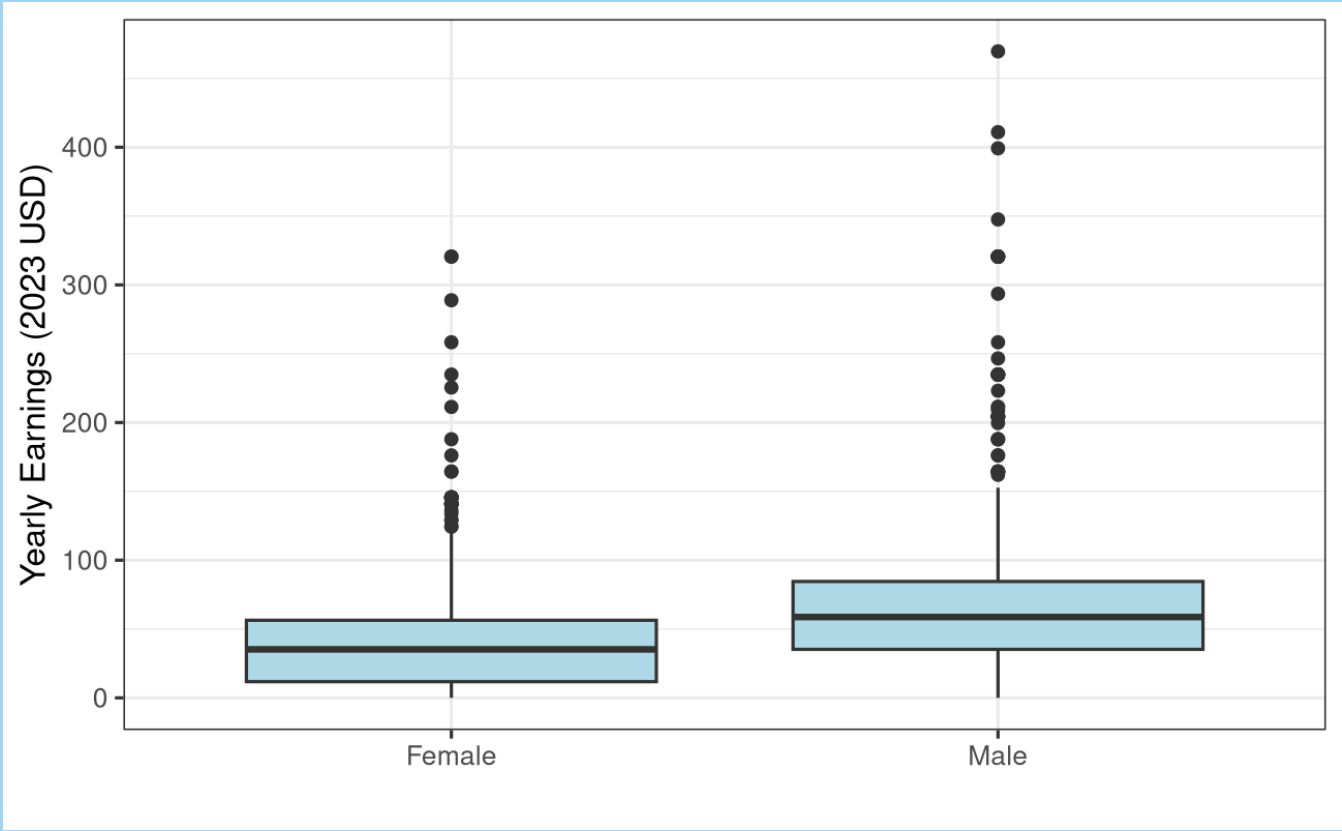
gender	Mean
Female	37.22
Male	69.41

Is there a significant gender difference in income?

Make a box plot of earnk_adj with separate boxes for each gender

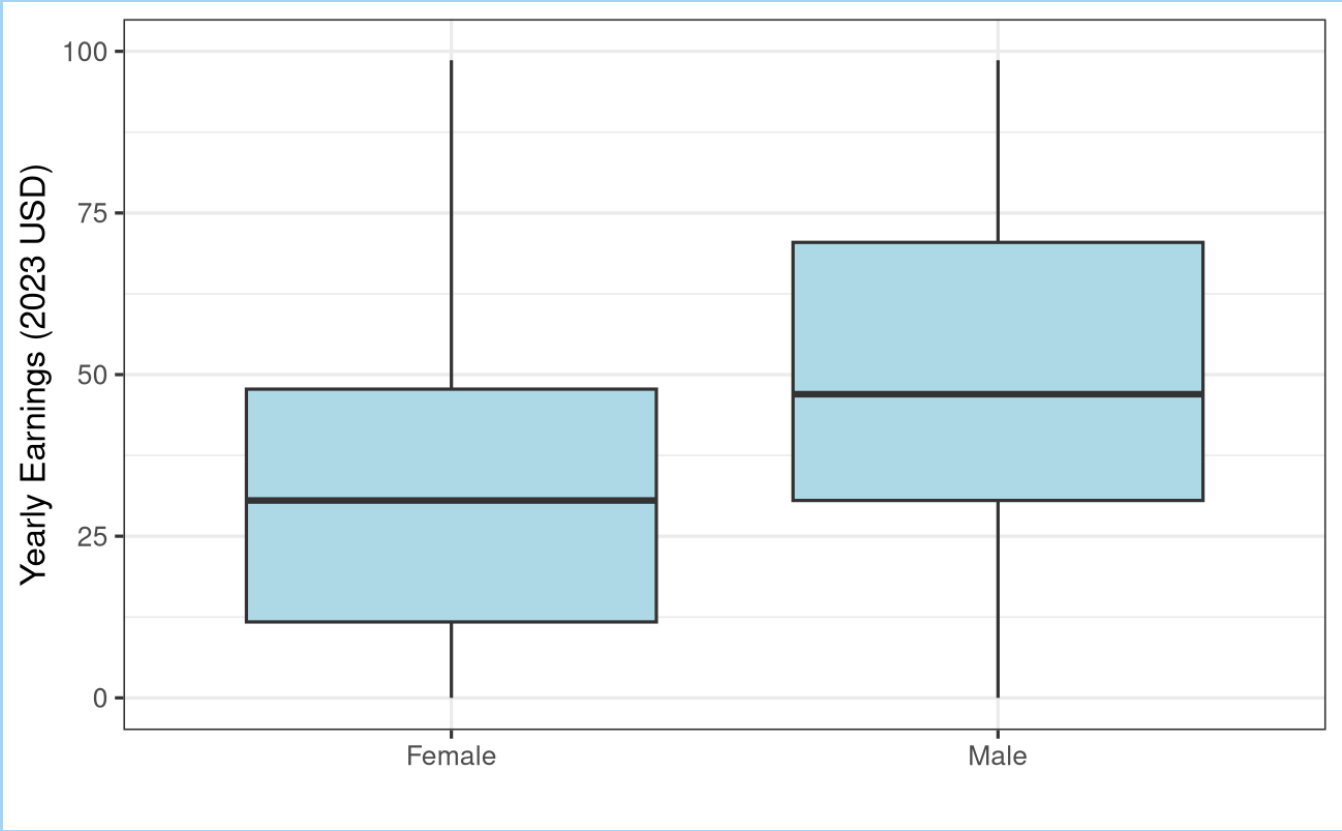
Is there a significant gender difference in income?

gender	Mean
Female	37.22
Male	69.41



Is there a significant gender difference in income?

gender	Mean
Female	37.22
Male	69.41



OLS Regressions with Categorical Variables

Excel cannot fit an OLS to categorical variables

1. Create dummy variables for each level of the categorical variable, but

2. Omit one level as a baseline (the intercept).

Convert Categorical Vars to Dummies

	A	B	C	D
1	**earnk_adj**	**gender**		
2	117.415	Male		
3	140.898	Female		
4	70.449	Female		
5	58.7075	Female		
6	117.415	Female		
7	145.5946	Female		
8	119.7633	Female		
9	21.1347	Female		
10	68.1007	Female		
11	75.1456	Male		
12	4.6966	Male		
13	82.1905	Male		
14	63.4041	Male		
15	15.334399	Male		
16	0	Female		

Convert Categorical Vars to Dummies

	A	B	C	D
1	**earnk_adj**	**gender**		
2	140.898	Female		
3	70.449	Female		
4	58.7075	Female		
5	117.415	Female		
6	145.5946	Female		
7	119.7633	Female		
8	21.1347	Female		
9	68.1007	Female		
10	0	Female		
11	28.1796	Female		
12	46.966	Female		
13	0	Female		
14	58.7075	Female		
15	51.6626	Female		

Convert Categorical Vars to Dummies

	A	B	C	D
1	**earnk_adj**	**gender**	**Male**	**Female**
2	140.898	Female	0	1
3	70.449	Female	0	1
4	58.7075	Female	0	1
5	117.415	Female	0	1
6	145.5946	Female	0	1
7	119.7633	Female	0	1
8	21.1347	Female	0	1
9	68.1007	Female	0	1
10	0	Female	0	1
11	28.1796	Female	0	1
12	46.966	Female	0	1
13	0	Female	0	1
14	58.7075	Female	0	1
15	51.6626	Female	0	1

Convert Categorical Vars to Dummies

	A	B	C	D
1	earnk_adj	gender	Male	Female
2	469.66	Male	1	0
3	410.9525	Male	1	0
4	399.211	Male	1	0
5	347.5484	Male	1	0
6	320.54295	Female	0	1
7	320.54295	Female	0	1
8	320.54295	Male	1	0
9	320.54295	Male	1	0
10	320.54295	Male	1	0
11	320.54295	Male	1	0
12	320.54295	Male	1	0
13	320.54295	Male	1	0
14	293.5375	Male	1	0
15	288.8409	Female	0	1

Convert Categorical Vars to Dummies

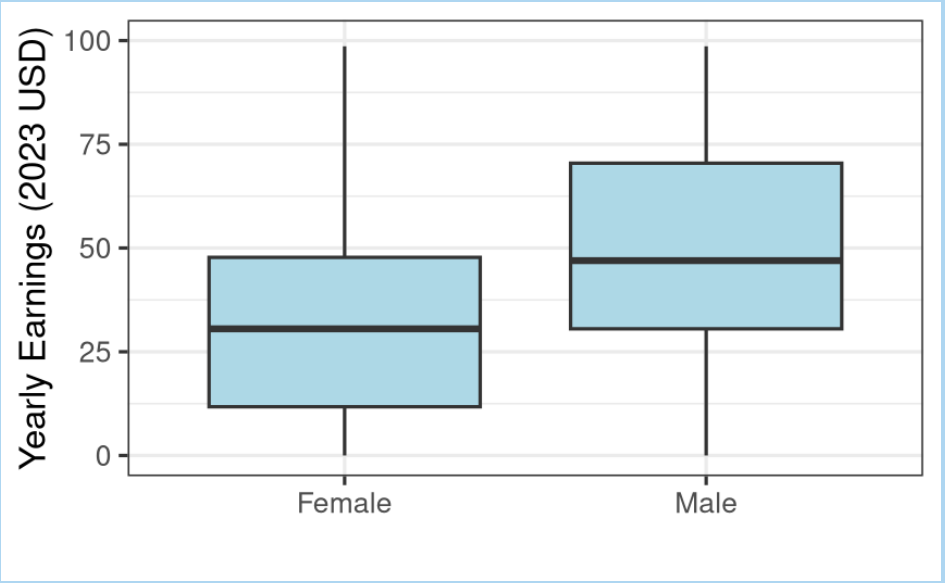
	A	B	C	D
1	earnk_adj	gender	Male	Female
2	469.66	Male	1	0
3	410.9525	Male	1	0
4	399.211	Male	1	0
5	347.5484	Male	1	0
6	320.54295	Female	0	1
7	320.54295	Female	0	1
8	320.54295	Male	1	0
9	320.54295	Male	1	0
10	320.54295	Male	1	0
11	320.54295	Male	1	0
12	320.54295	Male	1	0
13	320.54295	Male	1	0
14	293.5375	Male	1	0
15	288.8409	Female	0	1

Regression:

- Y = earnk_adj
- X = male

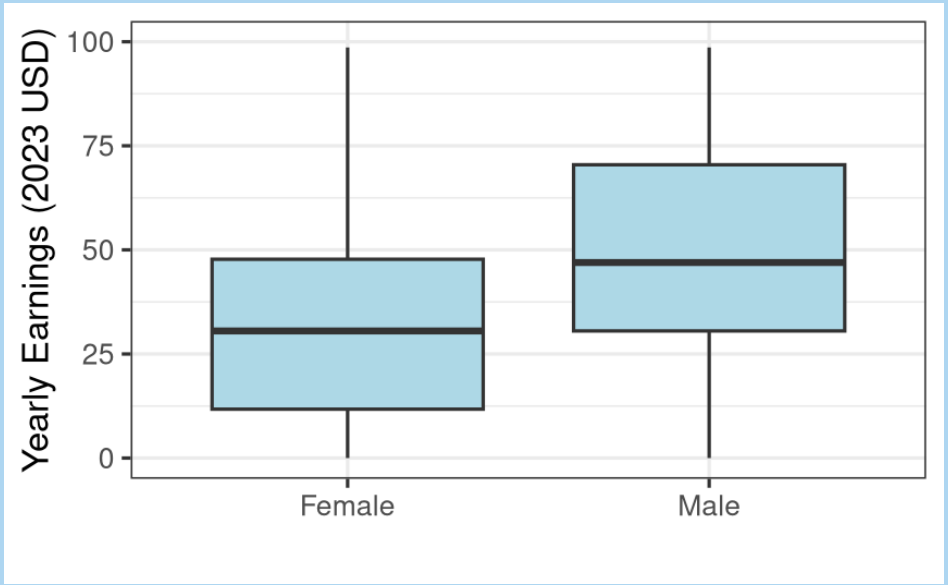
	Earnings (2023 USD)
Male	32.19*
	(2.24)
Constant	37.22*
	(1.36)
Observations	1,815
Adjusted R^2	0.10
Residual Std. Error	46.08 (df = 1813)
Note:	*p < 0.05

gender	Mean
Female	37.22
Male	69.41



	Earnings (2023 USD)
Male	32.19*
	(2.24)
Constant	37.22*
	(1.36)
Observations	1,815
Adjusted R^2	0.10
Residual Std. Error	46.08 (df = 1813)
Note:	*p < 0.05

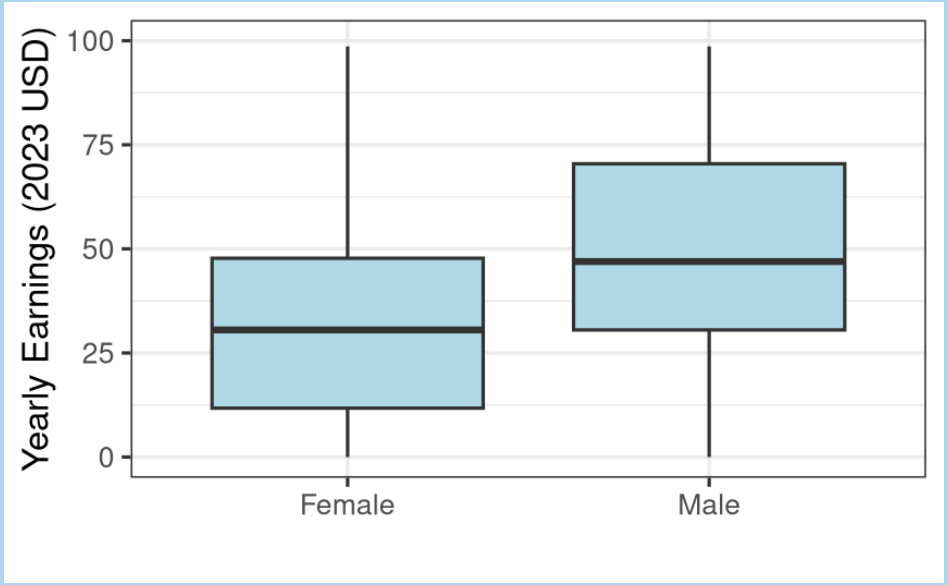
gender	Mean
Female	37.22
Male	69.41



Earnings = 37.22 + 32.19 * Male

	Earnings (2023 USD)
Male	32.19*
	(2.24)
Constant	37.22*
	(1.36)
Observations	1,815
Adjusted R^2	0.10
Residual Std. Error	46.08 (df = 1813)
F Statistic	206.76* (df = 1; 1813)
Note:	*p < 0.05

gender	Mean
Female	37.22
Male	69.41



Earnings = 37.22 + 32.19 * 0 = 37.22

Earnings = 37.22 + 32.19 = 69.41

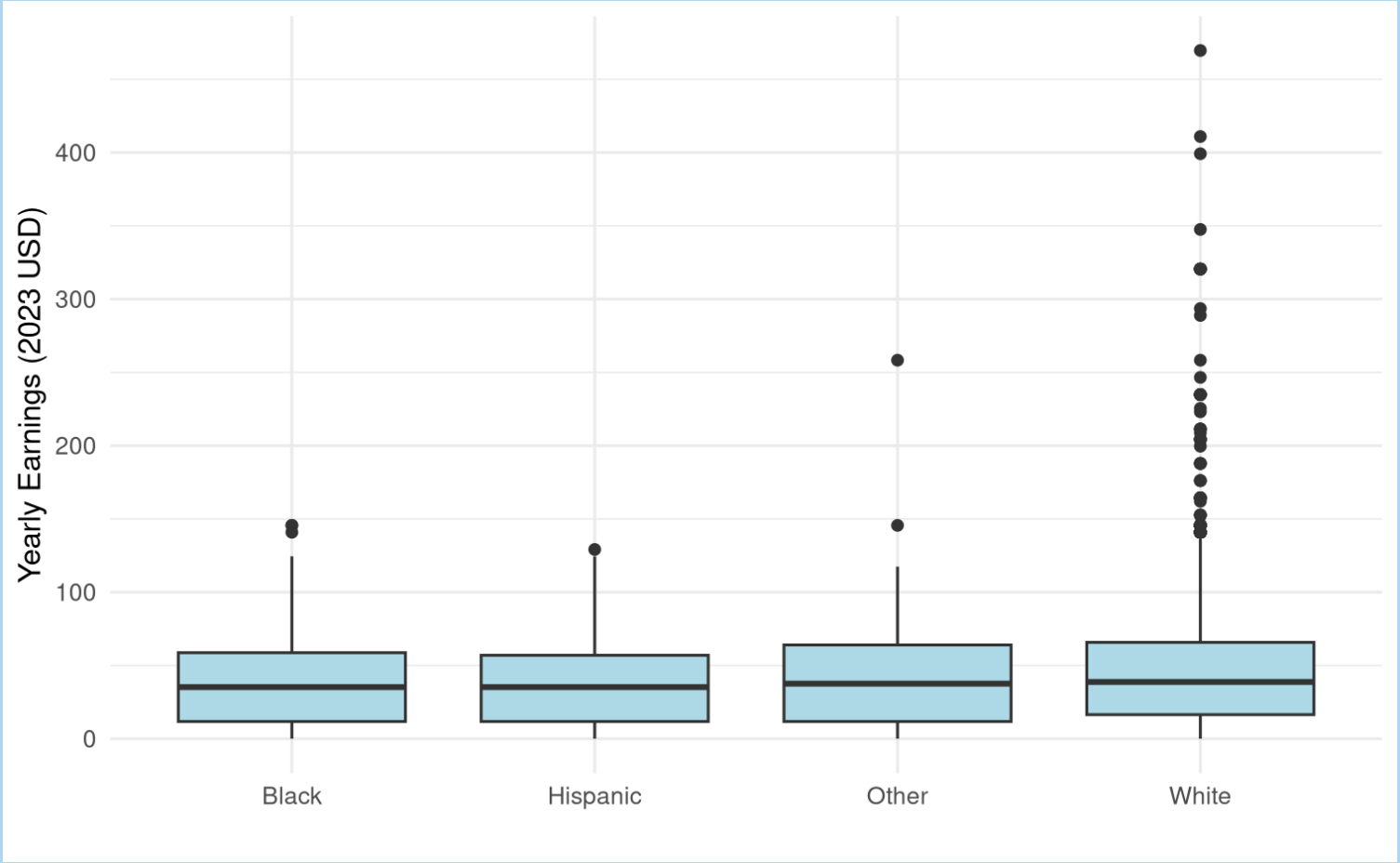
Is there a significant difference across reported ethnicities in yearly earnings?

1) Group means, and

2) OLS regression

Is there a significant difference across reported ethnicities in yearly earnings?

Ethnicity	Mean
Black	40.2
Hispanic	37.1
Other	49.7
White	51.1



Rule: Make dummies for all except one level of the categorical variable

	A	B	C	D	E
1	earnk_adj	ethnicity	Black	Hispanic	Other
2	117.415	White	0	0	0
3	140.898	White	0	0	0
4	70.449	White	0	0	0
5	58.7075	White	0	0	0
6	117.415	Other	0	0	1
7	145.5946	Black	1	0	0
8	119.7633	White	0	0	0
9	21.1347	White	0	0	0
10	68.1007	White	0	0	0
11	75.1456	White	0	0	0
12	4.6966	Hispanic	0	1	0
13	82.1905	White	0	0	0
14	63.4041	White	0	0	0
15	15.334399	White	0	0	0

Rule: Make dummies for all except one level of the categorical variable

	A	B	C	D	E
1	earnk_adj	ethnicity	Black	Hispanic	Other
2	117.415	White	0	0	0
3	140.898	White	0	0	0
4	70.449	White	0	0	0
5	58.7075	White	0	0	0
6	Y	Other	0	X	1
7	1 6	Black	1		0
8	119.7633	White	0	0	0
9	21.1347	White	0	0	0
10	68.1007	White	0	0	0
11	75.1456	White	0	0	0
12	4.6966	Hispanic	0	1	0
13	82.1905	White	0	0	0
14	63.4041	White	0	0	0
15	15.334399	White	0	0	0

Ethnicity	Mean
Black	40.2
Hispanic	37.1
Other	49.7
White	51.1

	Earnings (2023 USD)
Black	-10.83*
	(3.82)
Hispanic	-13.99*
	(4.92)
Other	-1.36
	(7.96)
Constant	51.07*
	(1.25)
Observations	1,815
Note:	*p < 0.05

$$\text{Earnings} = 51.1 + -10.8(\text{Black}) + -14(\text{Hispanic}) + -1.4(\text{Other})$$

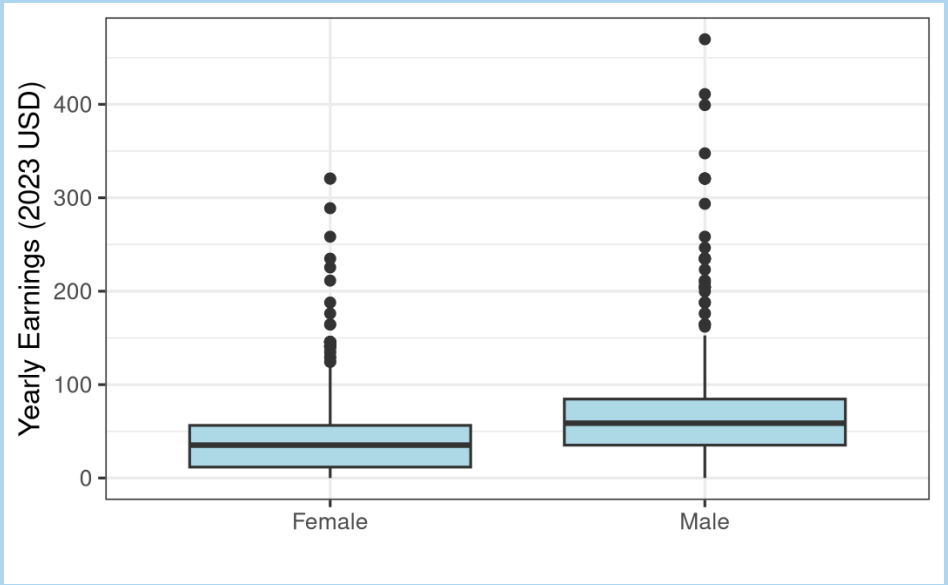
Simple linear regression formula

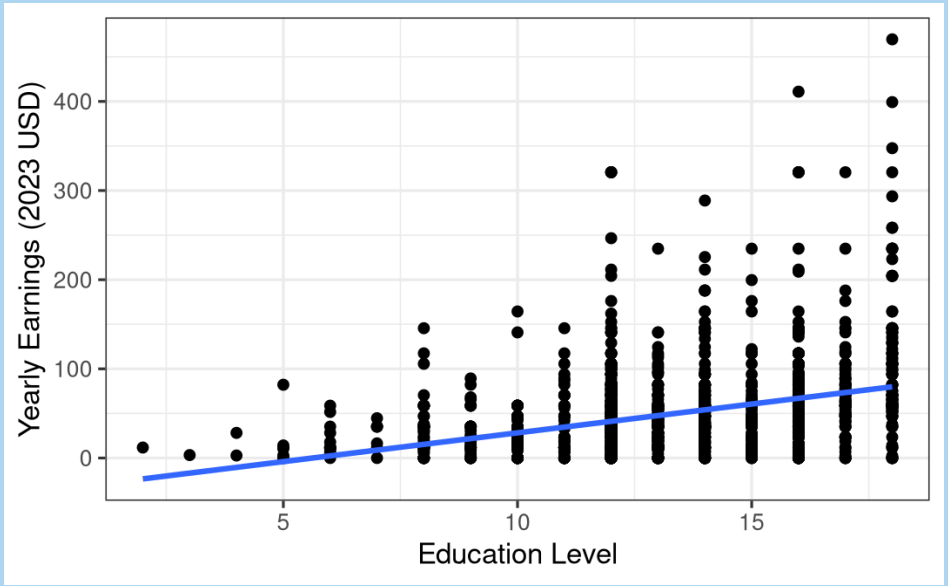
$$Y = \alpha + \beta X$$

Multiple linear regression formula

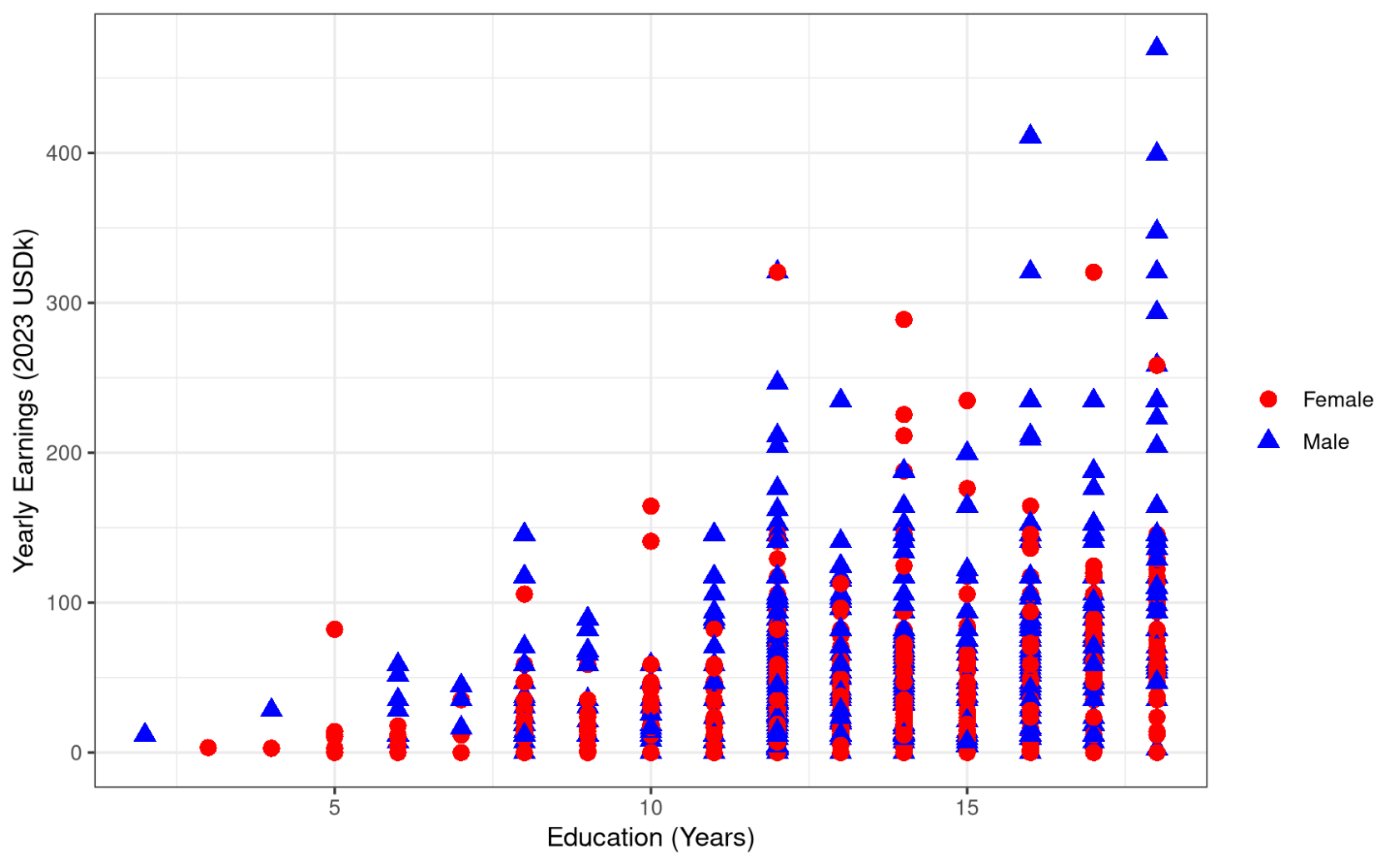
$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + + \beta_k X_k$$

	Earnings (2023 USD)	
	(1)	(2)
Male	32.19*	
	(2.24)	
Education		6.46*
		(0.42)
Constant	37.22*	-36.31*
	(1.36)	(5.67)
Observations	1,815	1,813
Adjusted R^2	0.10	0.11
Residual Std. Error	46.08 (df = 1813)	45.77 (df = 1811)
F Statistic	206.76* (df = 1; 1813)	235.79* (df = 1; 1811)
Note:		*p < 0.05

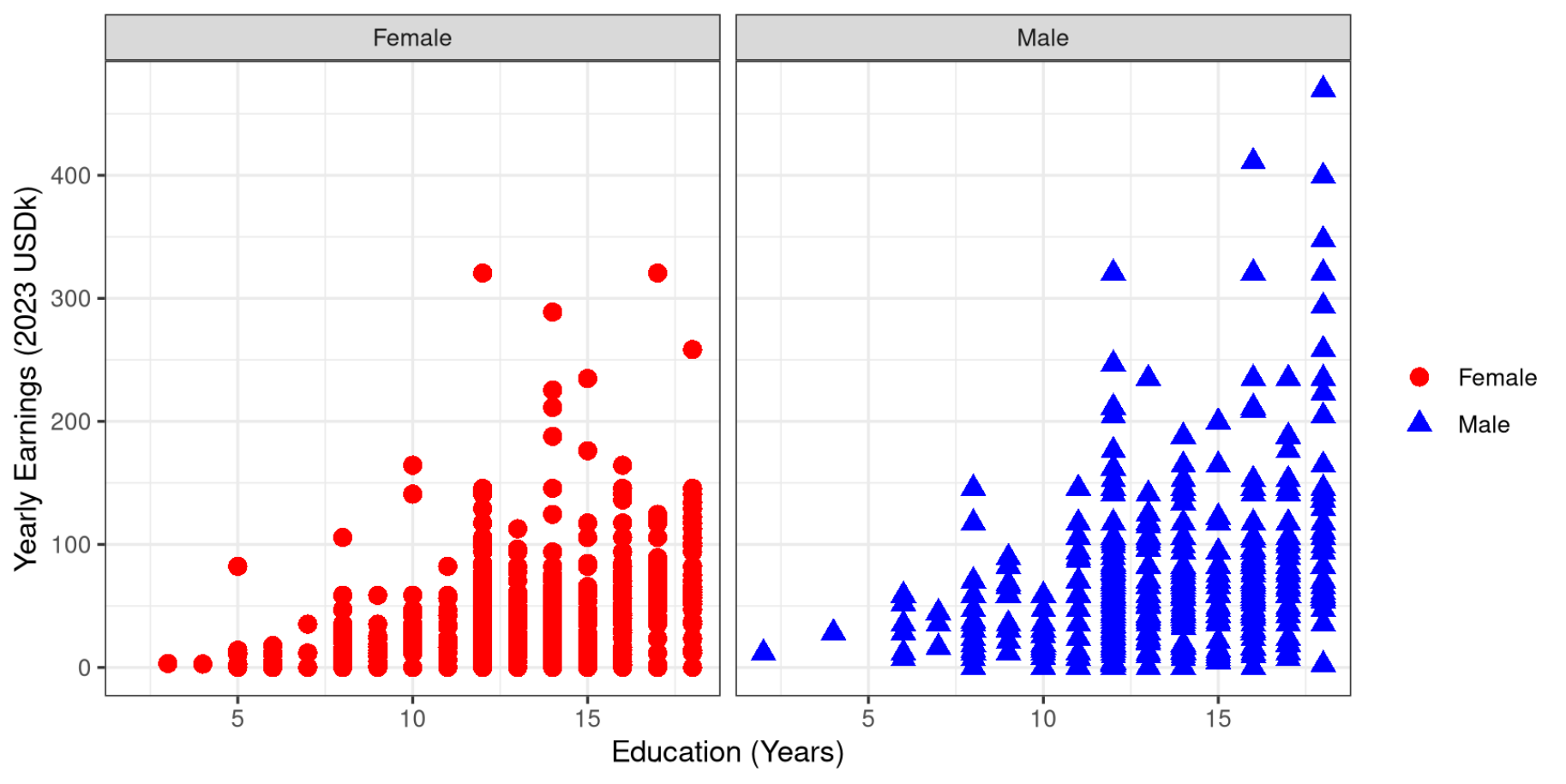




What explains the variation in yearly earnings?



What explains the variation in yearly earnings?



Multiple Linear Regression

Regress earnk_adj on male AND education

- Make sure to remove missing data!
- Include both predictors (male and education) in the 'X' of the regression selection window

	Earnings (2023 USD)		
	(1)	(2)	(3)
Male	32.19*		30.74*
	(2.24)		(2.11)
Education		6.46*	6.20*
		(0.42)	(0.40)
Constant	37.22*	-36.31*	-44.35*
	(1.36)	(5.67)	(5.39)
Observations	1,815	1,813	1,813
Adjusted R^2	0.10	0.11	0.21
Note:			*p < 0.05

Evaluation Step 1

Is it logical?

	Earnings (2023 USD)
Male	30.74*
	(2.11)
Education	6.20*
	(0.40)
Constant	-44.35*
	(5.39)
Observations	1,813
Adjusted R^2	0.21
Residual Std. Error	43.31 (df = 1810)
F Statistic	238.09* (df = 2; 1810)
Note:	*p < 0.05

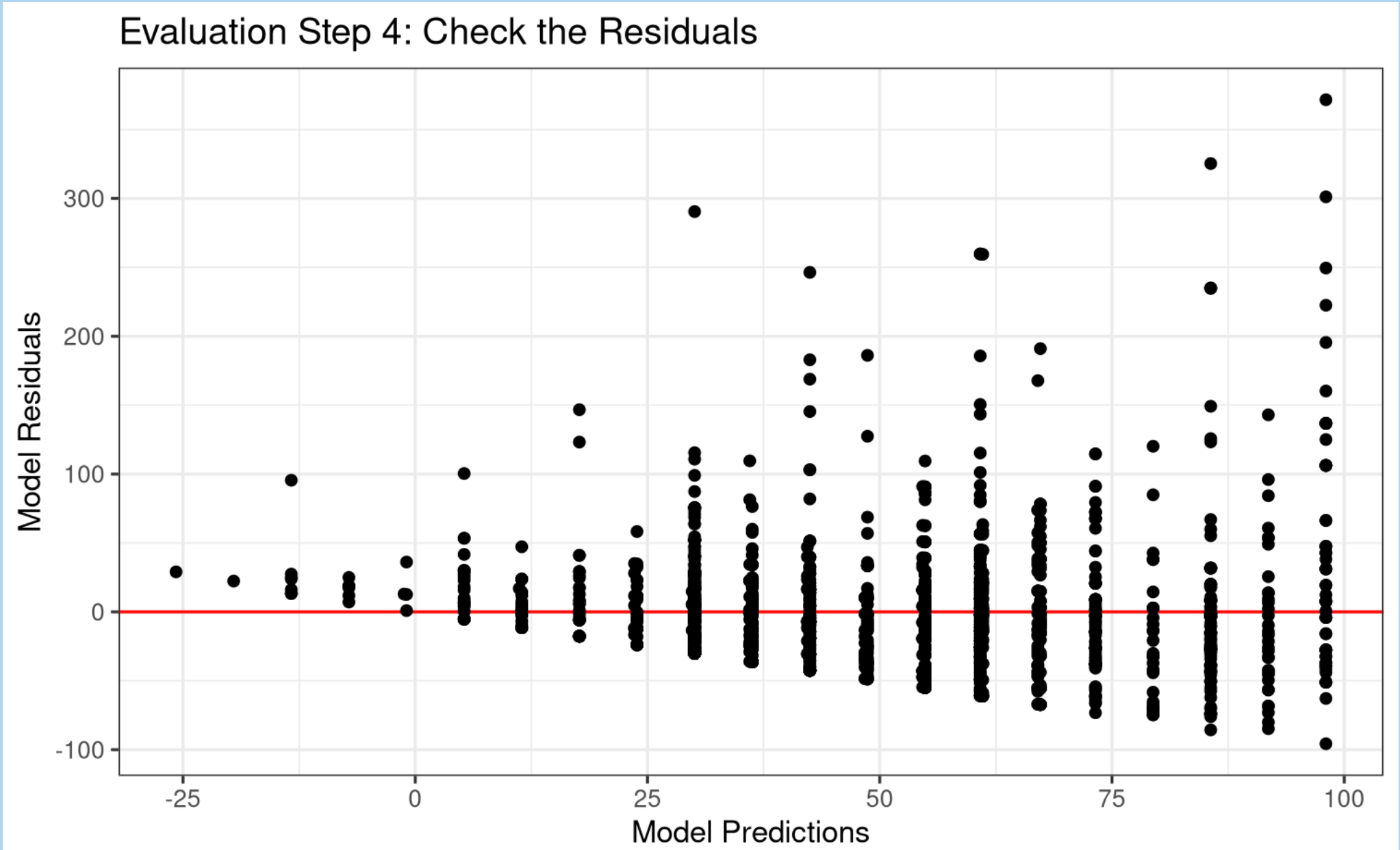
Evaluation Step 2

Is it significant?

	Earnings (2023 USD)
Male	30.74*
	(2.11)
Education	6.20*
	(0.40)
Constant	-44.35*
	(5.39)
Observations	1,813
Adjusted R^2	0.21
Residual Std. Error	43.31 (df = 1810)
F Statistic	238.09* (df = 2; 1810)
Note:	*p < 0.05

Step 3: Coefficient of Determination (R^2)

	Earnings (2023 USD)		
	(1)	(2)	(3)
Male	32.19*		30.74*
	(2.24)		(2.11)
Education		6.46*	6.20*
		(0.42)	(0.40)
Constant	37.22*	-36.31*	-44.35*
	(1.36)	(5.67)	(5.39)
Observations	1,815	1,813	1,813
Adjusted R^2	0.10	0.11	0.21
Residual Std. Error	46.08 (df = 1813)	45.77 (df = 1811)	43.31 (df = 1810)
F Statistic	206.76* (df = 1; 1813)	235.79* (df = 1; 1811)	238.09* (df = 2; 1810)
Note:			*p < 0.05


Evaluation Step 4: Check the Residuals
Model Residuals
300
200
100
0
-100
-25
0
25
50
75
100
Model Predictions

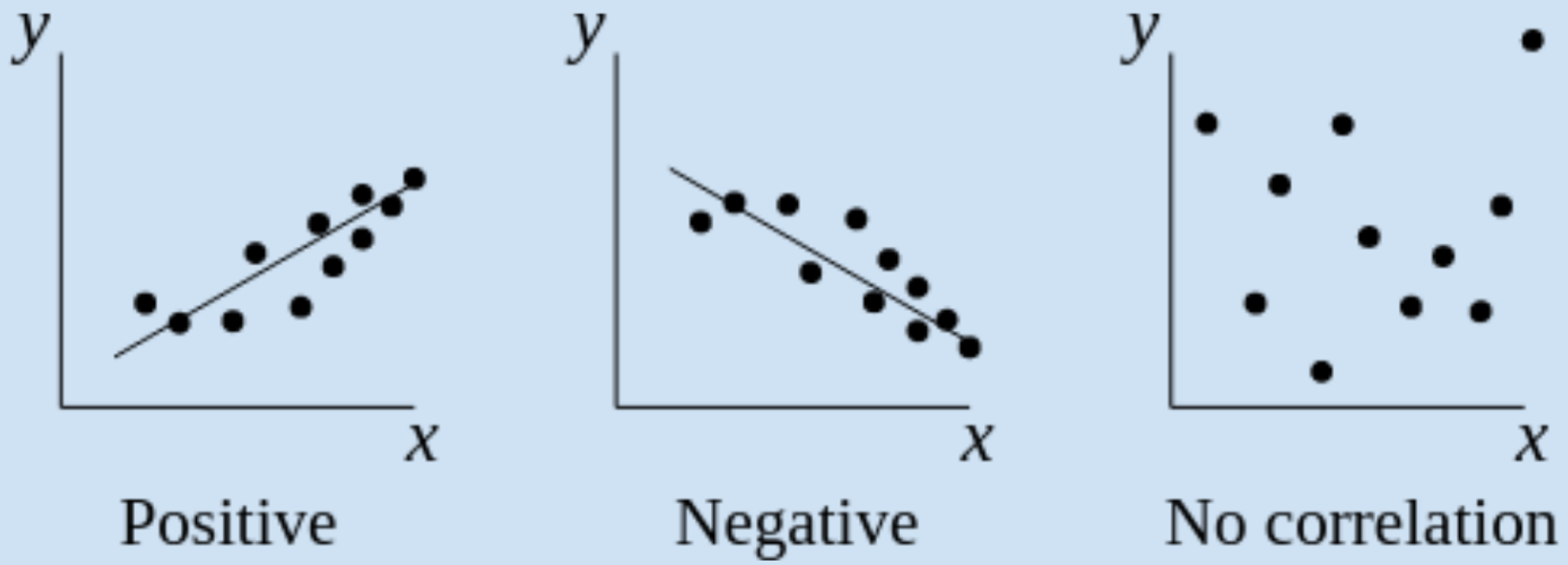
Step 5: Evidence of Multicollinearity?

	Education	Male
Education	1.00	0.04
Male	0.04	1.00

'Data' ribbon → 'Data Analysis' → 'Correlation'

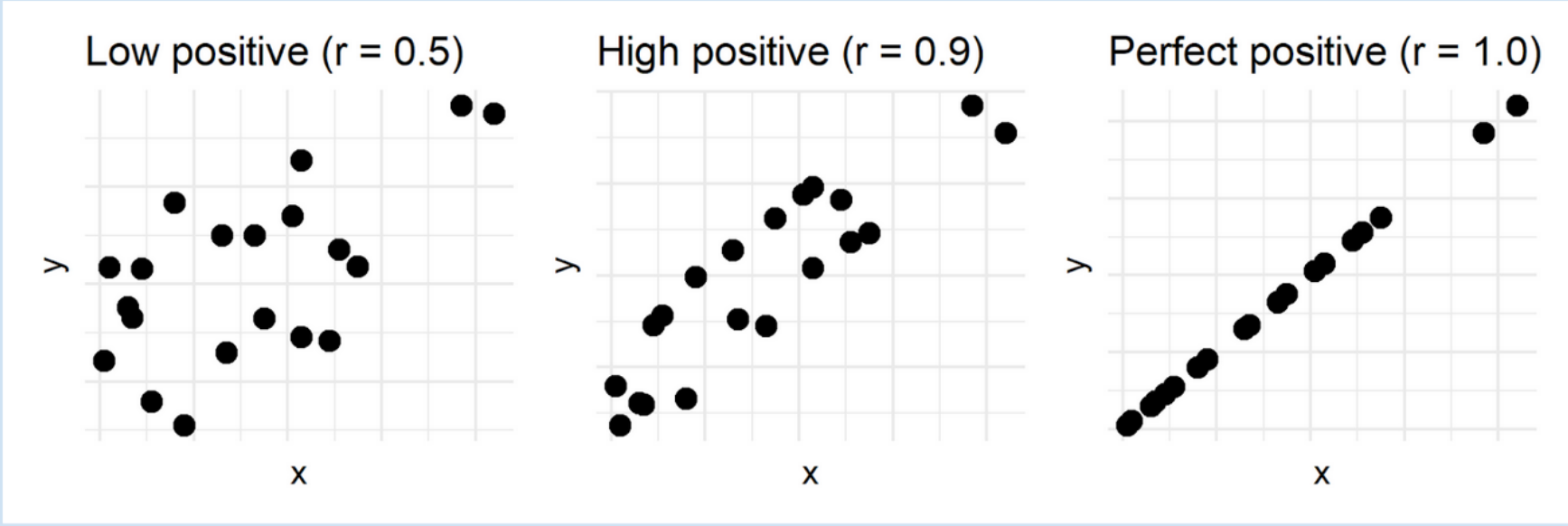
Evaluation Step 5: Multicollinearity

"*Correlation* is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate)" (JMP from SAS).

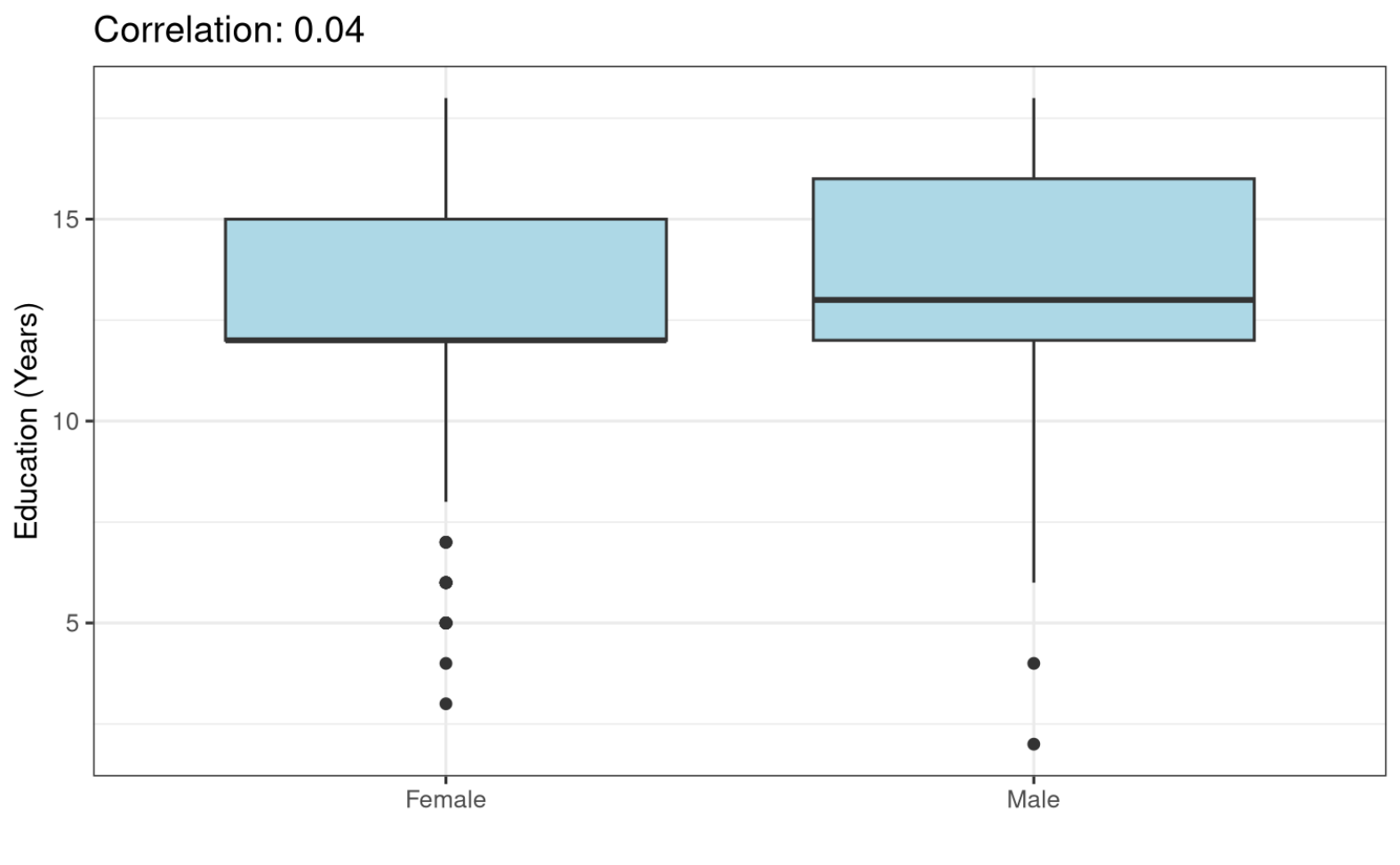


Evaluation Step 5: Multicollinearity

"Correlation strength is measured from -1 to +1. The correlation coefficient, often expressed as r, indicates a measure of the direction and strength of a relationship between two variables" (Verywellmind).



Step 5: Evidence of Multicollinearity?



	Earnings (2023 USD)
Male	30.74*
	(2.11)
Education	6.20*
	(0.40)
Constant	-44.35*
	(5.39)
Observations	1,813
Adjusted R^2	0.21
Residual Std. Error	43.31 (df = 1810)
F Statistic	238.09* (df = 2; 1810)
Note:	*p < 0.05

$$\text{Earnings} = -44.35 + 30.74\,(\text{Male}) + 6.2\,(\text{Education})$$

Earnings = -44.35 + 30.74 (Male) + 6.2 (Education)

Education	Male	Earnings
1	1	
9	1	
17	1	

Earnings = -44.35 + 30.74 (Male) + 6.2 (Education)

Education	Male	Earnings
1	1	-7.4k
9	1	42.2k
17	1	91.8k

Earnings = -44.35 + 30.74 (Male) + 6.2 (Education)

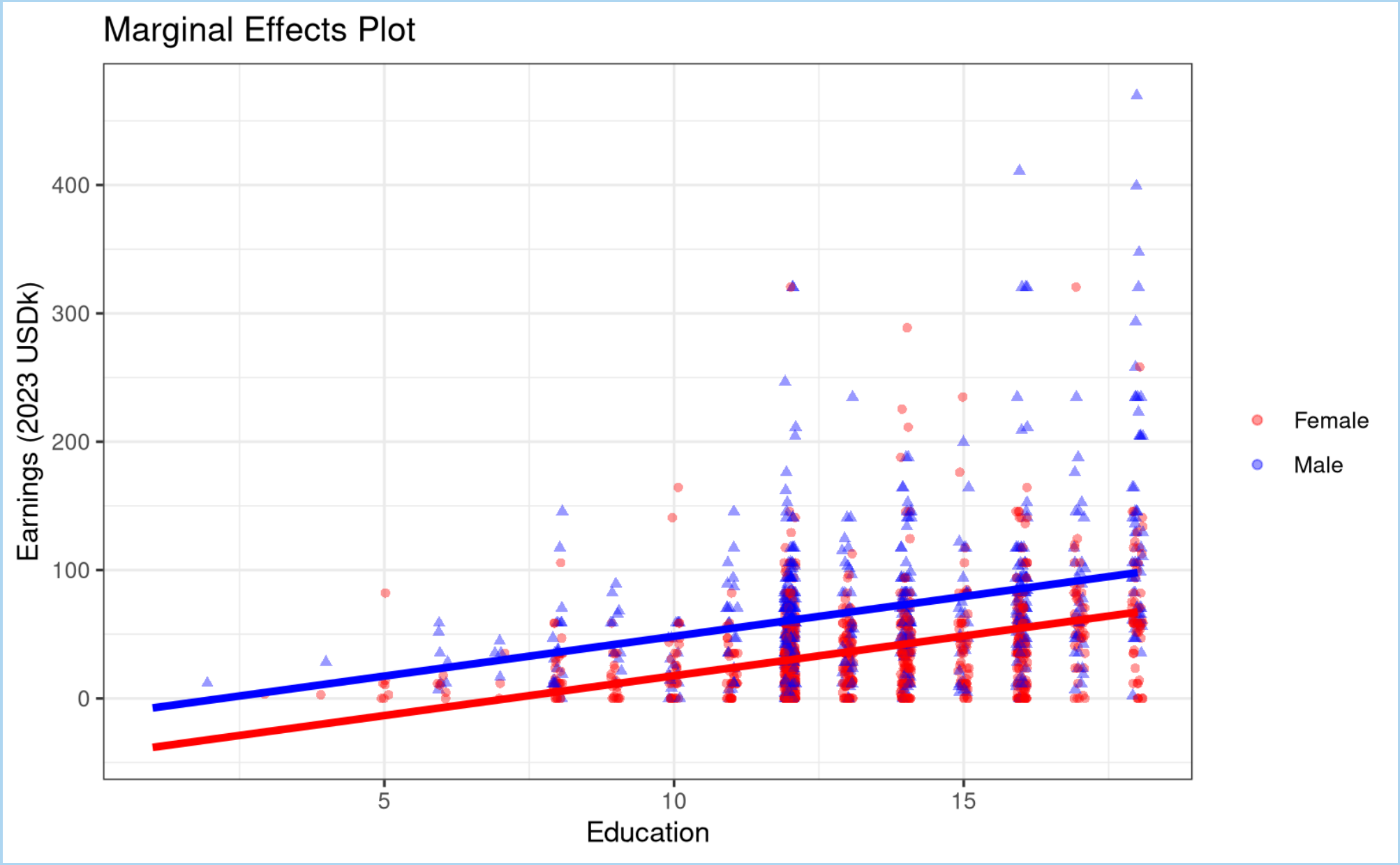
Education	Male	Earnings
1	0	
9	0	
17	0	

Earnings = -44.35 + 30.74 (Male) + 6.2 (Education)

Education	Male	Earnings
1	0	-38.2k
9	0	11.5k
17	0	61.1k

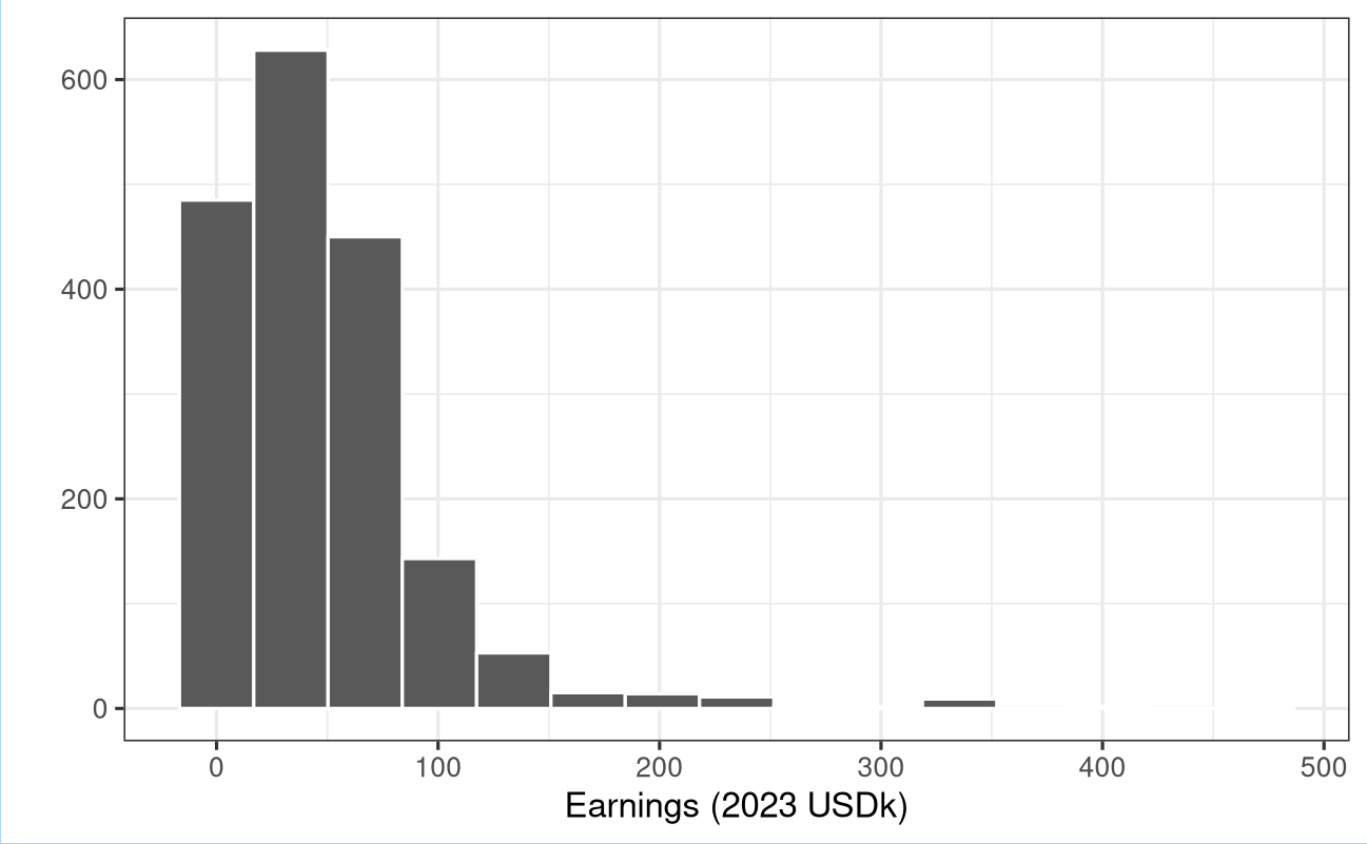
Earnings = -44.35 + 30.74 (Male) + 6.2 (Education)

Education	Male	Female
1	-7.4k	-38.2k
9	42.2k	11.5k
17	91.8k	61.1k


Marginal Effects Plot
Earnings (2023 USDk)
400
300
200
100
0
5
10
15
Education
Female
Male

One approach to building a "best" multiple regression model

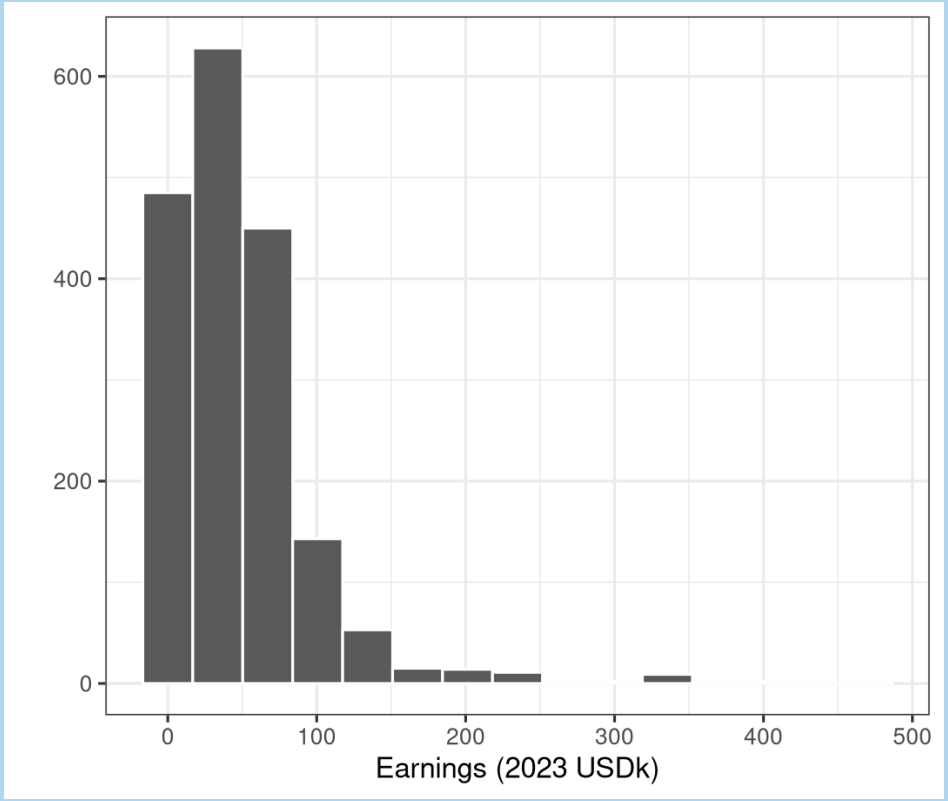
1. Choose the "logical" predictors from your options
2. Evaluate a simple OLS regression for each
3. Fit a multiple regression with the "best" of those
4. Evaluate the model using our five steps
5. Consider tweaks to improve model fit



Predictors to Consider

- Height, age, education, exercise and male

	Earnings (Thousands 2023 USD)				
	(1)	(2)	(3)	(4)	(5)
height	3.57*				
	(0.29)				
age		0.25*			
		(0.07)			
education			6.46*		
			(0.42)		
exercise				2.18*	
				(0.49)	
male					32.19*
					(2.24)
Constant	-188.34*	38.44*	-36.31*	42.51*	37.22*
	(19.08)	(3.06)	(5.67)	(1.88)	(1.36)
Observations	1,815	1,815	1,813	1,815	1,815
Adjusted R^2	0.08	0.01	0.11	0.01	0.10
Note:					*p < 0.05



For Next Class: What is the "best" model of bachelor's degree completion in the Session 1 data?

- Outcome:
  - Bachelors' Degrees

- Predictors to consider:
  - GDP (Rate), Homeownership, Minimum wage, State Tax Rate on Wages, Unemployment