STAT 311: Hypothesis Testing for Two Way Tables

Y. Samuel Wang

Summer 2016

Logistics

- Homework 6 posted
- Lab 6 posted as well
- Final Exam next Friday
- Practice Final will be posted Friday

Example: Weather Patterns

El Nino refers to a periodic phenomenon in which a band of warm water develops in the central Pacific Ocean. This change in water temperature can have far reaching consequences and can effect weather patterns across the globe. In particular, we might be interested in seeing how El Nino affects weather patterns in WA state.

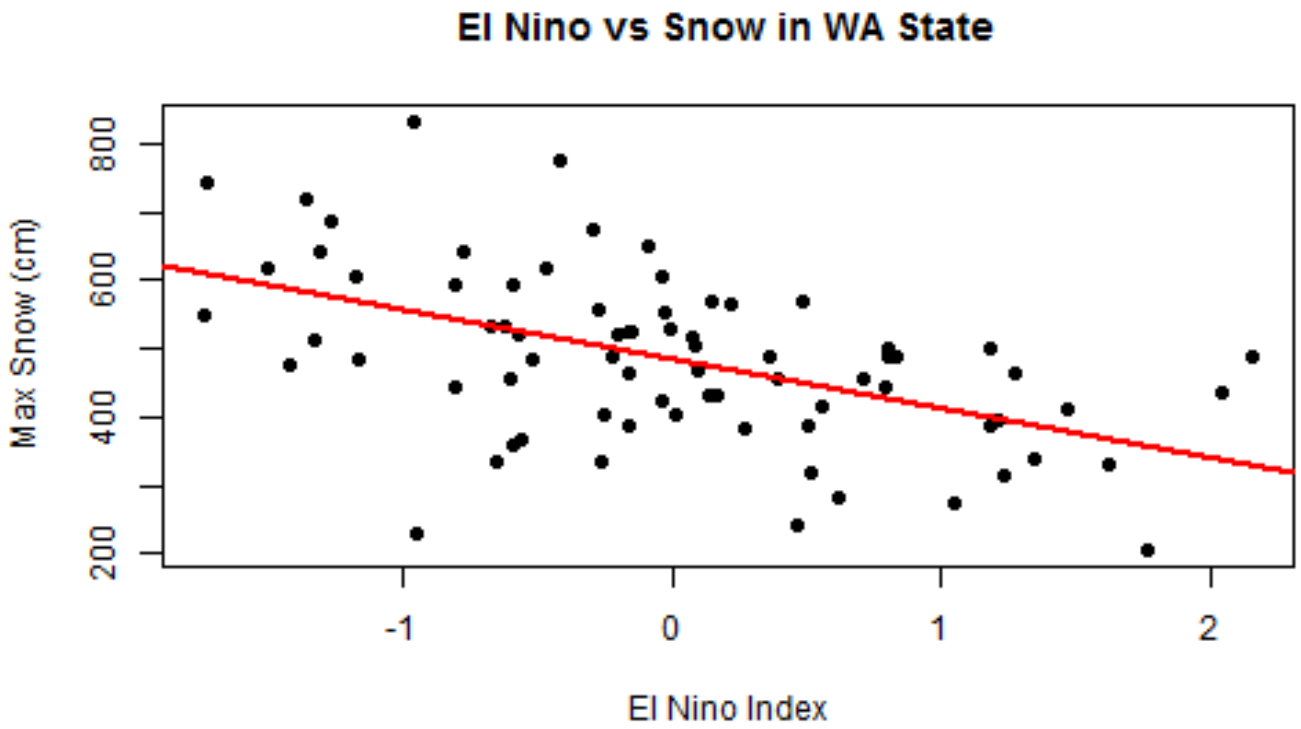
$$X = \text{Z-score of Central Pacific Sea Surface Temperature}$$

$$Y = \text{Max April Snow Depth at Paradise Station on Mt Ranier}$$

Example: Weather Patterns

Given measurements from 1920-2013, I estimate a relationship of

$$\text{Snow}_i = 484.23 - 71.62\text{Nino}_i + \epsilon_i$$



Questions of Interest

For a regression, I might be interested in a few things

- Is there actually a relationship between the two variables?
  - I see a relationship in the data I have gathered, but it is a sample
  - Is there a relationship in the theoretical population?
- Given some X value, what is the mean of the Y value?
  - Given an X value, I can estimate what the average Y value will be
  - What is a reasonable range for the mean?
- Given some X value, what is a reasonable range to expect for an individual's Y value?
  - Given an X value, I can estimate what the average Y value will be
  - What is a reasonable range for an individual with that X value?

Regression Review

A quick review of things you should remember for this lecture:

- We can measure the linear relationship between two variables through the covariance or correlation
- Given data, we can calculate the observed correlation

$$r_{xy} = \frac{1}{n-1} \sum_i \frac{(x_i - \bar{x})}{s_x} \frac{(y_i - \bar{y})}{s_y}$$

- This is simply an estimate of the true underlying population correlation
- Two parameters which govern a line: intercept (a) and slope (b) We select parameters which minimize

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

- Given some X value, what is a reasonable range to expect for an individual's Y value?

Regression Review

A quick review of things you should remember for this lecture:

- Two parameters which govern a line: intercept denoted by a; and slope denoted by b
- In linear regression, we select parameters estimates which minimize

$$SS_{error} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - (a + bx_i))^2$$

- Given these $\hat{a}$ and $\hat{b}$, we can form predictions for y when supplied with an x value-

$$\hat{Y}_i = \hat{a} + \hat{b}X_i$$

Regression Review

A quick review of things you should remember for this lecture:

- We can decompose the total sum of squares-

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

$$= SS_{regression} + SS_{error}$$

- We can also calculate the ratio $\frac{SS_{regression}}{SS_{total}} = r^2$

Regression Assumptions

In order to perform tests and confidence intervals with regression, we must assume

$$Y_i = a + bX_i + \epsilon_i$$

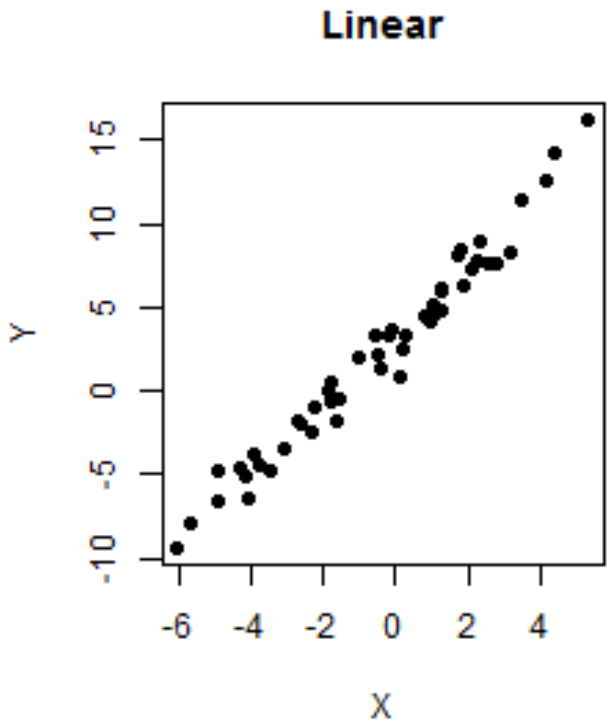
- Linearity- Y is a linear function of X
- Normally distributed errors-

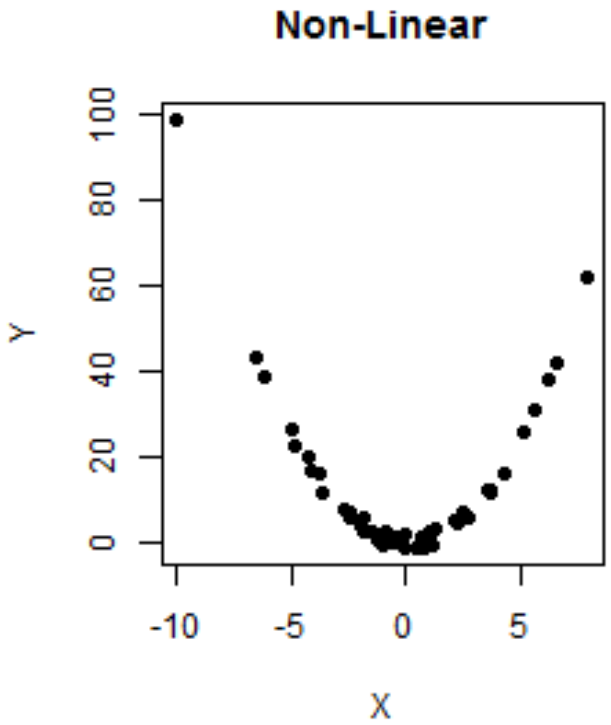
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- Homoskedasticity or equal variance- σ^2 does not change with X

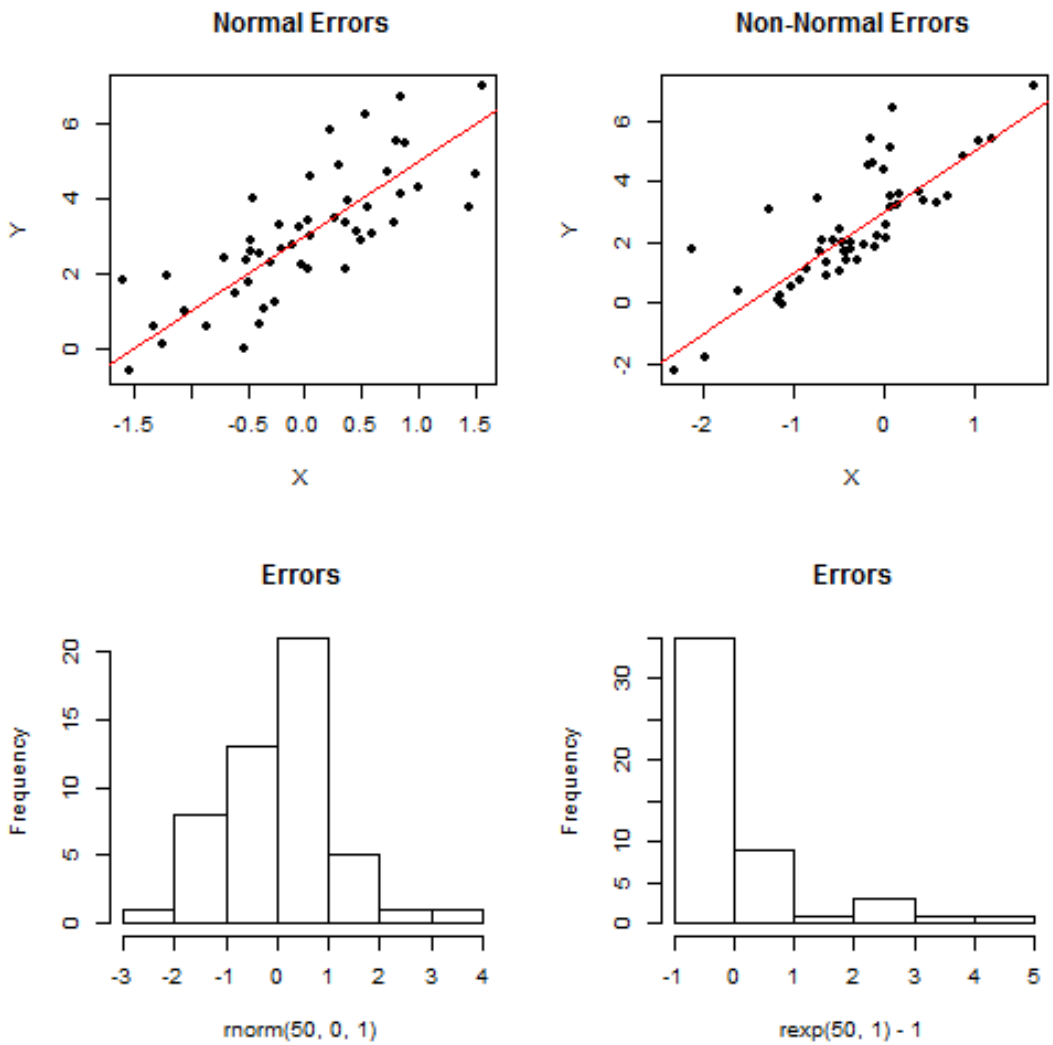
The least squares minimizing still holds if these assumptions do not, but any tests or confidence intervals will be invalid

Linearity

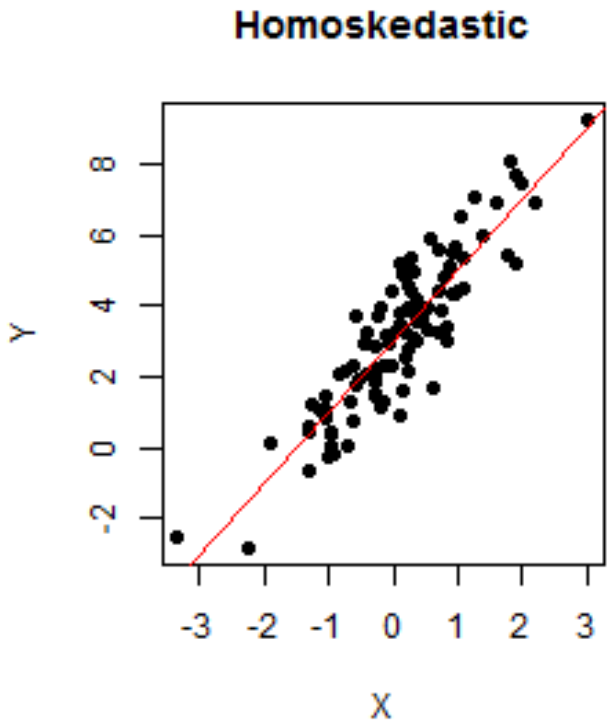

Linear
Y
X
-10
-5
0
5
10
15
-6
-4
-2
0
2
4

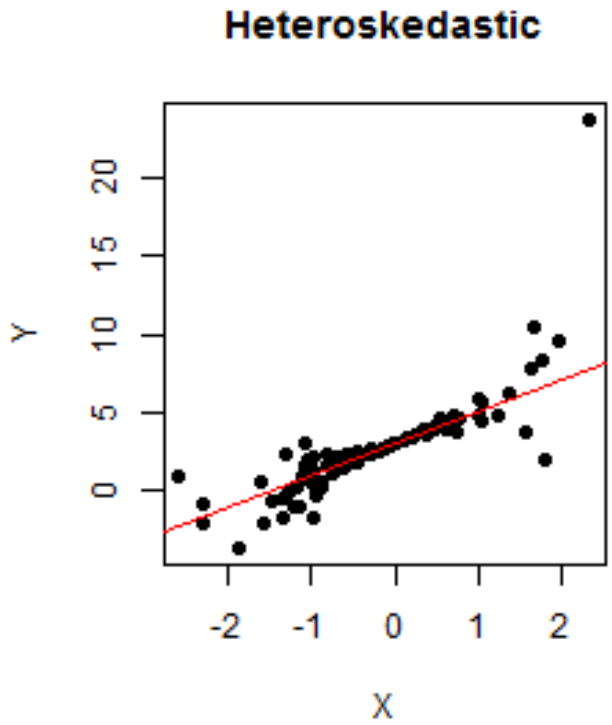

Non-Linear
Y
X
0
20
40
60
80
100
-10
-5
0
5

Normal Errors



Homoskedastic





Regression Standard Error

We assumed that $\epsilon_i \mathcal{N}(0, \sigma^2)$, so how can we estimate σ^2

$$s_{reg} = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}}$$

Notice that we divide by $n-2$ whereas before we divided by $n-1$. This is essentially because we are estimating two parameters in $\hat{y}_i$ instead of just one parameter of $\bar{y}$.

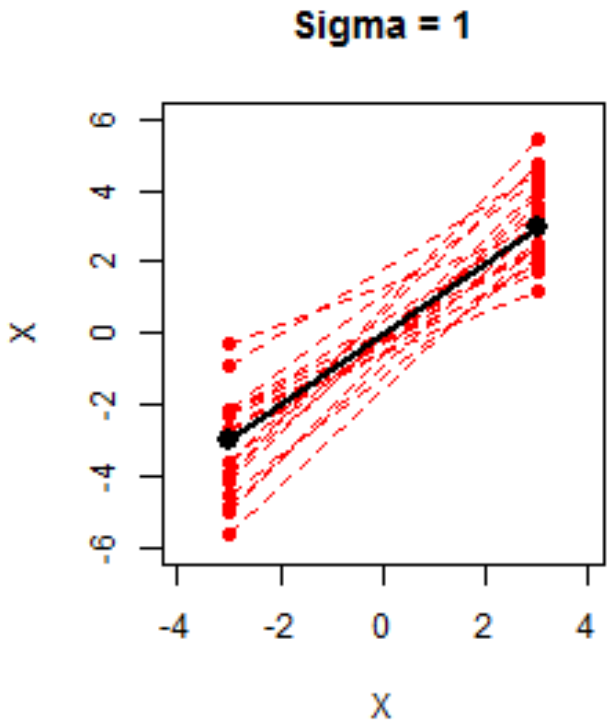
Standard Error for $\hat{b}$

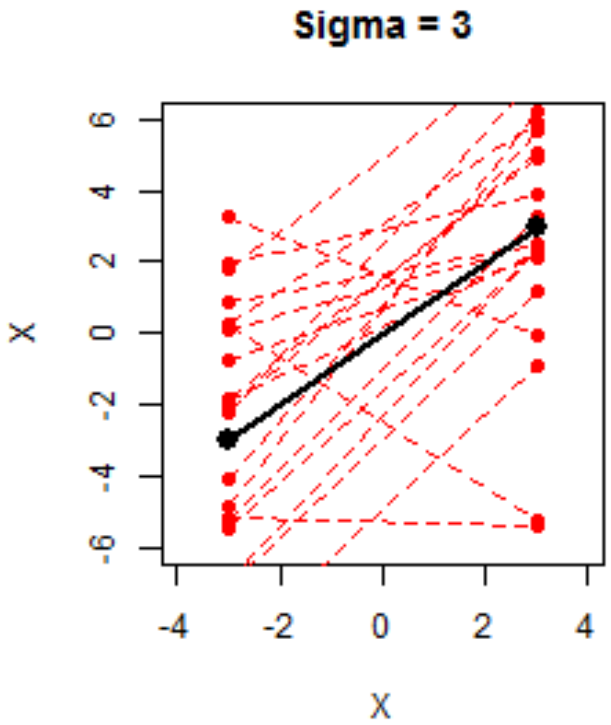
Since $\hat{b}$ is only an estimate of the "true population" slope, it will change each time we take new sample. How much $\hat{b}$ will change should depend on

- The variance of the errors ϵ_i.
- The variance of the X values.

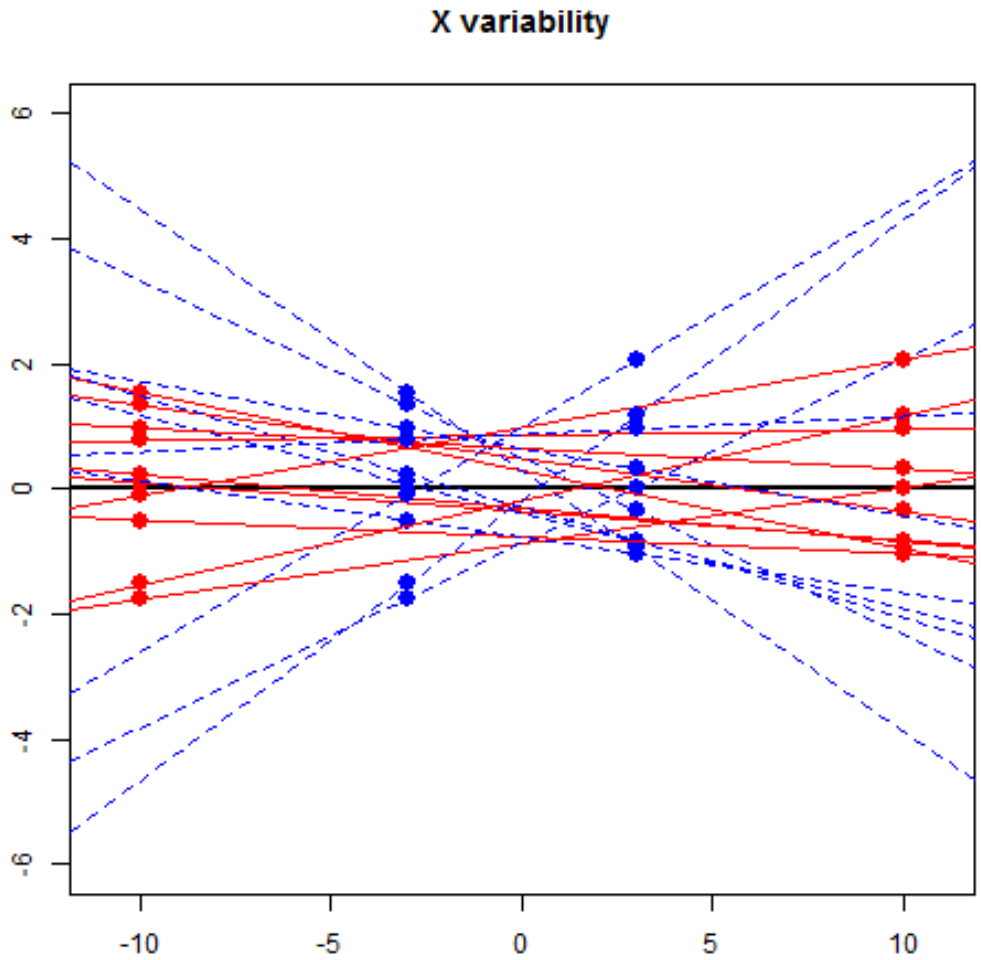
$$se(\hat{b}) = \frac{s_{reg}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

Effect of σ





Effect of X



Test Statistic for regression

We can form the following test statistic for regression

$$\frac{\hat{b} - b_0}{se(\hat{b})} \sim \mathcal{T}_{n-2}$$

In particular, we are typically interested in testing whether or not a relationship exists

$$H_0 : b = 0$$

$$H_A : b \neq 0$$

so b_0 is typically 0.

Confidence intervals for regression coefficients

We can form the confidence intervals in the normal way (and interpretation)

$$\hat{b} \pm t^{\star}_{n-2} se(\hat{b})$$

where $t^{\star}$ comes from a T distribution with $n-2$ degrees of freedom.

Example: El Nino Weather

For the El Nino data

$$n = 75$$

$$s_{reg} = 108.68$$

$$\sqrt{\sum_i (x_i - \bar{x})^2} = 7.76$$

$$se(\hat{b}) = 14.01$$

so to test whether or not there is a relationship between El Nino and snow pack,

$$\frac{-71.62}{14.01} = -5.1$$

which yields a p-value of 2.4×10^{-6} under a T_{73}

Example: El Nino Weather

Note that we tested whether or not a relationship exists, not necessarily the causal direction or whether the relationship is direct or indirect. We don't know what causes what, just that there is strong evidence for a linear relationship.

Example: El Nino Weather

Note that we tested whether or not a relationship exists, not necessarily the causal direction or whether the relationship is direct or indirect. We don't know what causes what, just that there is strong evidence for a linear relationship.

The reverse regression yields

$$\text{El Nino}_i = 1.78 - 0.0037\text{Snow}_i$$

and when testing the significance of the relationship, yields the exact same p-value

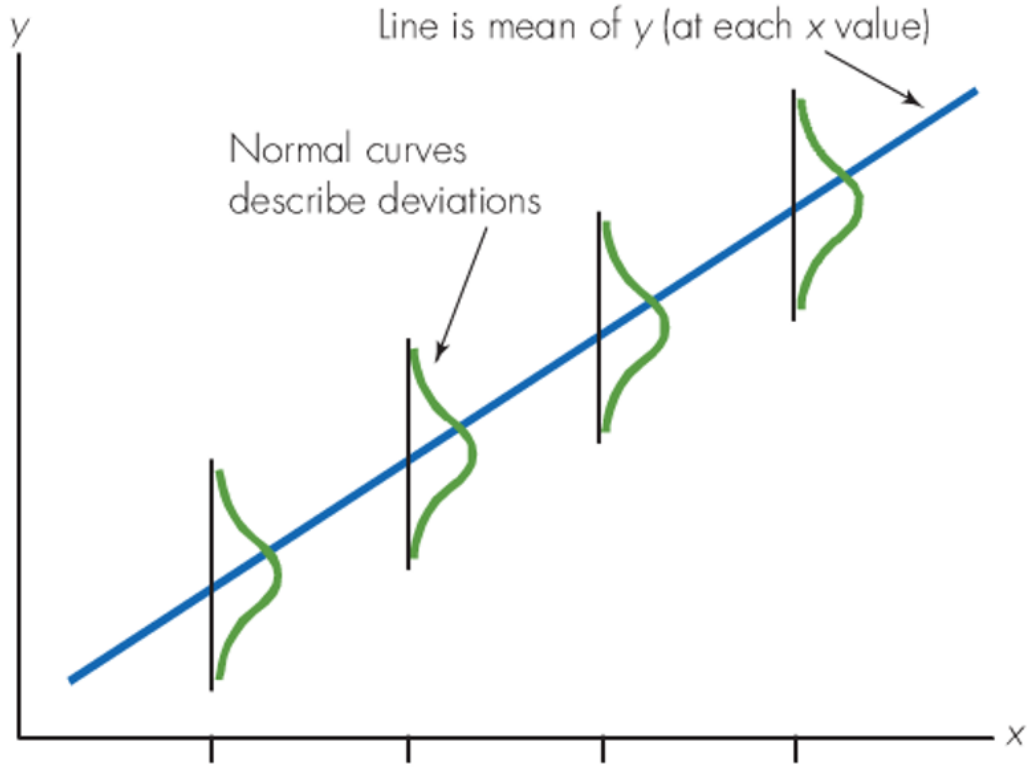
Prediction interval

Suppose, we are given an El Nino Index value for a given year, can we specify a reasonable range of values for the Max Snow pack?

Note that this is not a confidence interval. Confidence intervals are random intervals which try to cover an unchanging parameter. This is a **prediction interval** which seeks to describe the typical range of a the conditional random variable (snow pack given El Nino).

Predictive interval

This is somewhat like using the empirical rule for each distribution given X.



Predictive interval

The Predictive interval for a specific x value of x_j

$$\hat{y} \pm t^{\star} \sqrt{s_{reg}^2 + se(fit)^2}$$

with

$$se(fit) = s_{reg} \sqrt{\frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$$

Note that we typically are only able to make good predictions when the X value lies within our range of observed X values. This is called **interpolation**. When we try to predict Y for X values outside the range of observed X values, this is **extrapolation**.

Cofidence interval

Suppose, we are given an El Nino Index value for a given year, can we form a confidence interval for the average value of snow?

This is a proper confidence interval for the parameter $\mathbb{E}(Y|X)$. We form the confidence interval with the point estimate $\hat{y}$, standard error of $se(fit)$ and multiplier from a T_{n-2} distribution.

$$\hat{y} \pm t^{\star} se(fit)$$

Predictive interval: El Nino

What is an interval I would expect 95% of all snow packs to fall in when the El Nino index is 1.

$$s_{reg} = 108.68$$

$$se(fit) = 18.83$$

$$\hat{y} = 412.61$$

$$412.61 \pm 1.99\sqrt{108.68^2 + 18.83^2}$$

We expect 95% of all snow packs to fall within 193.11 and 632.11

Confidence interval: El Nino

What is a confidence interval for the mean snow pack when the El Nino Index is 1?

$$se(fit) = 18.83$$

$$\hat{y} = 412.61$$

$$412.61 \pm 1.99 \times 18.83$$

We are 95% confident that the mean snow pack when the El Nino Index is 1 is between 375.13 and 450.09