STAT 311: Regression

Y. Samuel Wang

Summer 2016

Logistics

- Resubmit lab
- Homework is posted
- Questions on material covered so far

Parameters which govern a line

The equation for a line can be put into the following form

$$Y = a + bX \tag{1}$$

Parameters which govern a line

The equation for a line can be put into the following form

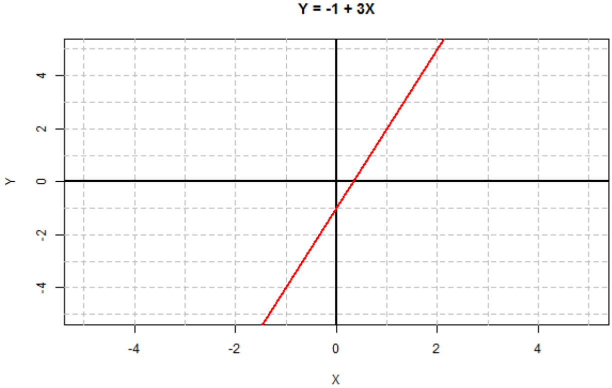
$$Y = a + bX \tag{1}$$

- X and Y are variables
- a is the **Y-intercept**. It is the value of the Y coordinate when $X = 0$
- b is the **slope**. It describes how Y changes as X changes.

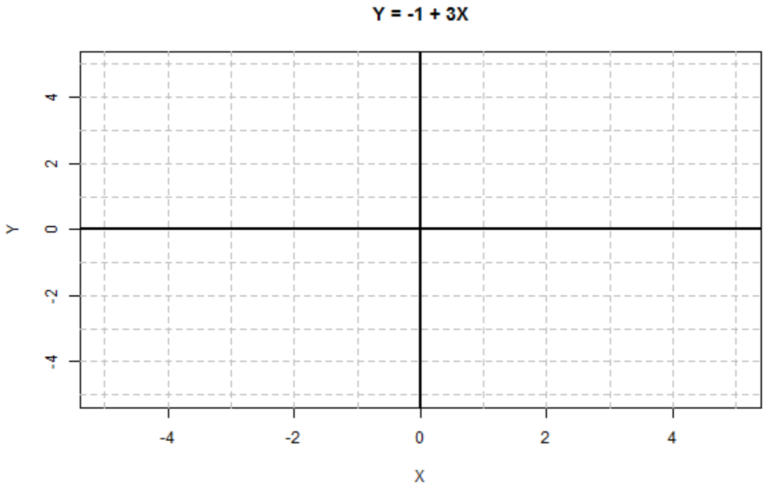
Parameters which govern a line

Consider the following equation

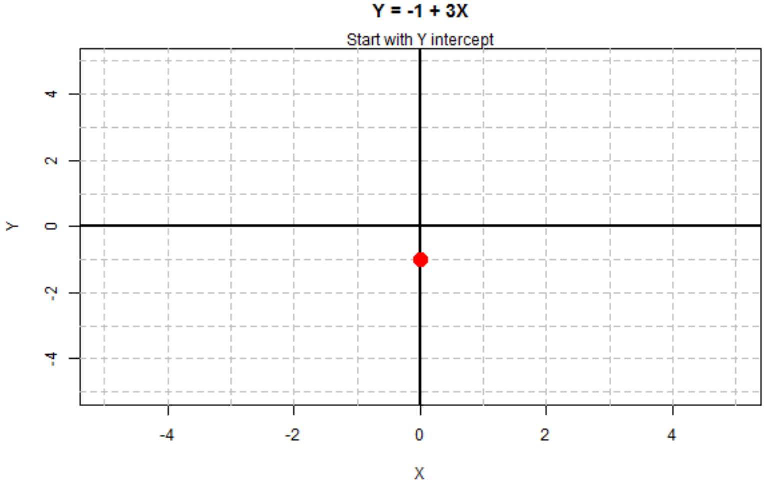
$$Y = -1 + 3X \tag{2}$$



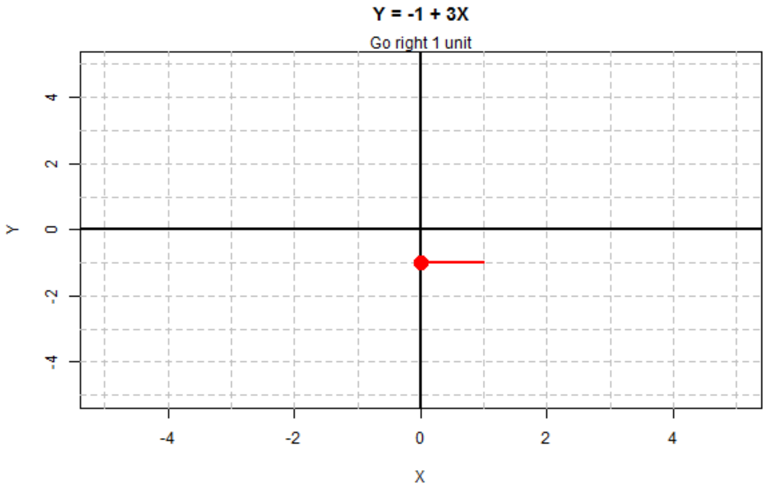
How to plot a line



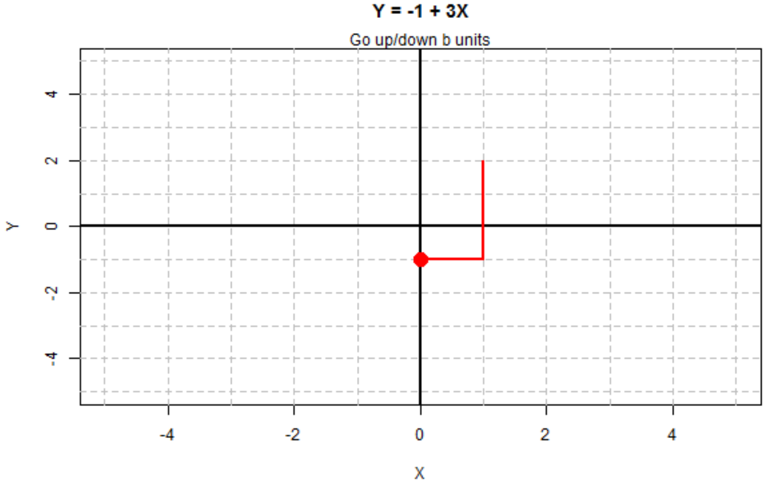
How to plot a line



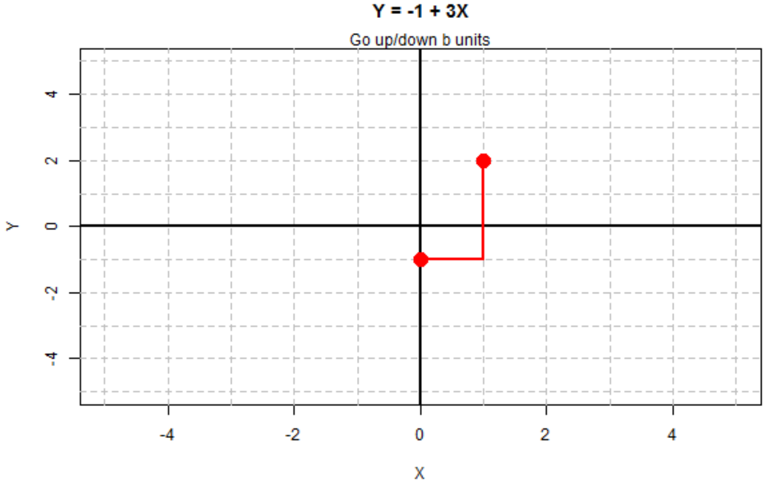
How to plot a line



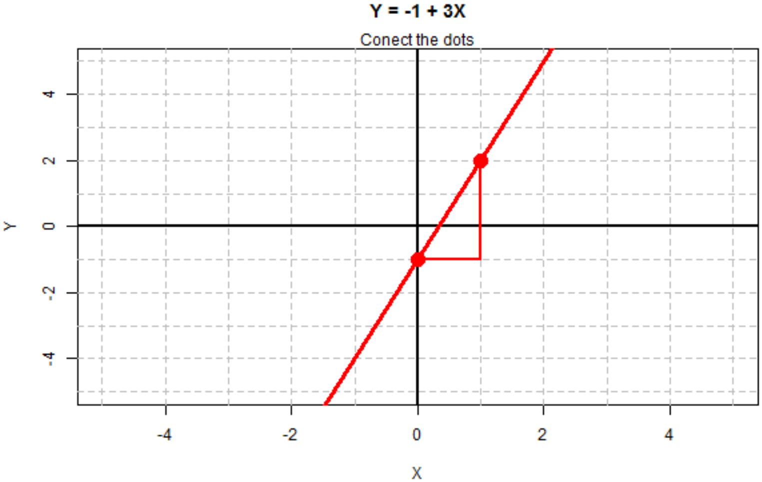
How to plot a line



How to plot a line



How to plot a line



How to describe data with a line

On the handout, draw a line that best describes the relationship between the data.

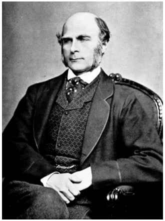
How to describe data with a line

On the handout, draw a line that best describes the relationship between the data.

- How did you decide where to put the line?
- How would you tell if your line is better than someone else's line?

Why is it called regression?

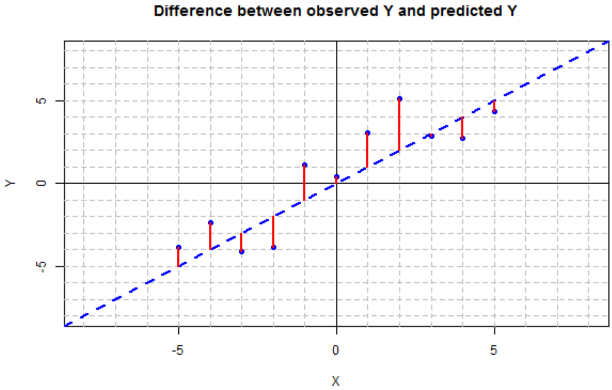
In the 1880's, Francis Galton was interested in studying the height of children, relative to the height of their parents



- In general, children with taller than average parents were also taller than average
- In general, children with shorter than average parents were also shorter than average
- But on average, the children were less extreme than their parents
- The child's height typically "regressed" back to the mean

Errors in Y

Consider the difference between the predicted (point on the line) and observed values of y. Use $\hat{y}_i$ to denote the predicted value for the ith observation.



Selecting Regression Coefficient

How can we select a slope and intercept to minimize the sum of squared errors?

$$SSE = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - (a + bx_i))^2 \tag{3}$$

Selecting Regression Coefficient

How can we select a slope and intercept to minimize the sum of squared errors?

$$SSE = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - (a + bx_i))^2 \tag{3}$$

Take a derivative and set equal to 0!

$$\frac{\partial SSE}{\partial b} = -2 \sum_i x_i (y_i - (a + bx_i)) = 0 \tag{4}$$

$$\frac{\partial SSE}{\partial a} = -2 \sum_i (y_i - (a + bx_i)) = 0 \tag{5}$$

Selecting Regression Coefficient: a

$$
\begin{aligned}
0 = \frac{\partial SSE}{\partial a} &= -2\sum_i (y_i - (\hat{a} + \hat{b}x_i)) \\
&= -2\sum_i y_i + 2n\hat{a} + \hat{b}\sum_i x_i
\end{aligned} \tag{6}
$$

$$
\Rightarrow \hat{a} = \bar{y} - \hat{b}\bar{x} \tag{7}
$$

Selecting Regression Coefficient: b

$$\begin{aligned}
\frac{\partial SSE}{\partial b} &= -2\sum_i x_i(y_i - (\hat{a} + \hat{b}x_i)) = 0 \\
&= -2\sum_i x_i(y_i - (\hat{a} + \hat{b}x_i))
\end{aligned} \tag{8}$$

$$\hat{b} = \frac{1}{N-1}\sum_i \frac{(x_i - \bar{x})(y_i - \bar{y})}{1/(n-1)\sum_i (x_i - \bar{x})^2} = \frac{cov(x,y)}{var(x)} = r_{xy}\frac{s_y}{s_x} \tag{9}$$

Ordinary least squares regression

This procedure is called Ordinary least squares (OLS)

$$\hat{y} = \hat{a} + \hat{b}x \tag{10}$$

- The best fit line passes through the centroid $(\bar{x}, \bar{y})$
- $y_i - \hat{y}_i$ is called the **residual**
- The sum of the residuals for the best fit line is 0
- We can use the output to either **predict** new values, or **explain** scientific phenomenon
- The estimated parameters are not symmetric. If we swap what is "x" and what is "y", the line will change.

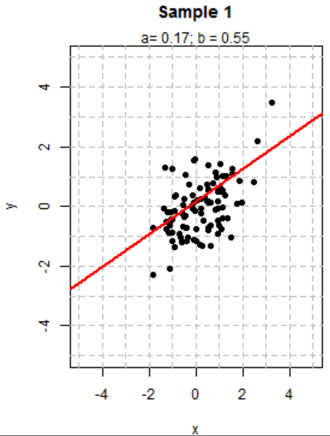
Cautions

Let's take a step back and consider what we have calculated

- Still have "hat's" on a and b because they are statistics
- There is a true a and b which minimize the SSE for the population
- What if the true relationship is not actually linear?
- What if we have grouped multiple sub-populations together?

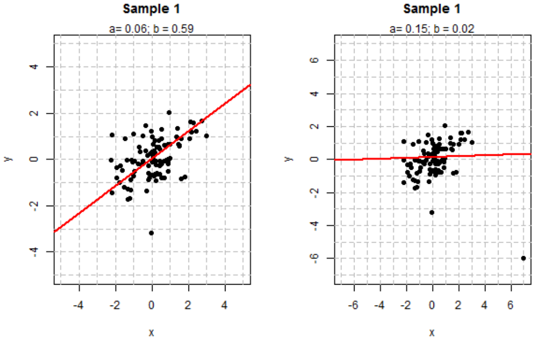
Statistic vs Population

If the sample is not the entire population, the estimated $\hat{a}$ and $\hat{b}$ can change from sample to sample.



Outliers

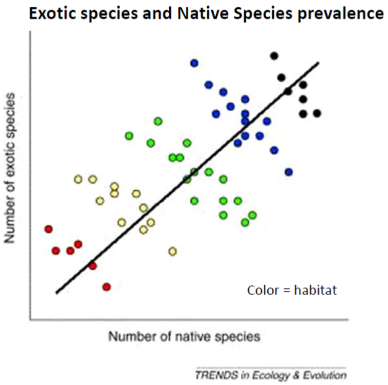
How can outliers influence regression estimates?



We'll talk more about this in the lab

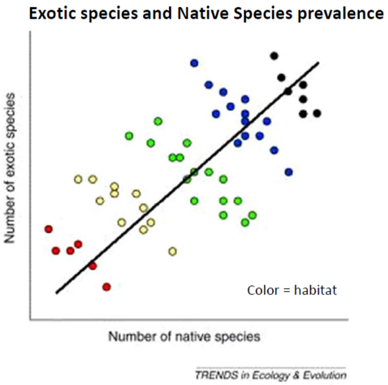
Multiple sub-populations

What happens if multiple sub-populations are included?



Multiple sub-populations

What happens if multiple sub-populations are included?



this is called Simpson's paradox. We'll talk more about it later.

Components of the squared error

Given a response and explanatory variable, we can estimate the best fit line for a response variable. But what if I want assess how useful the explanatory variable is for predicting the response variable?

$$\begin{aligned}(y_i - \bar{y}) &= (y_i - \hat{y}_i + \hat{y}_i - \bar{y}) \\ &= (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \\ &= \text{residual} + \text{Predicted deviation from mean}\end{aligned} \tag{11}$$

Components of the squared error

Given a response and explanatory variable, we can estimate the best fit line for a response variable. But what if I want assess how useful the explanatory variable is for predicting the response variable?

$$\begin{aligned}(y_i - \bar{y}) &= (y_i - \hat{y}_i + \hat{y}_i - \bar{y}) \\ &= (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \\ &= \text{residual} + \text{Predicted deviation from mean}\end{aligned} \tag{11}$$

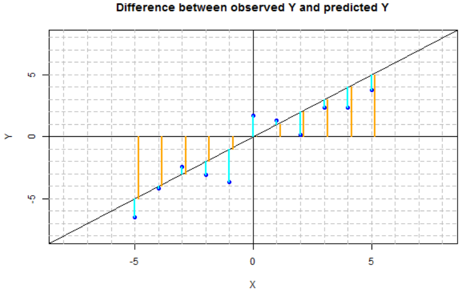
Using a bit of algebra, we can decompose the total sum of squares for Y into

$$SS_{total} = \sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2 = SS_{regression} + SS_{error} \tag{12}$$

Components of the squared error

If $SS_{regression}$ is large compared to SS_{error}, then the explanatory variable is a good predictor of the response variable

$$\frac{SS_{regression}}{SS_{total}} = \frac{\sum_i (\hat{y}_i - \bar{y})}{\sum_i (y_i - \bar{y})} = r^2 \tag{13}$$



Example: Components of the squared error

Back to Outliers

We saw in the lab yesterday, that an outlier can drastically effect our regression

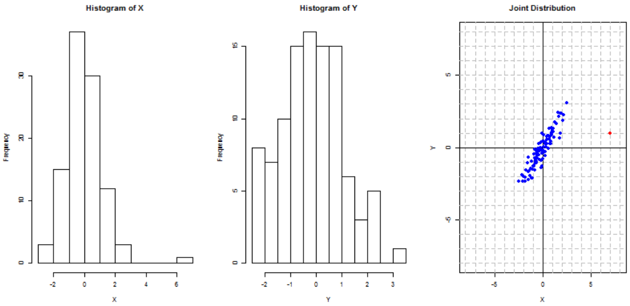
Outliers and Influential Points

Outliers are “unusual” observations. But what does it mean to be “unusual”

- Unusual X value (marginal)
- Unusual Y value (marginal)
- Unusual X and Y value together (joint)
- Might be consistent with the trend, might be inconsistent with the trend

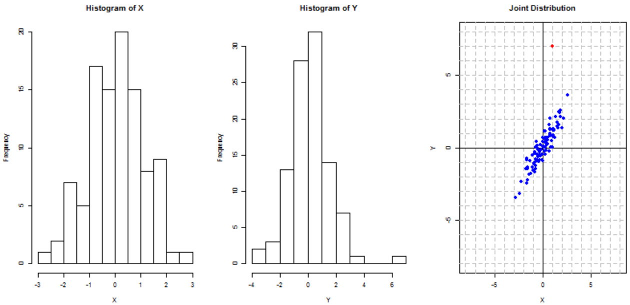
Outliers and Influential Points

Unusual X Value



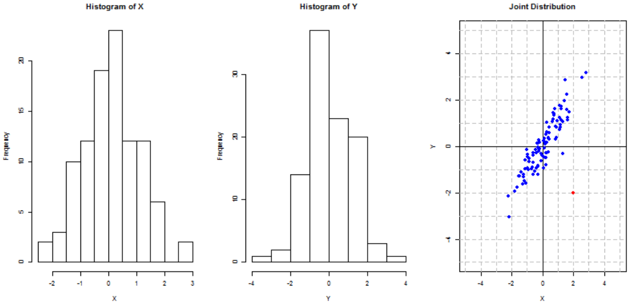
Outliers and Influential Points

Unusual Y Value



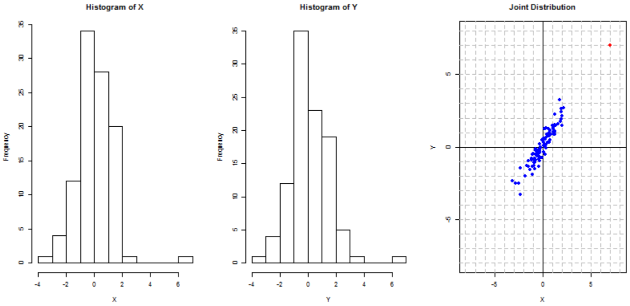
Outliers and Influential Points

Unusual X and Y Value



Outliers and Influential Points

Unusual X and Y Value, but consistent with the trend



Outliers and Influential Points

Typically, we are most interested in the slope of a regression (rather than the intercept). The type of outlier changes the affect of the outlier on the slope.

$$\hat{b} = cov(x, y)/var(x) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \tag{14}$$

Outliers and Influential Points

Typically, we are most interested in the slope of a regression (rather than the intercept). The type of outlier changes the affect of the outlier on the slope.

$$\hat{b} = cov(x,y)/var(x) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \qquad (14)$$

Does $\hat{b}$ change if I add a point at

- $(\bar{x}, \bar{y})$.

Outliers and Influential Points

Typically, we are most interested in the slope of a regression (rather than the intercept). The type of outlier changes the affect of the outlier on the slope.

$$\hat{b} = cov(x, y)/var(x) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \tag{14}$$

Does $\hat{b}$ change if I add a point at

- $(\bar{x}, \bar{y})$. No!
- $(\bar{x}, y)$.

Outliers and Influential Points

Typically, we are most interested in the slope of a regression (rather than the intercept). The type of outlier changes the affect of the outlier on the slope.

$$\hat{b} = cov(x, y)/var(x) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \tag{14}$$

Does $\hat{b}$ change if I add a point at

- $(\bar{x}, \bar{y})$. No!
- $(\bar{x}, y)$. No!
- $(x, \bar{y})$.

Outliers and Influential Points

Typically, we are most interested in the slope of a regression (rather than the intercept). The type of outlier changes the affect of the outlier on the slope.

$$\hat{b} = cov(x, y)/var(x) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \quad (14)$$

Does $\hat{b}$ change if I add a point at

- $(\bar{x}, \bar{y})$. No!
- $(\bar{x}, y)$. No!
- $(x, \bar{y})$. Yes!

Outliers in the X direction affect the slope much more than outliers in the Y direction

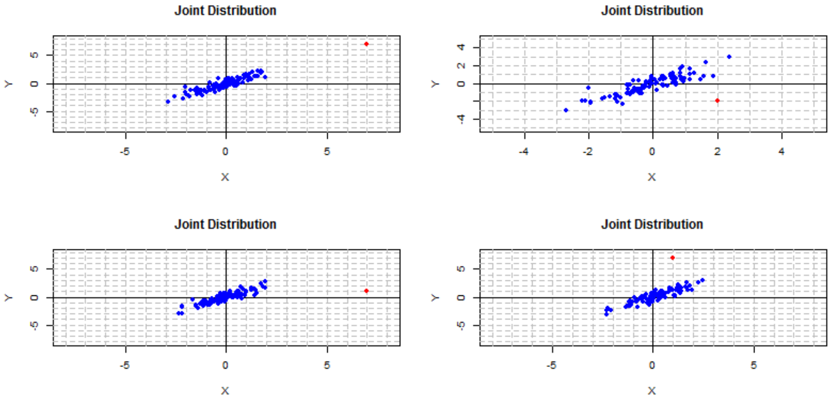
Outliers and Influential Points

Outliers in the X direction can affect the slope much more than outliers in the Y direction

- Leverage- The potential of an X value to affect the slope. High or low leverage only depends on X value
- Influence- How much a point changes the regression slope. Depends on both X and Y values

Outliers and Influential Points

Are the previous outliers we showed high leverage? high influence?



Outliers and Influential Points

So what should we do with outliers?

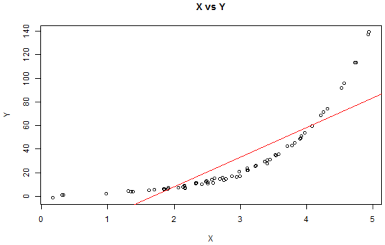
- As with most thing in statistics... it depends
- What do we know about the outlier? What trend are we trying to capture?
- Was the Palm County data point an outlier?
- Would you use that data point or not?

Non-linearity

What can we do about non-linearity? Does linear regression still work?

Non-linearity

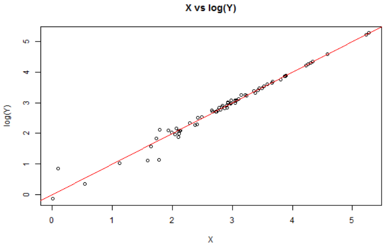
What can we do about non-linearity? Does linear regression still work? ...sort of



We can still estimate a best linear approximation to the underlying relationship. Looking at the sign and magnitude of the regression coefficients can be useful scientifically, but the interpretation is not always as clear.

Non-linearity

Can we do better? Transform the data instead



We can apply transformations to the X and Y variable to make the relationship roughly linear, but we need to be careful about interpretation

Log transform

One often used transformation is the log transform.

$$Y \Rightarrow log(Y) \tag{15}$$

Log transform

One often used transformation is the log transform.

$$Y \Rightarrow log(Y) \tag{15}$$

- Not a linear transformation, but is still **monotonic**, or always increasing
- Shrinks large values more than it shrinks small values

$$\begin{aligned} log(1000) &= 3 \\ log(100) &= 2 \\ log(10) &= 1 \end{aligned} \tag{16}$$

- Corresponds to % increase

Other commonly used transforms include $1/Y$ and $\sqrt{Y}$

How to interpret the log transform

An increase in 1 unit of the X variable, corresponds to a b percent increase in the Y variable

$$log(Y) = a + bX \tag{17}$$

This is most useful for things with exponential growth

How to interpret the log transform

An increase in 1 unit of the X variable, corresponds to a b percent increase in the Y variable

$$log(Y) = a + bX \tag{17}$$

This is most useful for things with exponential growth

- Populations
- GDP
- Stock prices (hopefully)

How to interpret the log transform