Inference for linear regression

Overview

Inference on simple linear regression: hypothesis tests

Regression diagnostics

Inference on simple linear regression: confidence intervals

Next class: diagnostic plots and statistics for unusual points

Inference for linear regression: underlying model

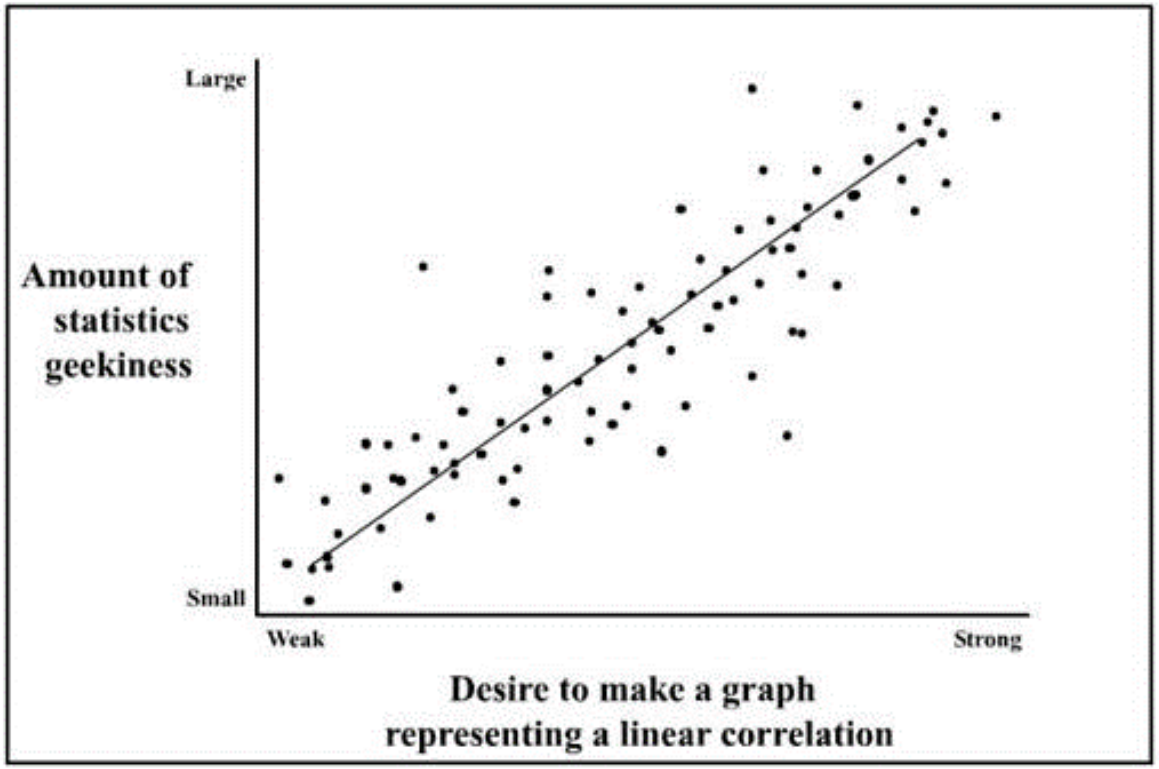
Linear regression

Regression is method of using one variable ***x*** to predict the value of a second variable ***y***

- i.e., $\hat{y} = f(x)$

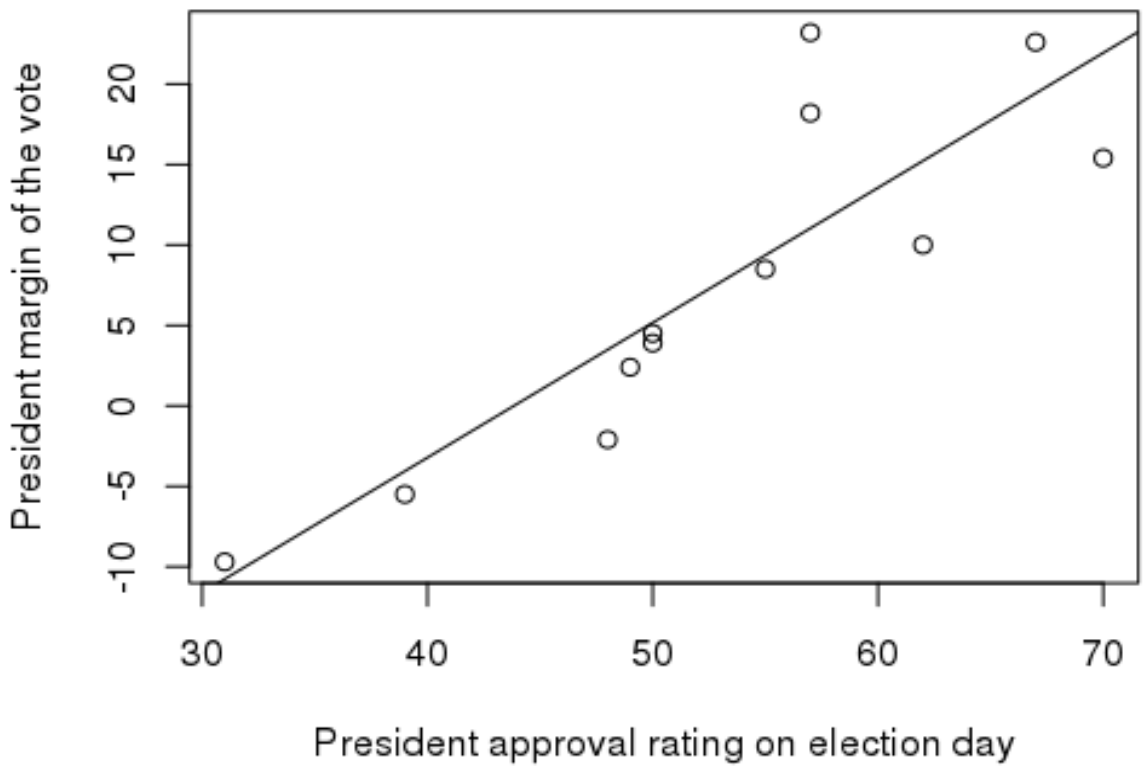
In **linear regression** we fit a line to the data, called the **regression line**

- In simple linear regression, we use a single variable x, to predict y



Approval rating vote margin regression line

From last 12 US president's running for reelection



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

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

R: `lm(y ~ x)`

$\hat{\beta}_0 = -36.76$

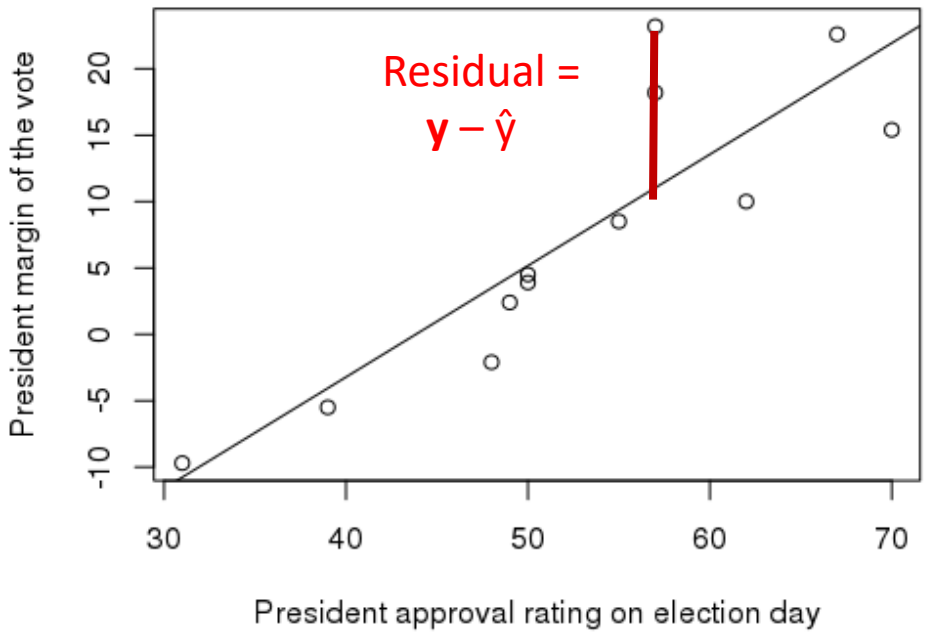
$\hat{\beta}_1 = 0.84$

$\hat{y} = -36.76 + .84 \cdot x$

Minimizing the sum of the squared residuals to find the regression coefficients

To find the regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ we minimize the **residual sum of squares (RSS)**

- The residual sum of squares is also called the **error sum of squares (SSE)**



$$residual = e_i = y_i - \hat{y}_i$$

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

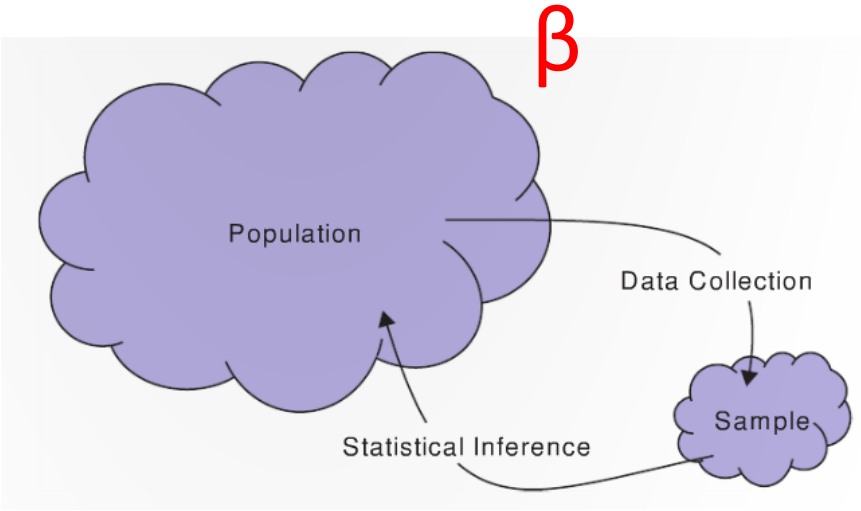
$$= \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x))^2$$

R: `lm(y ~ x)`

Inference for simple linear regression

The letter **b** or $\hat{\beta}$ is typically used to denote the slope ***of the sample***

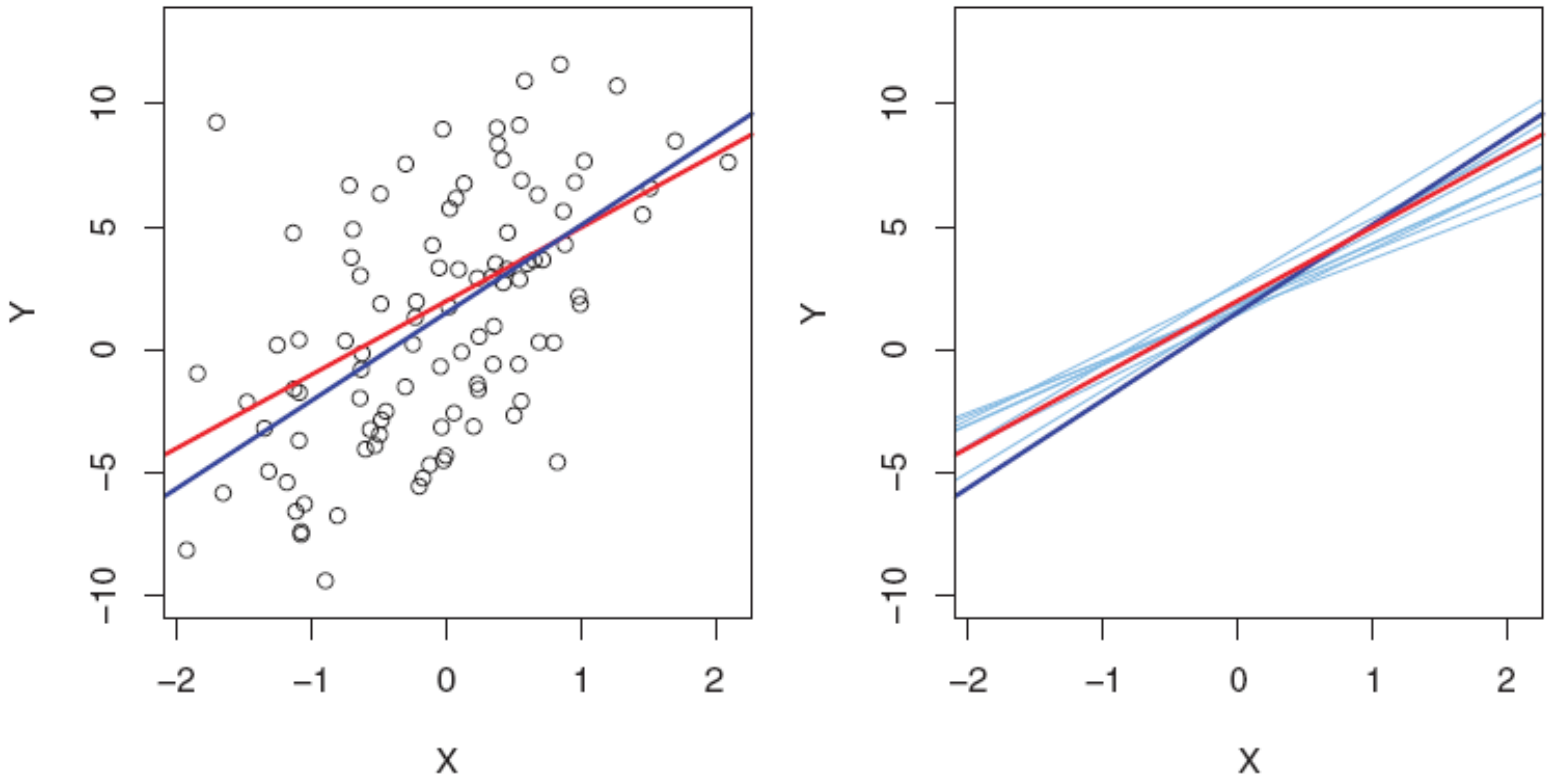
The Greek letter **β** is used to denote the slope ***of the population***



b $\hat{\beta}$

Population: β

Sample estimates: b $\hat{\beta}$



Linear regression underlying model

Intercept Slope } *Parameters*

True regression line: $\mu_Y = \beta_0 + \beta_1 x$

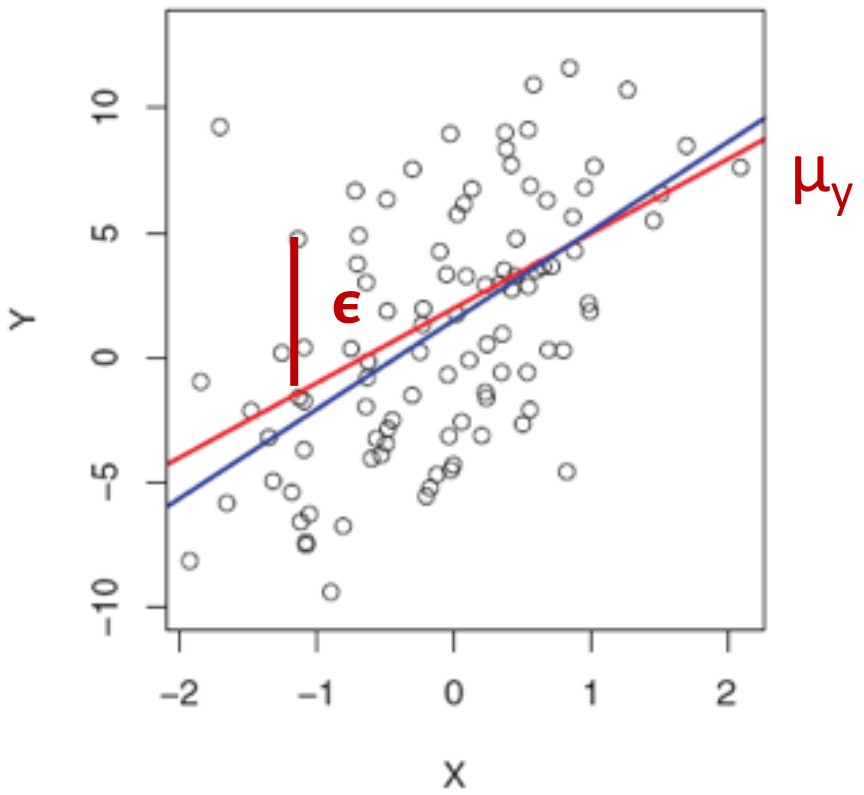
Error

Observed data point: $Y = \beta_0 + \beta_1 x + \epsilon$

$= \mu_Y + \epsilon$

Errors ϵ are the difference between the **true regression line** μ_y and observed data points Y

- **ϵ** = Y - μ_y



Linear regression underlying model

Intercept Slope } *Parameters*

True regression line: $\mu_Y = \beta_0 + \beta_1 x$

Error

Observed data point: $Y = \beta_0 + \beta_1 x + \epsilon$

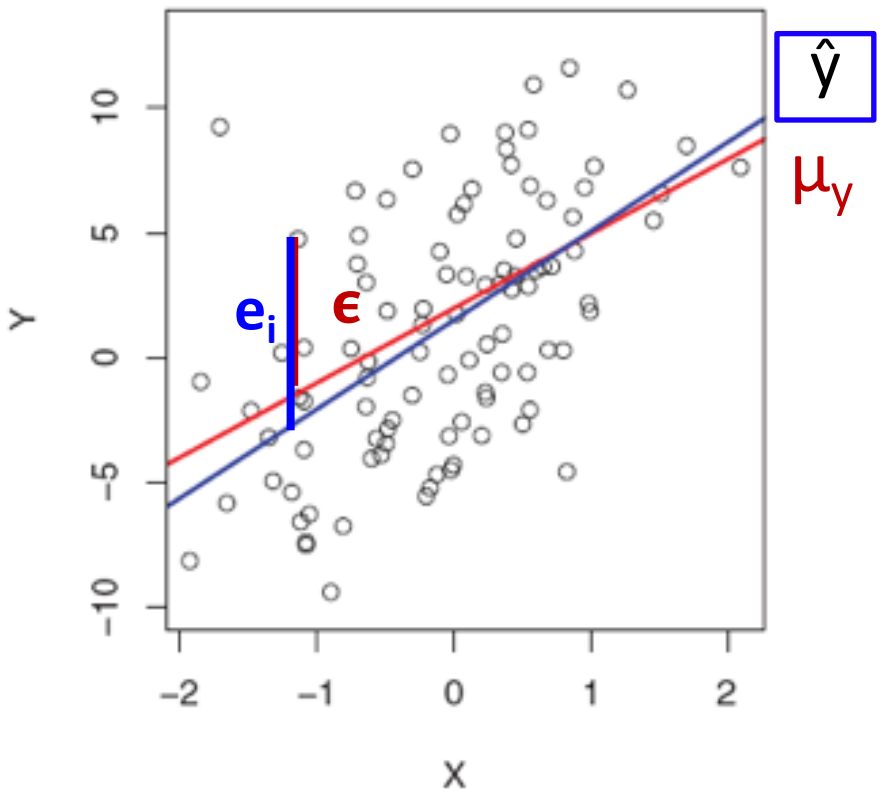
Estimated regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Errors ϵ are the difference between the **true regression line** μ_y and observed data points Y

- $\epsilon = Y - \mu_y$

Residuals e_i are the difference between the **estimated regression line** $\hat{y}$ and observed data points Y

- $e_i = Y - \hat{y}$



Linear regression underlying model

Intercept Slope } *Parameters*

True regression line: $\mu_Y = \beta_0 + \beta_1 x$

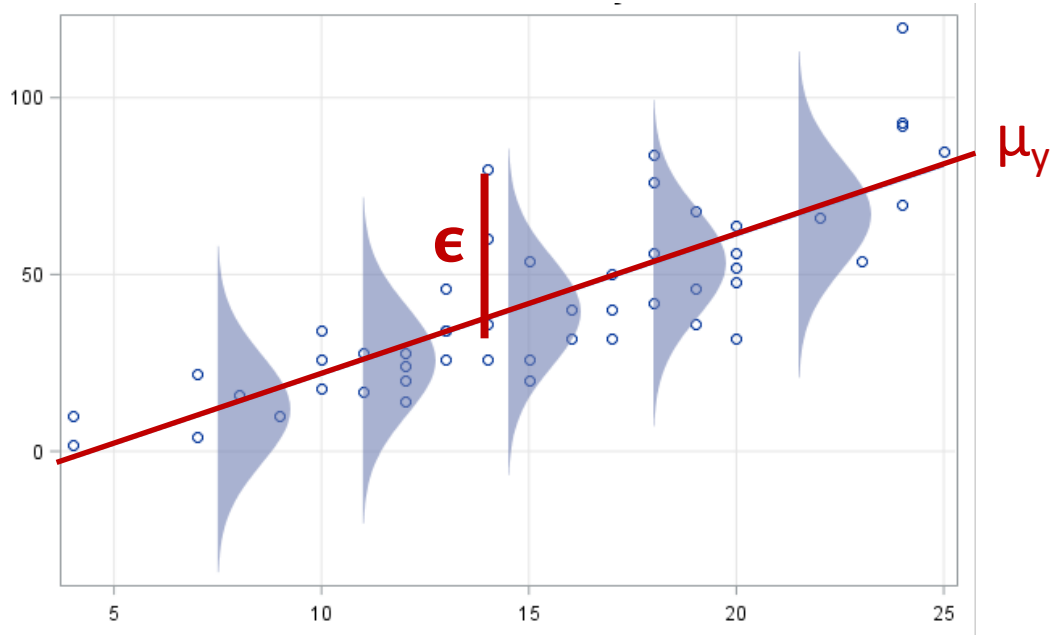
Error

Observed data point: $Y = \beta_0 + \beta_1 x + \epsilon$ $\epsilon \sim N(0, \sigma_\epsilon)$

Errors ϵ are the difference between the **true regression line** μ_y and observed data points Y

- $\epsilon = Y - \mu_y$

We will *assume* that the errors are **normally distributed**



Recap: Errors vs. residuals

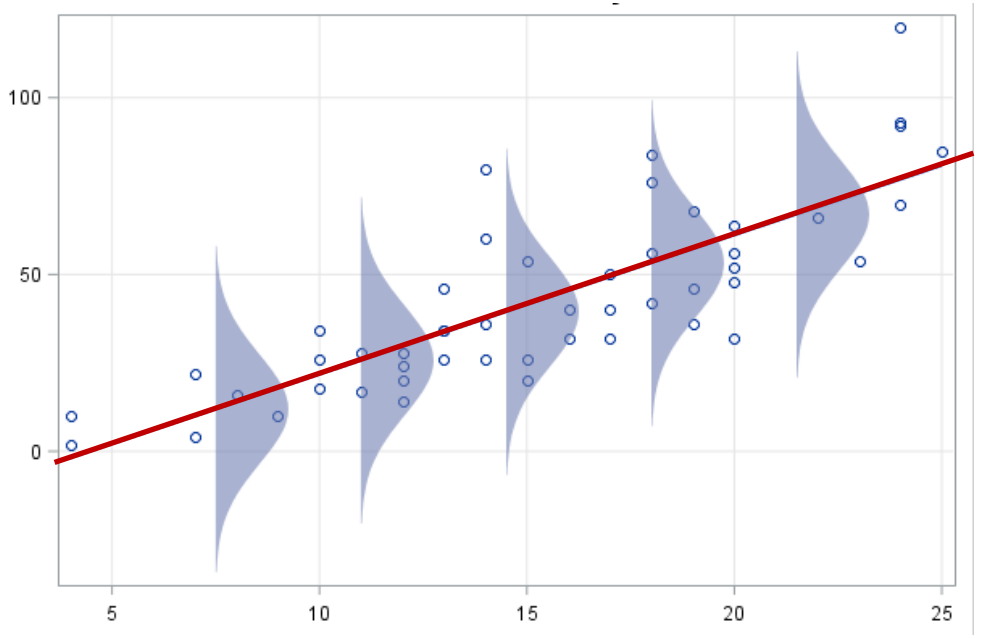
Data: $Y = \beta_0 + \beta_1 x + \epsilon \qquad \epsilon \sim N(0, \sigma_\epsilon)$

True model: $\mu_Y = \beta_0 + \beta_1 x$

- Errors: $\epsilon = Y - \mu_y$

Estimated model: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

- Residuals: $e_i = Y - \hat{y}$



Standard deviation of the errors: σ_ε

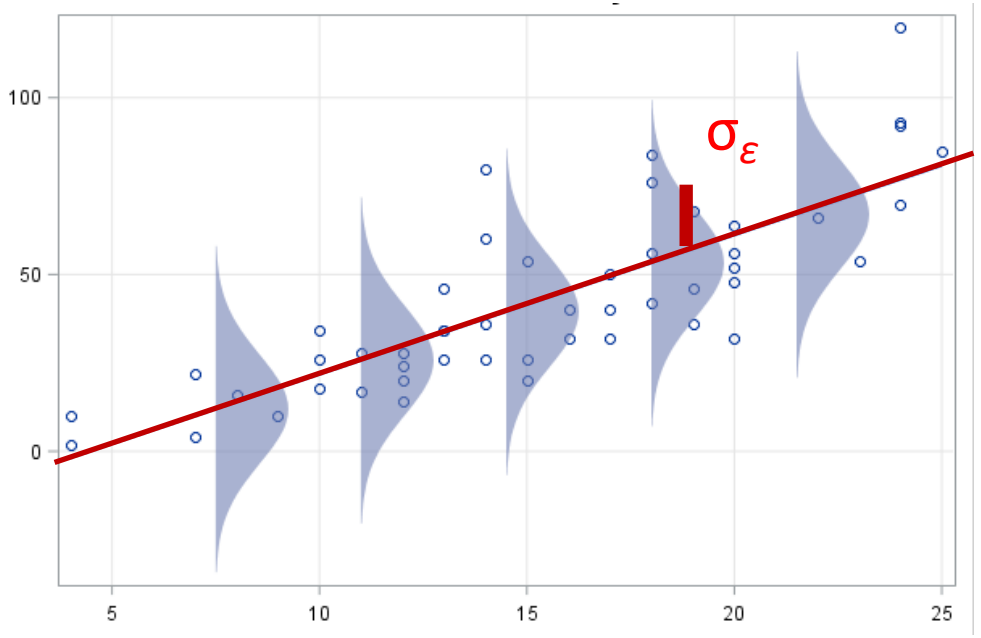
The standard deviation of the errors is denoted $\boldsymbol{\sigma_\varepsilon}$

We can use the **standard deviation of residuals** as an estimate standard deviation of the errors σ_ε. This is known as the…

- **regression standard error (RSE)**

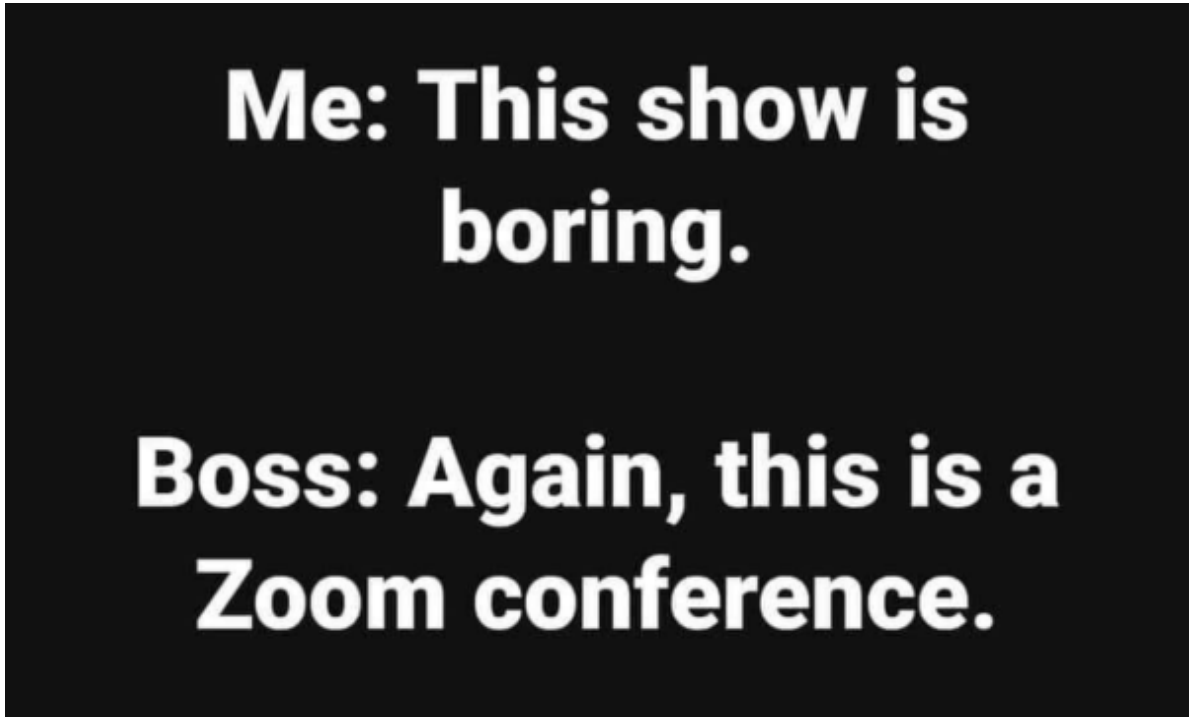
$$\hat{\sigma}_\epsilon = \sqrt{\frac{1}{n-2}RSS}$$

$$= \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$



How are we feeling?

Theoretical models are boring

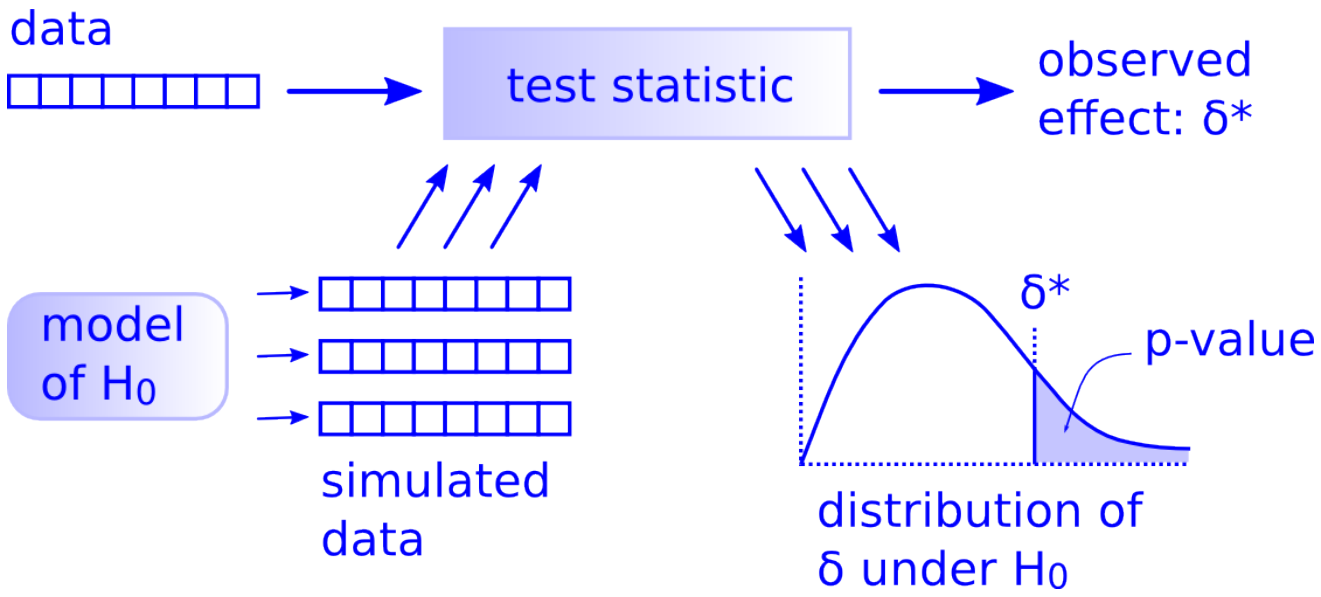


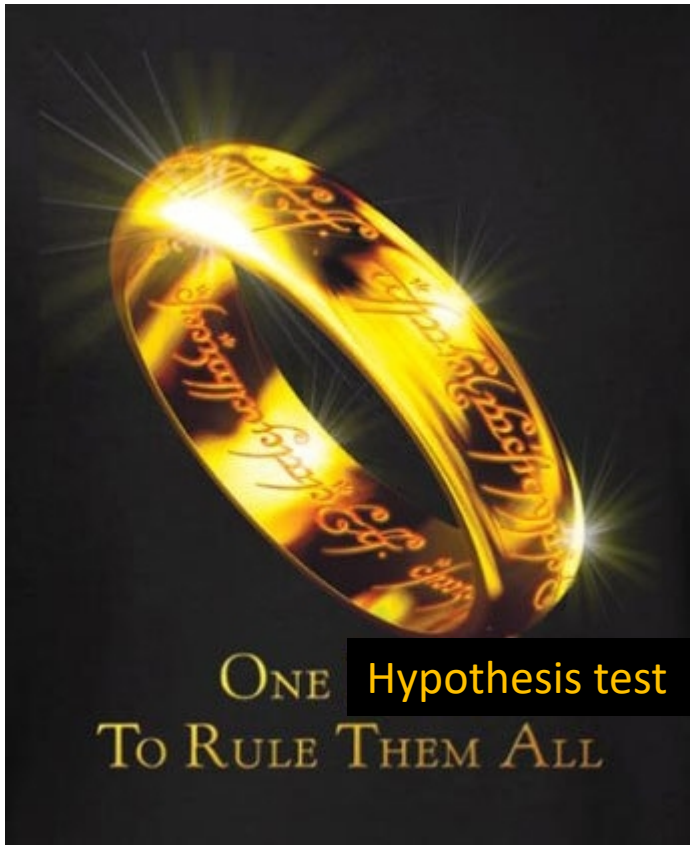
Let’s look at some real data on faculty salaries!

Inference for linear regression: hypothesis tests

Hypothesis test for regression coefficients

There is only one [hypothesis test](#)!





Hypothesis test for regression coefficients

We can run hypothesis tests to assess whether there is a relationship between y and x, and calculate p-values

- H_0: $\beta_1 = 0$ (slope is 0, so no relationship between x and y
- H_A: $\beta_1 \neq 0$

One type of hypothesis test we can run is based on a t-statistic: $t = \frac{\hat{\beta}_1 - 0}{SE_{\hat{\beta}_1}}$

- The t-statistic comes from a t-distribution with n - 2 degrees of freedom

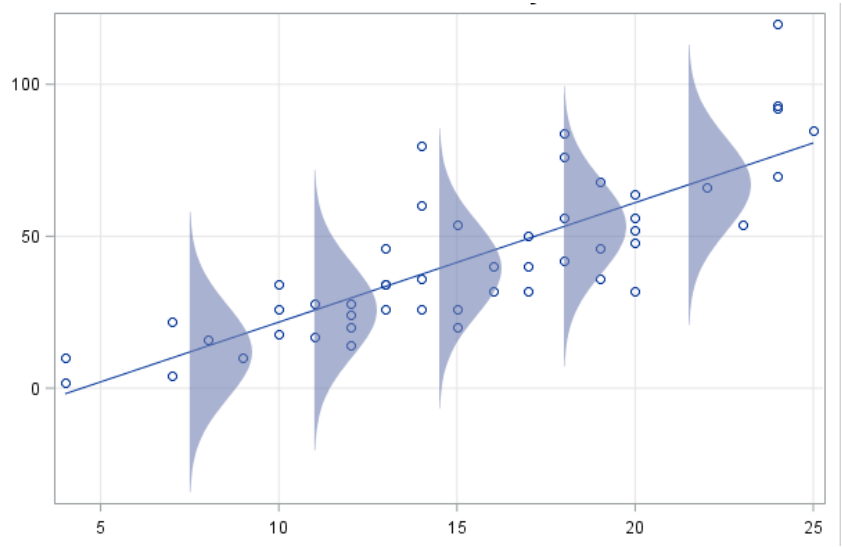
$$SE_{\hat{\beta}_1} = \frac{\sigma_\epsilon}{\sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$SE_{\hat{\beta}_0} = \sigma_\epsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2)}}$$

Inference using parametric methods

When using parametric methods, we usually make the following assumptions:

- **Normality**: residuals are normally distributed around the predicted value ŷ
- **Homoscedasticity**: constant variance over the whole range of x values
- **Linearity**: A line can describe the relationship between x and y
- **Independence**: each data point is independent from the other points



These assumptions are usually checked after the models are fit using ‘regression diagnostic’ plots.

Let’s look at inference for simple linear regression in R

Back to faculty salaries…

Inference for linear regression: confidence intervals

We can estimate three types of intervals for a regression:

1. Confidence intervals for the regression coefficients: β_0 and β_1
2. Confidence intervals for the full line μ_Y
3. Prediction intervals where most of the data is expected

Confidence intervals for regression coefficients

For the slope coefficient , the confidence interval is: $\hat{\beta}_1 \pm t^* \cdot SE_{\hat{\beta}_1}$

Where: $SE_{\hat{\beta}_1} = \frac{\sigma_\epsilon}{\sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}}$

t* is the critical value for the t_{n-2} density curve needed to obtain a desired confidence level

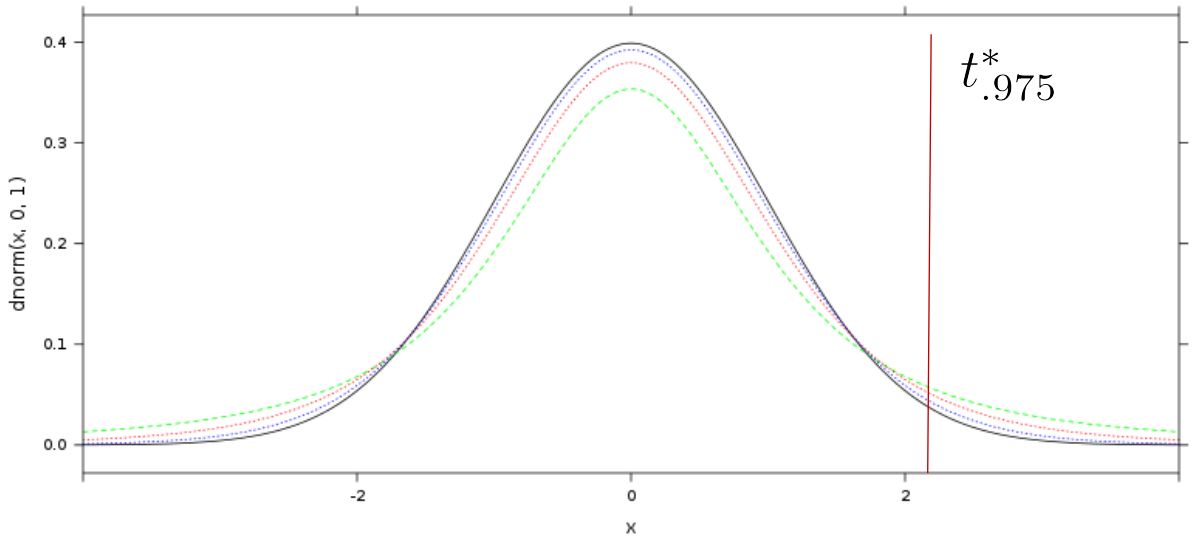
qt(.975, df)

N(0, 1)

df = 2

df = 5

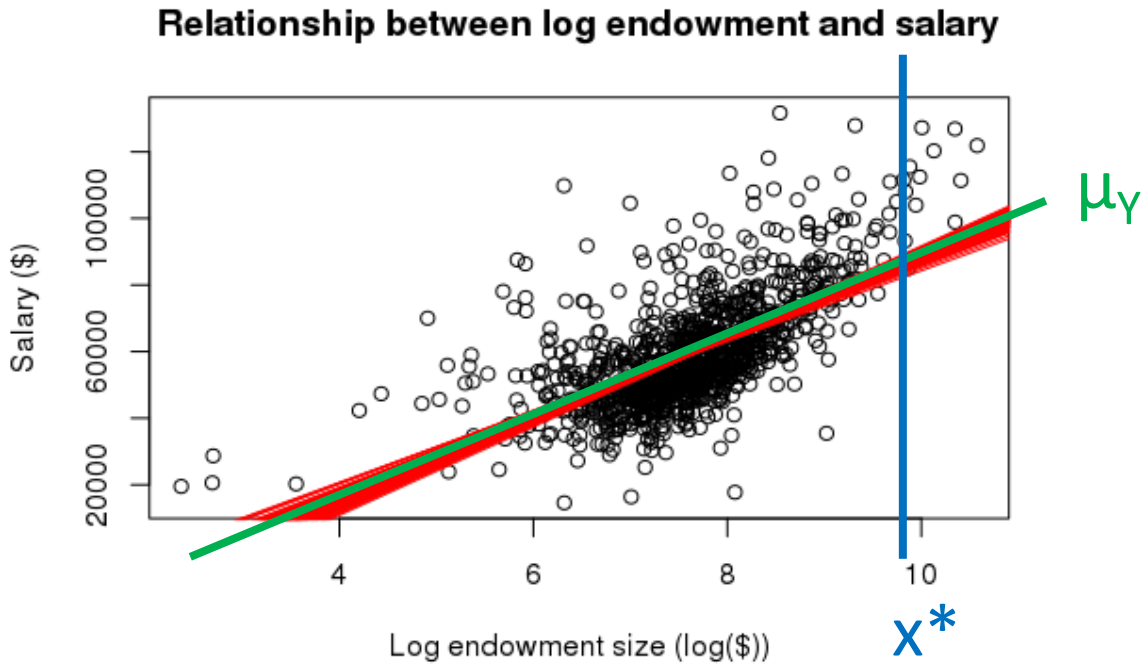
df = 15



Confidence intervals for the regression line μ_Y

A confidence interval for the mean response for the ***true regression line*** μ_Y when X = x* is:

$$\hat{y} \pm t^* \cdot SE_{\hat{\mu}} \quad \text{where} \quad SE_{\hat{\mu}} = \sigma_\epsilon \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}}$$



Note:

- There is more uncertainty at the ends of the regression line
- The confidence interval for the regression line μ_Y is different than the confidence interval for slope β_1

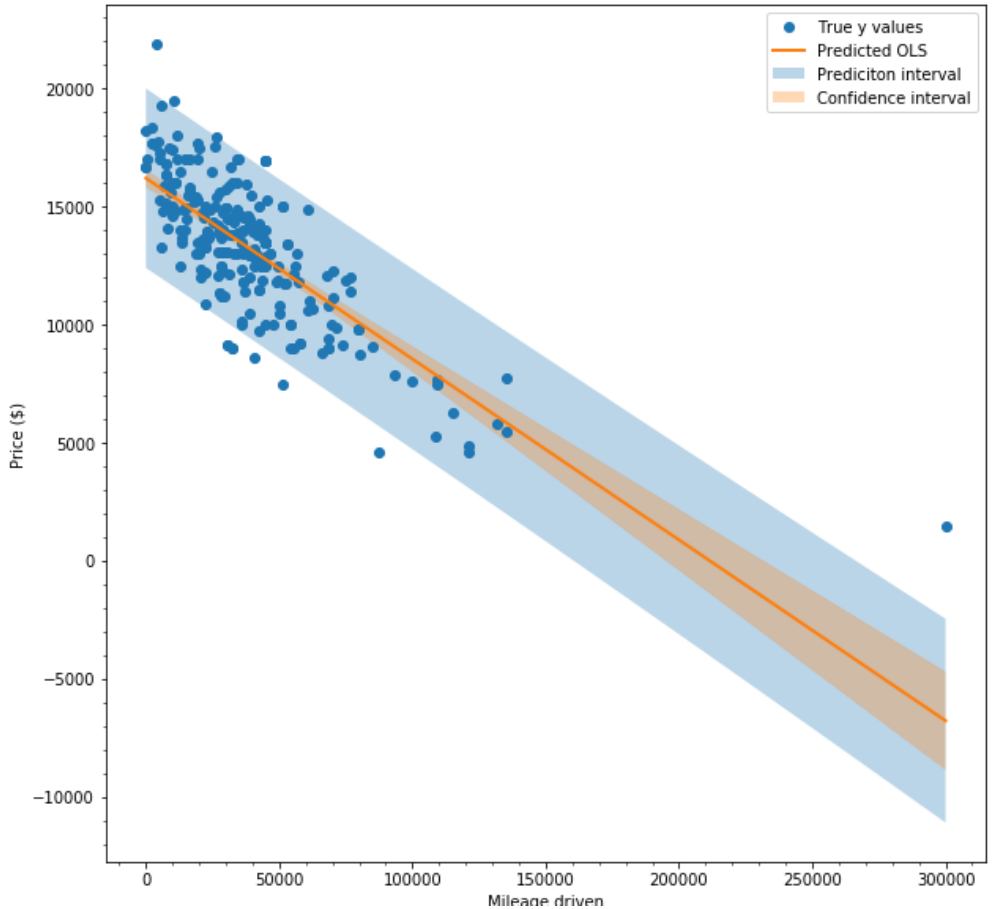
Prediction intervals

Confidence intervals give us a measure of uncertain about our the true relationship between x and y for:

- The true regression slope β_1
- The true regression line μ_Y

Prediction intervals give us a range of plausible values for y

- i.e., 95% of our y's with be within this range



Prediction intervals

A **prediction intervals** for the y can be calculated using:

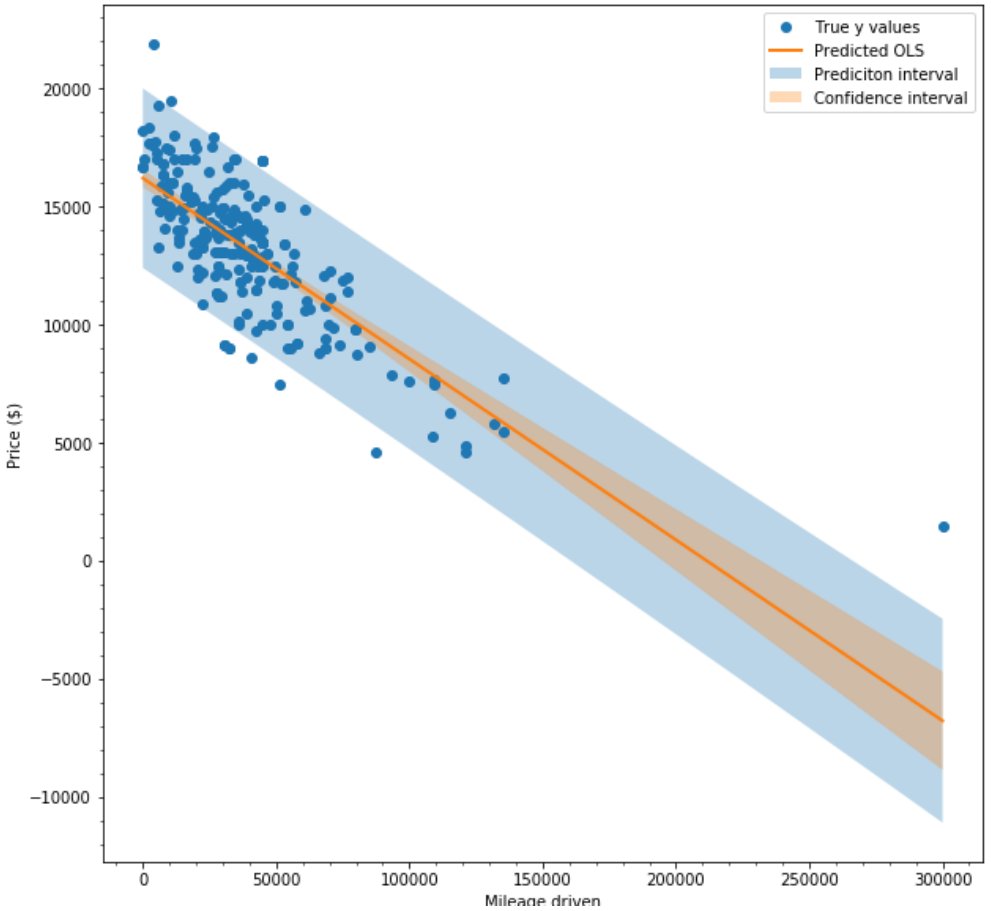
$$\hat{y} \pm t^* \cdot SE_{\hat{y}}$$

where

$$SE_{\hat{y}} = \sigma_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}}$$

Due to y's scattering around the true regression line

Due to uncertainty in where the true regression line is



Summary of confidence and prediction intervals

1. CI for Slope β $\hat{\beta}_1 \pm t^* \cdot SE_{\hat{\beta}_1}$ $SE_{\hat{\beta}_1} = \sigma_\epsilon \sqrt{\frac{1}{\Sigma_{i=1}^n (x_i - \bar{x})^2}}$

2. CI for regression line μ_Y at point x*

$$\hat{y} \pm t^* \cdot SE_{\hat{\mu}} \qquad SE_{\hat{\mu}} = \sigma_\epsilon \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\Sigma_{i=1}^n (x_i - \bar{x})^2}}$$

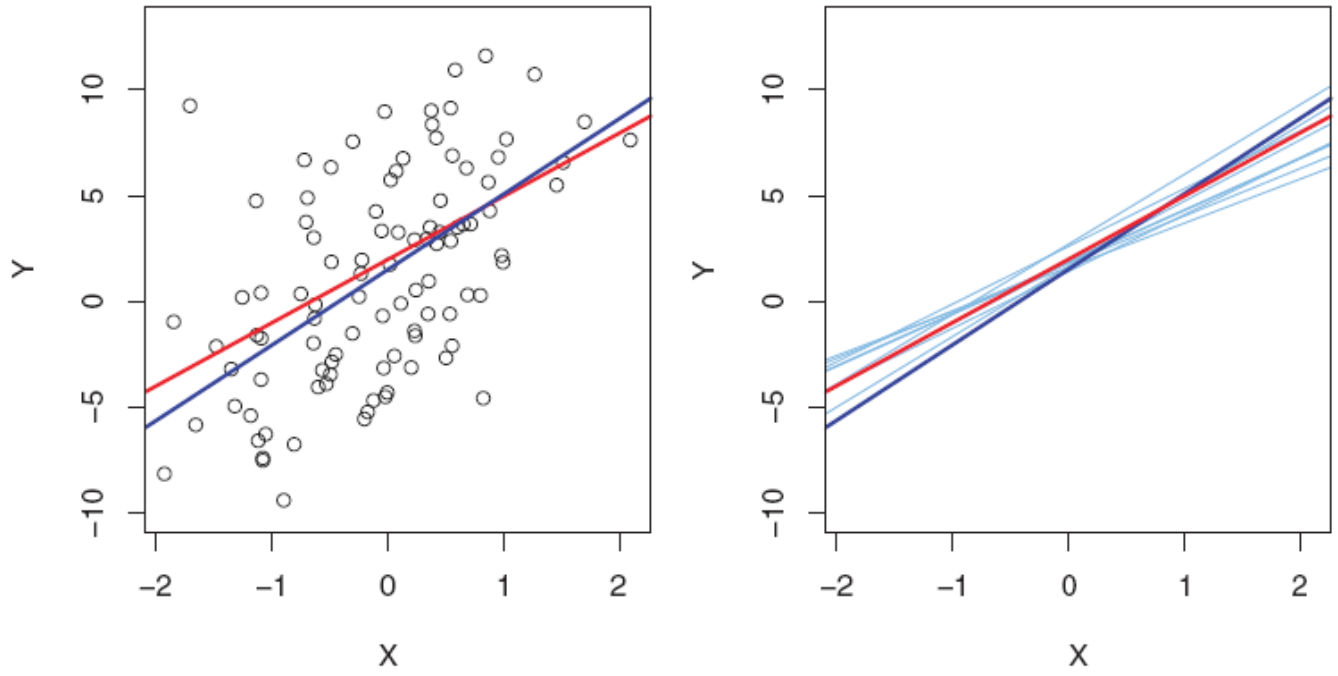
3. Prediction interval y

$$\hat{y} \pm t^* \cdot SE_{\hat{y}} \qquad SE_{\hat{y}} = \sigma_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\Sigma_{i=1}^n (x_i - \bar{x})^2}}$$

Resampling methods for inference in regression

We can also use resampling methods to estimate run hypothesis tests and create confidence intervals for the regression coefficients

- Bootstrap
- Permutation test



Let’s look at inference for simple linear regression in R

More faculty salary data