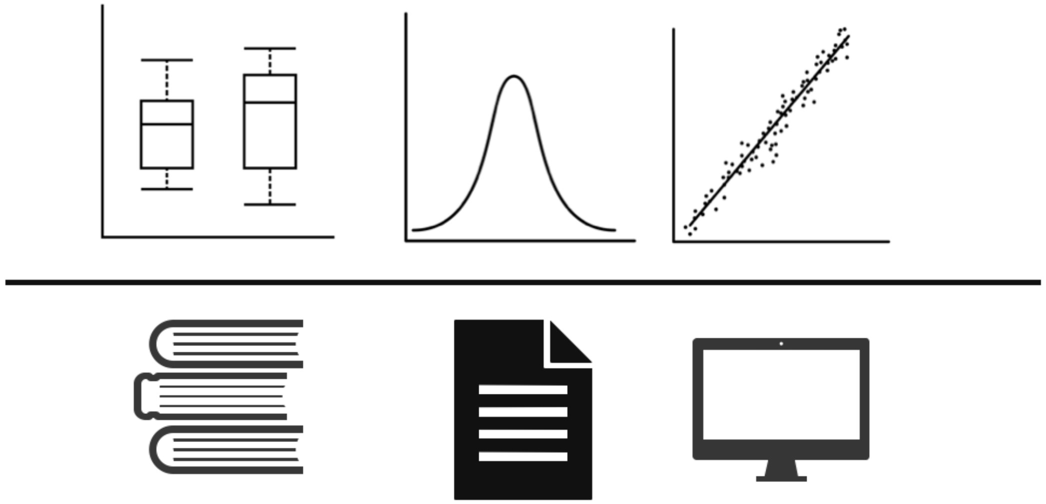
PSYC 2300

Introduction to Statistics



Lecture 15: Making predictions using regression

Outline for today

- Updates
- Review parts of last class
- Simple Linear Regression
- Multiple Regression
- Regressions in JASP
  - Download `Stats Class 17 Dataset (Regressions).jasp`

Updates

- **Final Exam** is scheduled for Thursday March 17, 2022 08:00-09:50am MT
- **Application Project** data and instructions will be posted this afternoon
  - Each focuses on descriptive statistics (mean, range), correlation, independent samples *t*-test, and factorial ANOVA
  - We will talk about these projects specifically during the next two class sessions

Review

Testing relationships using correlations

Calculating Correlations

Pearson's *r*

$$r_{xy} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

We just need six values:

$\sum X$

$\sum Y$

$\sum XY$

$\sum X^2$

$\sum Y^2$

n

Correlations: Hypothesis Testing

Null Hypothesis H_0

There is no relationship

$$H_0\colon \rho = 0$$

Alternative Hypothesis H_a

There is a relationship

$$H_a\colon \rho \neq 0$$

Note: ρ is "rho", the population parameter version of the sample statistic, r

t-test for correlation coefficient

_t_-test for correlation coefficient

$$t_{r_{xy}} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

$r_{xy} =$ the correlation between the two variables

$r_{xy}^2 =$ the coefficient of determination

$n =$ the sample size (number of paired observations)

t-test for correlation coefficient

Step 1: Calculate Pearson's *r*

$$r_{xy} = \frac{126}{\sqrt{(213)(192)}} = 0.62$$

Step 2: Calculate the coefficient of determination

$$r_{xy}^2 = 0.62^2 = 0.38$$

t-test for correlation coefficient

$$t_{r_{xy}} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

Step 3: Solve for the t-test statistic for correlations

$$t_{xy} = \frac{0.62\sqrt{6-2}}{\sqrt{1-0.38}}$$

Simple Linear Regression

Types of Models

- For ANOVAs, your factors need to be *nominal* or *ordinal*
  - These are called **mean structure models** because we seek to explain mean differences between groups or treatment conditions
- We also have **regression models**, however, which allows you to have *interval* or *ratio* variables on the x-axis

Simple Example

We want to predict **Total Cost** based on the **Number of drinks** you buy at a dance contest

Simple Example

Total cost = Cover Charge + $8.00 per drink

Simple Example

Total cost = Cover Charge + $8.00 per drink


Total cost = $15.00 + $8.00x

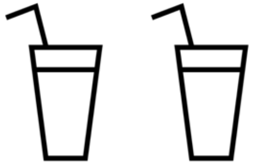
$15 + $8(1) = $23.00

Simple Example

Total cost = Cover Charge + $8.00 per drink

Total cost = $15.00 + $8.00x

$15 + $8(1) = $23.00



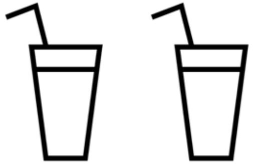
$15 + $8(2) = $31.00

Simple Example

Total cost = Cover Charge + $8.00 per drink

Total cost = $15.00 + $8.00x

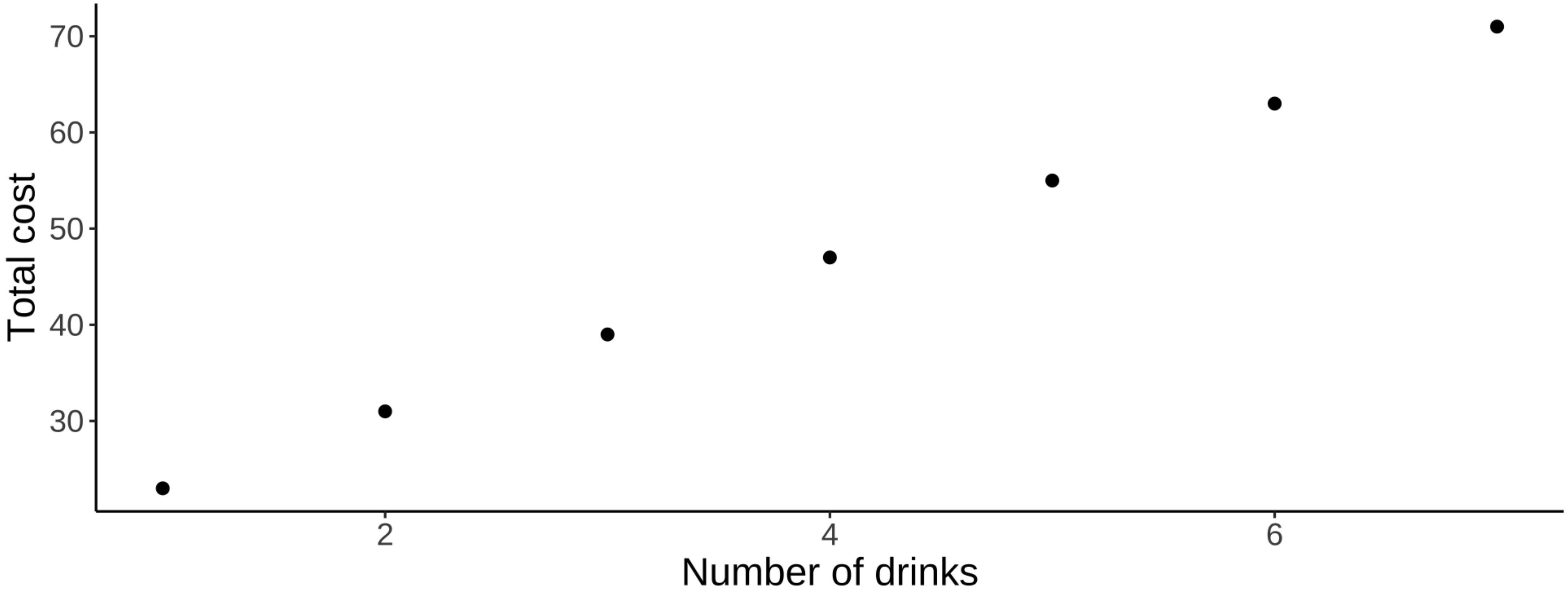
$15 + $8(1) = $23.00



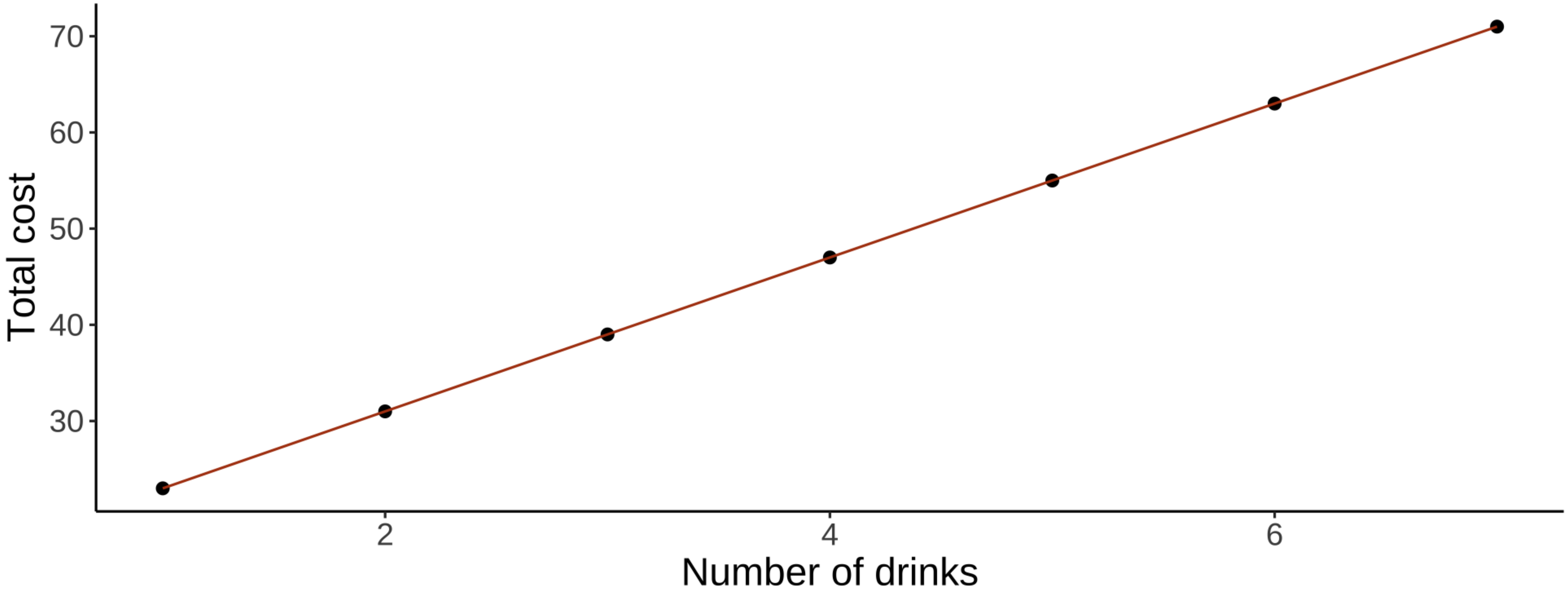
$15 + $8(2) = $31.00

$15 + $8(3) = $39.00

Simple Example

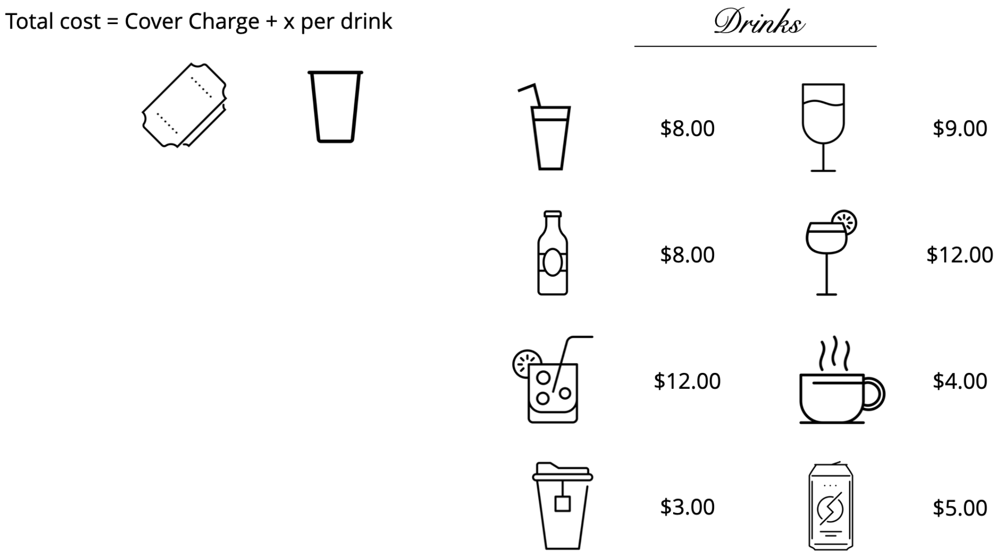


Simple Example



The Problem

- Typically, however, we will have error (which is not reflected in this model)
  - For example, different drinks cost different amounts

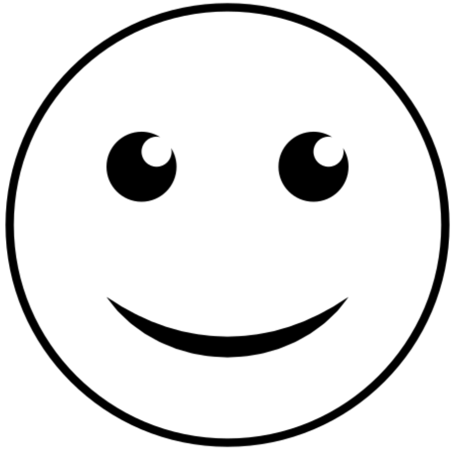


- Or if we have **sampling error** more generally

Another Example

Predicting **Final Grade** based on the number of **Hours Studied**

Two people who study the same amount may not get the same final grade



Studied 8 hours

83%

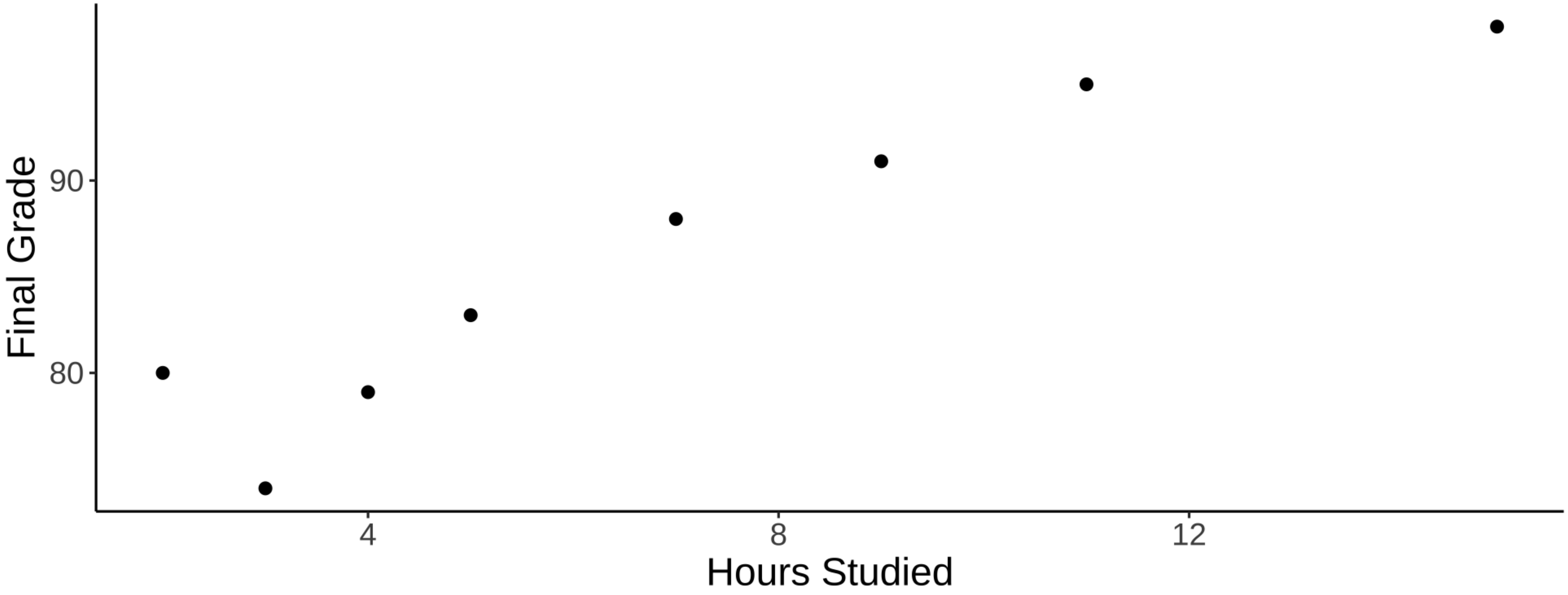
Studied 8 hours

62%

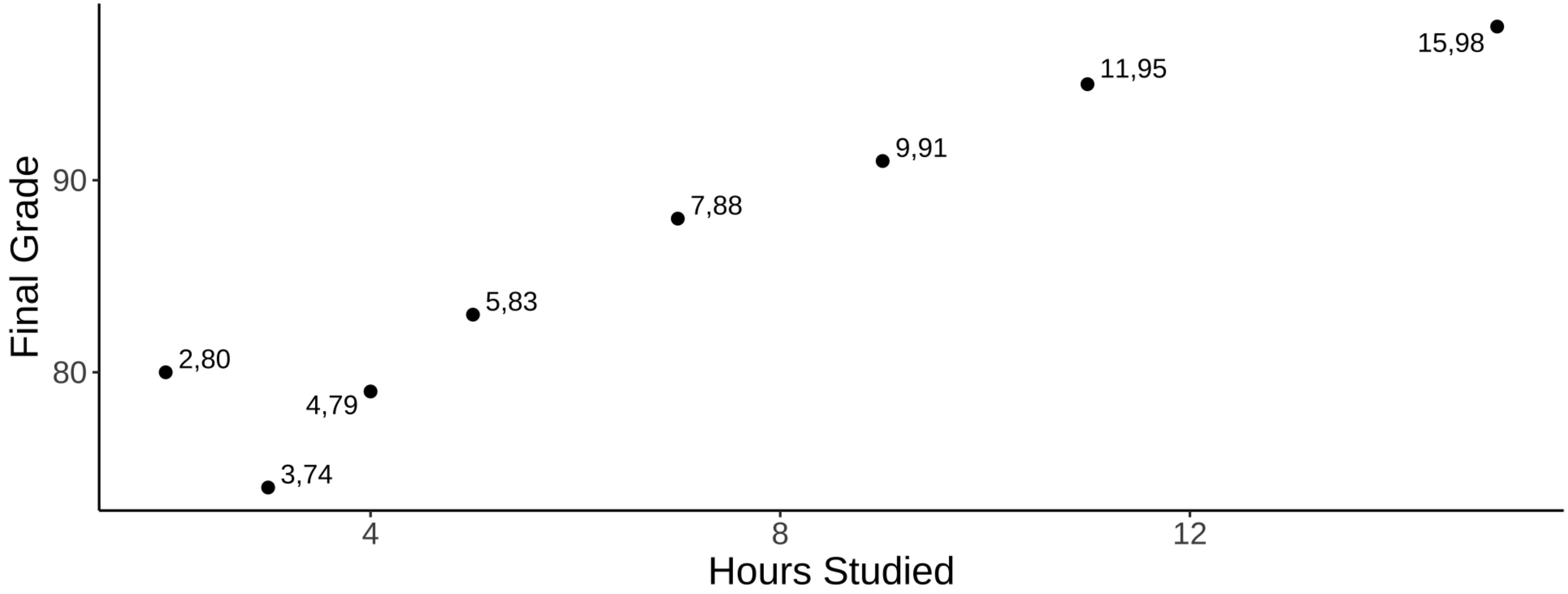
Predicting Grades from Hours Studied

Student	Hours_Studied	Final_Grade
Niwako	7	88
Kamiko	5	83
John	9	91
Ahmed	2	80
Jimothy	4	79
Jane	11	95
Jaleel	15	98
Steve	3	74

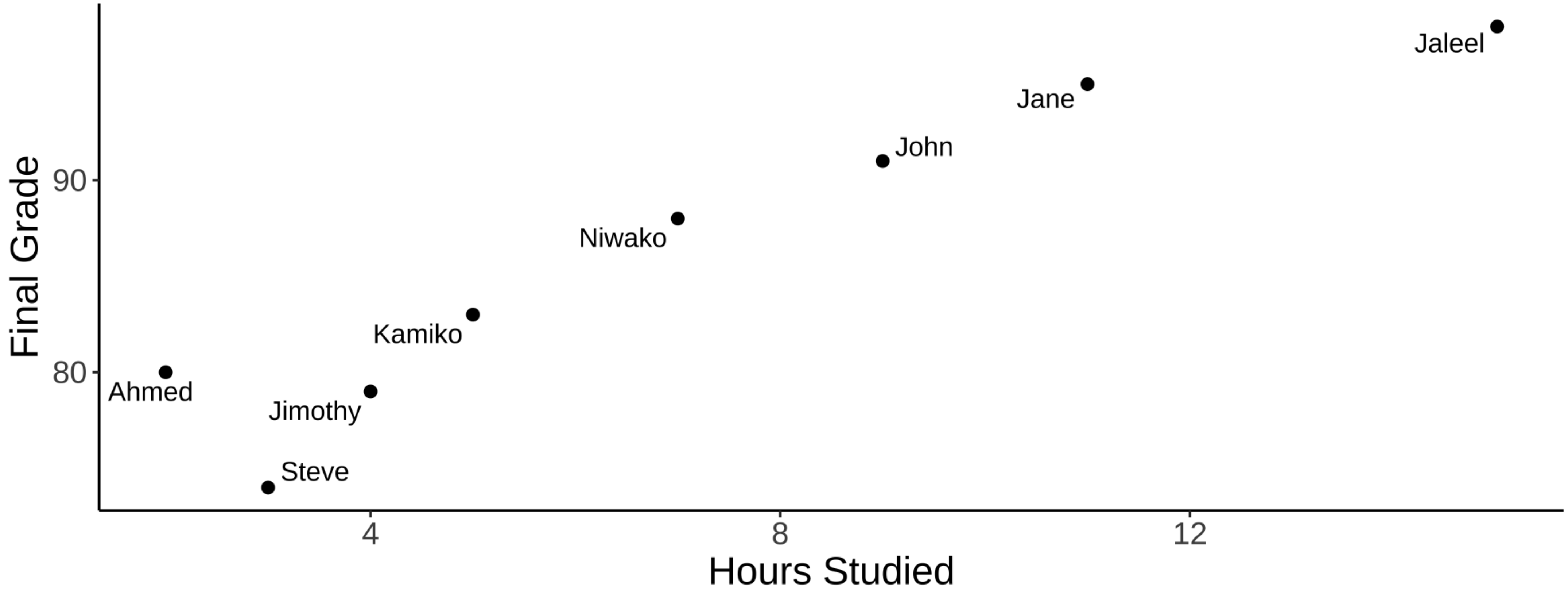
Predicting Grades from Hours Studied



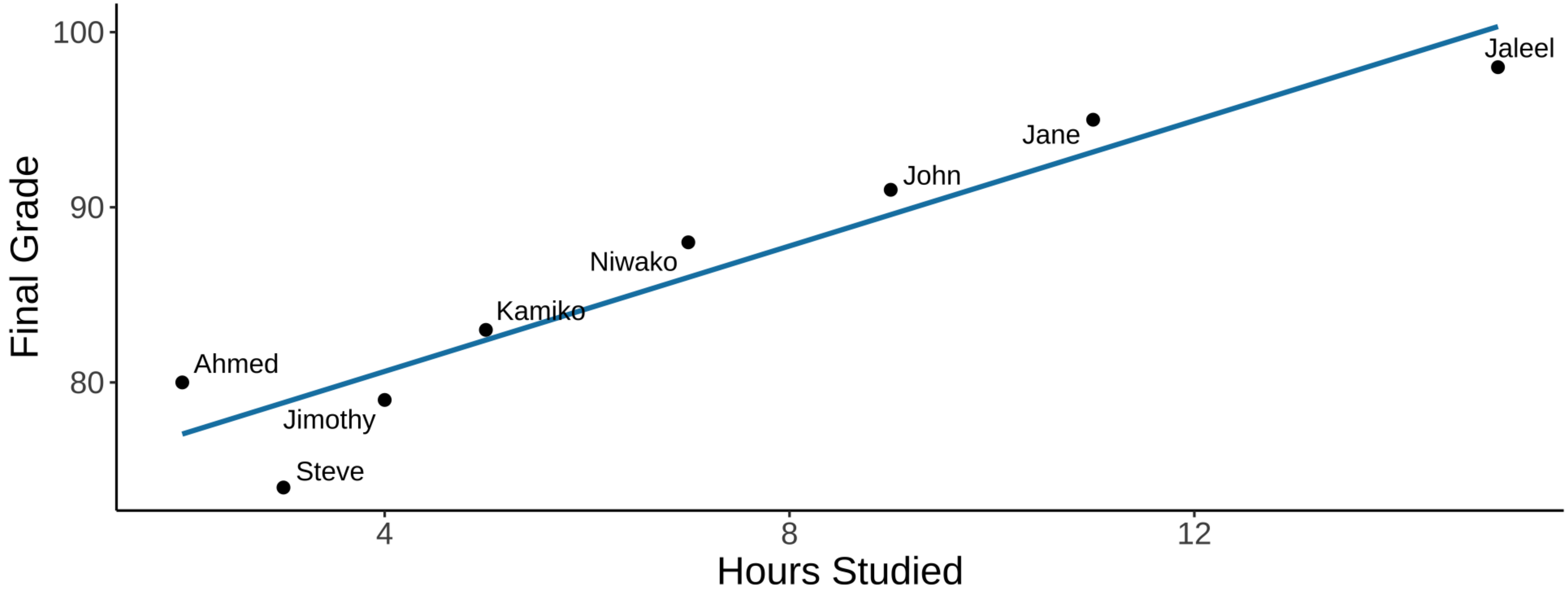
Predicting Grades from Hours Studied



Predicting Grades from Hours Studied



Predicting Grades from Hours Studied

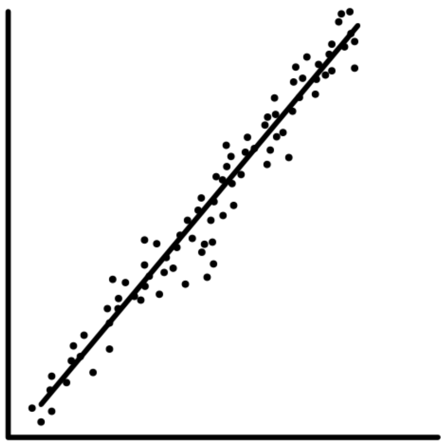


The blue line represents the **Regression Line**

Regression

Regression: A statistical technique for finding the best-fitting *line* (i.e., the regression line) for a set of data

Goal: To be able to model the quantitative relationship between two variables, allowing us to make *predictions* about one on the basis of the other



Statistical Hypotheses

Null Hypothesis H_0

The slope of the regression equation, b (or "beta") is zero

Alternative Hypothesis H_a

The slope of the regression equation, b (or "beta") is not zero

We'll see how this is applied when using JASP

Example

Research question: How much does each additional hour of studying, on average, help our final grade?

Simple Linear Regression

Simple linear regression

$$\hat{y} = bX + a$$

Slope

$$b = \frac{\Sigma XY - \frac{\Sigma x \Sigma y}{n}}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}$$

Intercept

$$a = \frac{\Sigma Y - b\Sigma X}{n}$$

Review: Correlation

Pearson's *r*

$$r_{xy} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

Once again, *these* are the values we need:

$\sum X$
$\sum Y$
$\sum XY$

$\sum X^2$
$\sum Y^2$
n

Calculating the regression line

Set up our table

i	x	y	x^2	y^2	xy
1	5	6			
2	9	11			
3	10	6			
4	3	4			
5	5	6			
6	7	9			

Calculating the regression line

Square our x values

i	x	y	x^2	y^2	xy
1	5	6	25		
2	9	11	81		
3	10	6	100		
4	3	4	9		
5	5	6	25		
6	7	9	49		

Calculating the regression line

Square our y values

i	x	y	x^2	y^2	xy
1	5	6	25	36	
2	9	11	81	121	
3	10	6	100	36	
4	3	4	9	16	
5	5	6	25	36	
6	7	9	49	81	

Calculating the regression line

Multiply x and y together row-wise

i	x	y	x^2	y^2	xy
1	5	6	25	36	30
2	9	11	81	121	99
3	10	6	100	36	60
4	3	4	9	16	12
5	5	6	25	36	30
6	7	9	49	81	63

Calculating the regression line

i	x	y	x^2	y^2	xy
1	5	6	25	36	30
2	9	11	81	121	99
3	10	6	100	36	60
4	3	4	9	16	12
5	5	6	25	36	30
6	7	9	49	81	63

Sum each of the columns

$\sum x$	$\sum y$	$\sum x^2$	$\sum y^2$	$\sum xy$
39	42	289	326	294

Calculating the regression line

Slope

$$b = \frac{\Sigma XY - \frac{\Sigma x \Sigma y}{n}}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}$$

$\sum x$	$\sum y$	$\sum x^2$	$\sum y^2$	$\sum xy$
39	42	289	326	294

$$b = \frac{294 - \frac{(39)(42)}{6}}{289 - \frac{(39)^2}{6}}$$

$$b = \frac{294 - 273}{289 - 253.5} = \frac{21}{35.5} = 0.59$$

Calculating the regression line

Intercept

$$a = \frac{\Sigma Y - b\Sigma X}{n}$$

$\sum x$	$\sum y$	$\sum x^2$	$\sum y^2$	$\sum xy$
39	42	289	326	294

Slope, $b = 0.59$

$$a = \frac{42 - (0.59)(39)}{6} = \frac{18.99}{6} = 3.16$$

Calculating the regression line

Simple linear regression

$$\hat{y} = bX + a$$

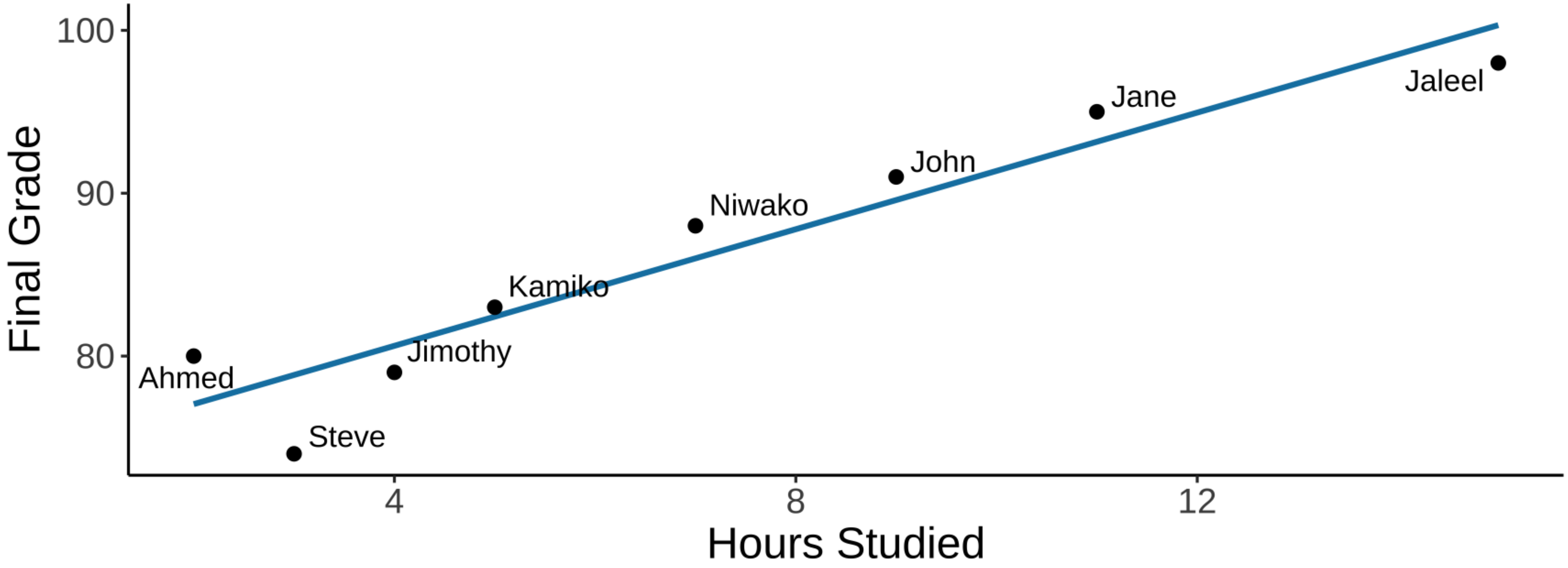
Slope: $b = 0.59$

Intercept: $a = 3.16$

$$\hat{y} = 0.59X + 3.16$$

Back to our previous example

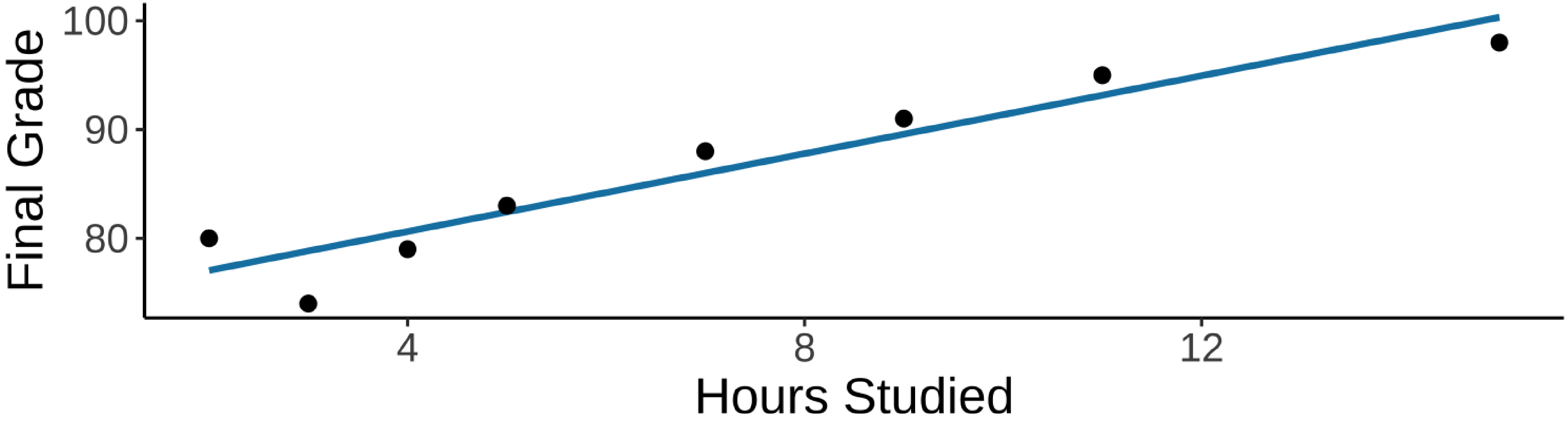
Research question: How much does each additional hour of studying, on average, help our final grade?



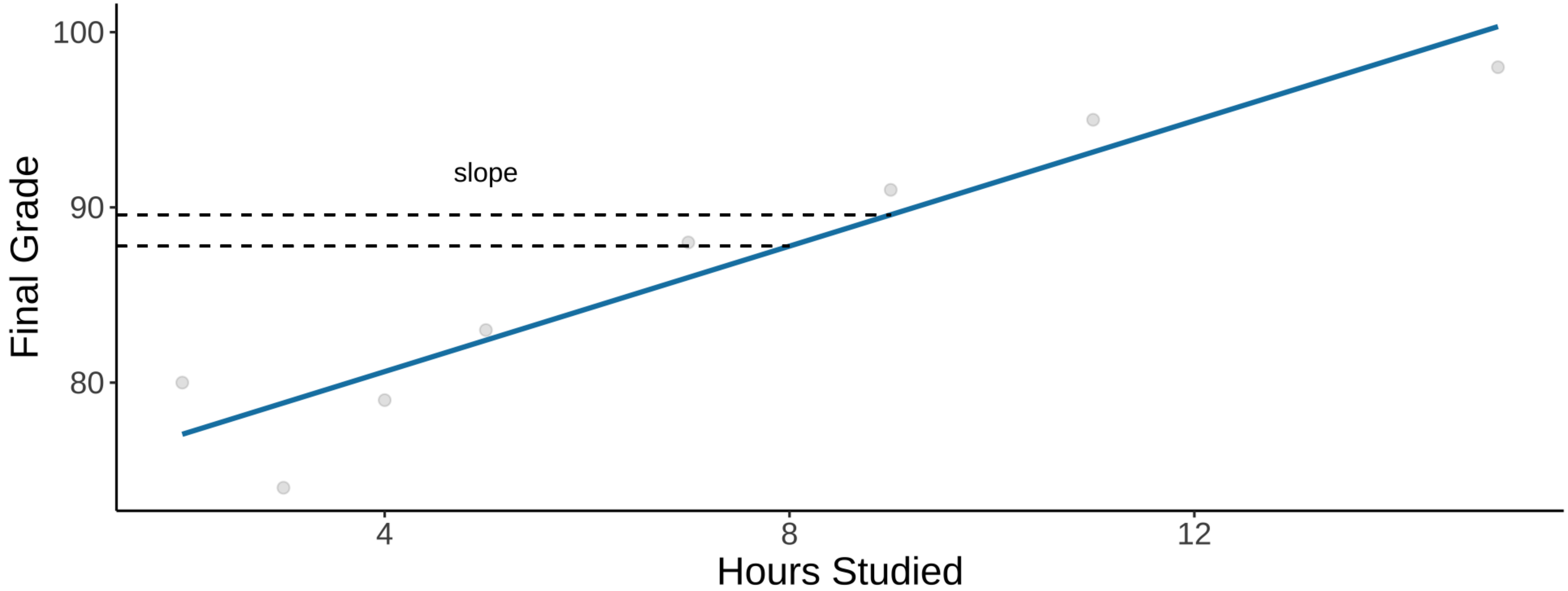
Visualizing the Regression Formula

$$\hat{y} = bX + a + \epsilon$$

$$\hat{y} = 1.79X + 73.47$$



Visualizing b in the Regression Formula

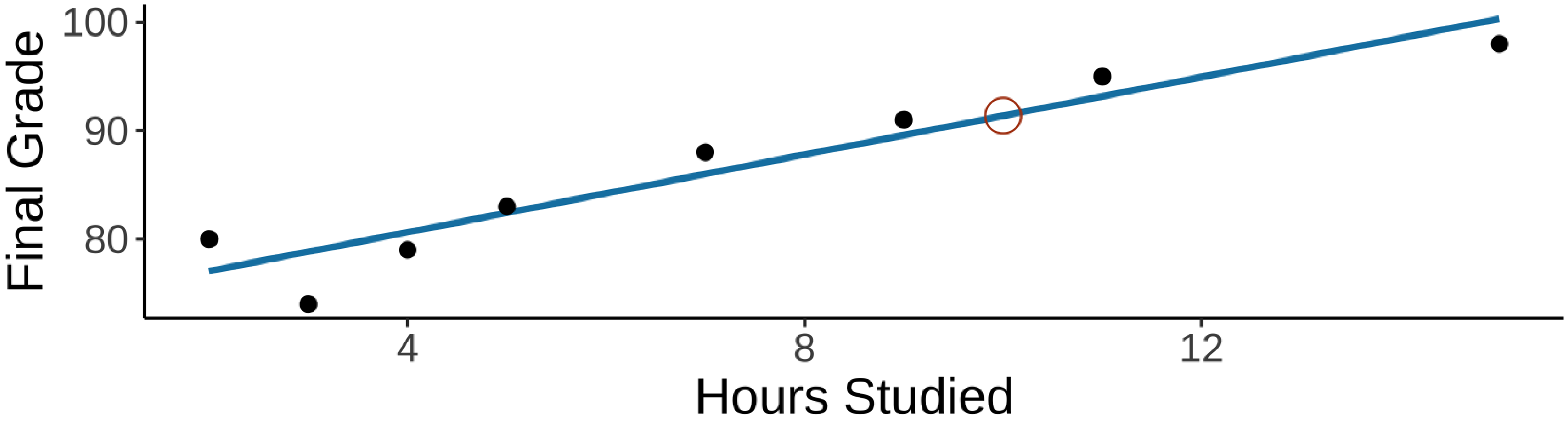


Slope determines how much the y variable changes when X is increased by 1 point (or unit)

Visualizing X in the Regression Formula

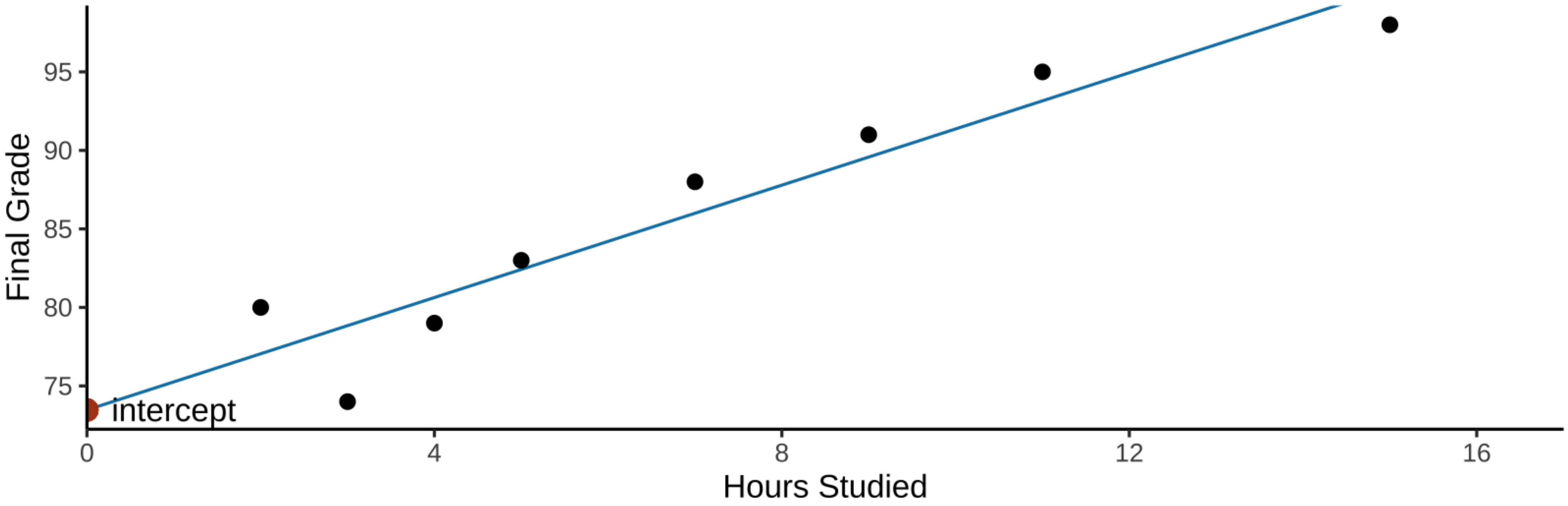
What if a student studied for 10 hours?

$$\hat{y} = 1.79(10) + 73.47 = 91.37$$



Visualizing a in the Regression Formula

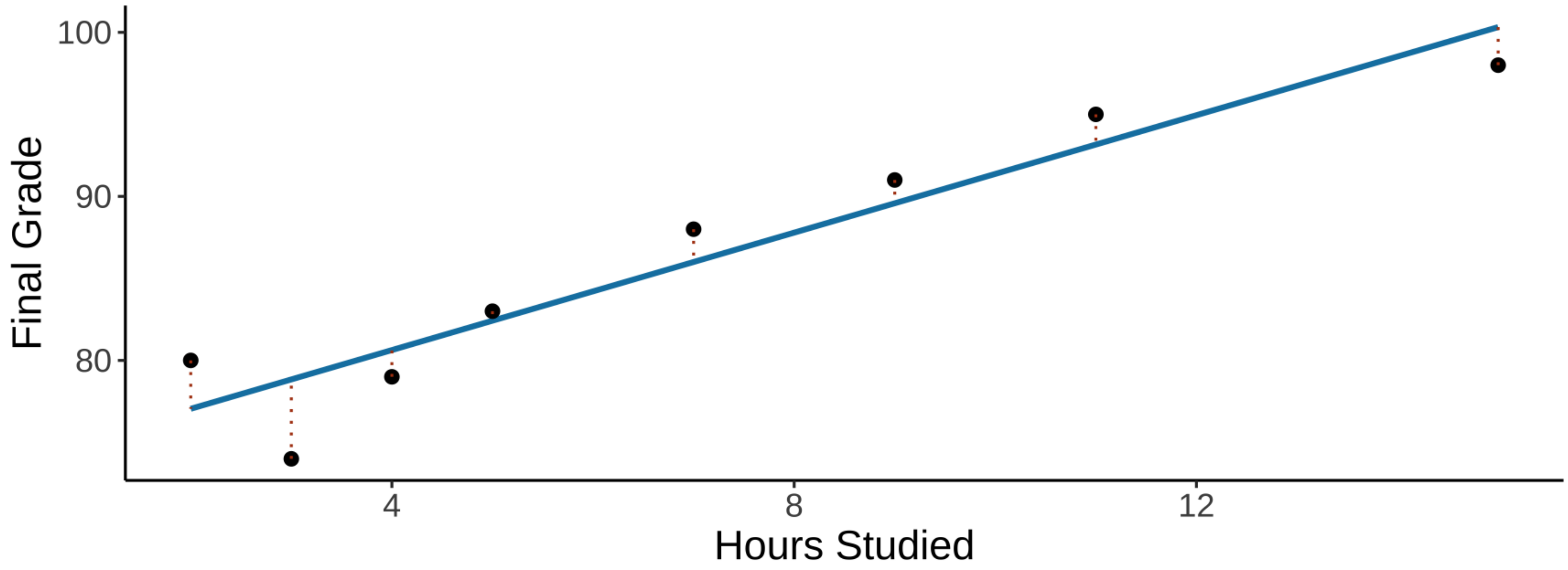
What if a student studied for 0 hours?



The y-intercept, a, is the value of y when X is 0

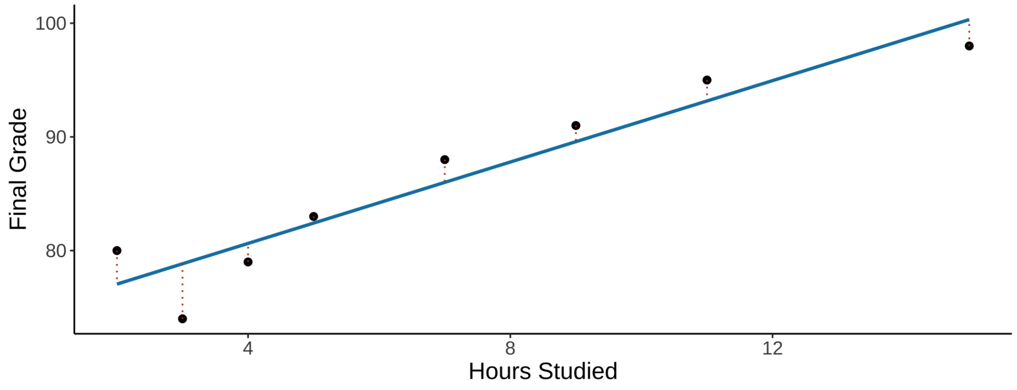
Visualizing ϵ in the Regression Formula

The model is not perfect; $residuals = y - \hat{y}$



Visualizing ϵ in the Regression Formula

Final_Grade	Prediced_Grade	Residual
88	86.00	2.00
83	82.42	0.58
91	89.58	1.42
80	77.05	2.95
79	80.63	-1.63
95	93.16	1.84
98	100.32	-2.32
74	78.84	-4.84

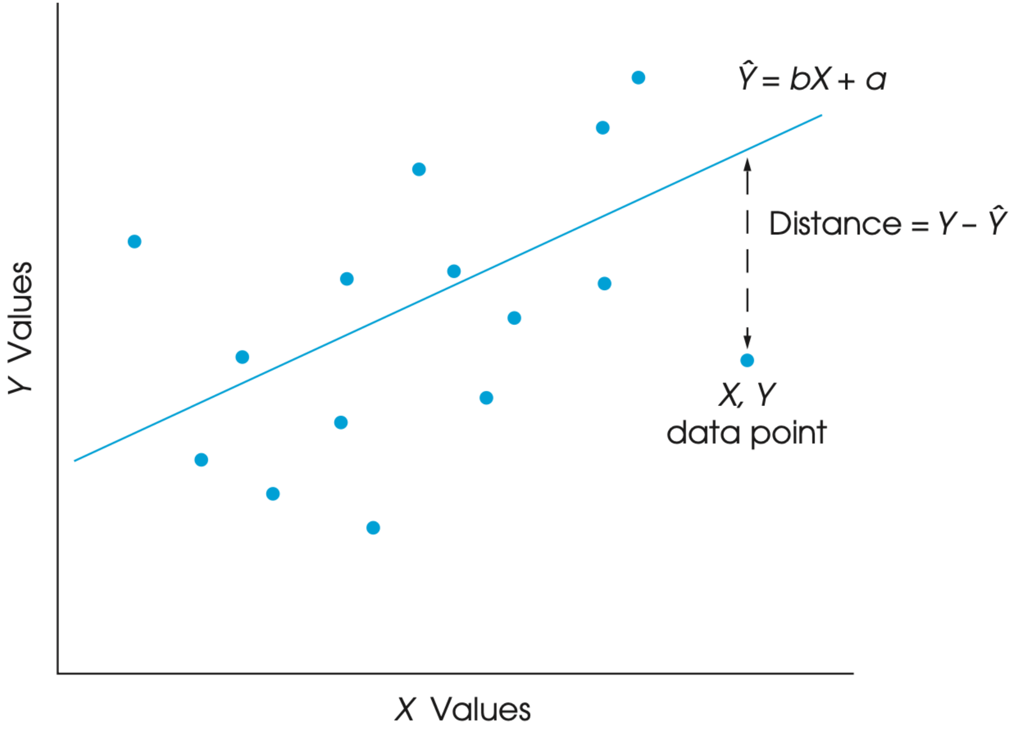


Model Error

- The best-fitting line is the one that has the *smallest* total squared error

$$Total\ Squared\ Error = \Sigma(y - \hat{y})^2$$

- The resulting line of our linear regression is commonly called the **least-squared-error solution**



For quizzes and exams, you do not need to know this equation

Simple Linear Regression

Simple linear regression

$$\hat{y} = bX + a$$

- $\hat{y}$ is the predicted value
- b determines how much the y variable changes when x is increased by 1 point (or unit of measurement)
- a is the value of y when x is zero

Slope

$$b = \frac{\Sigma XY - \frac{\Sigma x \Sigma y}{n}}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}$$

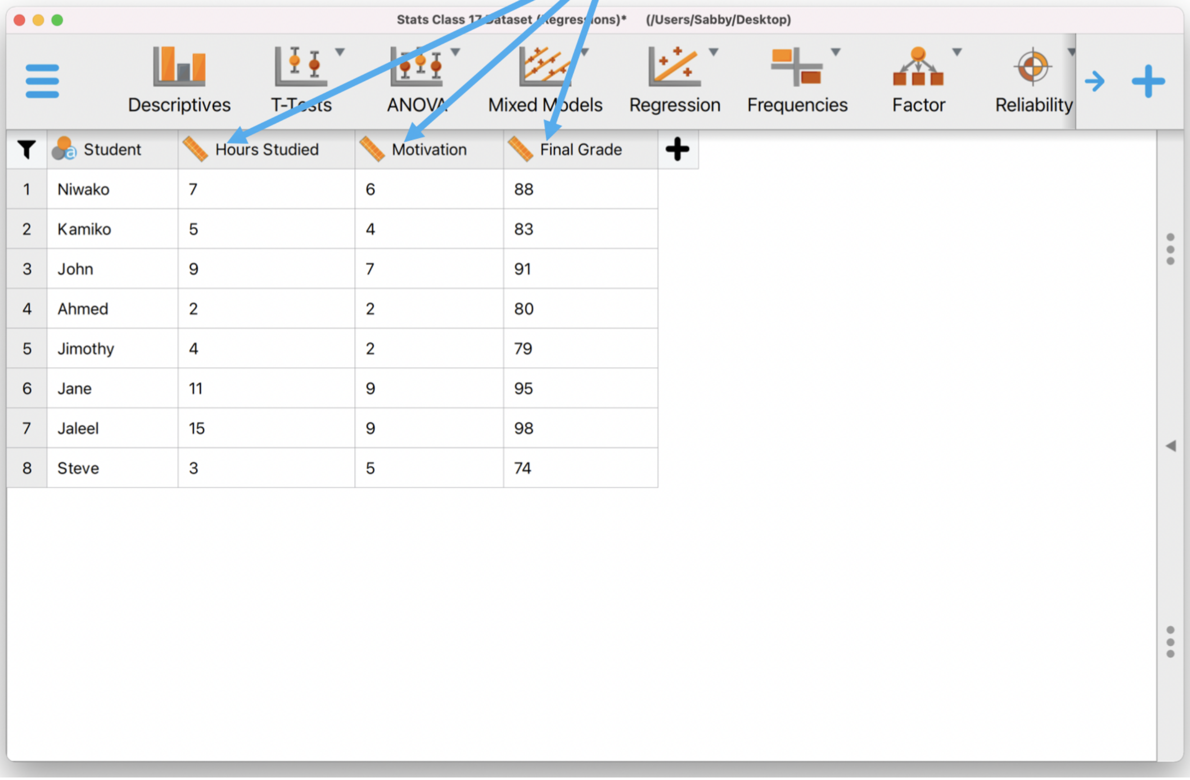
Intercept

$$a = \frac{\Sigma Y - b\Sigma X}{n}$$

JASP: Linear Regression

JASP: Linear Regression

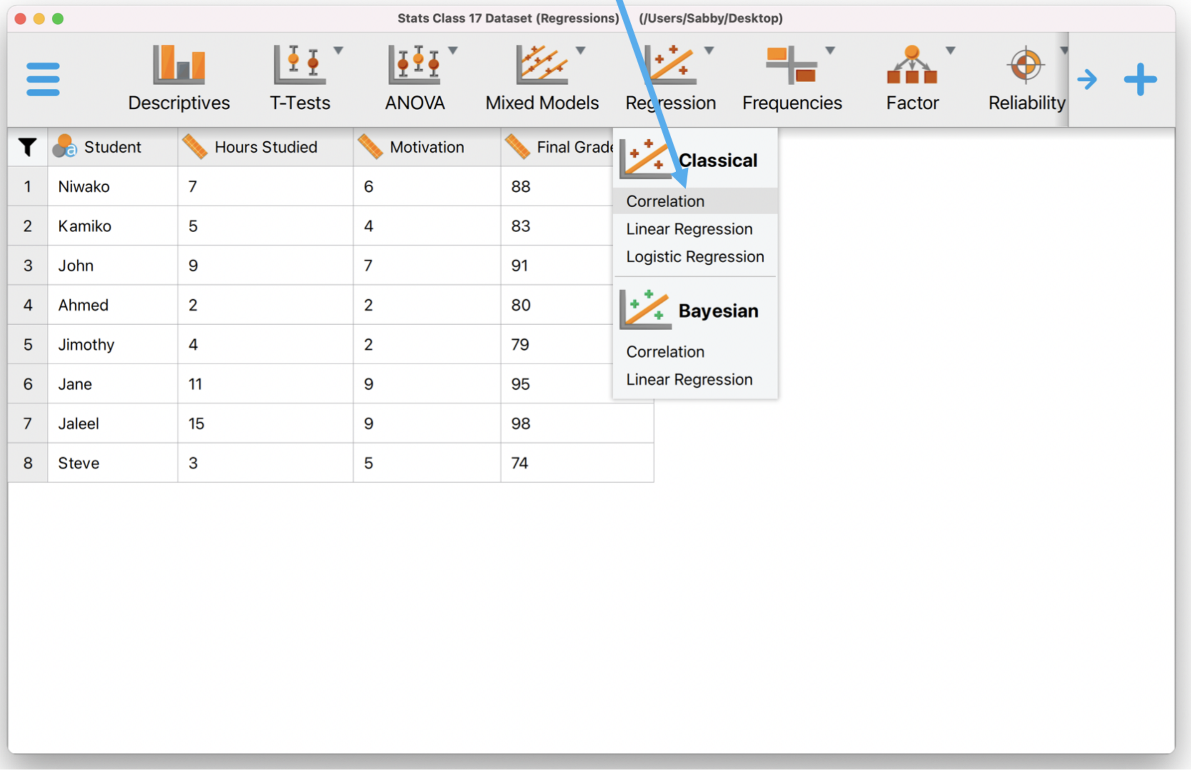
We want to know whether the variables Hours Studied and Motivation play a role in determining students' Final Grade.



	Student	Hours Studied	Motivation	Final Grade
1	Niwako	7	6	88
2	Kamiko	5	4	83
3	John	9	7	91
4	Ahmed	2	2	80
5	Jimothy	4	2	79
6	Jane	11	9	95
7	Jaleel	15	9	98
8	Steve	3	5	74

JASP: Linear Regression

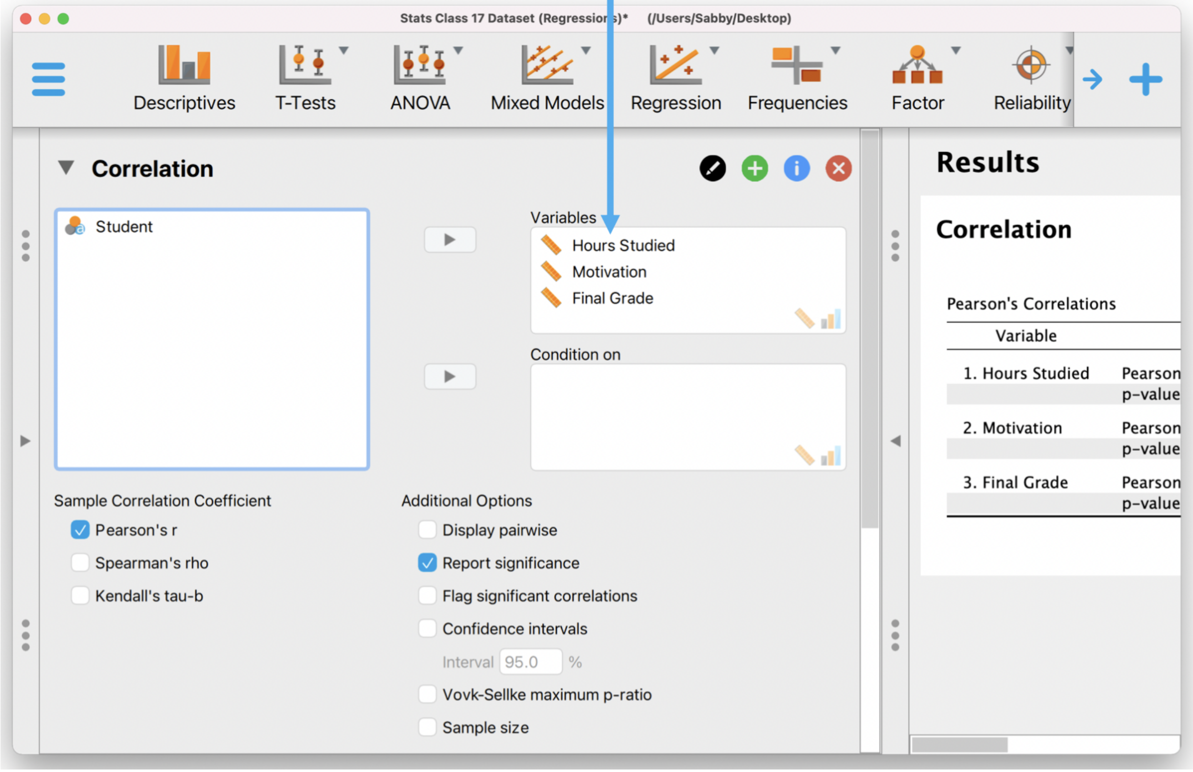
A great place to start is by looking at the correlations between these variables, so let's click on "Correlation" under "Regression."



2

JASP: Linear Regression

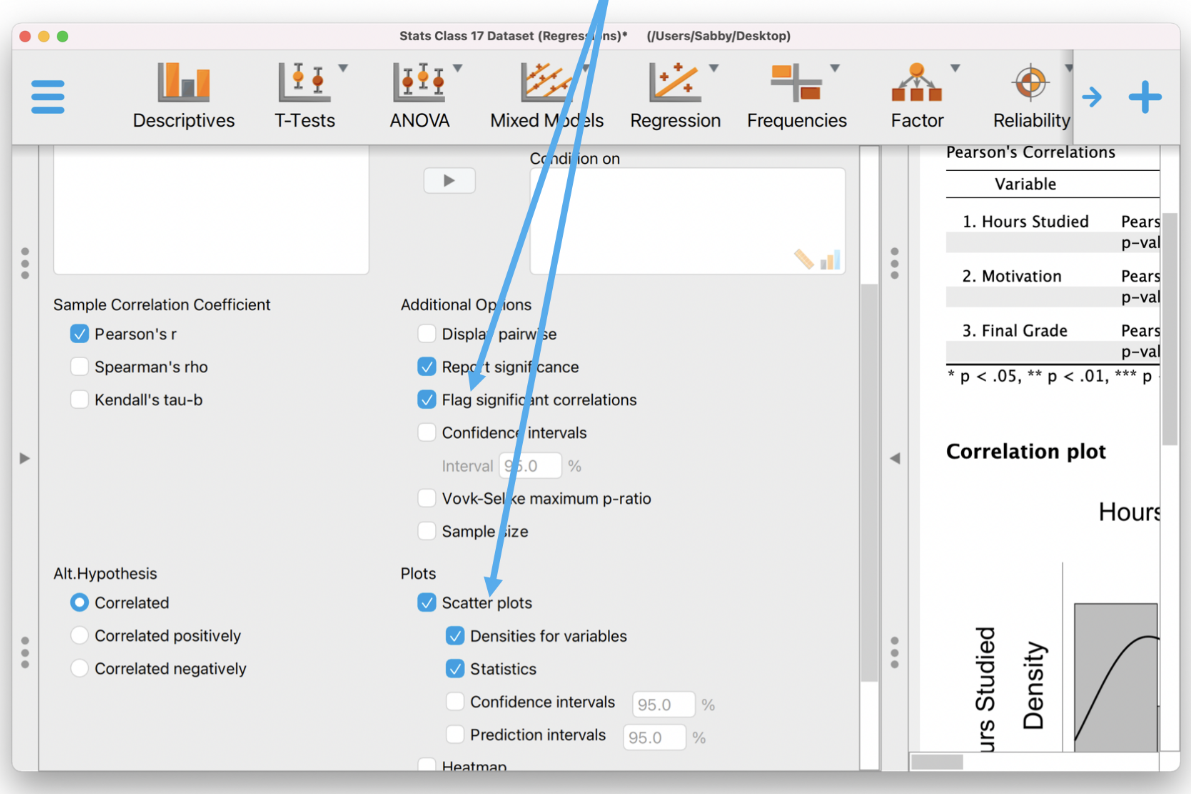
Place all three of our interval-ratio variables in the "Variables" box to begin constructing the correlation table.



3

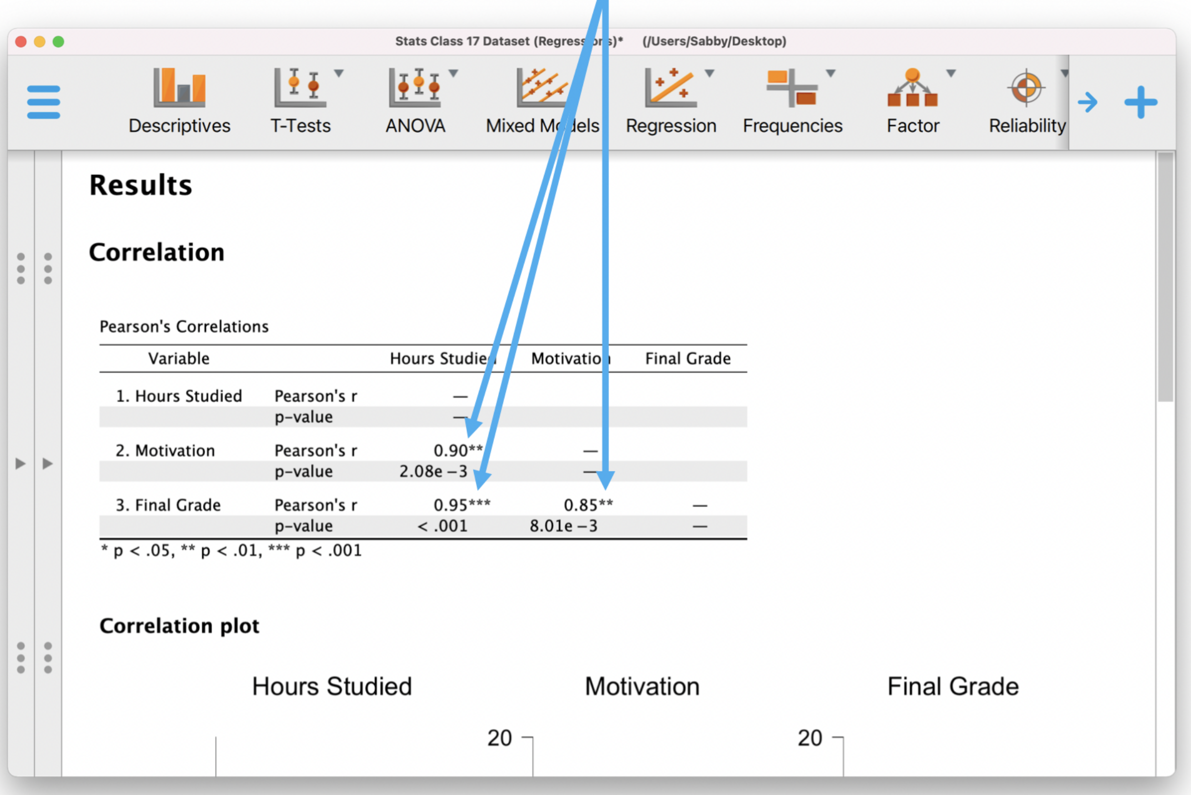
JASP: Linear Regression

Select “Flag significant correlations,” “Scatter plots,” “Densities for variables,” and “Statistics” as additional options.



JASP: Linear Regression

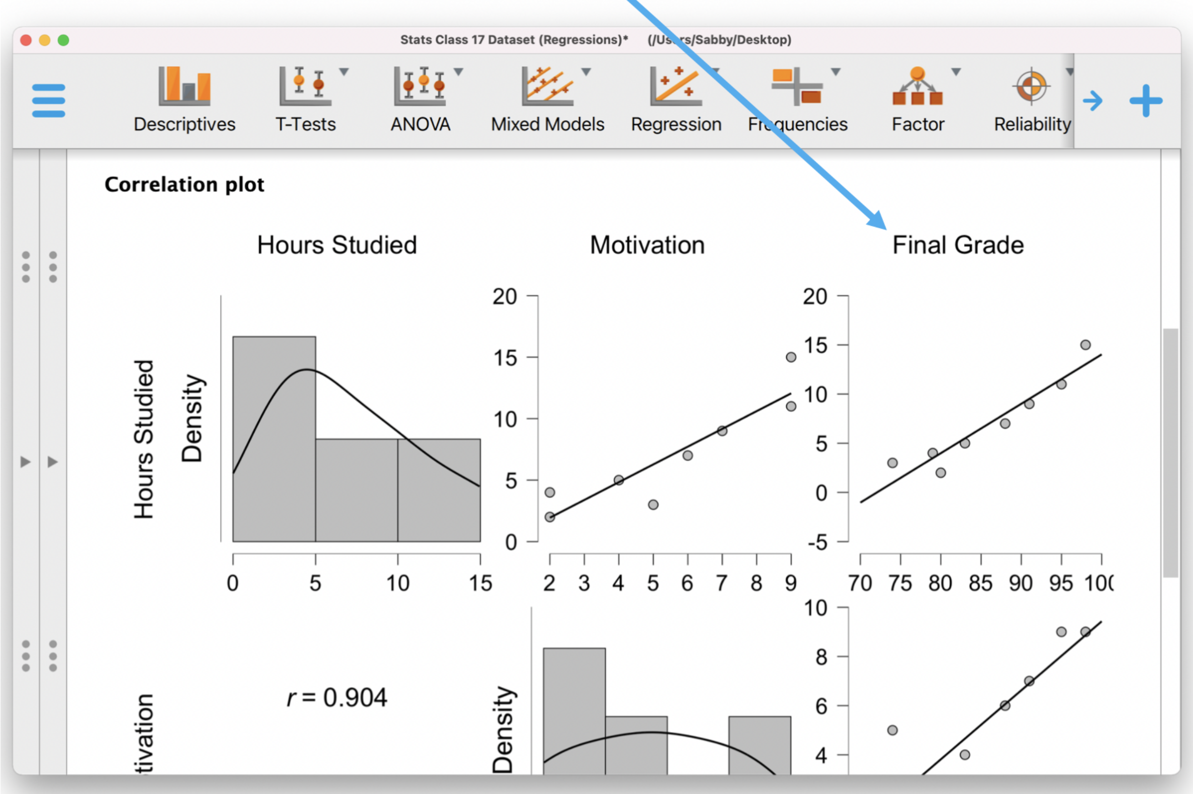
All the pairwise correlations came out significant (i.e., each *p* is less than .05), with each being a positive relationship.



5

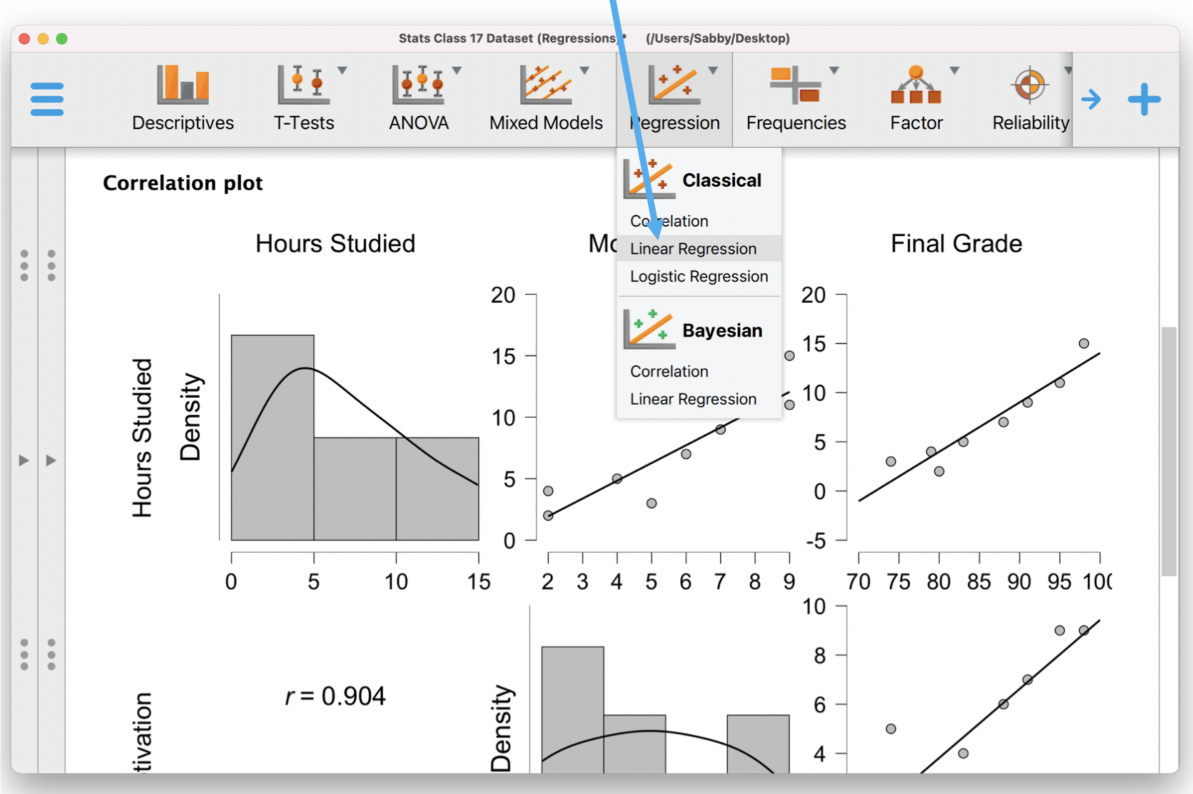
JASP: Linear Regression

We can see these positive correlations visually by looking at the scatterplot matrix. Notice, e.g., Hours Studied vs. Final Grade.



JASP: Linear Regression

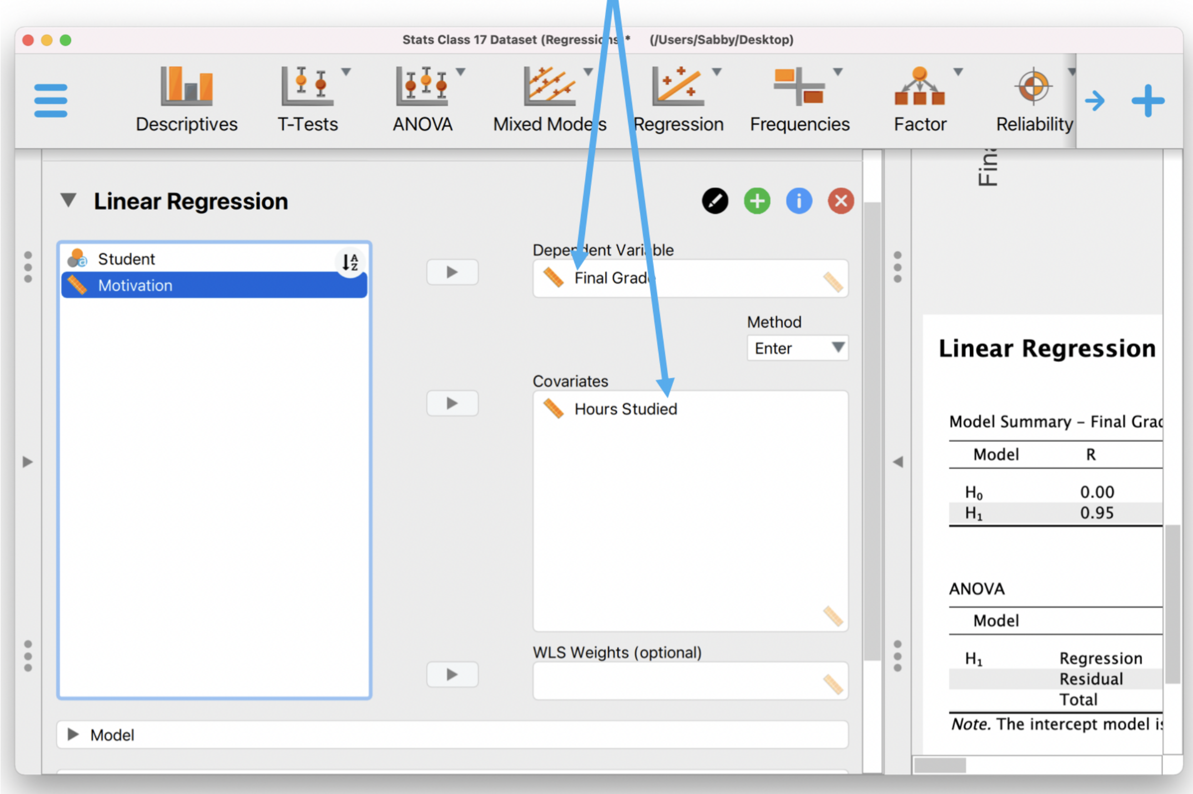
Okay, let's start with a simple linear regression of Hours Studied onto Final Grade by selecting the "Linear Regression" menu.



7

JASP: Linear Regression

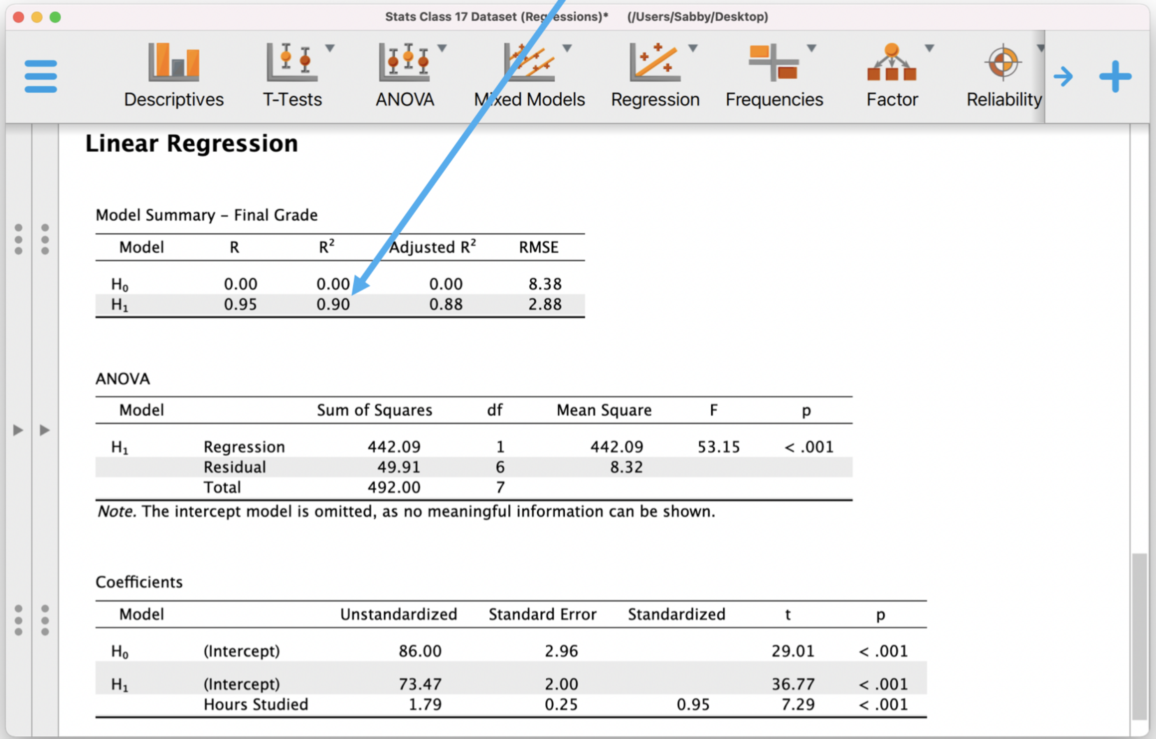
Put Final Grade in the "Dependent Variable" box and Hours Studied in the "Covariates" box. JASP runs the regression!



8

JASP: Linear Regression

R^2 is the coefficient of determination that tells you what percentage of the variance in the DV your model is capturing. 90% – very good!



Linear Regression

Model Summary – Final Grade

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.00	0.00	0.00	8.38
H_1	0.95	0.90	0.88	2.88

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H_1	Regression	442.09	1	442.09	53.15	< .001
	Residual	49.91	6	8.32		
	Total	492.00	7			

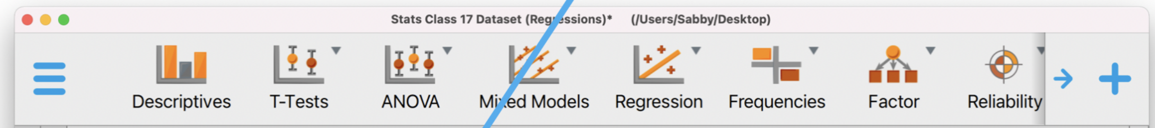
Note. The intercept model is omitted, as no meaningful information can be shown.

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	86.00	2.96		29.01	< .001
H_1	(Intercept)	73.47	2.00		36.77	< .001
	Hours Studied	1.79	0.25	0.95	7.29	< .001

JASP: Linear Regression

You'll want to ignore the ANOVA section for our purposes and focus on the "Coefficients" output, which gives us the values we need.



Linear Regression

Model Summary – Final Grade

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.00	0.00	0.00	8.38
H_1	0.95	0.90	0.88	2.88

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H_1	Regression	442.09	1	442.09	53.15	< .001
	Residual	49.91	6	8.32		
	Total	492.00	7			

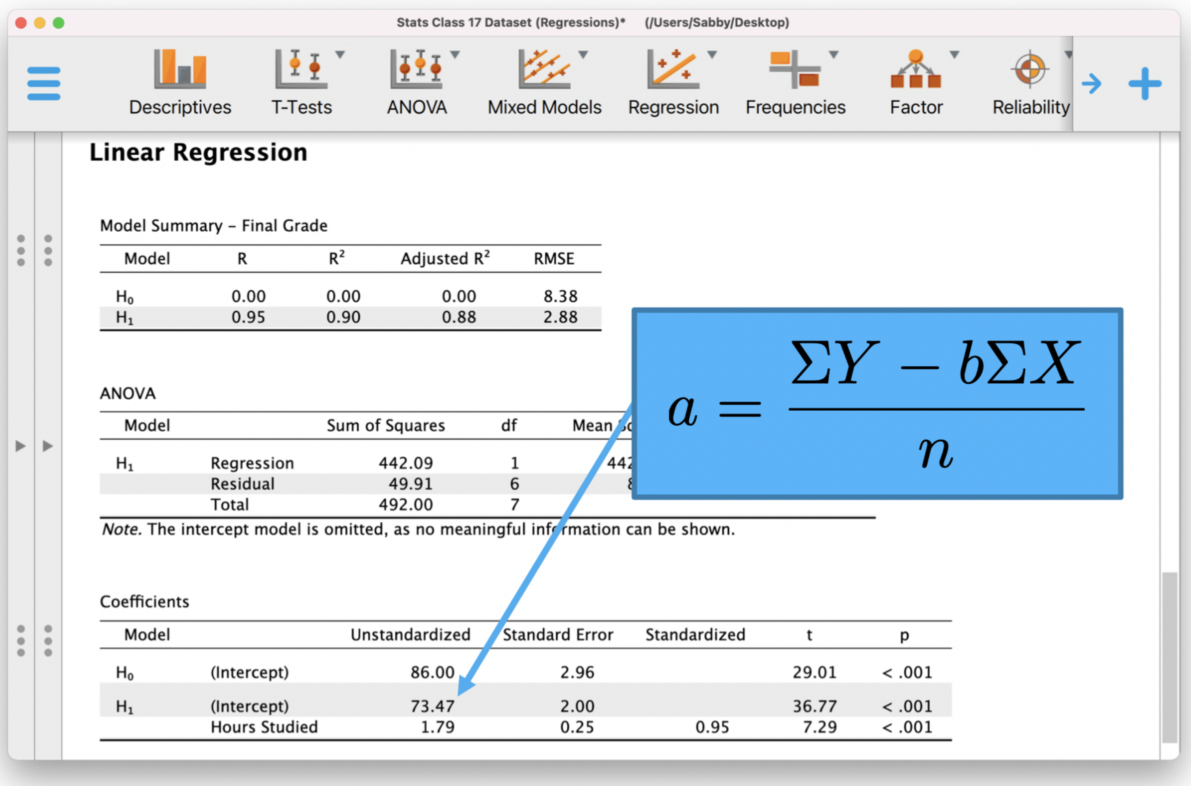
Note. The intercept model is omitted, as no meaningful information can be shown.

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	86.00	2.96		29.01	< .001
H_1	(Intercept)	73.47	2.00		36.77	< .001
	Hours Studied	1.79	0.25	0.95	7.29	< .001

JASP: Linear Regression

The unstandardized (Intercept) term represents the “a” term in our formula for the regression line of best fit.



Linear Regression

Model Summary – Final Grade

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.00	0.00	0.00	8.38
H_1	0.95	0.90	0.88	2.88

ANOVA

Model		Sum of Squares	df	Mean Sq
H_1	Regression	442.09	1	442
	Residual	49.91	6	8
	Total	492.00	7	

Note. The intercept model is omitted, as no meaningful information can be shown.

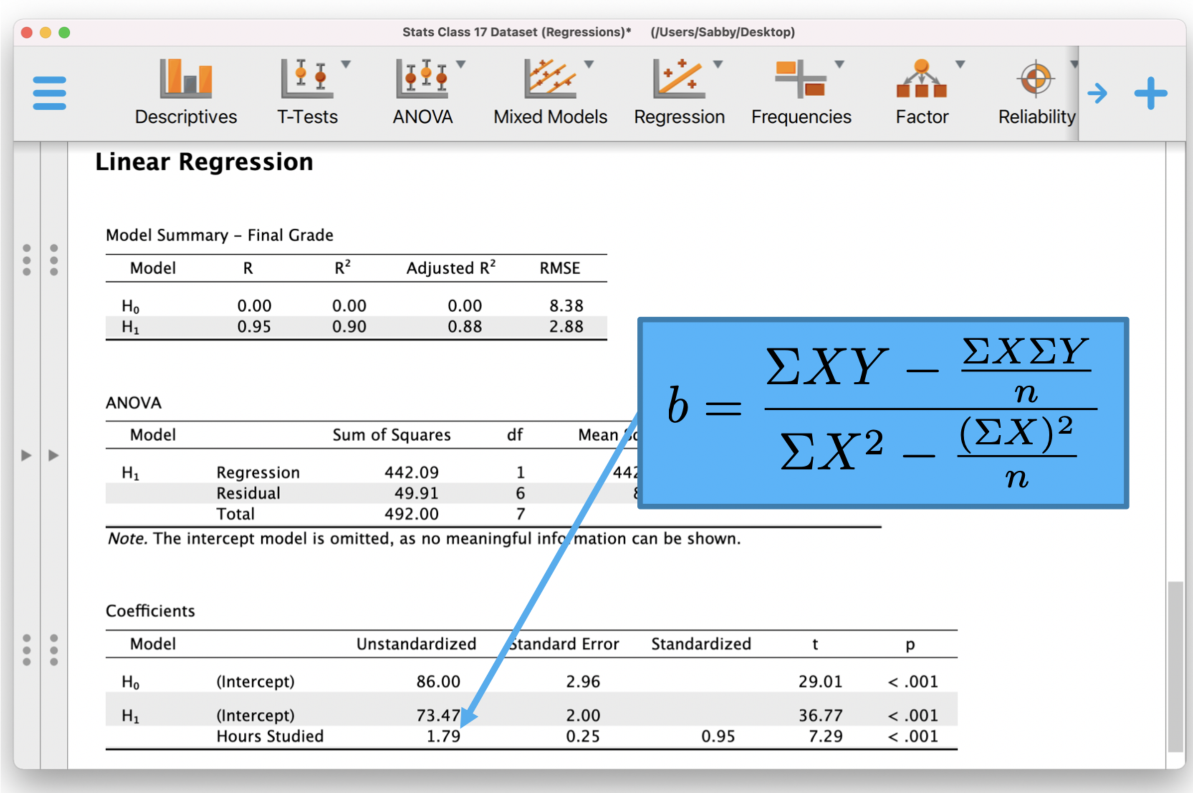
Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	86.00	2.96		29.01	< .001
H_1	(Intercept)	73.47	2.00		36.77	< .001
	Hours Studied	1.79	0.25	0.95	7.29	< .001

$$a = \frac{\Sigma Y - b\Sigma X}{n}$$

JASP: Linear Regression

The unstandardized Hours Studied coefficient represents “b,” which in this case is a significant effect (the slope’s p < .001).



Linear Regression

Model Summary – Final Grade

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.00	0.00	0.00	8.38
H_1	0.95	0.90	0.88	2.88

ANOVA

Model		Sum of Squares	df	Mean S
H_1	Regression	442.09	1	442
	Residual	49.91	6	8
	Total	492.00	7	

Note. The intercept model is omitted, as no meaningful information can be shown.

Coefficients

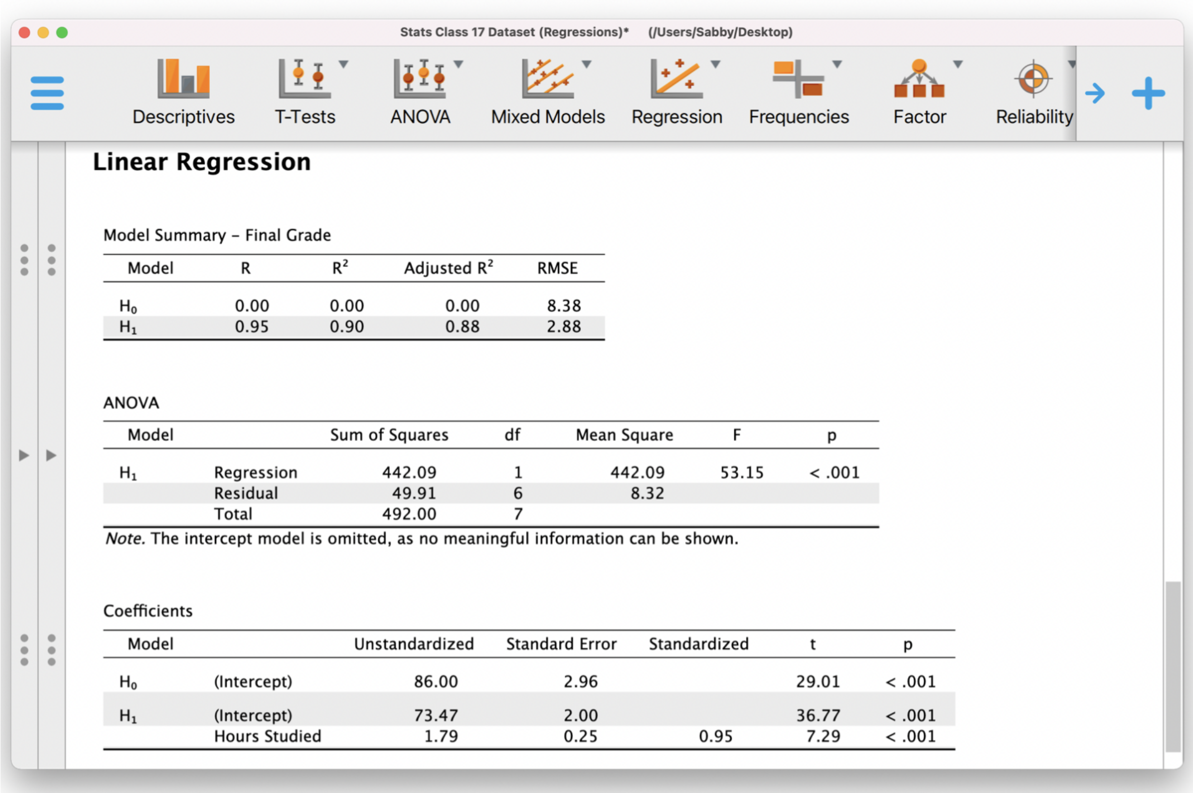
Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	86.00	2.96		29.01	< .001
H_1	(Intercept)	73.47	2.00		36.77	< .001
	Hours Studied	1.79	0.25	0.95	7.29	< .001

$$b = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{n}}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}$$

12

JASP: Linear Regression

So, the regression line of best fit for this particular regression model is Final Grade = 1.79*Hours Studied + 73.47.



Linear Regression

Model Summary – Final Grade

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.00	0.00	0.00	8.38
H_1	0.95	0.90	0.88	2.88

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H_1	Regression	442.09	1	442.09	53.15	< .001
	Residual	49.91	6	8.32		
	Total	492.00	7			

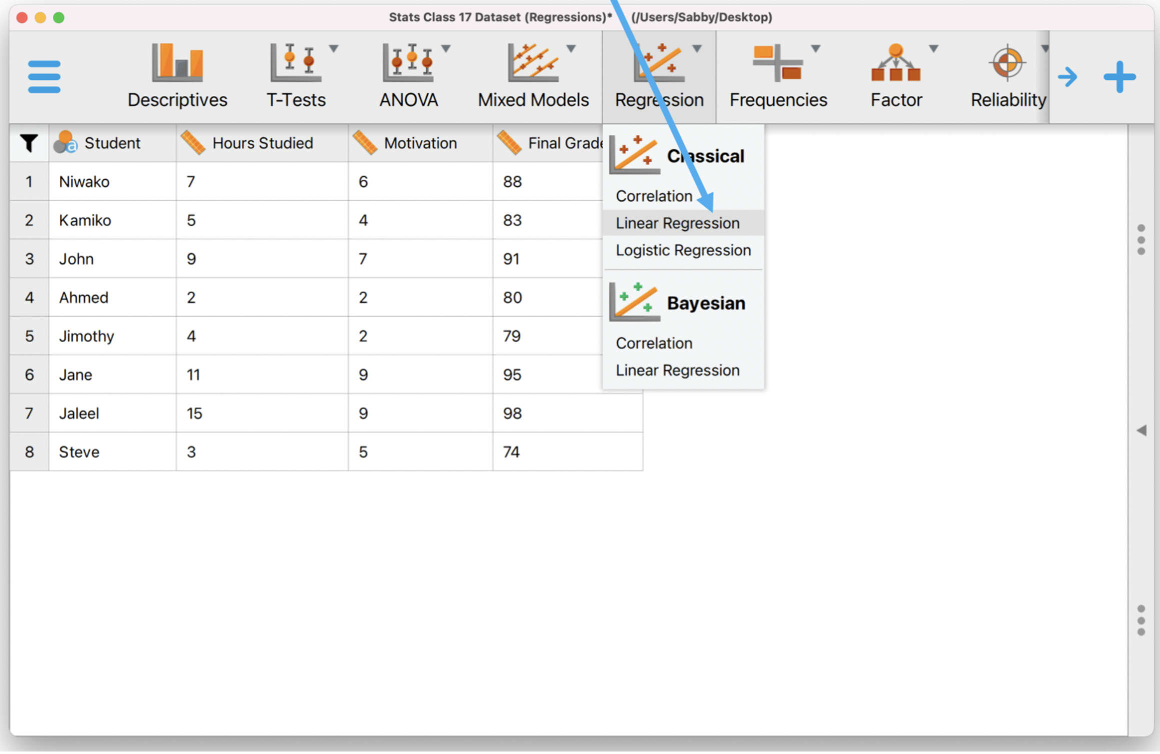
Note. The intercept model is omitted, as no meaningful information can be shown.

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	86.00	2.96		29.01	< .001
H_1	(Intercept)	73.47	2.00		36.77	< .001
	Hours Studied	1.79	0.25	0.95	7.29	< .001

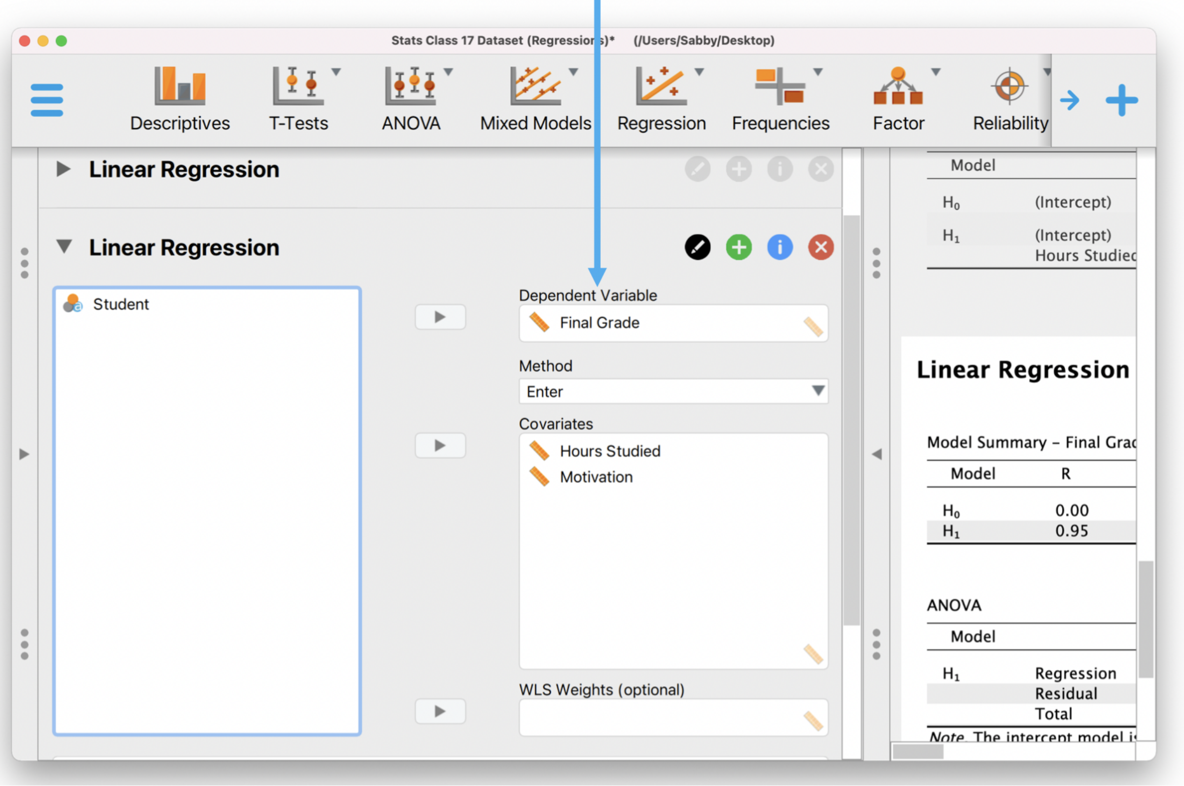
JASP: Linear Regression

Now let's do a multiple regression to assess the relative strength at which *both* of our variables predict Final Grade.



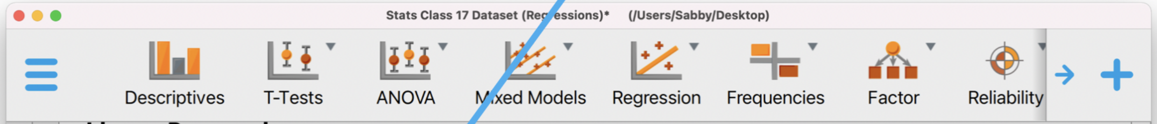
JASP: Linear Regression

Place Final Grade in the “Dependent Variable” box and *both* of our predictors (Hours Studied *and* Motivation) in the “Covariates” box.



JASP: Linear Regression

And that's it! Notice that R^2 hasn't changed, so adding Motivation hasn't improved our predictive power. Studying matters more!



Linear Regression

Model Summary – Final Grade

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.00	0.00	0.00	8.38
H_1	0.95	0.90	0.86	3.15

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H_1	Regression	442.36	2	221.18	22.28	3.23e −3
	Residual	49.64	5	9.93		
	Total	492.00	7			

Note. The intercept model is omitted, as no meaningful information can be shown.

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	86.00	2.96		29.01	< .001
H_1	(Intercept)	73.72	2.67		27.56	< .001
	Hours Studied	1.88	0.63	1.00	3.01	0.03
	Motivation	−0.16	1.00	−0.05	−0.16	0.88

Next time

Lecture

- Analyzing data using JASP I

Reading

- Chapter 16

Quiz 6

- Due Wednesday 3/2/2022 11:59pm MT
- Covers Ch.15-16