Today's Agenda

1. Explore the intuitions of OLS regression

2. Practice fitting and interpreting simple OLS regressions

Justin Leinaweaver (Spring 2022)

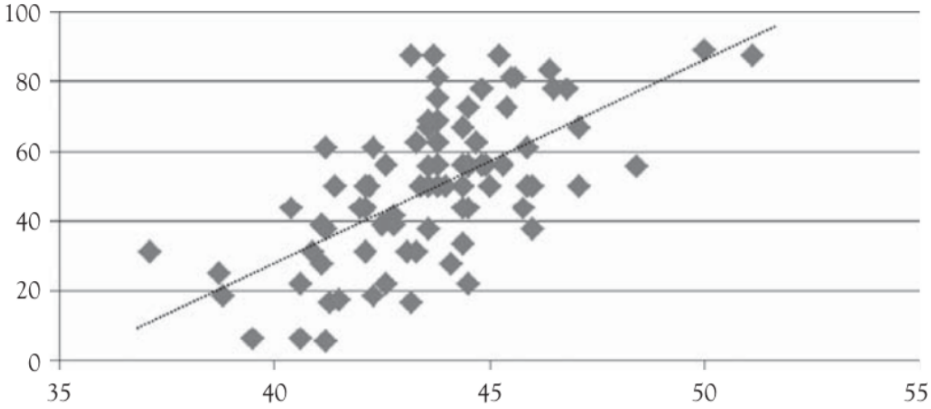
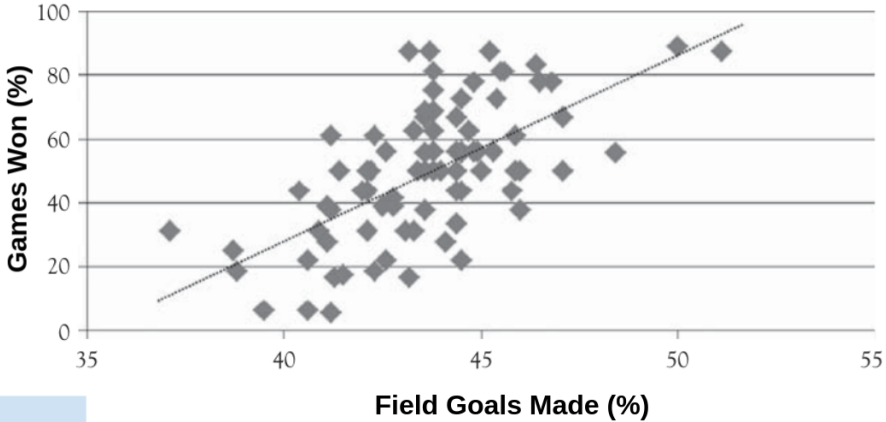


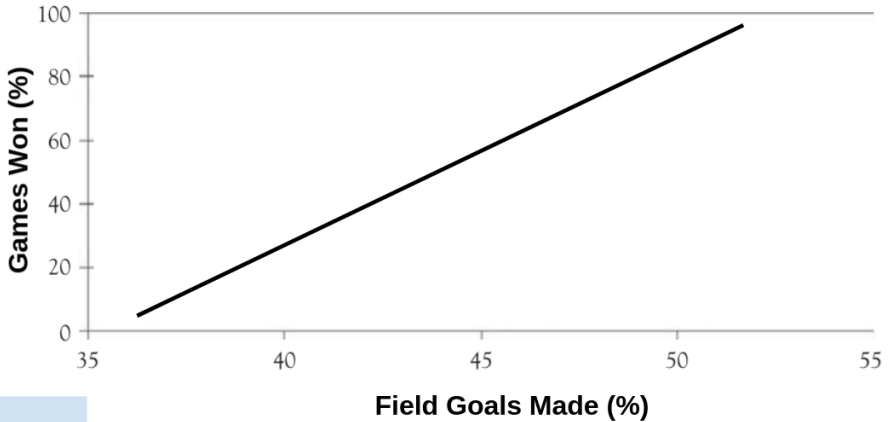
Figure 2.4. A scatterplot of winning percentage (vertical Y-axis) versus field goal percentage (horizontal X-axis).

Do more efficient college basketball teams win more games?



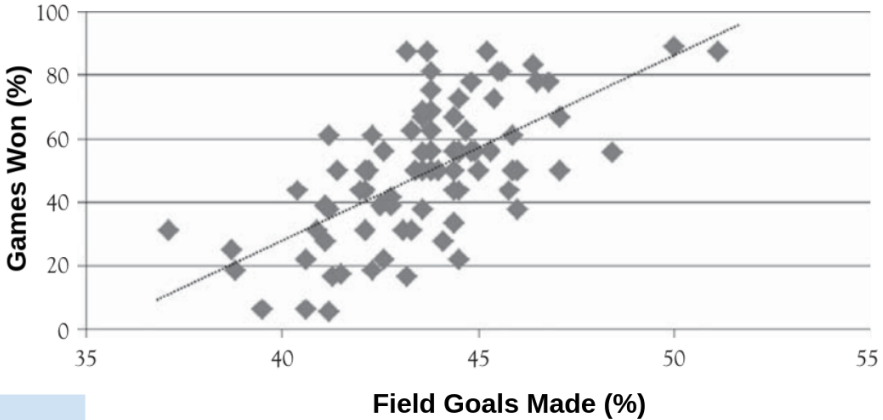
Source: Wilson, Keating, and Beal-Hodges 2012

Do more efficient college basketball teams win more games?



Source: Wilson, Keating, and Beal-Hodges 2012

Do more efficient college basketball teams win more games?



Source: Wilson, Keating, and Beal-Hodges 2012

The Formula for a Line

$$y = mx + b$$

is equivalent to

$$y = \alpha + \beta\, x$$

The Formula for a Line

$$y = \alpha + \beta x$$

- y is the outcome
- α is the constant
- β is the coefficient estimate
- x is the predictor

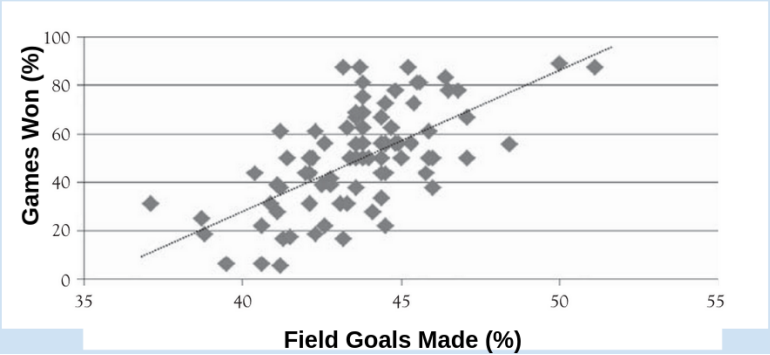
The Formula for a Line

$$y = \alpha + \beta x$$

- y is the outcome
- α is the constant (**the intercept**)
- β is the coefficient estimate (**the slope**)
- x is the predictor

$$y = \alpha + \beta x$$

Games Won = -198.9 + 5.7 (Field Goals Made)



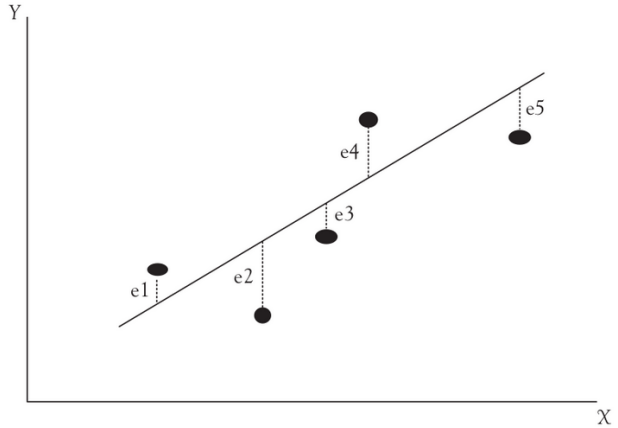
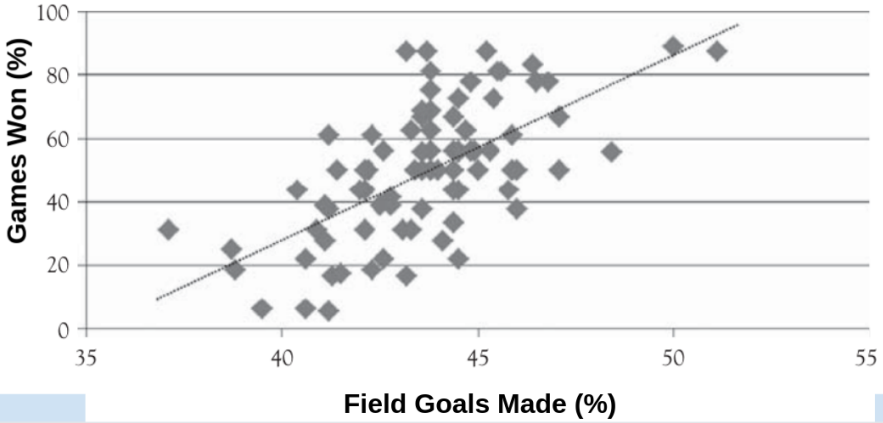


Figure 3.2. The ordinary least squares regression line for Y as a function of X. Residuals (or deviations or errors) between each point and the regression line are labeled e_i.

$$WP = -198.9 + 5.707(FG)$$



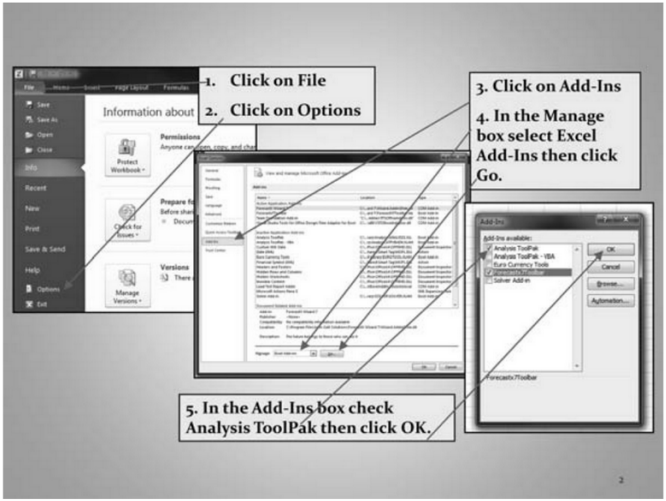


Figure 1.6. Getting "Data Analysis" in Excel 2010.

Work, Family, and Well-Being in the United States, 1990 (ICPSR 6666)

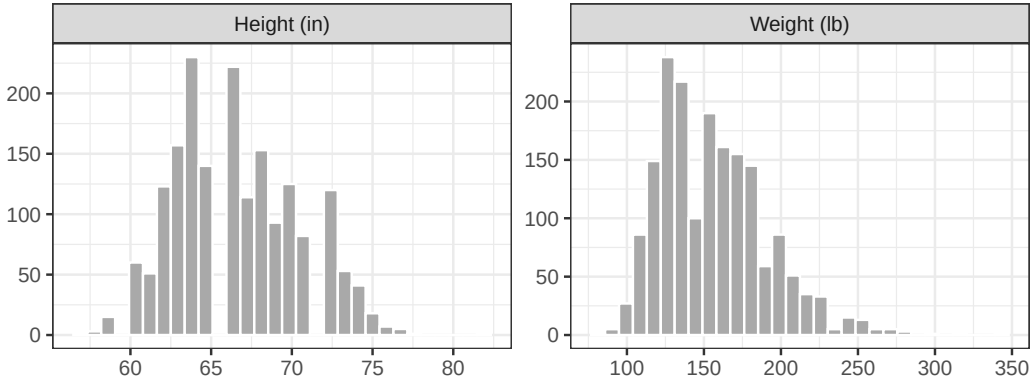
Version Date: Jun 10, 1996 Cite this study | Share this page

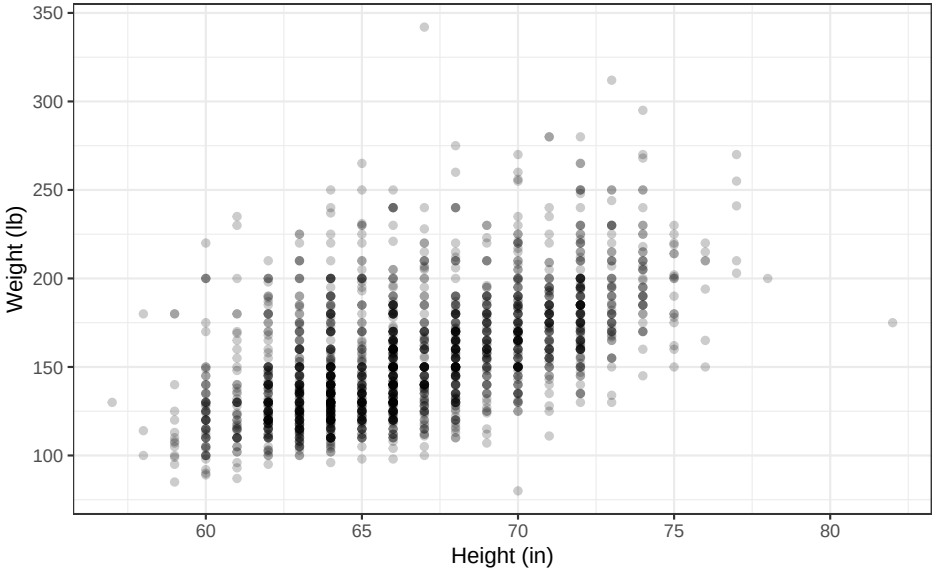
Principal Investigator(s):
Catherine E. Ross

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	height	weight	male	earn	earnk	ethnicity	education	mother_education	father_education	walk	exercise	smokenow	tense	angry	age
2	74	210	1	50000	50	White	16	16	16	3	3	2	0	0	45
3	66	125	0	60000	60	White	16	16	16	6	5	1	0	0	58
4	64	126	0	30000	30	White	16	16	16	8	1	2	1	1	29
5	65	200	0	25000	25	White	17	17	NA	8	1	2	0	0	57
6	63	110	0	50000	50	Other	16	16	16	5	6	2	0	0	91
7	68	165	0	62000	62	Black	18	18	18	1	1	2	2	2	54
8	63	190	0	51000	51	White	17	17	17	3	1	2	4	4	39
9	64	125	0	9000	9	White	15	15	15	7	4	1	4	4	26
10	62	200	0	29000	29	White	12	12	12	2	2	2	0	0	49
11	73	230	1	32000	32	White	17	17	17	7	1	1	0	0	46
12	72	176	1	2000	2	Hispanic	15	15	15	8	1	2	0	0	21
13	72	265	1	35000	35	White	NA	NA	NA	1	1	2	0	0	53
14	72	160	1	27000	27	White	12	12	12	1	2	2	1	1	26
15	70	225	1	6530	6.53	White	16	16	NA	4	1	2	0	0	65
16	63	107	0	0	0	White	14	14	14	7	4	2	2	2	50

Is height a useful model of weight in the Ross (1990) sample?

Univariate Analysis




Weight (lb)
350
300
250
200
150
100
60
65
70
75
80
Height (in)

Formatting a Simple OLS Regression Table

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.55
R Square	0.30
Adjusted R Square	0.30
Standard Error	28.96
Observations	1788.00

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	643873.73	643873.73	767.70	7.3198E-141
Residual	1786	1497935.78	838.71		
Total	1787	2141809.51			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-173.26	11.91	-14.54	2.27E-45	-196.63
height	4.95	0.18	27.71	7.3E-141	4.60

Formatting a Simple OLS Regression Table

	Model 1
Predictor	*Coefficient*
	(Standard Error)
Constant	*Coefficient*
	(Standard Error)
Observations	*# of Observations*
Adjusted R^2	*Adj R^2 value*
Residual Std Error	*Model standard error*
F Statistic	*F value and significance*

Add '*' next to any coefficient with a p-value less than or equal to 0.05

Formatting a Simple OLS Regression Table

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.55
R Square	0.30
Adjusted R Square	0.30
Standard Error	28.96
Observations	1788.00

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	643873.73	643873.73	767.70	7.3198E-141
Residual	1786	1497935.78	838.71		
Total	1787	2141809.51			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-173.26	11.91	-14.54	2.27E-45	-196.63
height	4.95	0.18	27.71	7.3E-141	4.60

	Model 1
Predictor	*Coefficient*
	(Standard Error)
Constant	*Coefficient*
	(Standard Error)
Observations	*# of Observations*
Adjusted R^2	*Adj R^2 value*
Residual Std Error	*Model standard error*
F Statistic	*F value and significance*

Formatting a Simple OLS Regression Table

SUMMARY OUTPUT

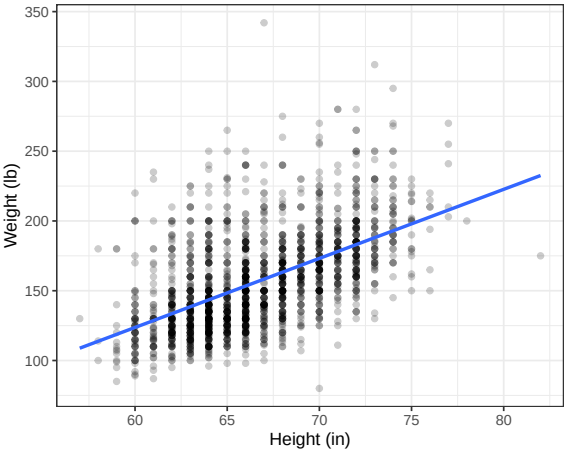
Regression Statistics	
Multiple R	0.55
R Square	0.30
Adjusted R Square	0.30
Standard Error	28.96
Observations	1788.00

ANOVA

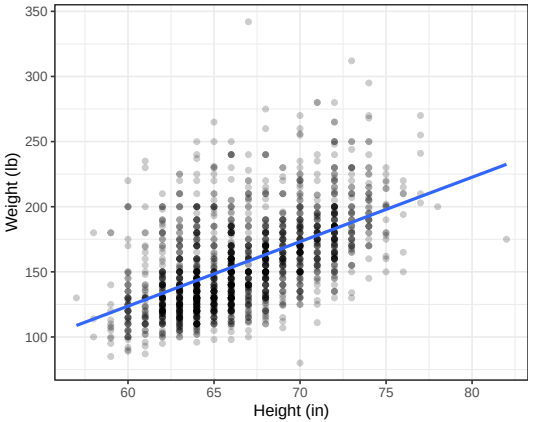
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	643873.73	643873.73	767.70	7.3198E-141
Residual	1786	1497935.78	838.71		
Total	1787	2141809.51			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-173.26	11.91	-14.54	2.27E-45	-196.63
height	4.95	0.18	27.71	7.3E-141	4.60

	Weight
Height	4.95*
	(0.18)
Constant	-173.26*
	(11.91)
Observations	1,788
Adjusted R2	0.30
Residual Std. Error	28.96 (df = 1786)
F Statistic	767.70* (df = 1; 1786)
Note:	*p < 0.05

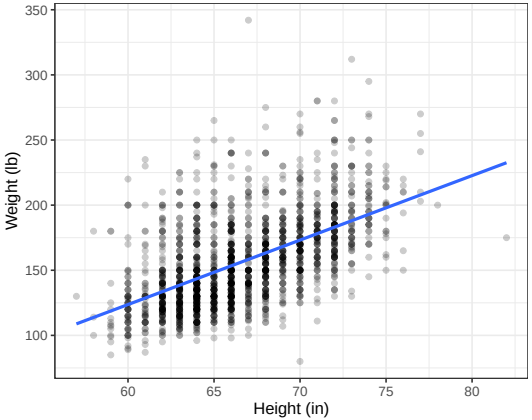


	Weight
Height	4.95*
	(0.18)
Constant	−173.26*
	(11.91)
Observations	1,788
Adjusted R^2	0.30
Residual Std. Error	28.96 (df = 1786)
F Statistic	767.70* (df = 1; 1786)
Note:	*p < 0.05



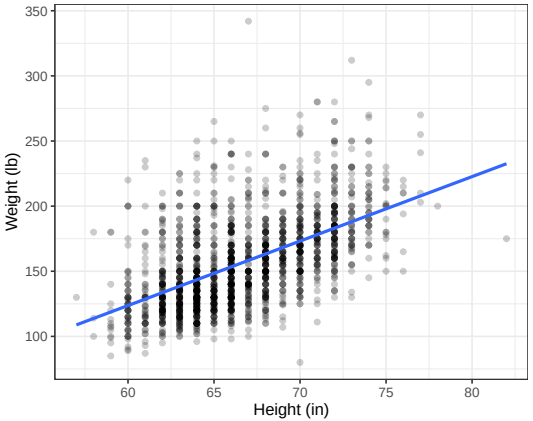
	Weight
Height	4.95*
	(0.18)
Constant	−173.26*
	(11.91)
Observations	1,788
Adjusted R^2	0.30
Residual Std. Error	28.96 (df = 1786)
F Statistic	767.70* (df = 1; 1786)
Note:	*p < 0.05

Outcome = Constant + Beta Coefficient * Predictor



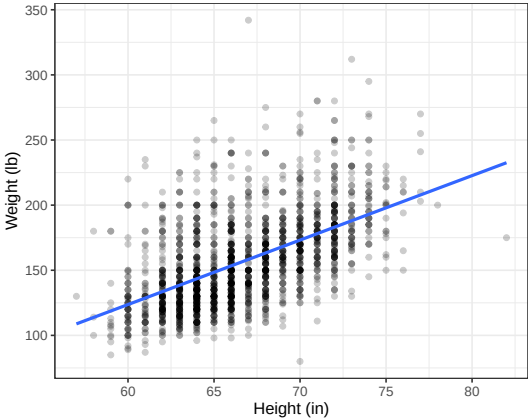
	Weight
Height	4.95*
	(0.18)
Constant	−173.26*
	(11.91)
Observations	1,788
Adjusted R^2	0.30
Residual Std. Error	28.96 (df = 1786)
F Statistic	767.70* (df = 1; 1786)
Note:	*p < 0.05

$$\text{Weight} = \text{-173.26} + 4.95 \text{ x Height}$$



	Weight
Height	4.95*
	(0.18)
Constant	−173.26*
	(11.91)
Observations	1,788
Adjusted R^2	0.30
Residual Std. Error	28.96 (df = 1786)
F Statistic	767.70* (df = 1; 1786)
Note:	*p < 0.05

$$\text{Weight} = \text{-173.26} + 4.95 \text{ x } 64 \approx 143.54 \text{ lb}$$



	Weight
Height	4.95*
	(0.18)
Constant	−173.26*
	(11.91)
Observations	1,788
Adjusted R^2	0.30
Residual Std. Error	28.96 (df = 1786)
F Statistic	767.70* (df = 1; 1786)
Note:	*p < 0.05

$$\text{Weight} = \text{-173.26} + 4.95 \text{ x } 69 \approx 168.29 \text{ lb}$$

Work, Family, and Well-Being in the United States, 1990 (ICPSR 6666)

Version Date: Jun 10, 1996 Cite this study | Share this page

Principal Investigator(s):
Catherine E. Ross

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	height	weight	male	earn	earnk	ethnicity	education	mother_education	father_education	walk	exercise	smokenow	tense	angry	age
2	74	210	1	50000	50	White	16	16	16	3	3	2	0	0	45
3	66	125	0	60000	60	White	16	16	16	6	5	1	0	0	58
4	64	126	0	30000	30	White	16	16	16	8	1	2	1	1	29
5	65	200	0	25000	25	White	17	17	NA	8	1	2	0	0	57
6	63	110	0	50000	50	Other	16	16	16	5	6	2	0	0	91
7	68	165	0	62000	62	Black	18	18	18	1	1	2	2	2	54
8	63	190	0	51000	51	White	17	17	17	3	1	2	4	4	39
9	64	125	0	9000	9	White	15	15	15	7	4	1	4	4	26
10	62	200	0	29000	29	White	12	12	12	2	2	2	0	0	49
11	73	230	1	32000	32	White	17	17	17	7	1	1	0	0	46
12	72	176	1	2000	2	Hispanic	15	15	15	8	1	2	0	0	21
13	72	265	1	35000	35	White	NA	NA	NA	1	1	2	0	0	53
14	72	160	1	27000	27	White	12	12	12	1	2	2	1	1	26
15	70	225	1	6530	6.53	White	16	16	NA	4	1	2	0	0	65
16	63	107	0	0	0	White	14	14	14	7	4	2	2	2	50

Is height a useful model of weight in the Ross (1990) sample?

Analyze Three OLS Models

For each model: Make a regression table, scatter plot and a prediction using the average value of the predictor.

- Model 1: Regress earnings (earnk2021) on height
- Model 2: Regress earnings (earnk2021) on age
- Model 3: Regress earnings (earnk2021) on education

For Thursday

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.