جامعـة نيويورك أبـوظـبي

NYU ABU DHABI

PSYCH-UH 1004Q: Statistics for Psychology

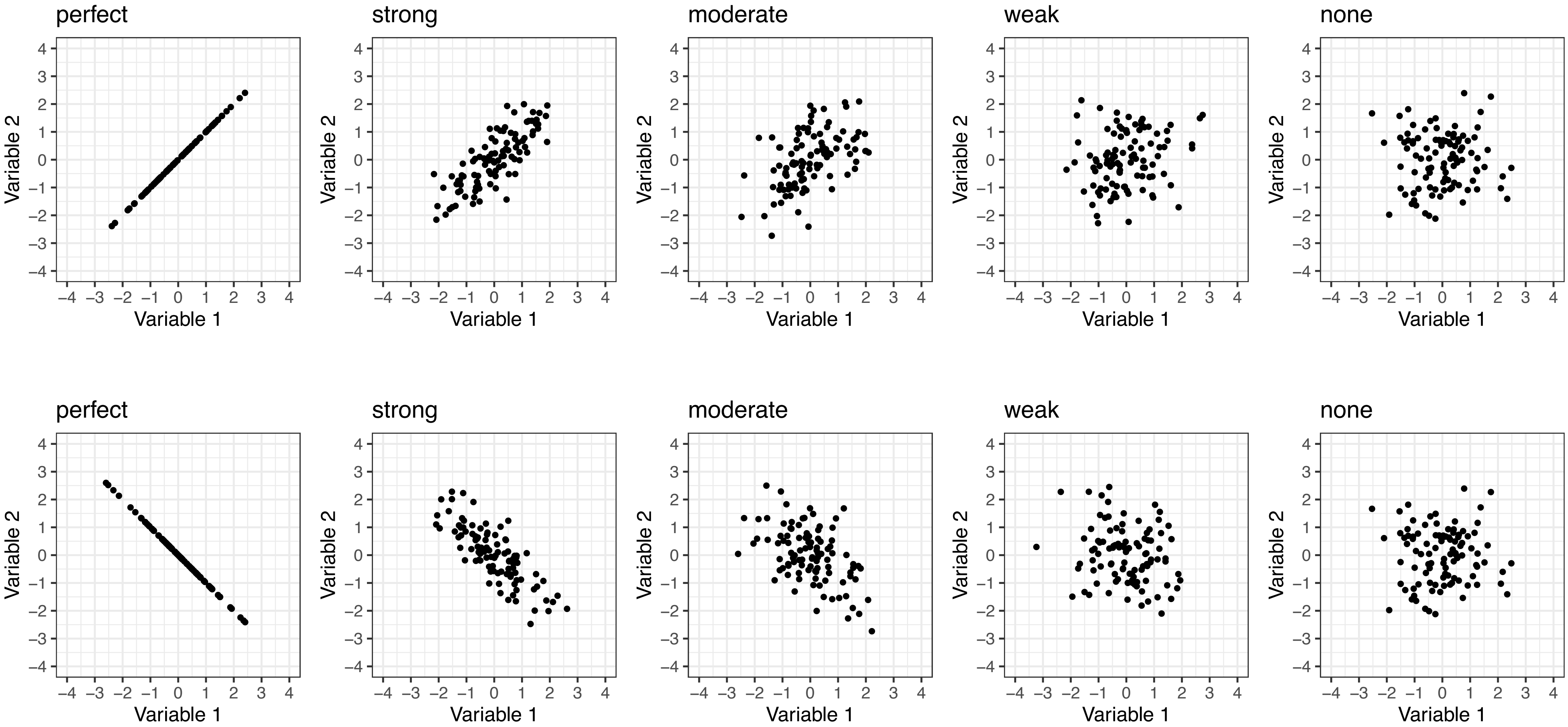
Class 16: Linear regression

Prof. Jon Sprouse
Psychology

It is all about mathematical modeling

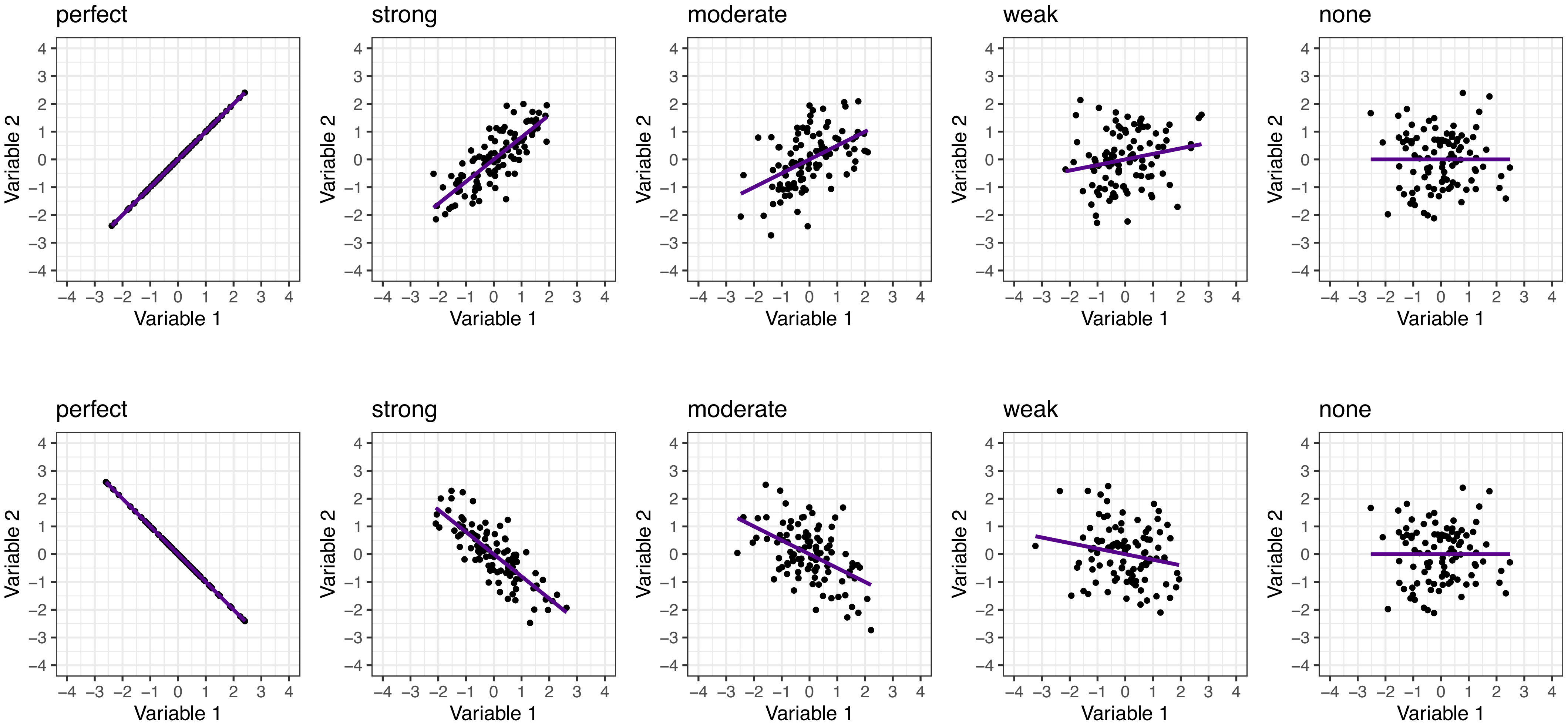
Correlation asks if there is a relationship

What we saw last time was (linear) correlation. Correlation asks: Is there a relationship between two variables? And, if so: How strong is it?



Regression actually models the relationship

Today we will talk about (linear) regression. Regression seeks to specify exactly what the relationship is between two variables. It does this by creating a mathematical model of the relationship that can be used to predict the value of one variable from the value of the other. The model is the line in these plots!



The goal of modeling in general

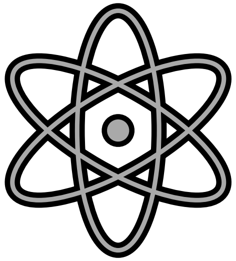
Modeling in general is a way of taking something that we can’t work with directly, and turning it into an object that we can work with:

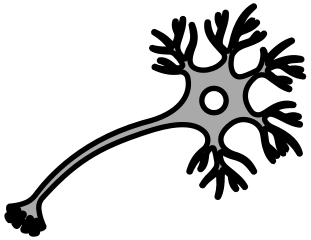
Perhaps students don’t live near a museum with an actual T-Rex skeleton. A model can help them learn about its biology in a classroom or at home.

Children typically can’t design or drive real trains. But model train sets can help them explore how to build track systems and control a train.

Scientific theories and mathematical models

Scientific theories are formulated in terms of real-world objects, like atoms or neurons:

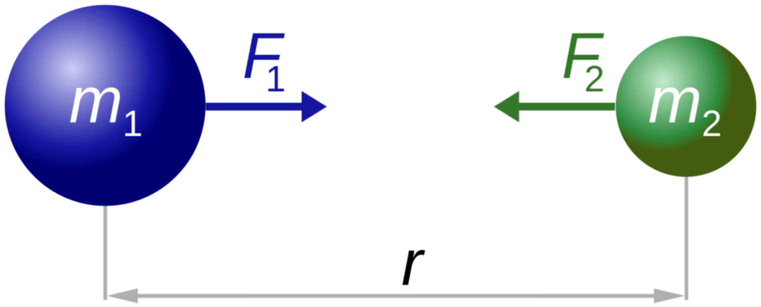




These are not things that we can work with directly when we want to evaluate a theory.

Mathematical modeling is a way to convert a scientific theory into something that we can work with directly. This allows us to do the things we want to do in science, like evaluate the theory against data points we've observed and make predictions about things we have not observed yet.

You have seen mathematical modeling in science many times in school. Every physics lesson you did that involved an equation, like this one for the gravitational force between two objects, was mathematical modeling. We can't work with gravity directly, but we can work with equations to evaluate the theory and make new predictions!



$$F_1 = F_2 = G\frac{m_1 \times m_2}{r^2}$$

A quick tutorial on lines

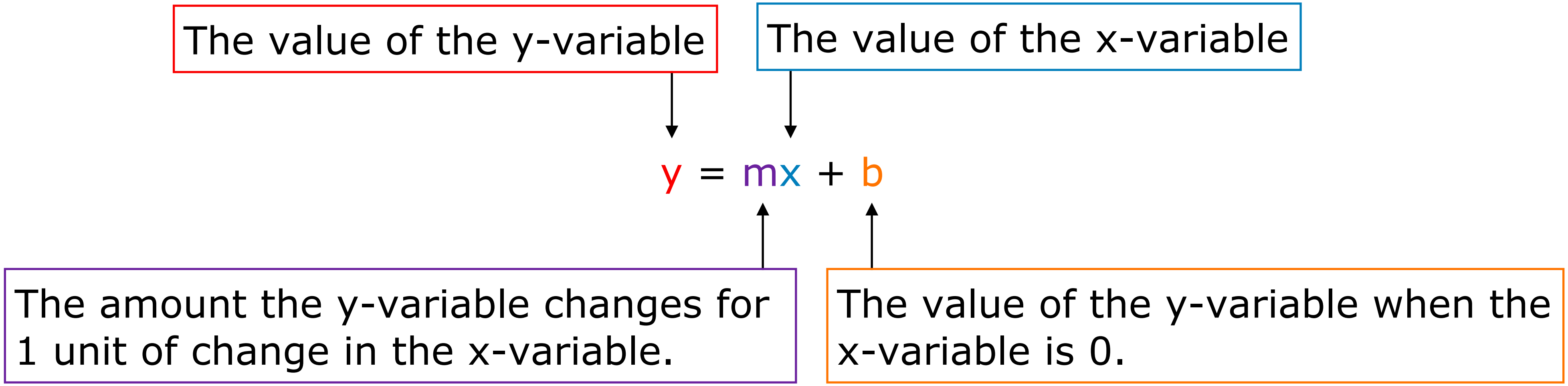
Linear models

One of the simplest types of mathematical models that we can construct are **linear models** - they posit a linear relationship between the variables.

The math for a linear model is very simple, and you already know it:

$$y = mx + b$$

Mathematical modeling is about finding an equation that models your scientific theory. Linear regression is when we use the **equation for a line** as the mathematical model. All models in linear regression will have this form:



The value of the y-variable

The value of the x-variable

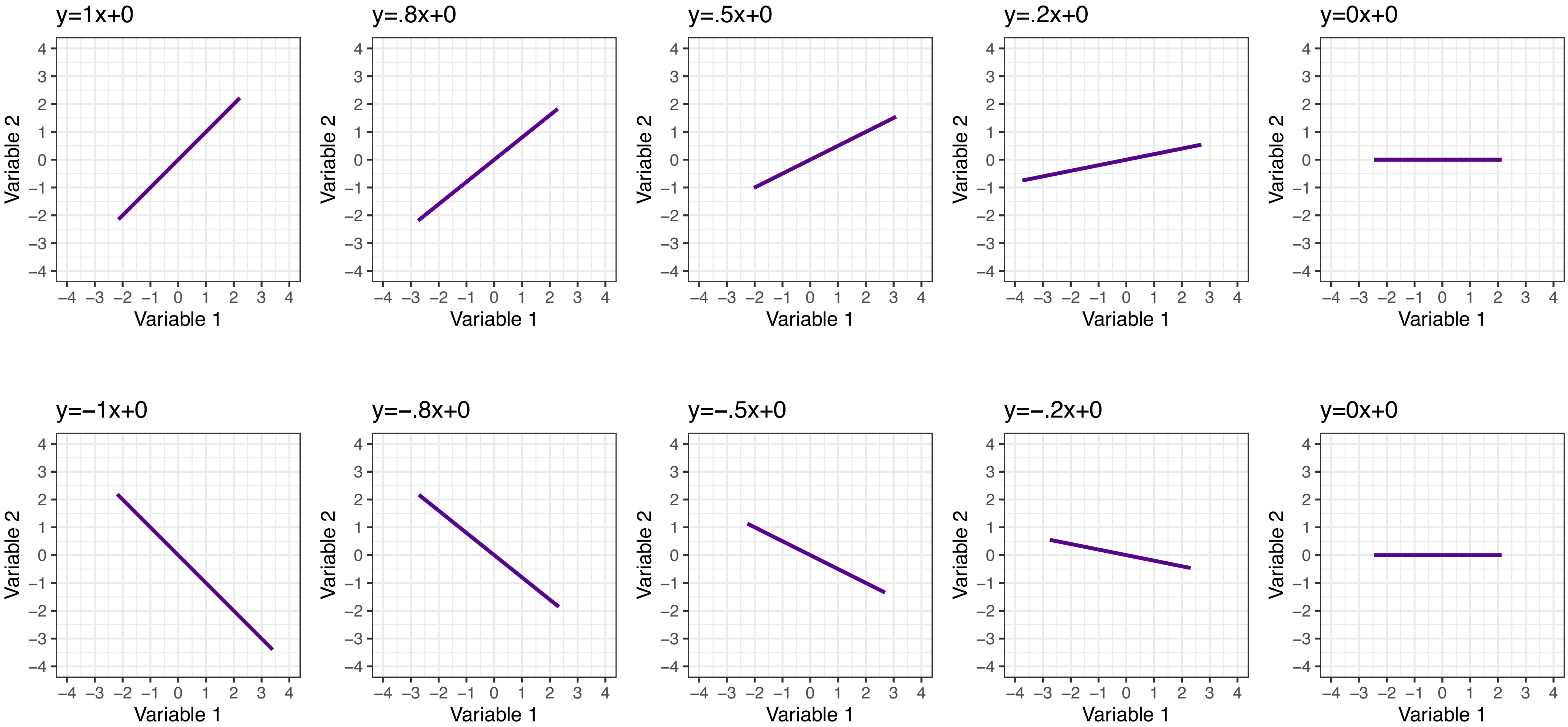
$$y = mx + b$$

The amount the y-variable changes for 1 unit of change in the x-variable.

The value of the y-variable when the x-variable is 0.

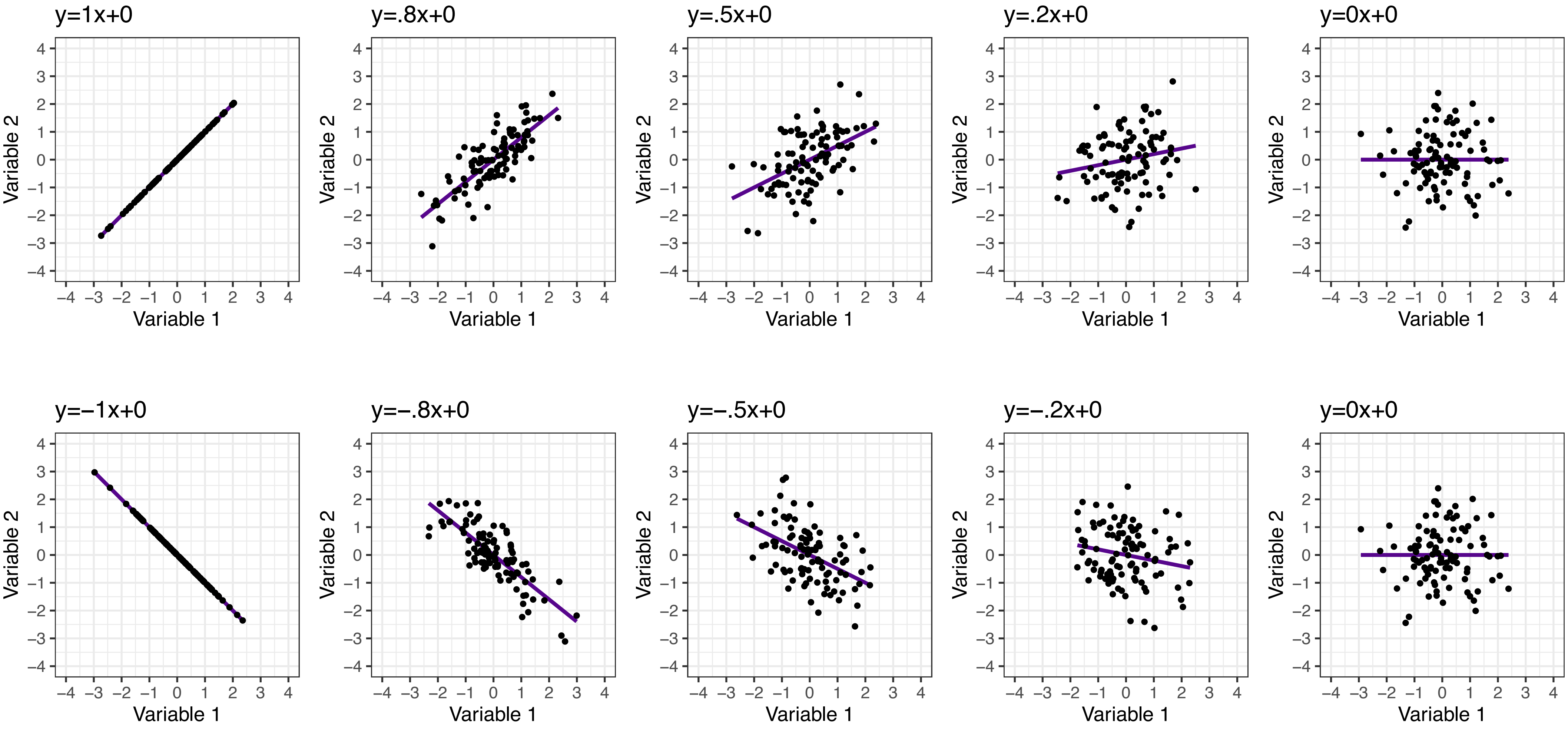
A quick tutorial: y=mx+b

Here is a quick look at how changing the m term, the slope, changes a line. Notice that it represents how much the y-variable will change when there is a 1 unit change in the x-variable. It can be positive (increasing) or negative (decreasing).



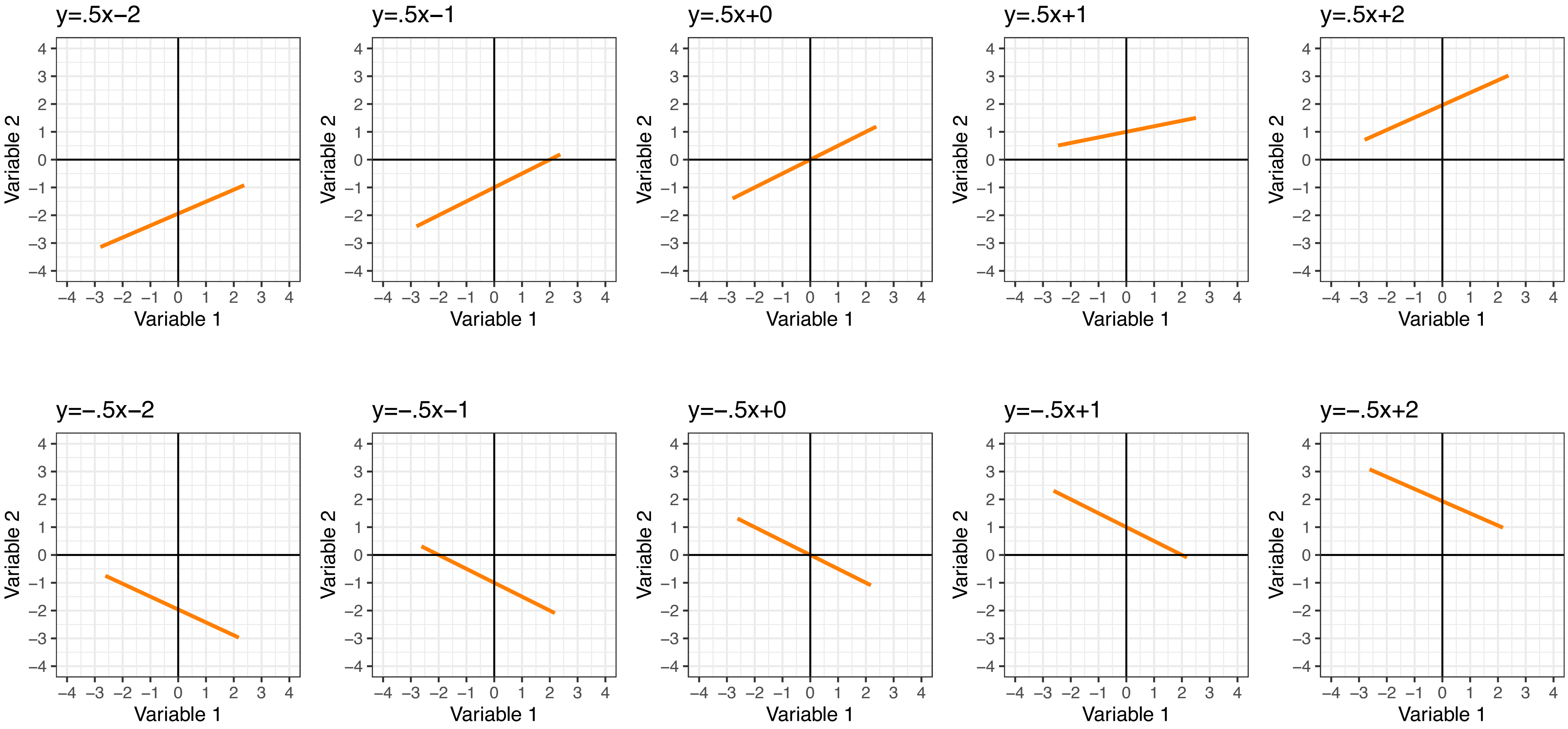
With data points!

This is the same manipulation of the slope, but with the data points there, so you can see that this is about modeling real data!



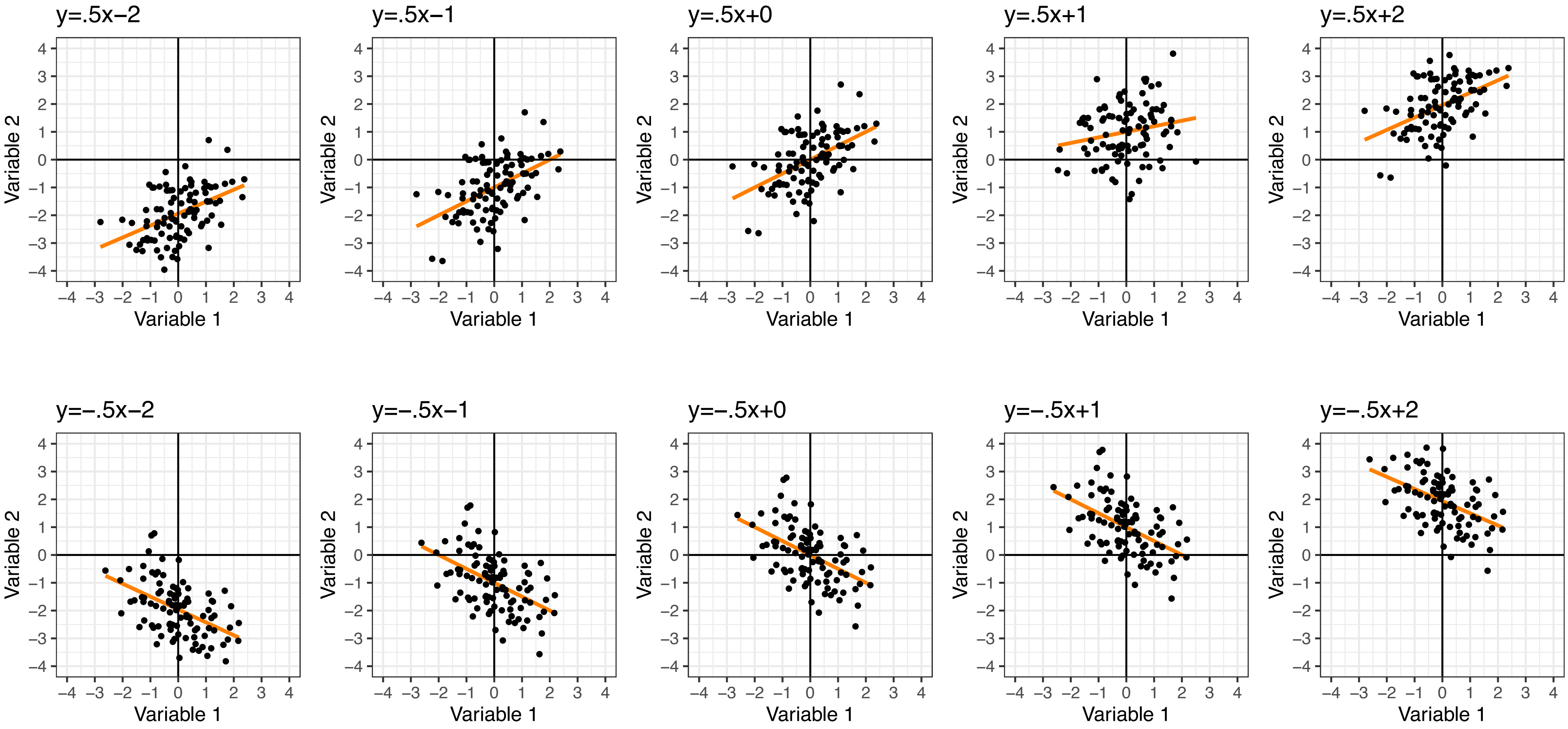
A quick tutorial: y=mx+b

Here is a quick look at how changing the b term, the y-intercept, changes a line. The slope is held constant at either .5 or -.5. Notice that the b term represents the value of the y-variable when the x-variable is 0. We call that the y-intercept because it is where the line crosses the y-axis.



With data points!

This is the same manipulation of the y-intercept, but with the data points there, so you can see that this is about modeling real data!



A quick note on line equation notation

In algebra, we use this: $y = mx + b$

But in statistics, we make some small notation changes: $\hat{y} = b_1 x + b_0$

We use $\hat{y}$ ("y hat") to indicate that we are trying to make predictions. $\hat{y}$ is the predicted value of y, not the actual value of y. So, y is the real value, and $\hat{y}$ is the predicted value.

We use b_1 for m and b_0 for b. Even though these both use the letter b, they are different values! The subscript is important! (We use the same letter because they are both coefficients for the line equation; it saves other letters for us for other things!).

Also, in statistics, we often switch the order around: $\hat{y} = b_0 + b_1 x$

Your book uses a slightly different notation, but it is also just a line: $\hat{y} = b_{yx} x + a_{yx}$

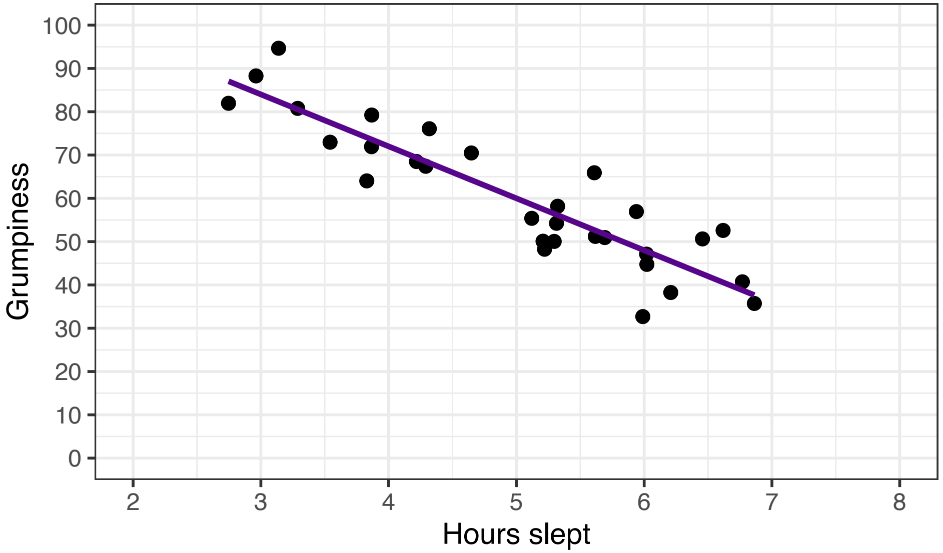
An example to get an intuition for linear regression as modeling

A plausible theory

In this example, we theorize that the number of hours that people sleep will influence how grumpy they are the next day. We collect some data, and plot it:

First, notice that our independent variable is hours of sleep. We always place our IV on the x-axis. It is x by convention.

Second, notice that our dependent variable is grumpiness. We always place our DV on the y-axis. It is y by convention.



Looking at this plot, I am sure you can see that there is a relationship between these two variables. So, stepping back to last lecture, if we looked for a correlation, we would find a strong negative correlation. (r = -.9)

But what we want to do here is actually model the relationship. We want to find an equation (in the form of a line) that will predict a specific quantity of grumpiness based on the number of hours of sleep someone gets.

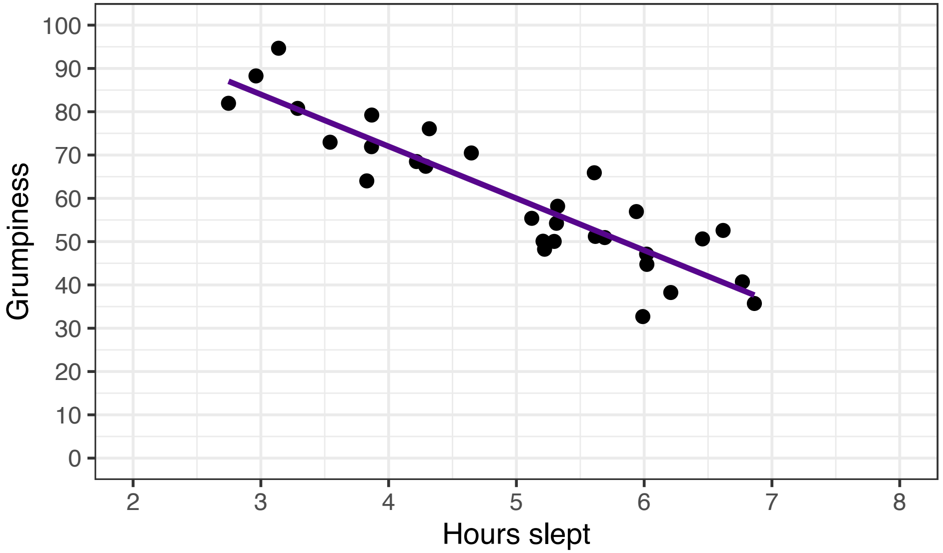
An example to get the intuition

Here is the equation for the **line of best fit** for this data set:

I'll use the statistics variables, but in the familiar order, slope first then intercept:

$$\hat{y} = b_1x + b_0$$

$$\hat{y} = -12x + 120$$



What this equation tells us is that we can predict how grumpy someone is by plugging the number of hours that they slept into x, and calculating $\hat{y}$.

So, if I slept 8 hours last night, this equation predicts that my grumpiness score will be 24 on this scale (which runs from 0 to 100, with 100 the most grumpy).

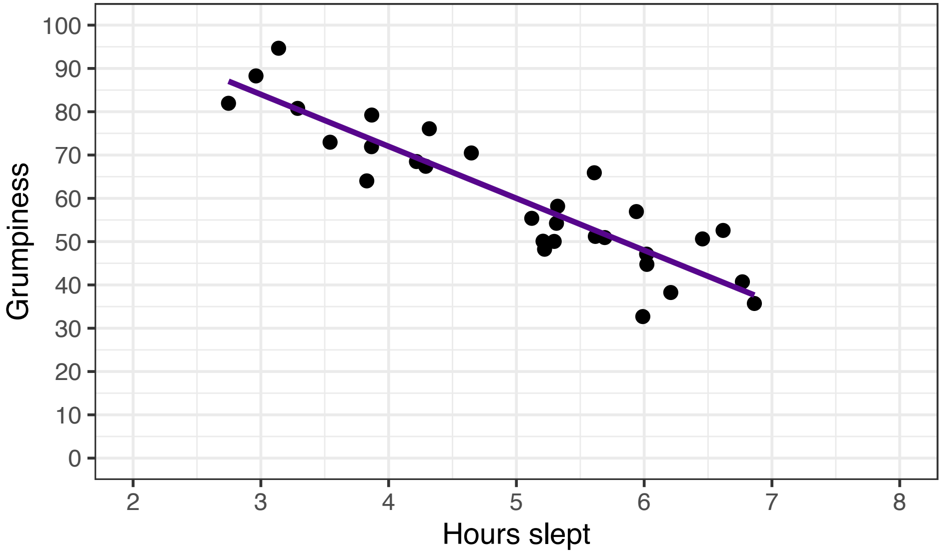
Questions to ask about our model

Obviously, there are bunch of questions we can ask about our model:

1. How did we find the values for the slope (b_1) and the intercept (b_0)?

$$\hat{y} = b_1 x + b_0$$

$$\hat{y} = -12x + 120$$



2. How well does the model fit the actual data? The dots are not perfectly on the line… so it is not perfect… but how good is it?

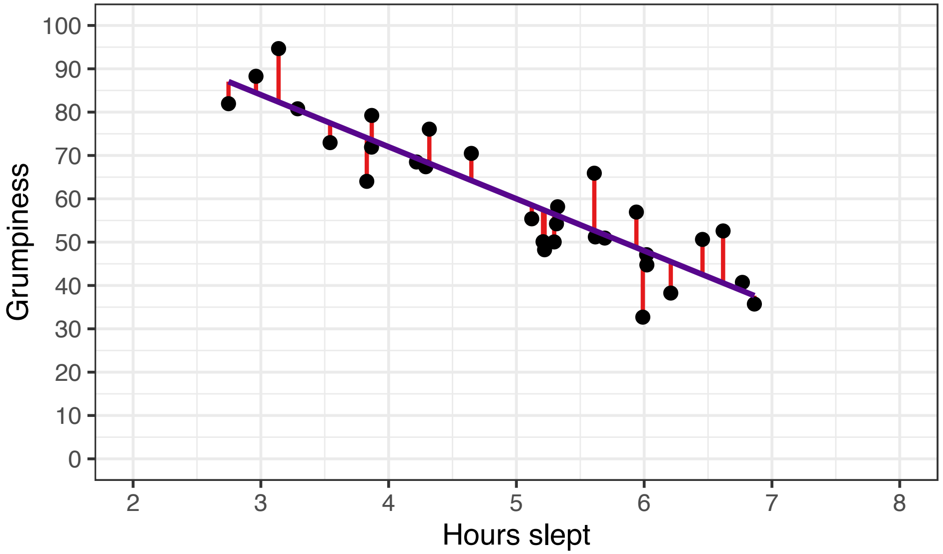
3. Can we perform a hypothesis test on our model against a null hypothesis?

(The book also shows you how to calculate variability, like standard error, and confidence intervals. But we can let R worry about those details.)

The answers to our questions build on one critical idea: **Residuals**

Residuals are easy to grasp - they are the difference between the actual values (y) and the predicted value (ŷ):

residual = y-ŷ



There is a residual for every data point in our data set. I have plotted them here as vertical lines between the points and the predicted value (the purple line of the model!).

It is worth taking a moment to think about what residuals represent in our theory. They are the bit that we **failed** to explain with our theory.

Our theory says that sleep will predict grumpiness. But sometimes people are more or less grumpy than sleep alone would predict. **Something else** is affecting grumpiness (like stress at school). And we don’t have a theory for that, so it is not in our model. Residuals capture the **failures** of our theory.

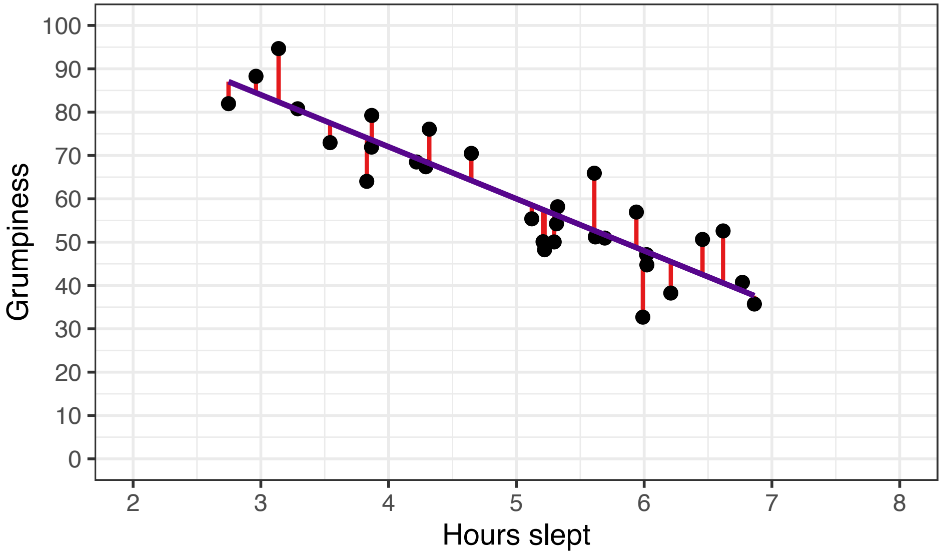
How do we find the line of best fit?

Ordinary Least Squares Regression

As with everything in statistics, the answer to the question of how to find the **line of best fit** for any given data set is calculus.

But, the logic of that process is simple. We want the model that fits **best**, so we have to define what best means.

And we already have a way to define that - the best model will have the smallest residuals. In other words, it will explain the most of all models, and fail to explain the least of all models.



So, all we need is a measure of how big the residuals are. We know how to do this. We know we can't sum them, because some are positive and some are negative. Instead, we square them just like we are used to:

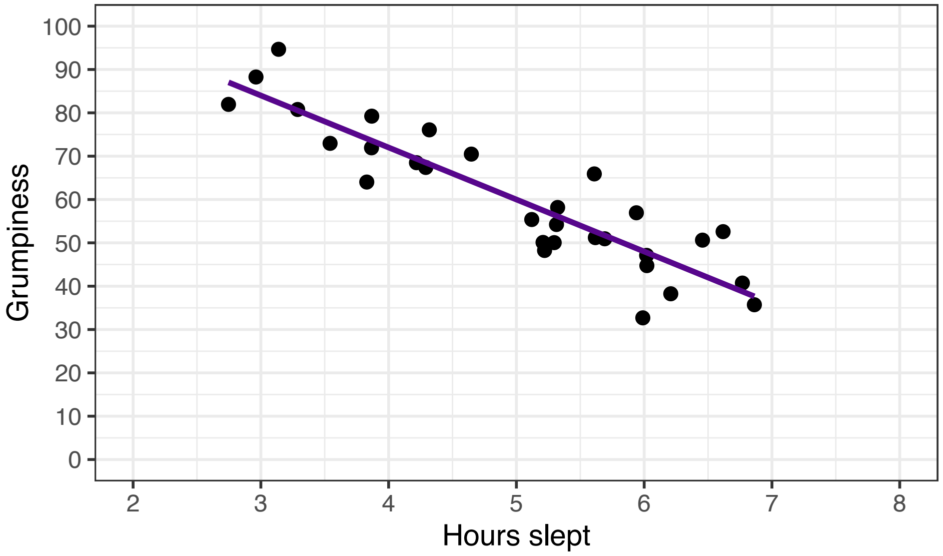
sum of squared residuals $= \Sigma(y-\hat{y})^2$

We then use calculus to find the line that minimizes this number!

Arithmetic shortcuts instead of calculus!

The book provides you with arithmetic equations for finding the line of best fit without using calculus.

For the slope term (b_1), the book gives us two options:



$$b_1 = \frac{cov_{xy}}{s_x^2} \qquad b_1 = \frac{s_y}{s_x} r$$

This is because: $r = \frac{cov_{xy}}{s_x s_y}$

For the intercept term (b_0):

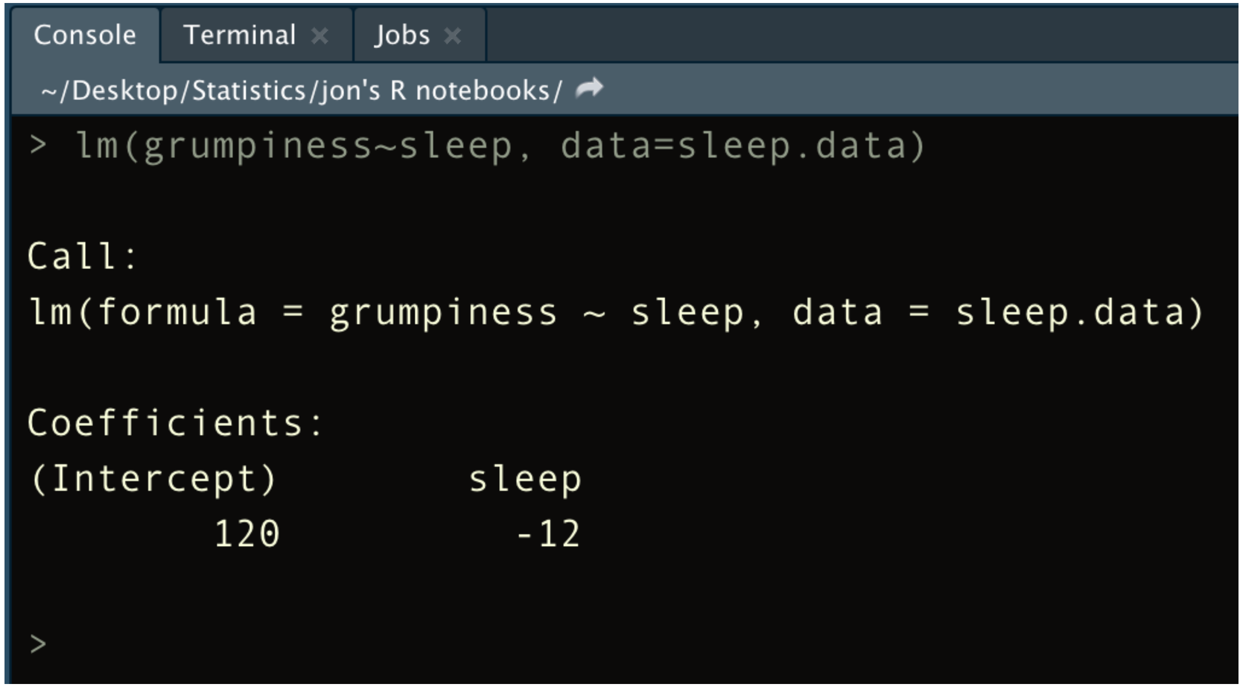
$b_0 = \bar{y} - b_1\bar{x}$

These are fun to know. But you will **never** calculate them by hand as a scientist. Never. So, please note them, squirrel away in your memory that these exist. And focus instead on letting R calculate the values.

Using the R function lm()

The R function for linear regression is lm(). It stands for linear model.

The function asks you to specify the model you want using a formula. Here it is grumpiness ~ sleep.



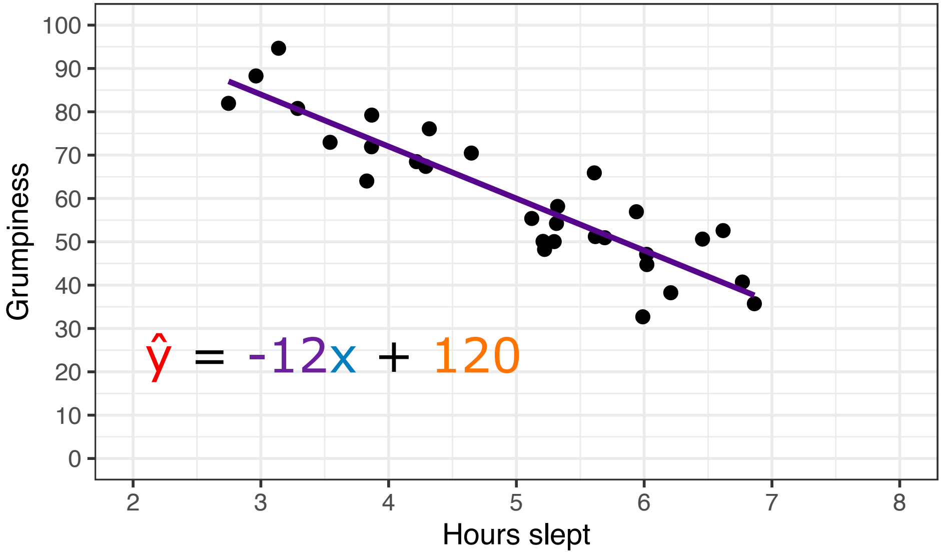
```
> lm(grumpiness~sleep, data=sleep.data)

Call:
lm(formula = grumpiness ~ sleep, data = sleep.data)

Coefficients:
(Intercept)        sleep
        120          -12

>
```

The idea is that the tilde (~) is like the equal sign in the line equation:

grumpiness ~ sleep

$$\hat{y} = b_1 x + b_0$$

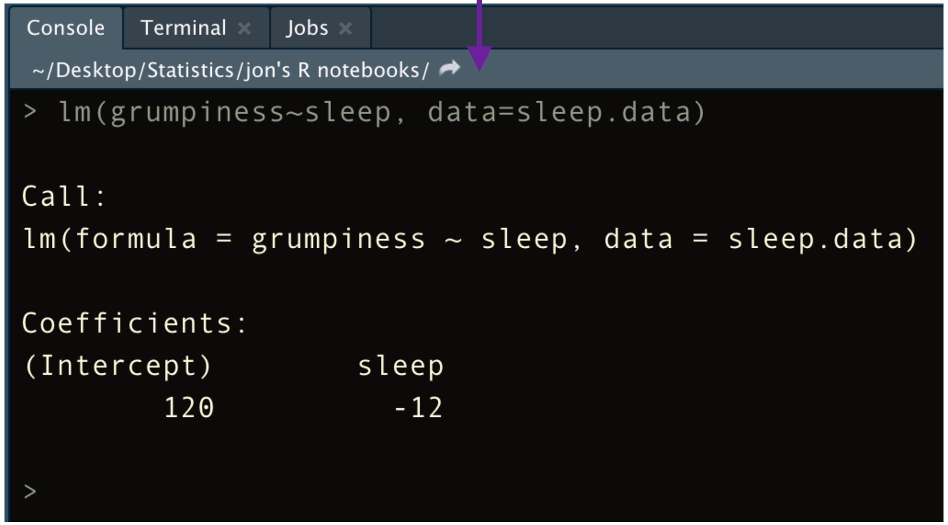
R doesn't need you to tell it to find an intercept. It knows to do that. So you don't type anything for that.



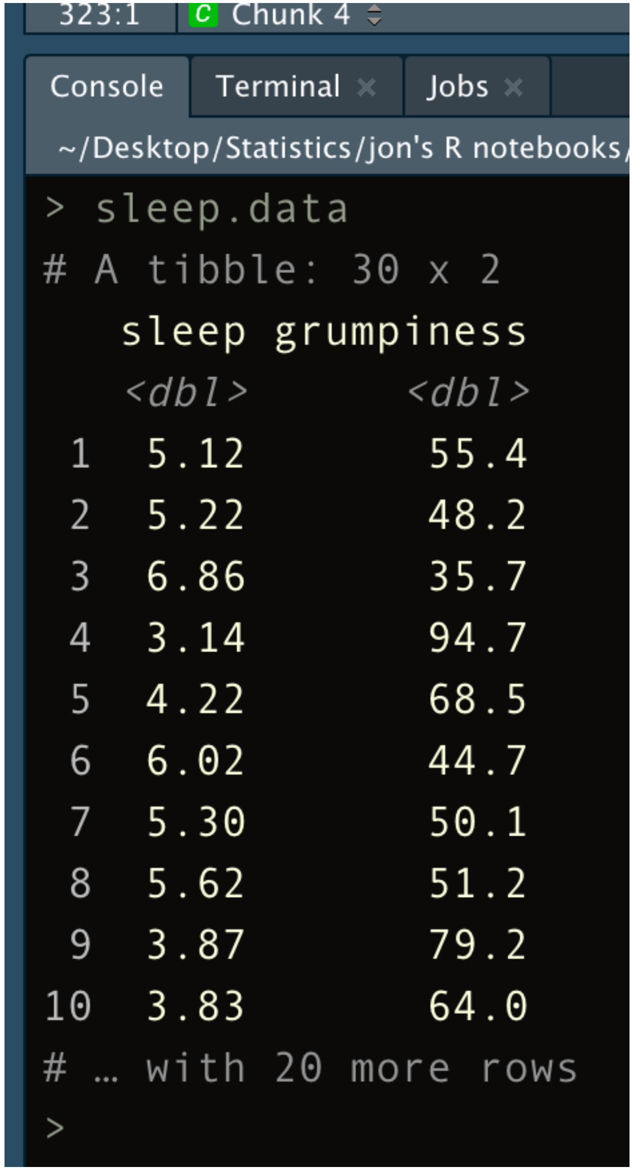
Using the R function lm()

The R function lm() works with data that has a variable (a column) for y and a variable (a column) for x, like this:

And after you specify the formula, you have to tell the function to look for that data set with the data argument.



```
> lm(grumpiness~sleep, data=sleep.data)

Call:
lm(formula = grumpiness ~ sleep, data = sleep.data)

Coefficients:
(Intercept)        sleep
        120          -12

>
```



```
> sleep.data
# A tibble: 30 x 2
   sleep grumpiness
   <dbl>      <dbl>
 1  5.12       55.4
 2  5.22       48.2
 3  6.86       35.7
 4  3.14       94.7
 5  4.22       68.5
 6  6.02       44.7
 7  5.30       50.1
 8  5.62       51.2
 9  3.87       79.2
10  3.83       64.0
# … with 20 more rows
>
```

How well does the model fit the data?

What does it mean to be a good theory?

This is a philosophical question. But we have basically been answering it all semester. A good theory explains the data.

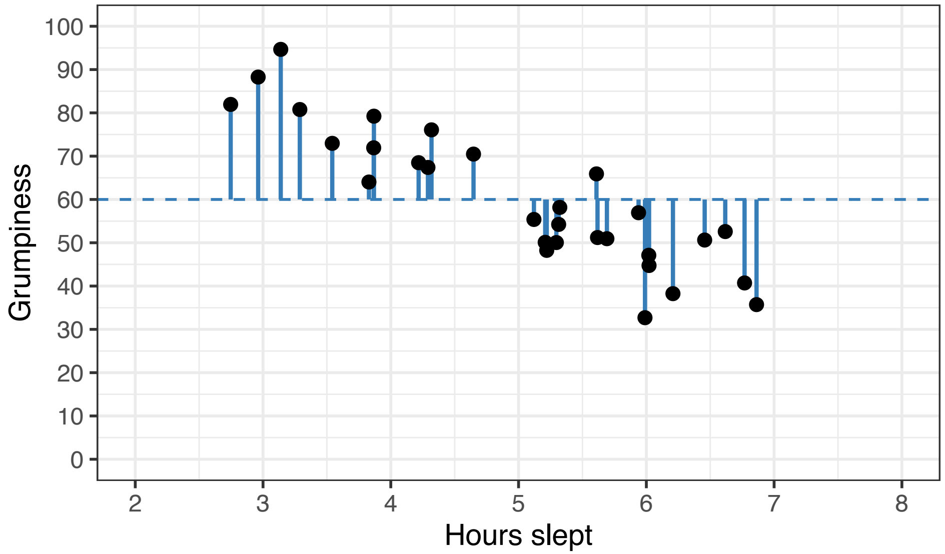
One way we can measure this using the tools we already have is to ask how much variability is in the data set, and then also ask: how much of the total variability in the data does our theory explain?.

Step 1: What is the total variability in the data set?

We can use sum of squares, from way back in the beginning of the semester, for this. It is a measure of variability. We focus on the y variable because it is the DV. It is the one we care about.

$SS_{total} \quad = \Sigma(y-\bar{y})^2$

I will call it SS_{total} because we will introduce other measures in a moment. The idea is that this is a measure of all of the variability in the data set.



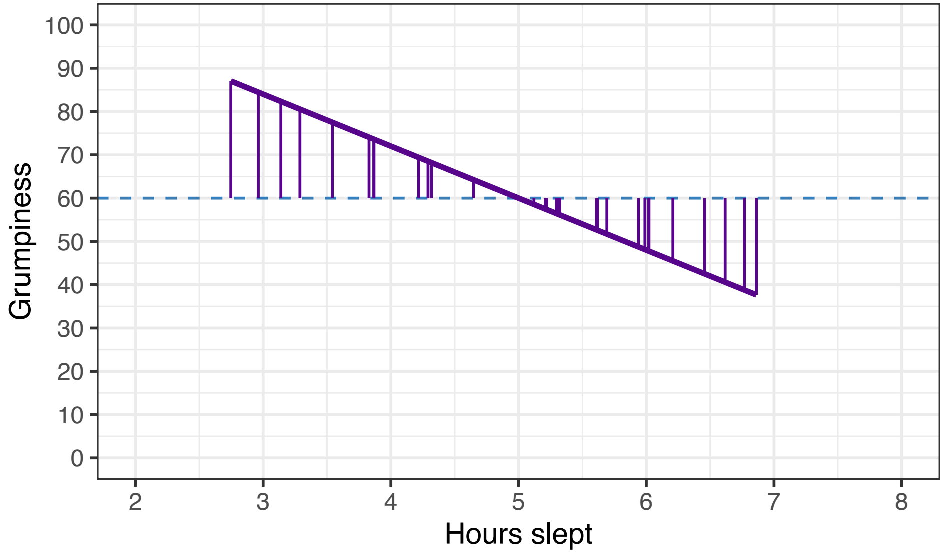
What does it mean to be a good theory?

Step 2: How much of this variability does our theory explain?

This number is new, but the concept is easy. Our theory makes a prediction. It is the line we found.

To know how much of the variance our theory explains, we need to compare it to what we would be doing if we had "no theory".

So what value do we predict when we have no theory? The answer is the mean



Remember, the central tendency of a data set is also the expected value of the data set. So the mean represents "no theory". Our line represents "a theory". So if we want to measure the effect of our theory, we have to compare it to the mean. We can do that by calculating the difference between our predicted values and the "no theory" value, the mean. We have to square it:

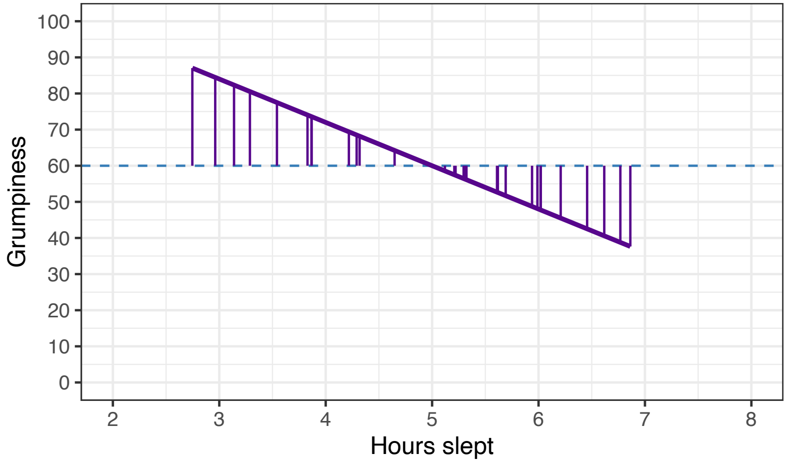
$$SS_{explained} = \Sigma(\hat{y}-\bar{y})^2$$

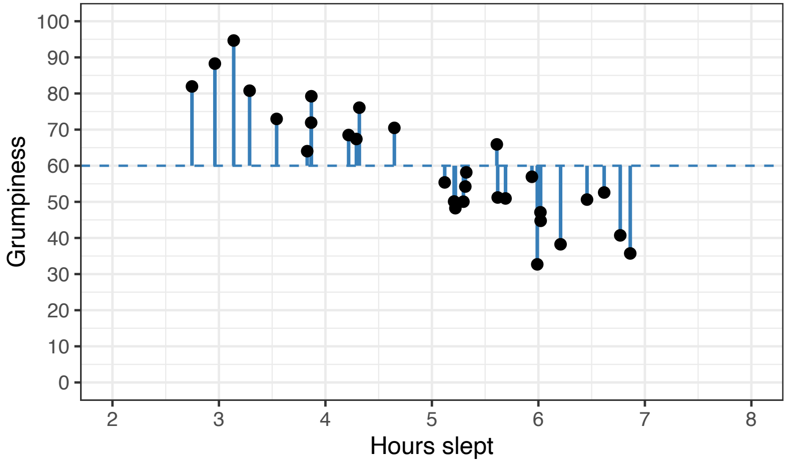
What does it mean to be a good theory?

Step 3: Calculate the proportion explained out of the total

We can put these two concepts together. We can measure the proportion of the total variability that is explained by our theory. We call it **r²**.

$$r^2 = \frac{SS_{explained}}{SS_{total}} = \frac{\Sigma(\hat{y}-\bar{y})^2}{\Sigma(y-\bar{y})^2}$$





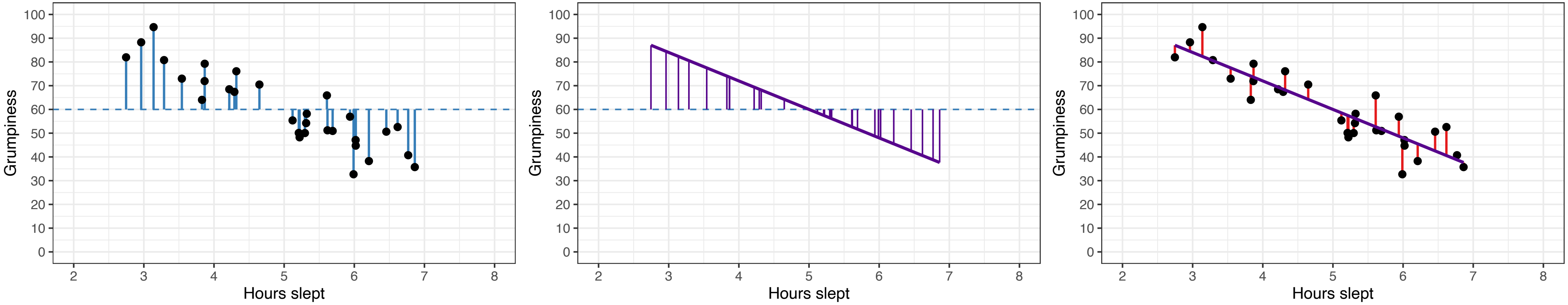
Take a moment to let this sink in. The measure r² is the amount of variability explained by the theory expressed as a proportion of the total variability in the data set.

How do residuals fit in?

Remember when I said residuals are the critical concept? Well, they are here too. Residuals are the variability that is not explained by our theory.

$$SS_{total} = SS_{explained} + SS_{unexplained}$$

$$\Sigma(y-\bar{y})^2 = \Sigma(\hat{y}-\bar{y})^2 + \Sigma(y-\hat{y})^2$$



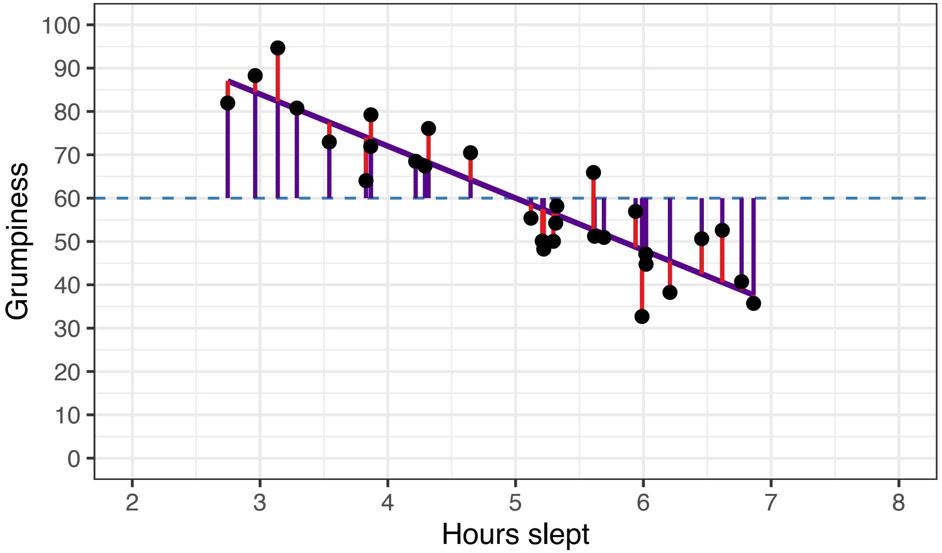
The idea is that there is a total amount of variability, and that this can be split into two types: the variability that is explained, and the variability that is not explained.

The apportionment of variability in one plot

Sometimes you will see plots like this that combine all of the ideas.

$$SS_{total} = SS_{explained} + SS_{unexplained}$$

$$\Sigma(y-\bar{y})^2 = \Sigma(\hat{y}-\bar{y})^2 + \Sigma(y-\hat{y})^2$$



r^2 is called the coefficient of determination

The fancy name for r^2 is the **coefficient of determination**.

Because it is a proportion, it ranges from 0 to 1.

$$r^2 = \frac{SS_{explained}}{SS_{total}} = \frac{\Sigma(\hat{y}-\bar{y})^2}{\Sigma(y-\bar{y})^2}$$

It has a very intuitive interpretation. It is the proportion of the variance explained by the model. So, higher is better!

r^2 = .10 means that 10% of the variance is explained.

r^2 = .50 means that 50% of the variance is explained.

Whoa. Why did I switch from SS to variance right there? Because they are the same. The (n-1) term divides out when we put them in a ratio:

total variance

$$\frac{\Sigma(y-\bar{y})^2}{n-1}$$

explained variance

$$\frac{\Sigma(\hat{y}-\bar{y})^2}{n-1}$$

$$r^2 = \frac{\Sigma(\hat{y}-\bar{y})^2}{\cancel{n-1}} \times \frac{\cancel{n-1}}{\Sigma(y-\bar{y})^2}$$

r^2 is r squared!

The similarity in names between r and r^2 is not accidental. r^2 is really r squared:

$$r = \frac{\frac{1}{n-1} \Sigma\,(x_i - \bar{x})\,(y_i - \bar{y})}{s_x s_y}$$

$$r^2 = \frac{SS_{explained}}{SS_{total}} = \frac{\Sigma(\hat{y}-\bar{y})^2}{\Sigma(y-\bar{y})^2}$$

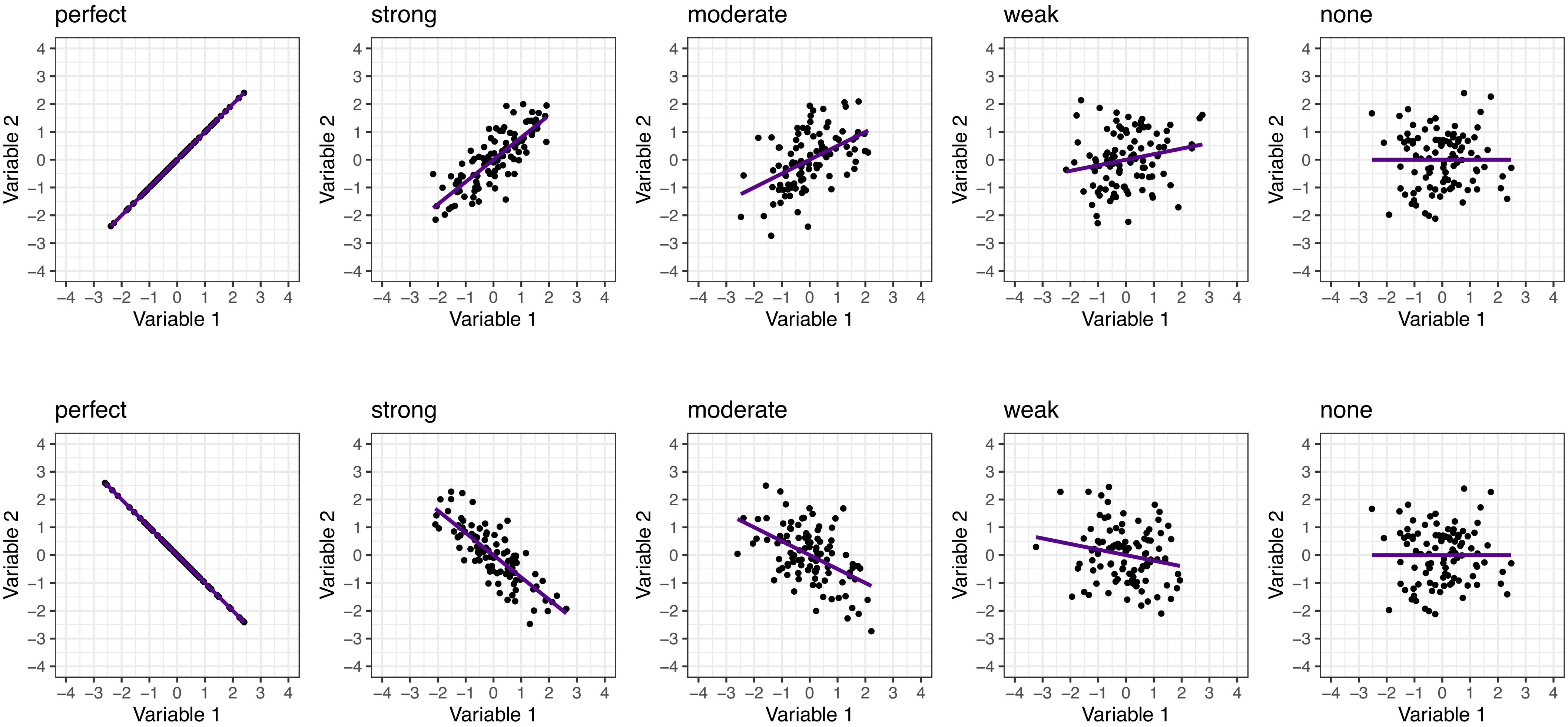
Proving this relationship is not as straightforward as we might like. It requires some mathematical facts we have not covered, and some fancy algebra.

That said, you can prove it to yourself inductively. Simply use R to calculate both the r and r^2 for data sets. You will see that they are always in a squared relationship with each other!.

Can we run a statistical test?

There will **always** be a line

One important thing to remember is that you can draw a line of best fit for any data set. That is what it means to be “best”. There is always a “best. We already saw this earlier — even when there is no relationship (“none” below), there is a line!



Testing the slope against the null hypothesis

Since there will always be a line, one thing you will always want to do is test the slope of the line of best fit against the null hypothesis that the slope is zero. There is a *t*-test for slopes. It looks very similar to the other t-tests that we've seen in this course:

one sample *t*-test

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}$$

***t*-test for Pearson's r**

$$t = \frac{r - \rho_0}{\sqrt{\frac{1-r^2}{n-2}}}$$

***t*-test for slope**

$$t = \frac{b_1 - 0}{\frac{s_y}{s_x}\sqrt{\frac{1-r^2}{n-2}}}$$

As always, the numerator simply compares our slope to the slope for the null hypothesis, which is usually 0. (I can't use greek letters here because sometimes people use greek letters for coefficients in regression.)

As always, the denominator is simply a measure of standard error for the slope. There is some nice discussion of how to derive this from the standard error of r in the book, but really, in the end, it just is the best equation for the standard deviation of the sampling distribution of slopes. That's it. It is the math that gives us the right value.

Statistical significance and Practical significance

Remember, we want both statistical significance and practical significance.

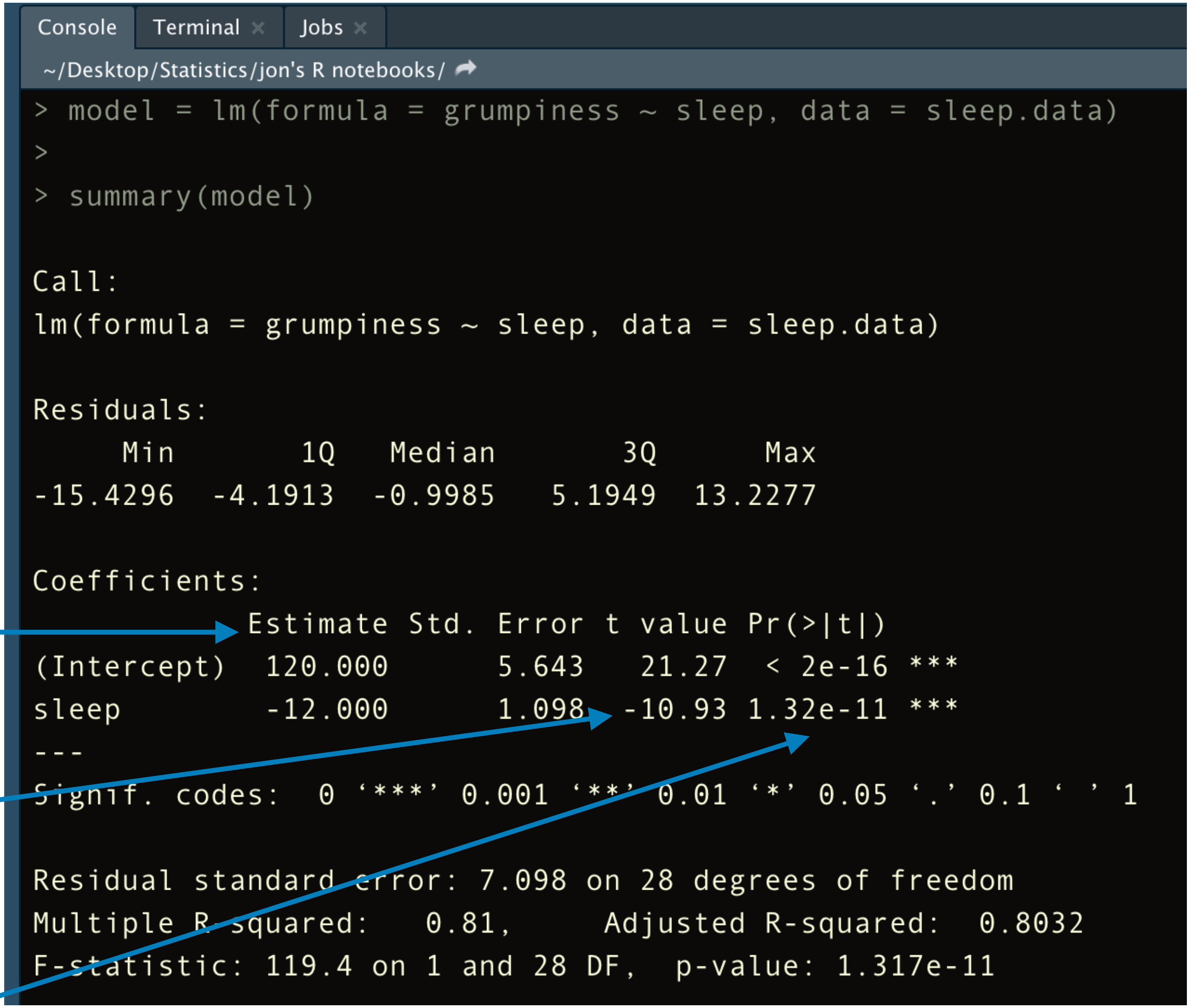
_t_-test on the slope: This will tell you statistical significance. It tells you the probability of observing a slope like this (or one more extreme) if the population slope were truly 0.

the slope coefficient (b_1): You should look at this. This is your raw effect size. Is it the size you would expect given your theory?

model fit (r^2): You should look at this. This is how well your theory predicts the data. It is possible to see a significant slope that is nonetheless not a very good fit to the data.

Using the summary() function with lm()

We have already seen that we can use lm() to create our linear model. All we have to do to run a significance test and calculate r^2 is save that lm to a variable, and run the summary() function on the variable.



```
> model = lm(formula = grumpiness ~ sleep, data = sleep.data)
>
> summary(model)

Call:
lm(formula = grumpiness ~ sleep, data = sleep.data)

Residuals:
     Min       1Q   Median       3Q      Max
-15.4296  -4.1913  -0.9985   5.1949  13.2277

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  120.000      5.643   21.27  < 2e-16 ***
sleep        -12.000      1.098  -10.93 1.32e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.098 on 28 degrees of freedom
Multiple R-squared:   0.81,     Adjusted R-squared:  0.8032
F-statistic: 119.4 on 1 and 28 DF,  p-value: 1.317e-11
```

save as model →

run summary() → on model

summary() reminds → us of the model

“Estimate” tells us the coefficients for the intercept and the slope (named after the x variable)

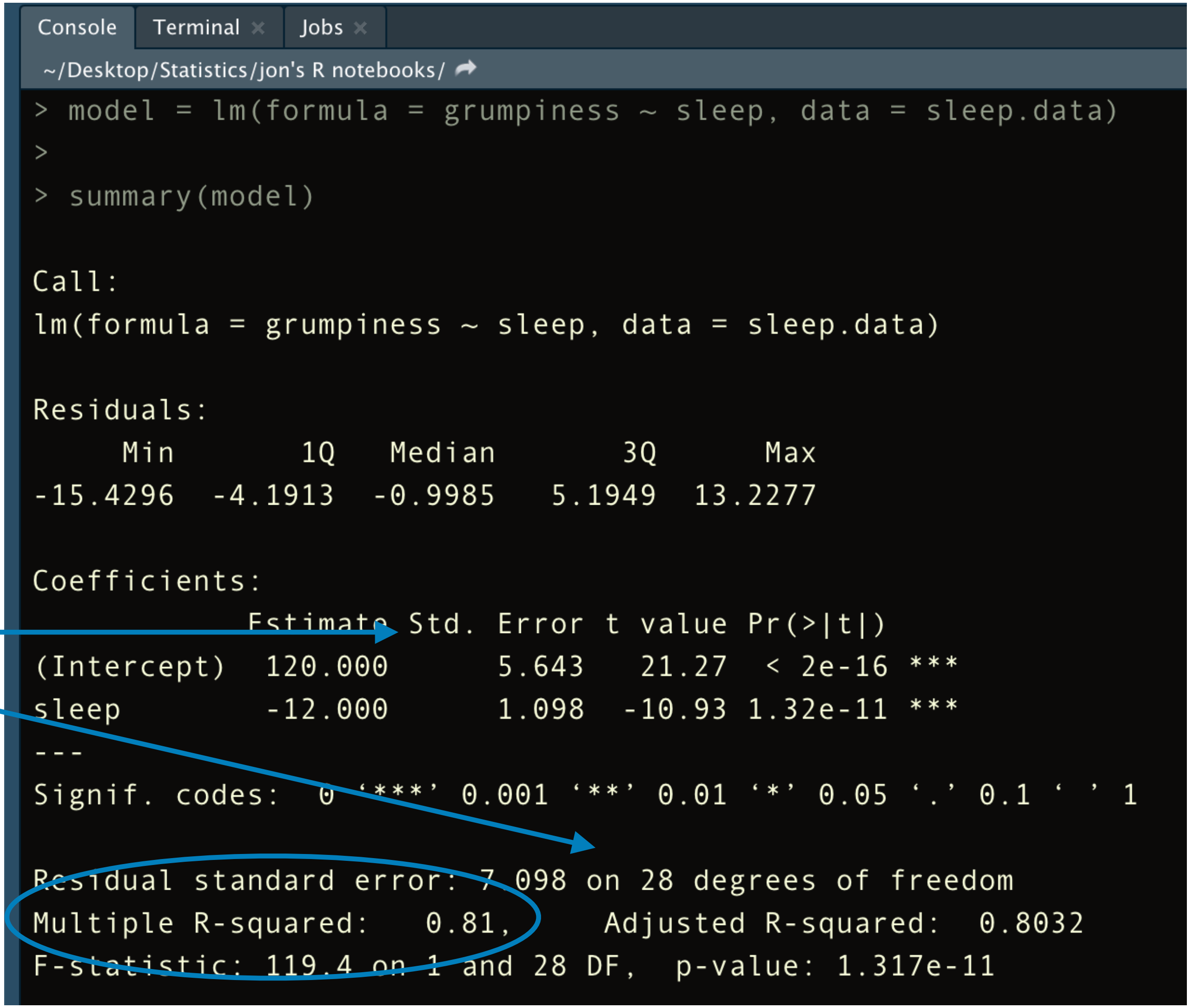
“t value” tells us the *t*. We only need the one for slope.

“Pr” tells us the *p*-value.

Using the summary() function with lm()

Summary() also tells you the r^2. It provides two. The one you want is called "multiple R-squared".

I should point out that it also tells you all of the quantities that you would calculate on your own if you were doing this by hand, like degrees of freedom, and the standard error of the coefficients and the standard error of the residuals.



```
> model = lm(formula = grumpiness ~ sleep, data = sleep.data)
>
> summary(model)

Call:
lm(formula = grumpiness ~ sleep, data = sleep.data)

Residuals:
     Min       1Q   Median       3Q      Max
-15.4296  -4.1913  -0.9985   5.1949  13.2277

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  120.000      5.643   21.27  < 2e-16 ***
sleep        -12.000      1.098  -10.93 1.32e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.098 on 28 degrees of freedom
Multiple R-squared:   0.81,     Adjusted R-squared:  0.8032
F-statistic: 119.4 on 1 and 28 DF,  p-value: 1.317e-11
```