Today's Agenda

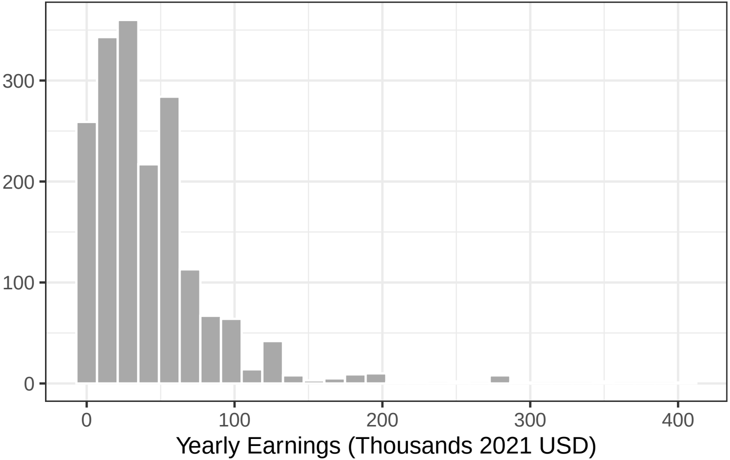
Evaluating simple OLS regressions

Justin Leinaweaver (Spring 2022)

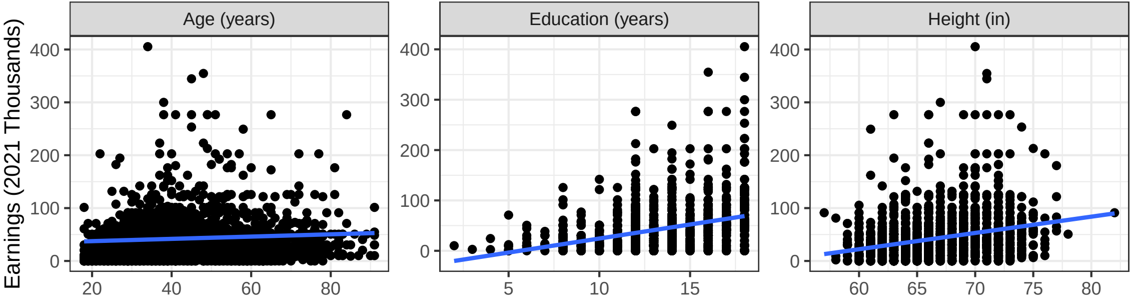
For Today

1. Finish the model building work from class today
2. Use the four steps outlined in Wilson, Keating, and Beal-Hodges (2012) chapters 4 and 5 to evaluate the fit of our models of earnings.

Can we build a useful model of yearly earnings in the Ross (1990) dataset?



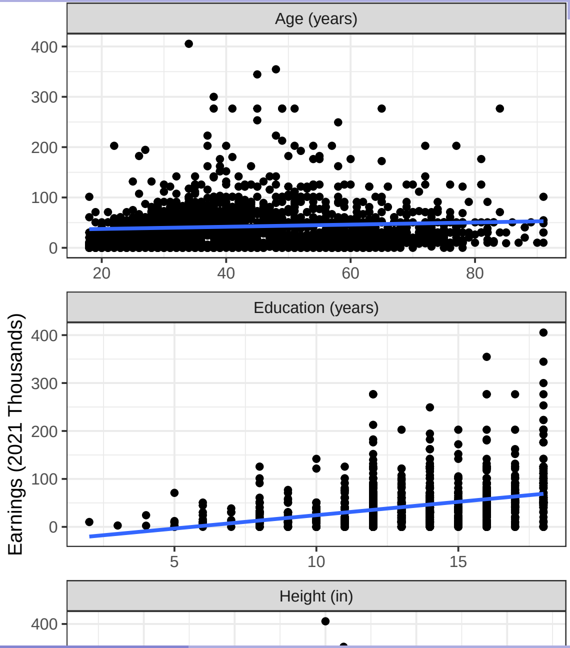
Can we build a useful model of yearly earnings in the Ross (1990) dataset?



	Earnings (2021)		
	(1)	(2)	(3)
Height	3.08*		
	(0.25)		
Age		0.22*	
		(0.06)	
Education			5.57*
			(0.36)
Constant	−162.54*	33.17*	−31.34*
	(16.46)	(2.64)	(4.89)
Observations	1,815	1,815	1,813
Adjusted R^2	0.08	0.01	0.11
Residual Std. Error	40.28 (df = 1813)	41.81 (df = 1813)	39.50 (df = 1811)
F Statistic	155.52* (df = 1; 1813)	14.22* (df = 1; 1813)	235.79* (df = 1; 1811)
Note:			*$p < 0.05$

	Earnings (2021)		
	(1)	(2)	(3)
Height	3.08*		
	(0.25)		
Age		0.22*	
		(0.06)	
Education			5.57*
			(0.36)
Constant	−162.54*	33.17*	−31.34*
	(16.46)	(2.64)	(4.89)

Note: $^*p < 0.05$



What are the predicted earnings. . .

For someone of average height in the sample?

- Avg height of 66.6 = $42.6k

For someone of average age in the sample?

- Avg age of 42.9 = $42.6k

For someone of average education in the sample?

- Avg education of 13.2 = $42.2k

What are the predicted earnings. . .

Model 1: Height

- Avg = \$42.6k vs Maximum = ?

Model 2: Age

- Avg = \$42.6k vs Maximum = ?

Model 3: Education

- Avg = \$42.2k vs Maximum = ?

What are the predicted earnings. . .

Model 1: Height

- Avg = \$42.6k vs Maximum (82) = ?

Model 2: Age

- Avg = \$42.6k vs Maximum (91) = ?

Model 3: Education

- Avg = \$42.2k vs Maximum (18) = ?

What are the predicted earnings...

Model 1: Height

- Avg = \$42.6k vs Maximum (82) = \$90k

Model 2: Age

- Avg = \$42.6k vs Maximum (91) = \$53.2k

Model 3: Education

- Avg = \$42.2k vs Maximum (18) = \$68.9k

	Earnings (2021)		
	(1)	(2)	(3)
Height	3.08*		
	(0.25)		
Age		0.22*	
		(0.06)	
Education			5.57*
			(0.36)
Constant	−162.54*	33.17*	−31.34*
	(16.46)	(2.64)	(4.89)
Observations	1,815	1,815	1,813
Adjusted R^2	0.08	0.01	0.11
Residual Std. Error	40.28 (df = 1813)	41.81 (df = 1813)	39.50 (df = 1811)
F Statistic	155.52* (df = 1; 1813)	14.22* (df = 1; 1813)	235.79* (df = 1; 1811)
Note:			*$p < 0.05$

What is statistical significance?

	Earnings (2021)
Education	5.57^{*}
	(0.36)
Constant	-31.34^{*}
	(4.89)
Observations	1,813
Adjusted R^2	0.11
Residual Std. Error	39.50 (df = 1811)
F Statistic	235.79^{*} (df = 1; 1811)
Note:	*p $<$ 0.05

What is statistical significance?

	Earnings (2021)
Education	5.57*
	(0.36)
Constant	−31.34*
	(4.89)
Observations	1,813
Adjusted R^2	0.11
Residual Std. Error	39.50 (df = 1811)
F Statistic	235.79* (df = 1; 1811)
Note:	*$p < 0.05$

Alternative Hypothesis (H_A)

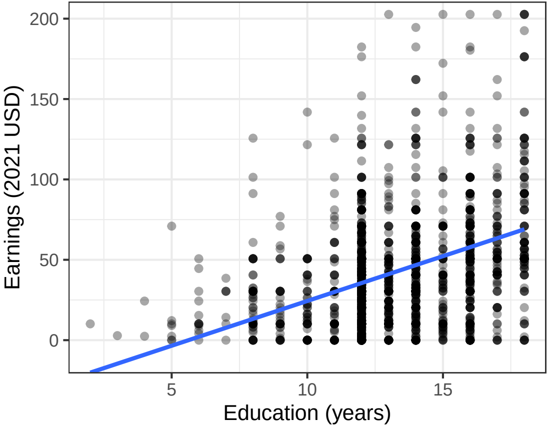
- Higher levels of education are associated with larger incomes.

Null Hypothesis (H_0)

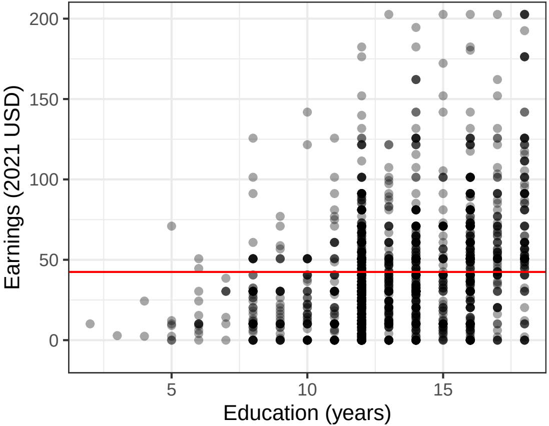
- Level of education is not associated with income.

What is statistical significance?

Alternative Hypothesis



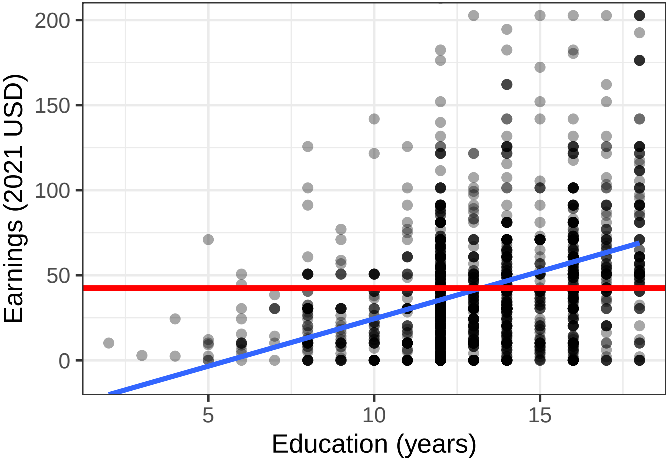
Null Hypothesis



P-Values: A shortcut to determining statistical significance

"The p-value is the probability of observing another computed test statistic ... that is more extreme (either positive or negative) than the one computed for your sample. ... Therefore, the smaller the p-value, the more support for the alternative hypothesis" (p82).

	Earnings (2021)
Education	5.57*
	(0.36)
Constant	−31.34*
	(4.89)
Observations	1,813
Adjusted R^2	0.11
Residual Std. Error	39.50 (df = 1811)
F Statistic	235.79* (df = 1; 1811)
Note:	*p < 0.05

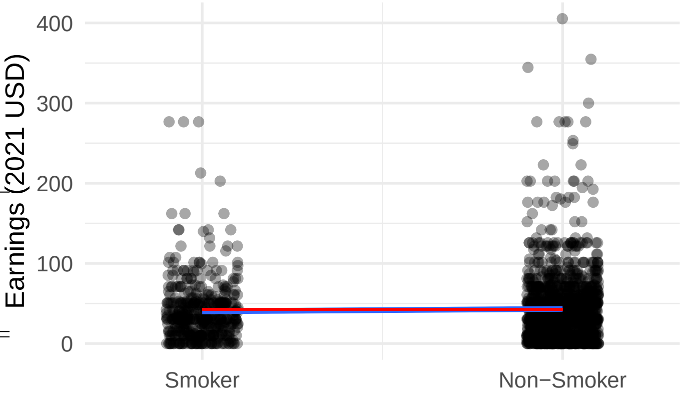


What is statistical significance?

	Earnings (2021)
Non-Smoker	3.01
	(2.26)
Constant	37.17*
	(4.07)
Observations	1,814
Adjusted R^2	0.0004
Residual Std. Error	41.97 (df = 1812)
F Statistic	1.77 (df = 1; 1812)
Note:	*p $<$ 0.05

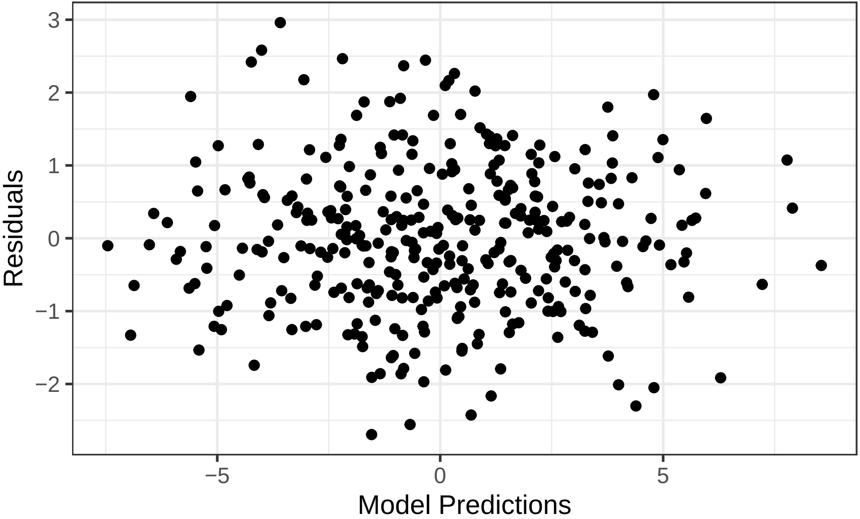
	Earnings (2021)
Non-Smoker	3.01
	(2.26)
Constant	37.17*
	(4.07)
Observations	1,814
Adjusted R^2	0.0004
Residual Std. Error	41.97 (df = 1812)
F Statistic	1.77 (df = 1; 1812)
Note:	*$p < 0.05$

Alternative vs the Null Hypothesis

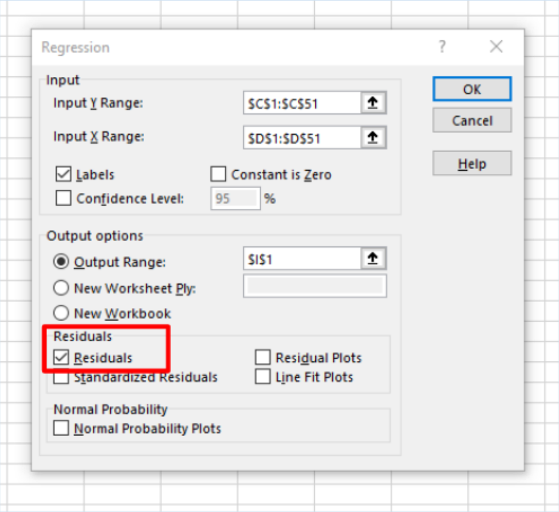


	Earnings (2021)		
	(1)	(2)	(3)
Height	3.08*		
	(0.25)		
Age		0.22*	
		(0.06)	
Education			5.57*
			(0.36)
Constant	−162.54*	33.17*	−31.34*
	(16.46)	(2.64)	(4.89)
Observations	1,815	1,815	1,813
Adjusted R^2	0.08	0.01	0.11
Residual Std. Error	40.28 (df = 1813)	41.81 (df = 1813)	39.50 (df = 1811)
F Statistic	155.52* (df = 1; 1813)	14.22* (df = 1; 1813)	235.79* (df = 1; 1811)
Note:			*p < 0.05

Step 4: Include a plot of the model's residuals



Step 4: Plot the Model's Residuals x Predictions



Regression

Input

Input Y Range: $C$1:$C$51

Input X Range: $D$1:$D$51

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

● Output Range: $I$1

○ New Worksheet Ply:

○ New Workbook

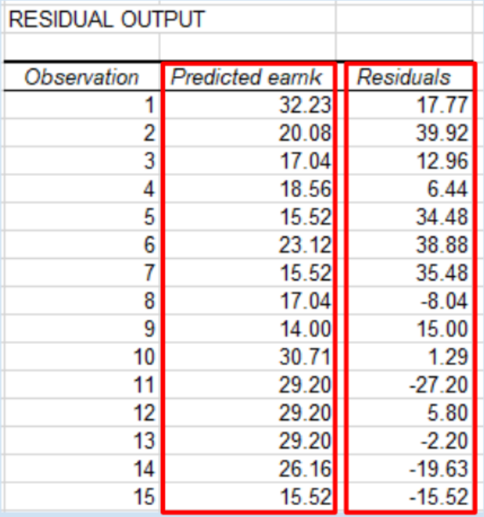
Residuals

☑ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

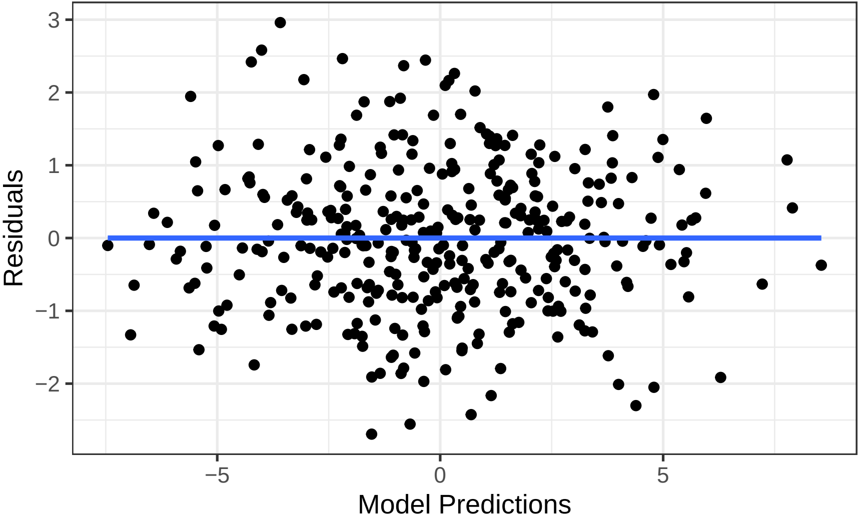
☐ Normal Probability Plots



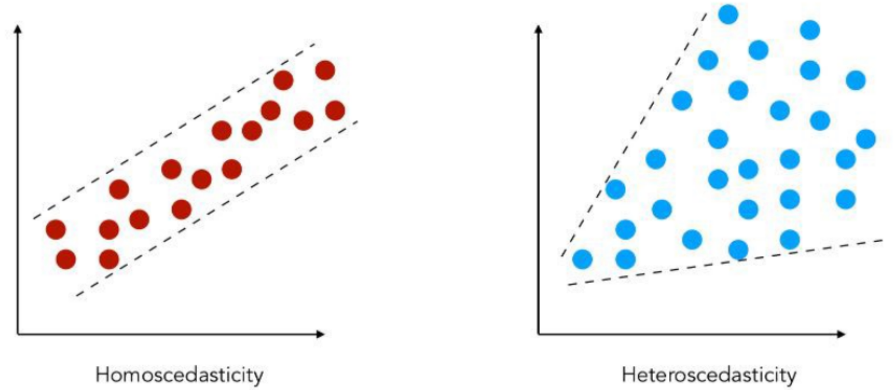
RESIDUAL OUTPUT

Observation	*Predicted earnk*	*Residuals*
1	32.23	17.77
2	20.08	39.92
3	17.04	12.96
4	18.56	6.44
5	15.52	34.48
6	23.12	38.88
7	15.52	35.48
8	17.04	-8.04
9	14.00	15.00
10	30.71	1.29
11	29.20	-27.20
12	29.20	5.80
13	29.20	-2.20
14	26.16	-19.63
15	15.52	-15.52

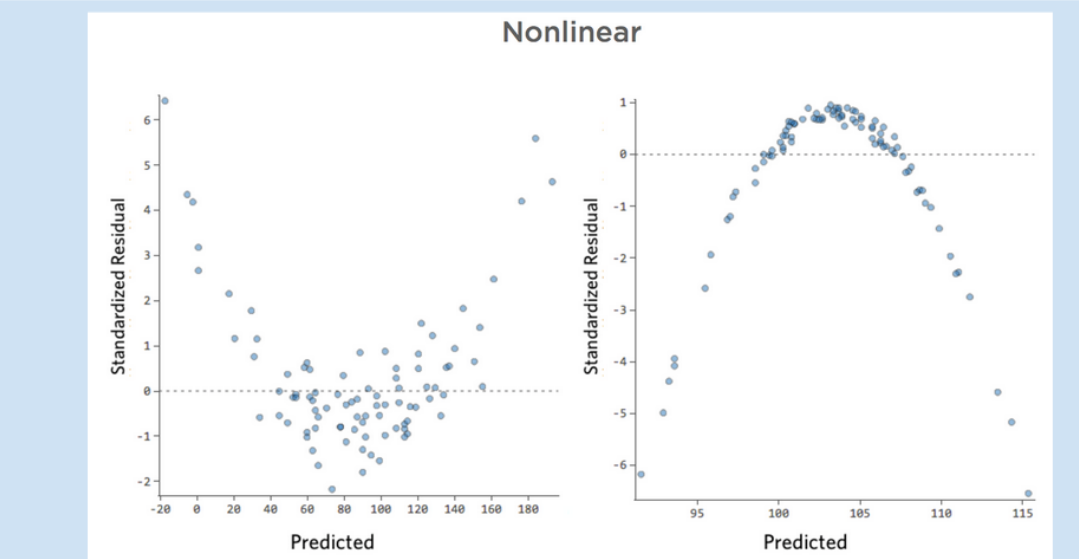
Step 4 Goal: Homoscedastic Errors



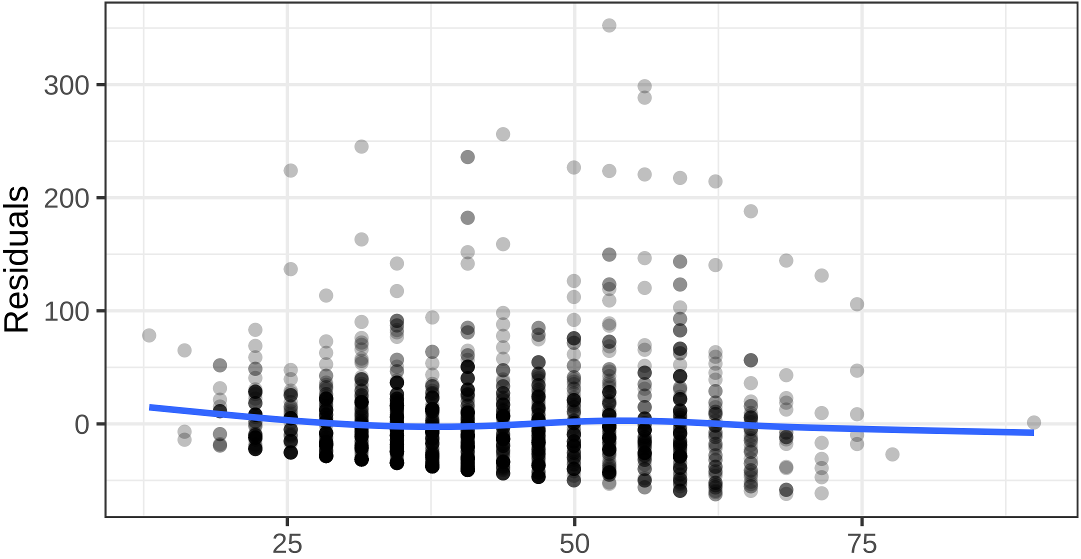
Bad: Heteroscedasticity



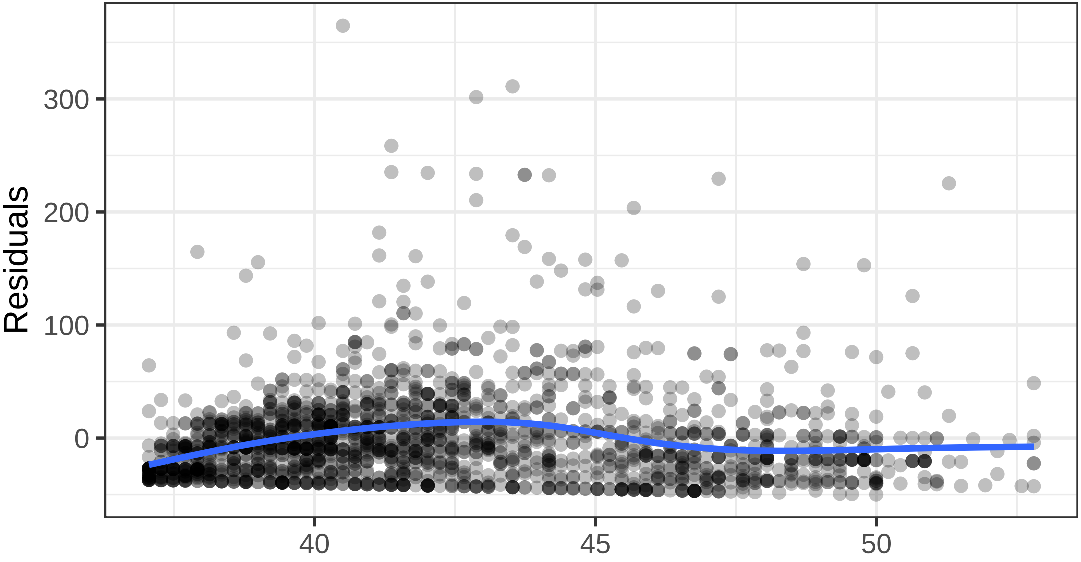
Bad: Nonlinear Residuals



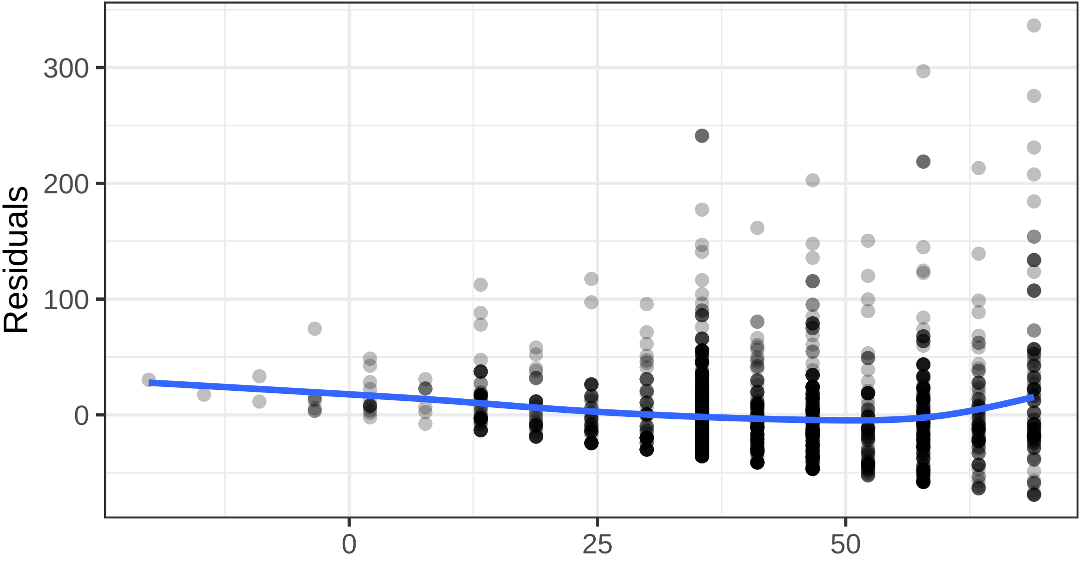
Residuals: Regressing Earnings on Height



Residuals: Regressing Earnings on Age



Residuals: Regressing Earnings on Education



	Earnings (2021)		
	(1)	(2)	(3)
Height	3.08*		
	(0.25)		
Age		0.22*	
		(0.06)	
Education			5.57*
			(0.36)
Constant	−162.54*	33.17*	−31.34*
	(16.46)	(2.64)	(4.89)
Observations	1,815	1,815	1,813
Adjusted R^2	0.08	0.01	0.11
Residual Std. Error	40.28 (df = 1813)	41.81 (df = 1813)	39.50 (df = 1811)
F Statistic	155.52* (df = 1; 1813)	14.22* (df = 1; 1813)	235.79* (df = 1; 1811)
Note:			*$p < 0.05$

Which is a better model of personal income (earnk):

1. Mother's education level, or
2. Personal exercise

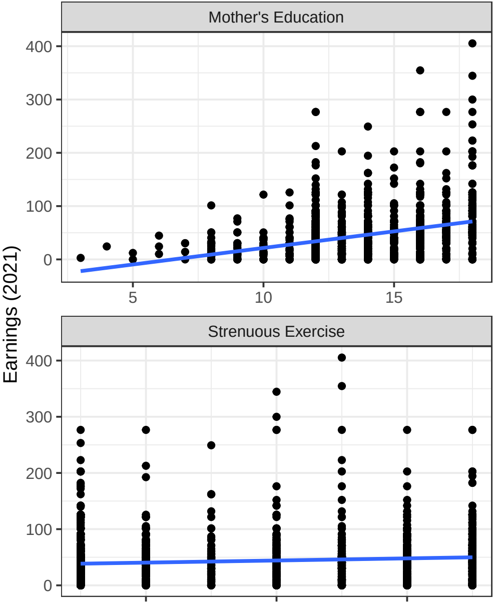
Which is a better model of personal income (earnk):

1. Mother's education level, or
2. Personal exercise

Step 1: Fit and evaluate both models

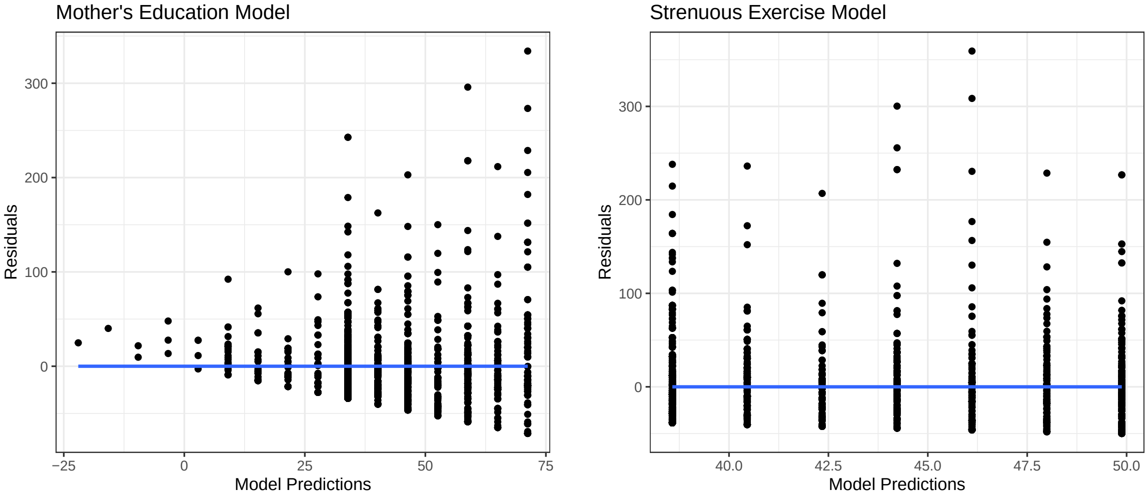
- Scatterplots, regression tables and residuals plots

	Earnings (2021)	
	(1)	(2)
Mother's Education	6.21* (0.43)	
Exercise		1.88* (0.42)
Constant	−40.64* (5.87)	36.69* (1.62)
Observations	1,570	1,815
Adjusted R^2	0.12	0.01
Residual Std. Error	40.45 (df = 1568)	41.75 (df = 1813)
F Statistic	212.09* (df = 1; 1568)	19.83* (df = 1; 1813)
Note:		*p < 0.05



	Earnings (2021)	
	(1)	(2)
Mother's Education	6.21*	
	(0.43)	
Exercise		1.88*
		(0.42)
Constant	−40.64*	36.69*
	(5.87)	(1.62)
Observations	1,570	1,815
Adjusted R^2	0.12	0.01
Note:		*$p < 0.05$

Step 4 - Check the Residuals



Step 2: Make four predictions

Mother’s Education Model

1. Mother completed high school (12)
2. Mother completed college (16)

Strenuous Exercise Model

1. No strenuous exercise (1)
2. Strenuous exercise $>$ 3x per week (7)

Step 2: Make four predictions

Mother's Education Model

1. Mother completed high school (12) $=$ \$33.9k
2. Mother completed college (16) $=$ \$58.7k

Strenuous Exercise Model

1. No strenuous exercise (1) $=$ \$38.6k
2. Strenuous exercise $>$ 3x per week (7) $=$ \$49.8k

For Tuesday

Use chapter 5 of the textbook to add confidence intervals to our four model predictions.