Linear Regression

NRES 710

Last compiled: 2024-09-03

Review quiz

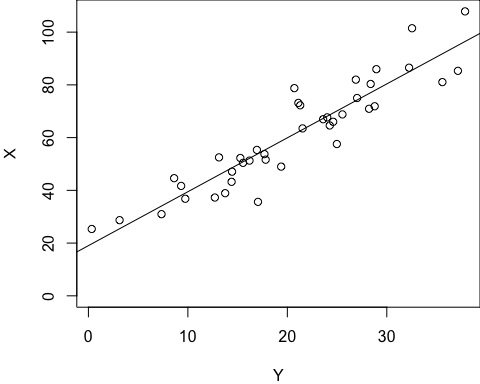
We will spend ~3 weeks learning about linear regression. This is an important topic! That is foundational to everything we will do this semester.

## Linear regression

**Regression – Continuous X; Continuous Y**

Linear regression is a common statistical analysis when you have both a **continuous x-variable and a continuous y-variable**. The idea is that the x-variable is meant to cause changes in the y-variable. X is the ‘predictor variable’, and y is the ‘response variable’. X causes changes in Y. (Although, again, we can’t demonstrate causation without an experiment.)

Graphically, this looks like:



* The purpose of regression is to fit a line to the data and determine the slope of the line.
* Linear regression is **not** meant to determine if the relationship is ‘significant’ – although that is something that we can do.

**Q:** Thinking back to algebra class from high school: what is the ‘equation’ for this line?

**y = mx + b**

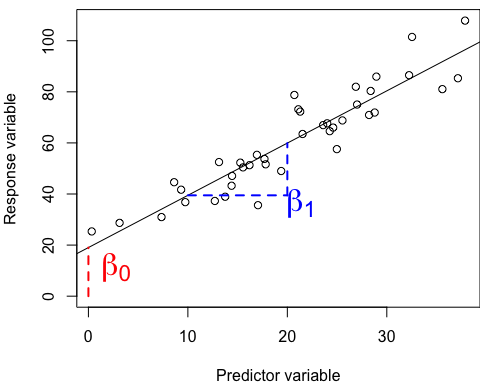
* **M = slope = =**
* **b = the y-intercept**

In statistics class, we change this equation a bit:

\*\*, where:

* is the y-intercept
* is the slope.
* The carrot-hat on top of Y () means it refers to an equation.

Graphically, this looks like:



This formula is good for the regression line, but it isn’t exactly right for statistics. What we want to do is to *associate every single X value with every single Y value*. We can adjust the equation slightly:

where the Greek letter is error, that is normally distributed, with a mean of 0 and a standard deviation of .

This means that (1) we have a line that we are trying to estimate, and (2) there is error around this line as well. The error is normally distributed.

**How might we visualize that?**

This formula explains the relationship between all X values and all Y values. It tells us that we have a *line* that we are trying to estimate, and around this line is some *noise* (error). That error is normally distributed with a mean of 0 and standard deviation of .

**Draw bell curves around points on graph**

The error can be visualized with bell curves around the line. Most of the points will be close the line, but some will be farther away – out on the tails of these bell curves.

**Draw vertical lines between points and regression line on the graph**

The distance between every point and the line on our graph is our **error**. Some people call this deviance, other people call it residuals – but it is error. The mean of the residuals will be zero, and the standard deviation will be .

We use linear regression to estimate three parameters:

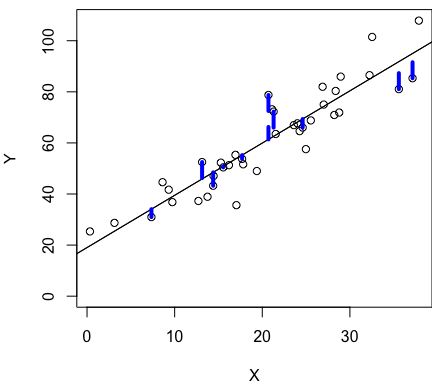
1. an intercept
2. a slope
3. a standard deviation.

If is small, all the points will be close to the line. If is large, the points will be far from the line.

### Plotting a regression line

**How does one plot or ‘fit’ a regression line to data?**

Let’s plot two continuous variables on the board again, similar to above, and have a student draw what they think is a best-fit line to the data.

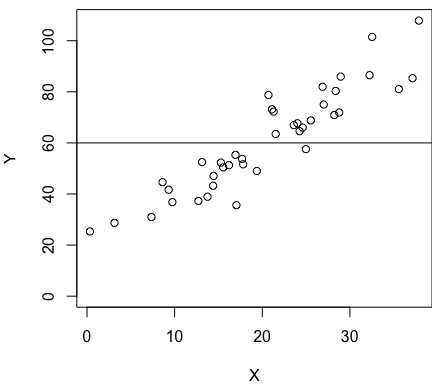


What is it that makes that the ‘best-fit line’?

* Does it… approximately go through the middle of the data?
* Are half of the points above and below the line?

If half of the points are above and below the line, then the average error is approximately zero.

However, the average error rule alone can fail us. Here’s another example:



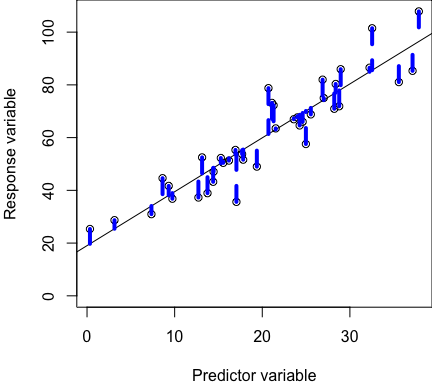
We have the same data as before, but now we have fit a horizontal line?

**Q:** What’s the average error for this case? **This creates here also!**

So it’s not enough to just say the line creates *global* average error of zero… we also want to create *local* average error of zero.

When we fit a regression line, we want to **minimize** the **sum of squared error**.

**Sum of Squared Error (SSE) =**  is the sum of squared error across all points! Or, the sum of squared residuals.



**Linear regression seeks to minimize SSE**, and the below regression line has the smallest sum of all of the blue residual lines possible.

Take-home message – the best fit line will have two attributes:

1. Minimize SSE.

## Significance testing

The goal of regression is to estimate and , so we can understand how changes in X influence changes in Y. How much does each change in X cause change in Y?

But we also want to know if this is ‘real’ relationship – this linear relationship is statistically significant.

To put this another way: if we assume there is no relationship between X and Y (i.e., the null hypothesis, , is true), what is the chance of getting the observed relationship, given that assumption and the data we have? This is what the p-value captures. If this p-value is really small (less than 0.05), then we can reject the null hypothesis and support the idea that there is a real relationship between X and Y.

### General Rules of Regression

I want to emphasize a few general rules about significance testing using regression.

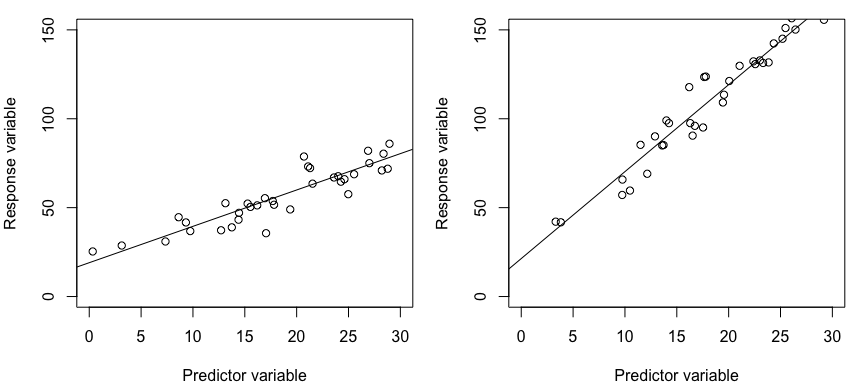
Let’s consider some **scenarios** given the null hypothesis, : there is no relationship between two continuous variables, X and Y.

**: no slope, or slope = 0.**

#### Scenario 1

*Brian make two large empty graphs, that will be re-used three times*

In which of these two cases would you be more likely to reject the null?



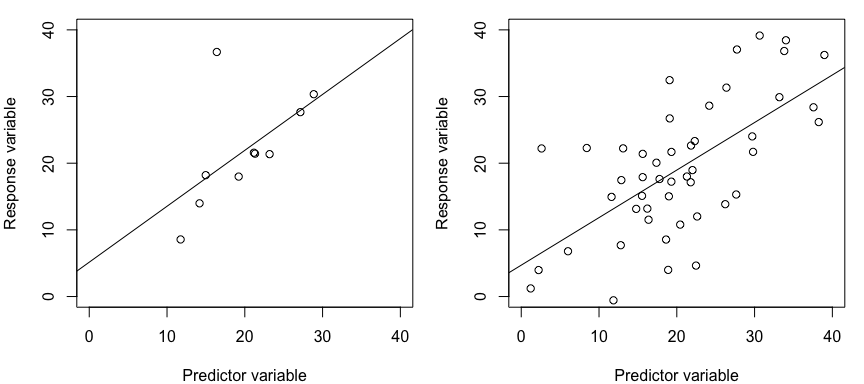
**Q:** In which dataset would you be more likely to conclude that you would reject the null that slope = 0?

The right graph, because the slope is a lot steeper. The effect of X on Y is larger.

**As slope increases, p-value decreases**.

#### Scenario 2

In which of these two cases would you be more likely to reject the null?



**Q:** In which dataset would you be more likely to conclude that you would reject the null that slope = 0?

**Q:** Why are we less likely to reject the null in the case of the left graph? Small sample size!

There are so few data in the left graph, it’s possible the slope might be zero (horizontal line). For the right graph, it’s unlikely that these data would give a horizontal line

These two plots have the same slope, but one has much more data than the other.

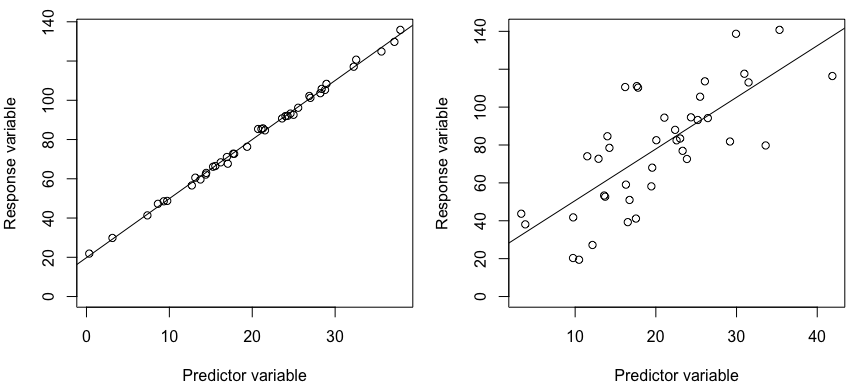
**As sample sizes increase, p-values decrease.**

So the first two scenarios we have considered reinforce the patterns we talked about last week: p-values are driven by (1) effect size () and (2) sample size (*n*).

But there is actually one more important feature influencing p-values in regression.

#### Scenario 3

In which of these two cases would you be more likely to reject the null?



**Q:** These graphs have the same slope. For which one would you be more likely to say the slope ‘definitely is not zero’?

* Left! The error on the left graph is quite small, which makes it very *clear* what the relationship is. The right graph is also statistically significant, but it has much more noise and is not not as ‘clear’.
* **As error increases, p-values increases**

So we have learned three ‘rules of thumb’ about our ability to detect statistically significant effects using linear regression:

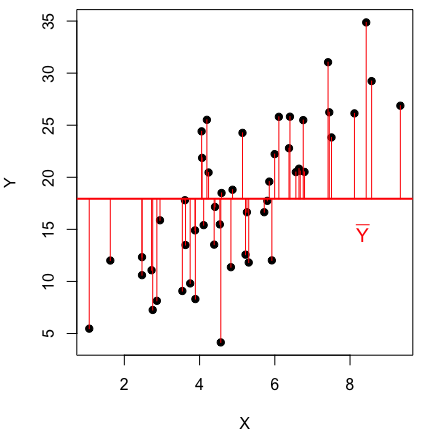
* **As slope increases, p-value decreases**
* **As *n* increases, p-value decreases**
* **As error increases, p-value increases**

### How do we calculate p-values…?

I am not going to teach you how to do that. Computers do this for us nowadays. But I will briefly mention it because this is a graduate student class so we should be aware of fundamentals.

The error in the data is partitioned into a few categories.

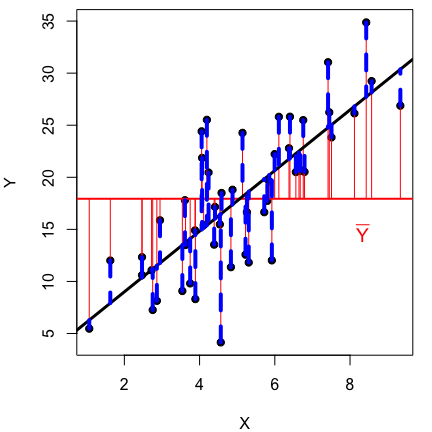
**Total Sum of Squares** = = the total variability in the response (Y) variable.



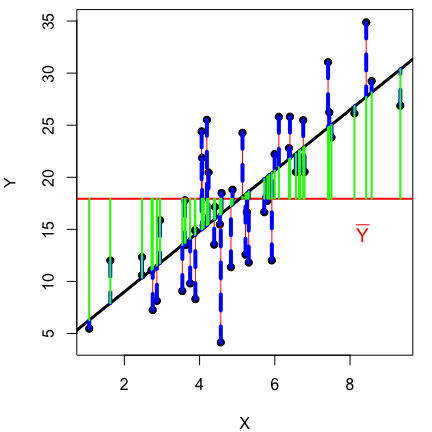
This can be partitioned into two other variables…

We already discussed **SSE = Sum of Squared Error (SSE)** = = the sum of the distance from each point to the regression line. We already covered this.

Add a line to the graph and a few blue lines for residuals:



**SSR = Sum of squares due to regression** = = the sum of the distance from each predicted y-value for the regression line to the average y-value.



**TSS = SSE + SSR**

Takeway: The smaller the error in the SSE, the smaller the p-value.

P-values are calculated by partitioning the **Total Sum of Squares** into the constituent **Sum of Squared Errors** and **Sum of Squared Errors due to Regression**. A ratio of SSE and SSR gives us the mean squared error, which gives us an ‘F-statistic’, we then check a table in a book, which gives us a p-value. But computers calculate this for us now, which is a vast improvement. Onward!

**We have just covered the theory of linear regression in ~45 minutes! Quick review:**

* The goal of regression is to estimate the slope, intercept, and (standard deviation).
* We can use regression to test the null hypothesis that the slope = 0.
* It does that by partitioning the TSS into the SSE and the SSR.
* The regression line is fit by minimizing the SSE and making it as small as possible.
* We should be aware of three patterns in regression:

1. Slope goes up, p-value goes down
2. As sample size goes up, p-value goes down
3. As error goes up, p-value goes up

We will discuss the assumptions of linear regression next Tuesday.

## Regression in R

Let’s see what this looks like in Program R.

Let’s consider two continuous variables, precipitation and biomass. Our hypothesis is that precipitation influences biomass.

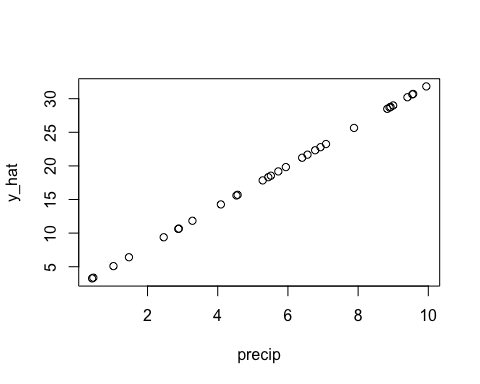
**Q:** Which is the predictor variable, and which is the response variable?

**X - Precipitation** **Y - Biomass**

Let’s now simulate data, so that we know what ‘Truth’ is.

**Truth:**

### To learn regression in R, let's simulate data so we know what 'Truth' is  
  
# Set the seed for reproducibility  
set.seed(123)  
  
# Simulate a continuous predictor variable, precipitation  
n <- 30  
precip <- runif(n, min = 0, max = 10)  
  
# Simulate the true, predicted response of biomass to precip (y-hat)  
y\_hat <- 2 + 3 \* precip  
  
# What does this look like?  
plot(y\_hat ~ precip)

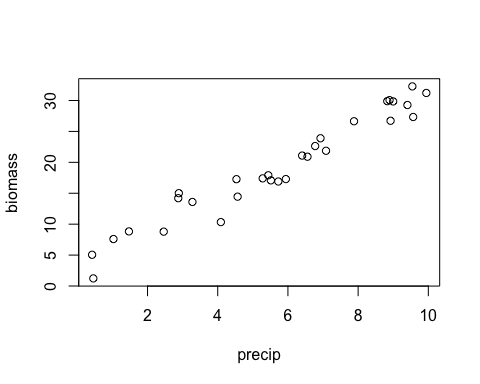


There is no error or noise in y-hat… but biomass is influenced by many things in nature! What are some things that might influence biomass? Sunlight, soil type, aspect, predators, etc.

# Simulate error for the response variable  
# Let's say the standard deviation = 2, which means that around our line there  
# is a 2 kg/ha standard deviation.  
# Rule of thumb: 66% of samples within 1 SD, 95% of samples within 2 SD, 99% within SD  
# With a SD of 2, 66% of samples will be within 2 units of the line, etc.  
error <- rnorm(n, mean = 0, sd = 2)  
  
# Create the response variable, biomass  
# biomass = beta0 + beta1 \* precip + epsilon  
biomass <- y\_hat + error  
  
# Create a data frame  
datum <- data.frame(precip = precip, y\_hat = y\_hat, error = error, biomass = biomass)  
  
# Save this dataset  
write.csv(datum, "lecture\_3\_dataset1.csv")

We have now simulated and saved our data. Let’s shift over to analyze it!

### Analyze the data  
  
# Read in the dataset  
datum <- read.csv("lecture\_3\_dataset1.csv")  
  
# Observe the first few rows of our data  
# head(datum)  
  
# First thing we should do when starting an analysis is.. look at our data!  
plot(biomass ~ precip, data = datum)



Does it look like there is a relationship? Yea! It passes my eyeball test.

The function we will be using for the rest of the semester to fit linear regression and many other statistical tests is… the linear model, ‘lm()’!

You can learn more about ‘lm()’ in R by reviewing the help files:

# Ask R for information about how 'lm()' works  
help(lm)  
?lm()

Let’s examine the ‘lm()’ documentation in R. Take a look at that.

‘lm()’ usually a formula, where the response variable comes and the predictor variable second: **Y ~ X**.

R will automatically put an intercept into your model, so you don’t have to specify this as an extra variable in your formula.

* Note: you can take the intercept out with Y ~ X - 1, or you can ask for the intercept specifically with Y~1.
* In general, specifying Y ~ X gives you the equation: beta0 + beta1X + error

Since we have defined all of our variables in R using code, we can fit the model just by calling those variable objects:

# Fit the linear model  
results <- lm(biomass ~ precip, data = datum)  
  
# Examine the results  
summary(results)

##   
## Call:  
## lm(formula = biomass ~ precip, data = datum)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.286 -1.441 0.039 1.411 3.762   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.113 0.806 3.86 6e-04 \*\*\*  
## precip 2.814 0.126 22.35 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.98 on 28 degrees of freedom  
## Multiple R-squared: 0.947, Adjusted R-squared: 0.945   
## F-statistic: 500 on 1 and 28 DF, p-value: <2e-16

In this class, I will refer to analysis outputs as ‘results’ and analysis data as ‘datum’. I do this to be consistent, avoid confusion with code in our analysis, and avoid confusion with other functions in R (e.g., the ‘data()’ function).

### The call

This describes the formula that we specified to R

### Residuals

Information that nobody ever looks at that describes the distribution of residuals around the line.

### Coefficients

#### The intercept

**Q:** What parameter from our statistical model does the ‘intercept’ represent, and how does it compare to ‘Truth’?

* **Rule of thumb:** truth should be within 2 standard errors of the estimate. 95% of all data is within two standard errors. Confidence intervals is basically two standard errors in either direction from the mean.
* We are given a ‘t-statistic’ and ‘p-value’ for the intercept. We don’t really care about these; they don’t tell us anything ecologically relevant about our system. We leave it in the model because it makes the line fit better, which we use to estimate slope.

#### The effect

**Q:** How does the estimate of the effect of precipitation on biomass compare to truth?

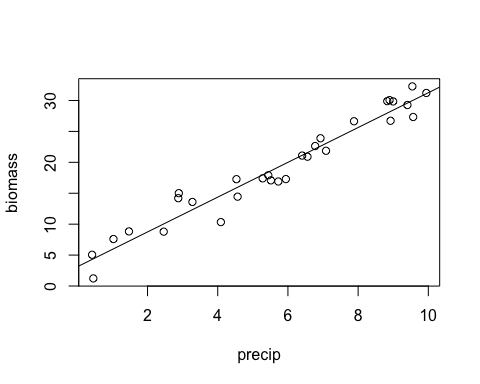
* For each 1 unit increase in precipitation, we get a […] unit increase in biomass.
* Again, the estimate is within two Standard Errors of truth (!)
* We also get a t-statistic and a p-value for this parameter. The p-value is 2x10^-16… Which is very small and suggests a very clear results.
  + This is the probability of getting our data or data more extreme, given the null hypothesis is true.
  + Because this probability is very small, we are learning that it’s extremely unlikely that we would get data like ours if the null hypothesis is true. Therefore, it seems like the null hypothesis is not true. So, we can reject the idea that the null hypothesis is truth, and infer that there is a real relationship between precipitation driving biomass.

### Other information

There is other information below that people often don’t pay attention to, but there is good stuff down here too.

* **Residual standard error**: this is our estimate of , or the standard deviation in the error around the line!
* and confidence intervals; we will discuss these more in future classes.
* P-value for the whole-model. Since there is only one parameter in the statistical model, the p-value for the entire model is the same as the p-value for the precipitation parameter.

# Plot the data with the line of best fit  
plot(biomass ~ precip)  
abline(results)



# Maybe you are used to seeing ANOVA tables...  
anova(results)

## Analysis of Variance Table  
##   
## Response: biomass  
## Df Sum Sq Mean Sq F value Pr(>F)   
## precip 1 1949 1949 500 <2e-16 \*\*\*  
## Residuals 28 109 4   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# This should look similar to the bottom line from our regression output

Question?!

[–go to next lecture–](lecture_4.html)