Simple linear regression

Overview

Simple linear regression

Inference for simple linear regression

- Errors and residuals
- Hypothesis tests for regression coefficients
- If there is time: confidence and prediction intervals

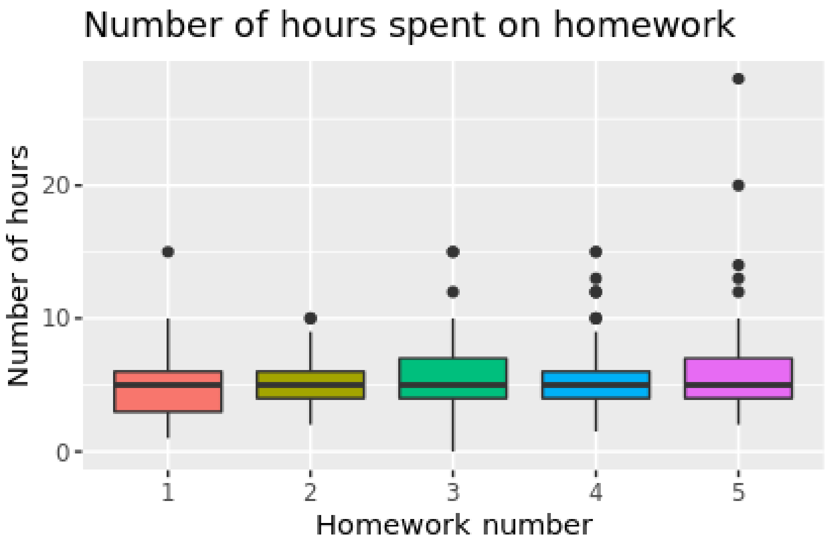
Announcements

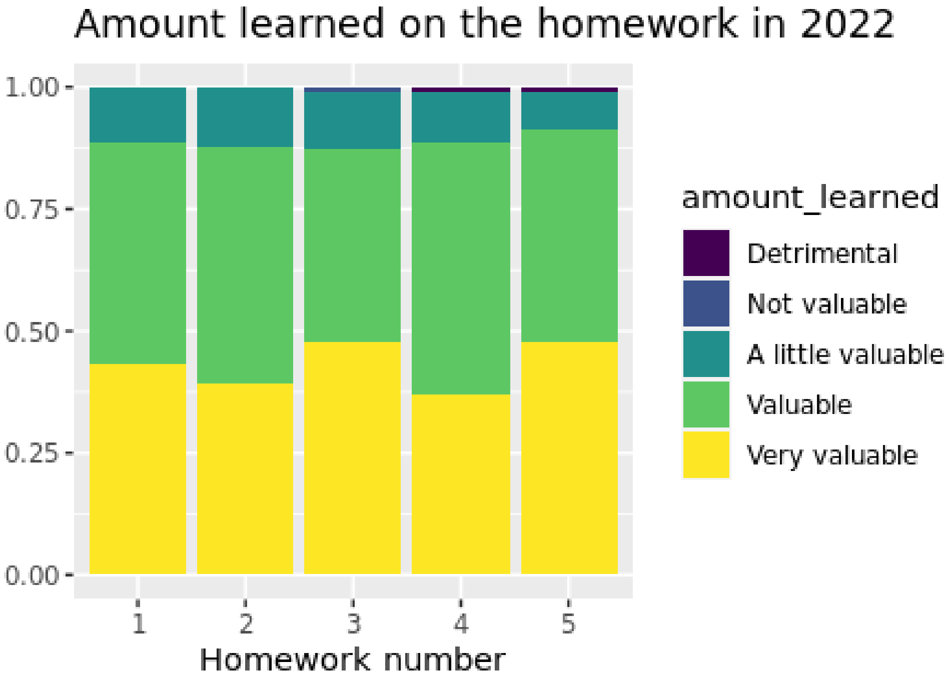
Homework 5 has been graded

- You have a week for regrade requests
- Midterm exams should be graded soon

Homework 6 is out

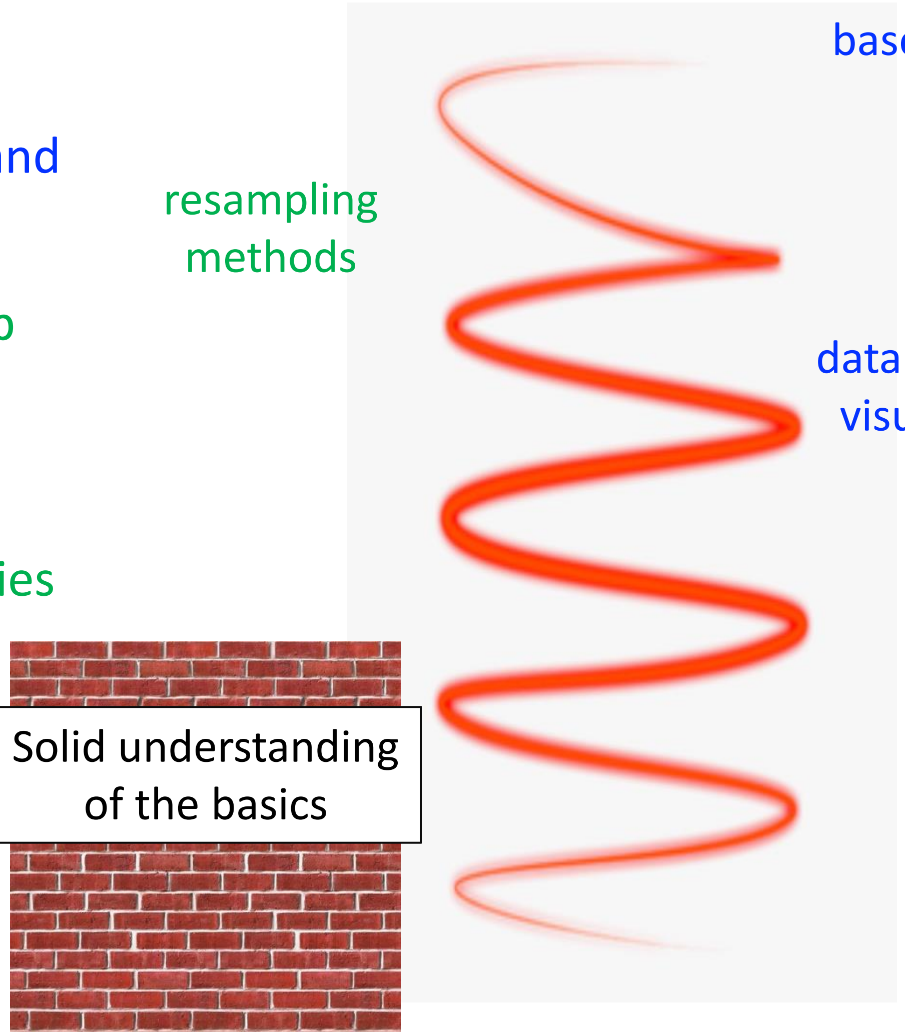
- Due Sunday October 30th at 11pm





Where we are: completed

1	Sep 1	Course overview, introduction to R, descriptive statistics
2	Sep 6-8	Review of central statistical concepts and exploratory analysis using R
3	Sep 13-15	Confidence Intervals and the bootstrap
4	Sep 20-22	Review of hypothesis tests and permutation tests in R
5	Sep 27-29	Parametric, non-parametric and theories of hypothesis testing
6	Oct 4-6	Data manipulation and visualization
7	Oct 11-14	Review and midterm exam
8	Oct 18-21	Odds and ends, October break



Analysis

R

base R

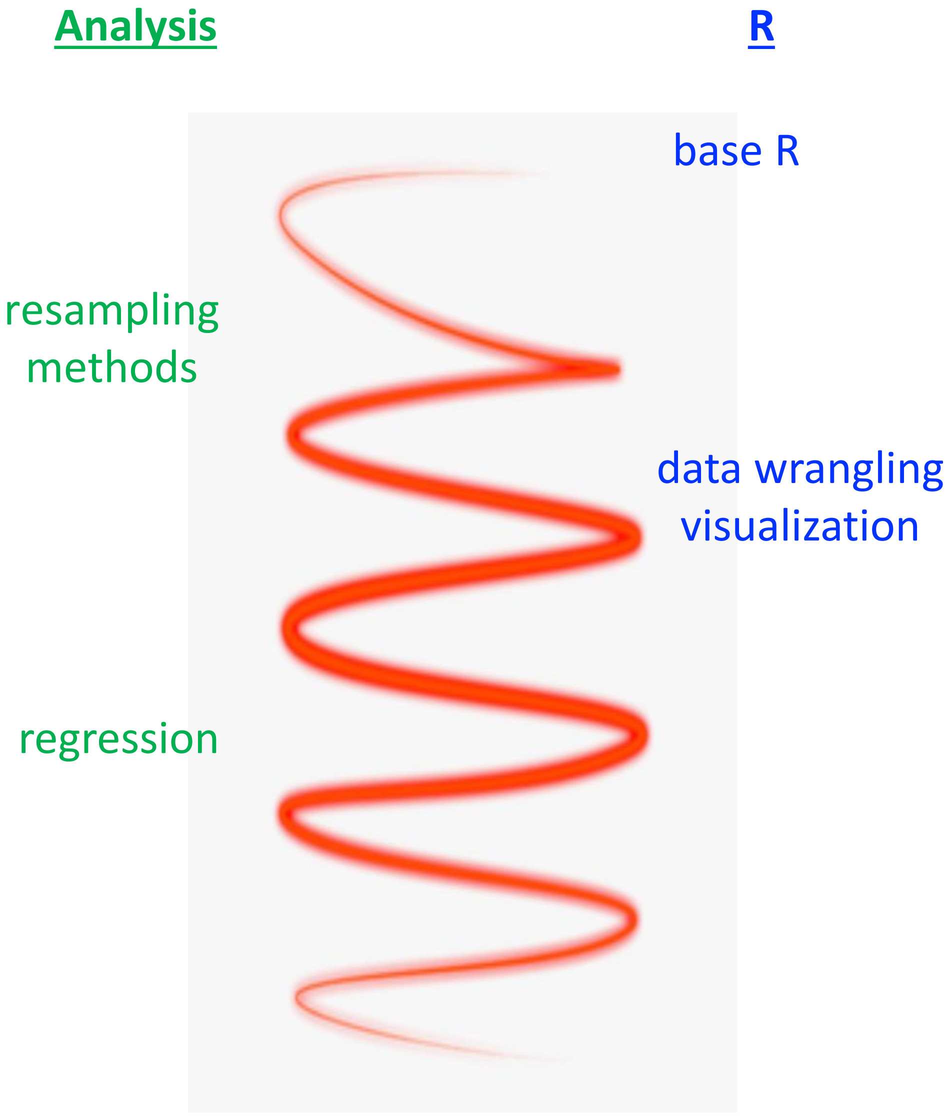
resampling methods

data wrangling visualization

Solid understanding of the basics

Where we are: up next

9	Oct 25-27	Simple linear regression
10	Nov 1-3	Multiple regression
11	Nov 8-10	Model selection and logistic regression



Analysis

R

base R

resampling methods

data wrangling visualization

regression

Next: building linear models

We will use these models to:

1. Make accurate predictions
2. Understand the relationship between explanatory variables x_i's and a response variable y

Linear regression

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

$$\hat{y} = f(x)$$

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

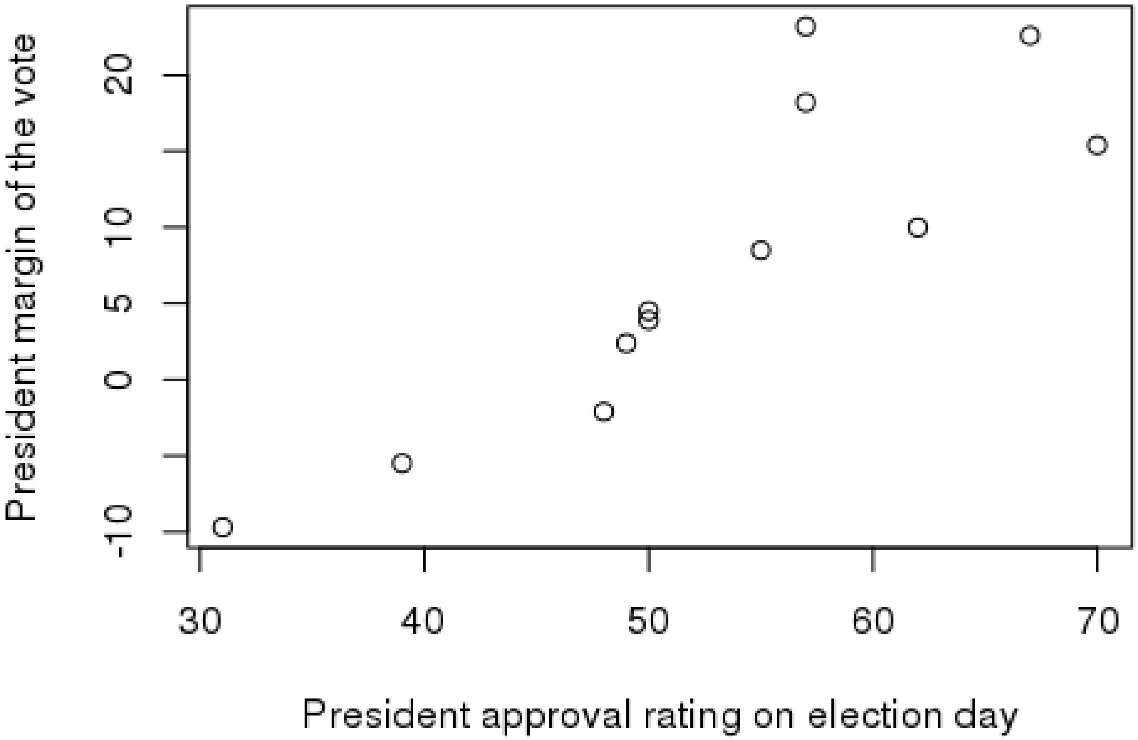
Motivation: Predicting the 2020 election

Predict the margin of the popular vote based on the president's approval rating

Data from an article on the 2012 election on the Five Thirty Eight website

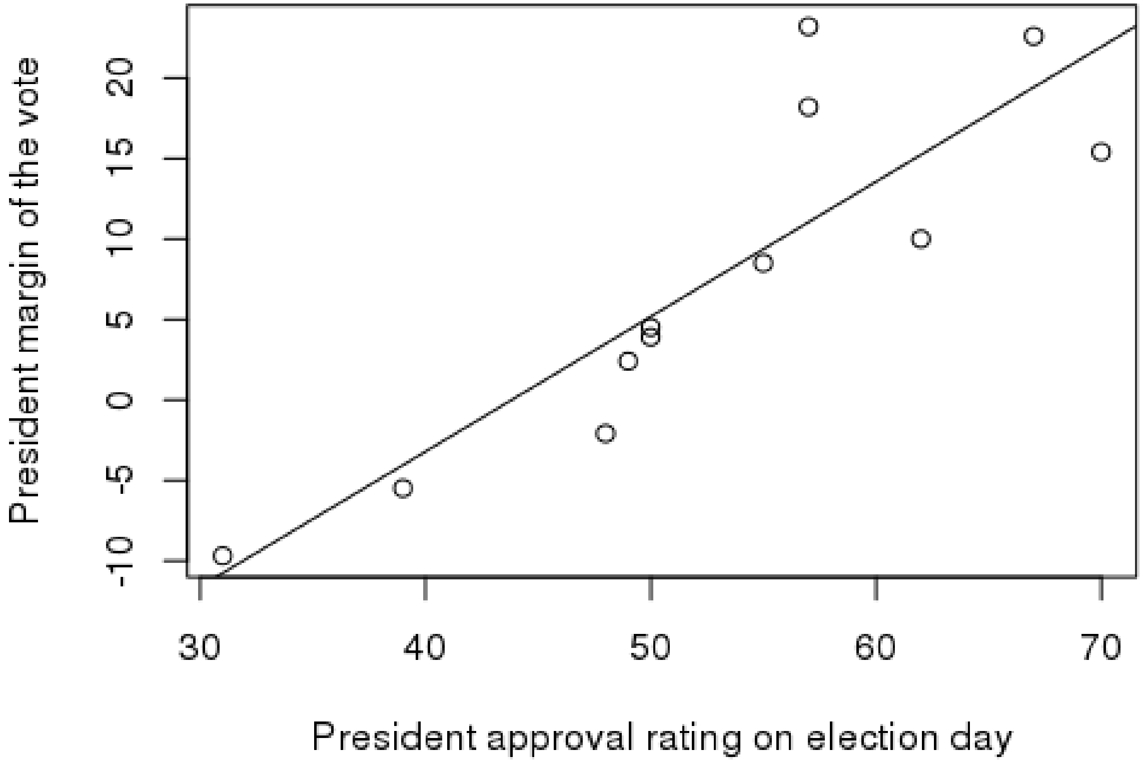
Approval rating vote margin regression line

From previous 12 US president’s running for reelection



Approval rating vote margin regression line

From previous 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

$\hat{y} = -36.76 + 0.84 \cdot x$

Approval rating vote margin regression line

1. If a president had a 0% approval rating, what percent of the vote margin does this model predict the president would get?

2. If a president's approval rating increased by 1%, how much of would the president's margin of the vote increase by?

3. At what presidential approval level would there be an exactly even split of the vote?

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

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

```
R: lm(y ~ x)
```

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

$$\hat{\beta}_1 = 0.84$$

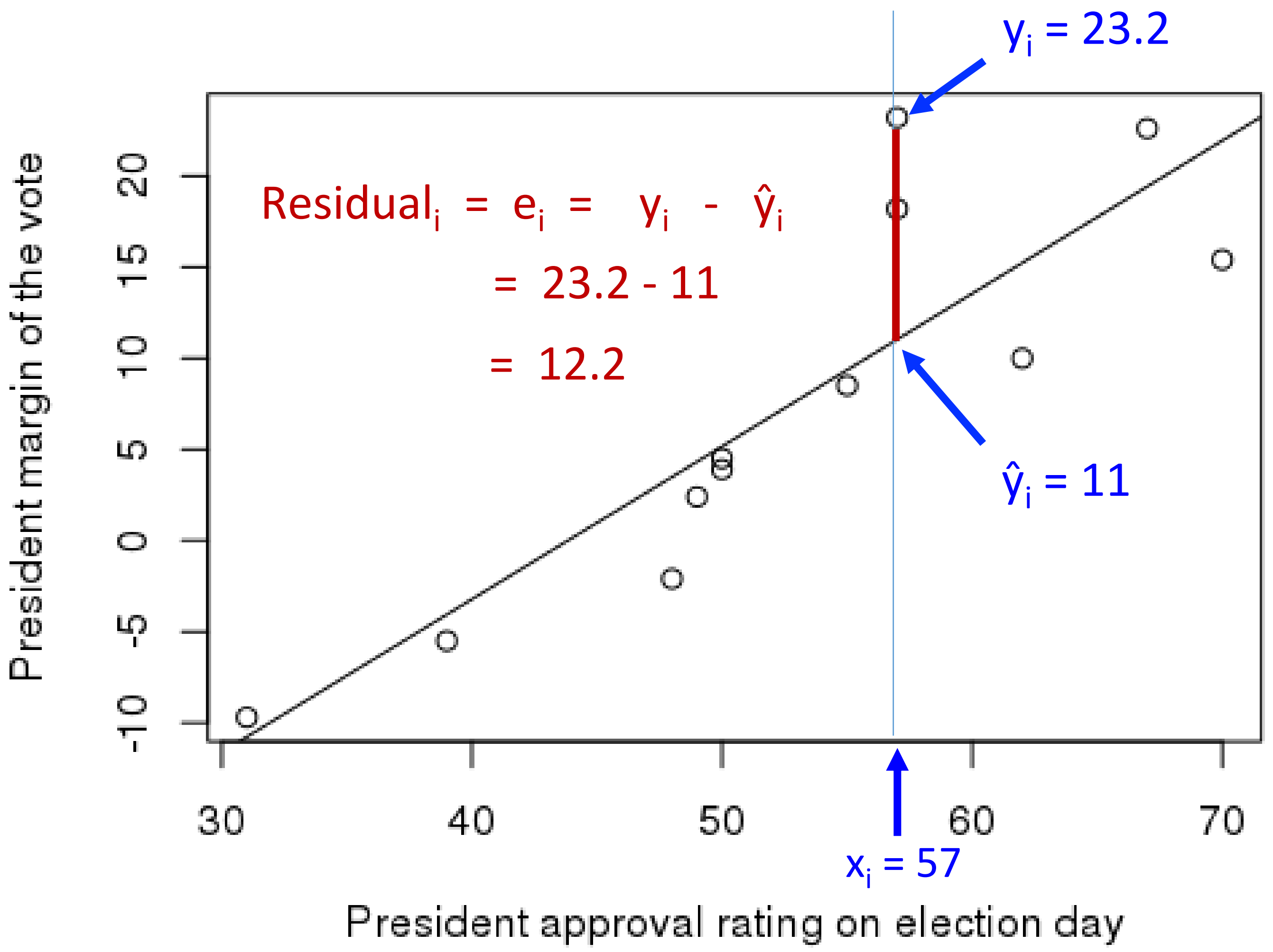
$$\hat{y} = -36.76 + 0.84 \cdot x$$

Residuals

The **residual** at a data value is the difference between the observed (y) and predicted value of the response variable

$$\begin{tikzcd}
\mathit{Residual}_i \arrow[d, leftrightarrow] & = & \mathit{Observed}_i \arrow[d, leftrightarrow] & - & \mathit{Predicted}_i \arrow[d, leftrightarrow] \\
e_i & = & y_i & - & \hat{y}_i
\end{tikzcd}$$

Approval rating vote margin regression line



$\text{Residual}_i = e_i = y_i - \hat{y}_i$

$= 23.2 - 11$

$= 12.2$

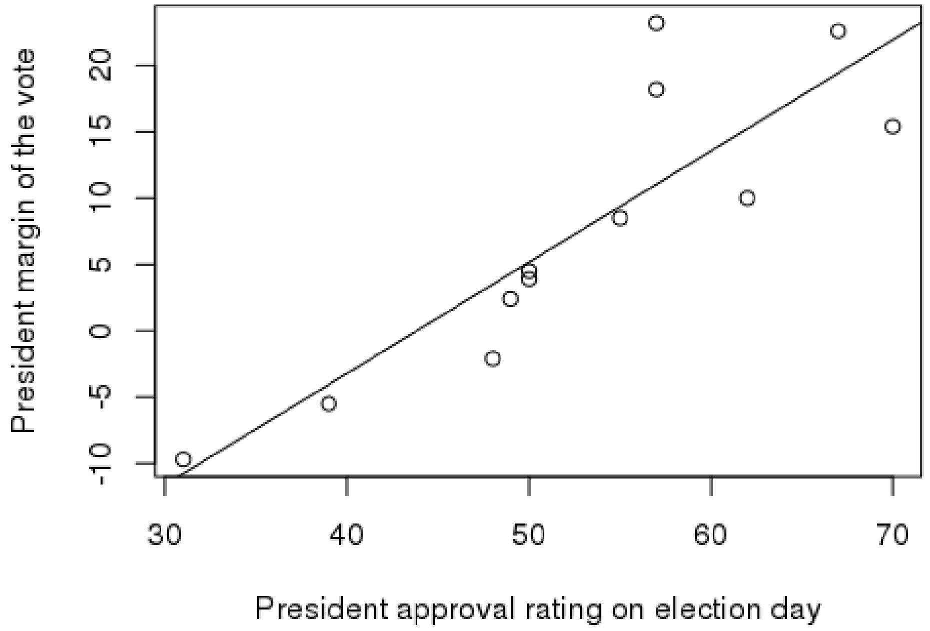
Approval rating vote margin regression line

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

Approval ***x***	**Margin obs** ***y***	**Margin pred** ***ŷ***	**Residuals** ***e = y - ŷ***
62	10	15.23	-5.23
50	4.5	5.17	-0.67
70	15.4	21.94	-6.54
67	22.6	19.43	3.17
57	23.2	11.04	12.16
48	-2.1	3.49	-5.59
31	-9.7	-10.76	1.06
57	18.2	11.04	7.16

Line of ‘best fit’

The **least squares line**, also called **‘the line of best fit’**, is the line which minimizes the sum of squared residuals



Try to find the line of best fit

Approval rating vote margin regression line

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

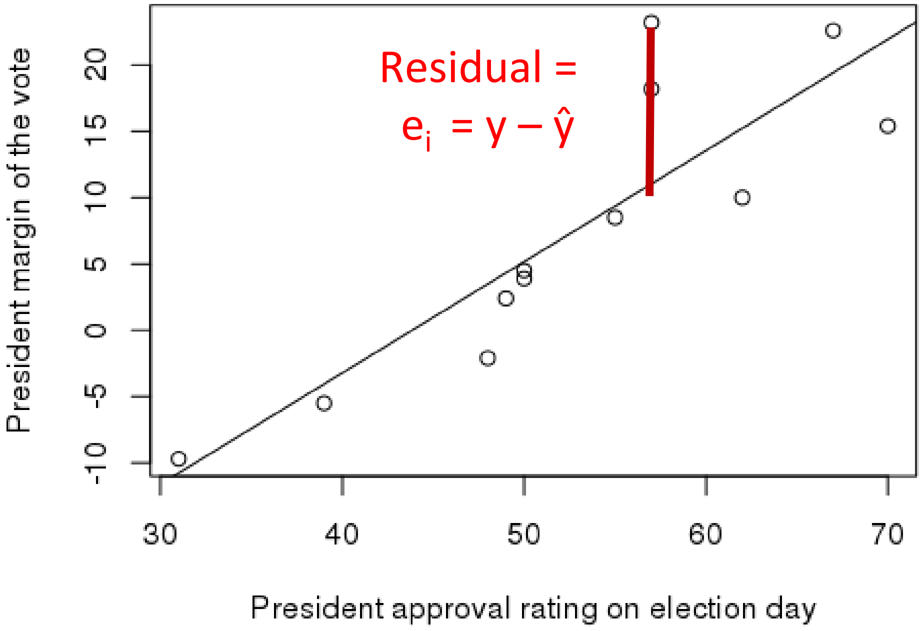
Approval ***x***	**Margin obs** ***y***	**Margin pred** ***ŷ***	**Residuals** ***e = y - ŷ***	**Residuals²** ***e² = (y - ŷ)²***
62	10	15.23	-5.23	27.40
50	4.5	5.17	-0.67	0.45
70	15.4	21.94	-6.54	42.81
67	22.6	19.43	3.17	10.07
57	23.2	11.04	12.16	147.84
48	-2.1	3.49	-5.59	31.29
31	-9.7	-10.76	1.06	1.13
57	18.2	11.04	7.16	51.25

Q: Why do we minimize the sum of ***squared*** residuals rather than just the sum of residuals?

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 **sum of squared residuals**

- We will use the notation **SSRes** to denote the some of squared residuals
  - (The residual sum of squares is also called the **"error sum of squares" (SSE)**)



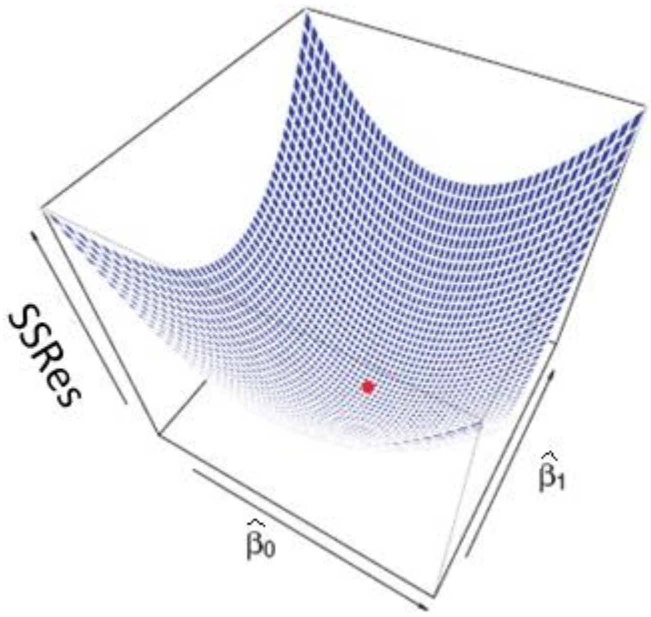
$$residual = e_i = y_i - \hat{y_i}$$

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

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

R: `lm(y ~ x)`

How do we minimize the SSE?



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

How do we find $\hat{\beta}_0, \hat{\beta}_1$?

Calculus and linear algebra:

- Take the derivative, set to 0 and solve
- This mathematical convenience is why the squared loss is so commonly used

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Basic regression caution # 1

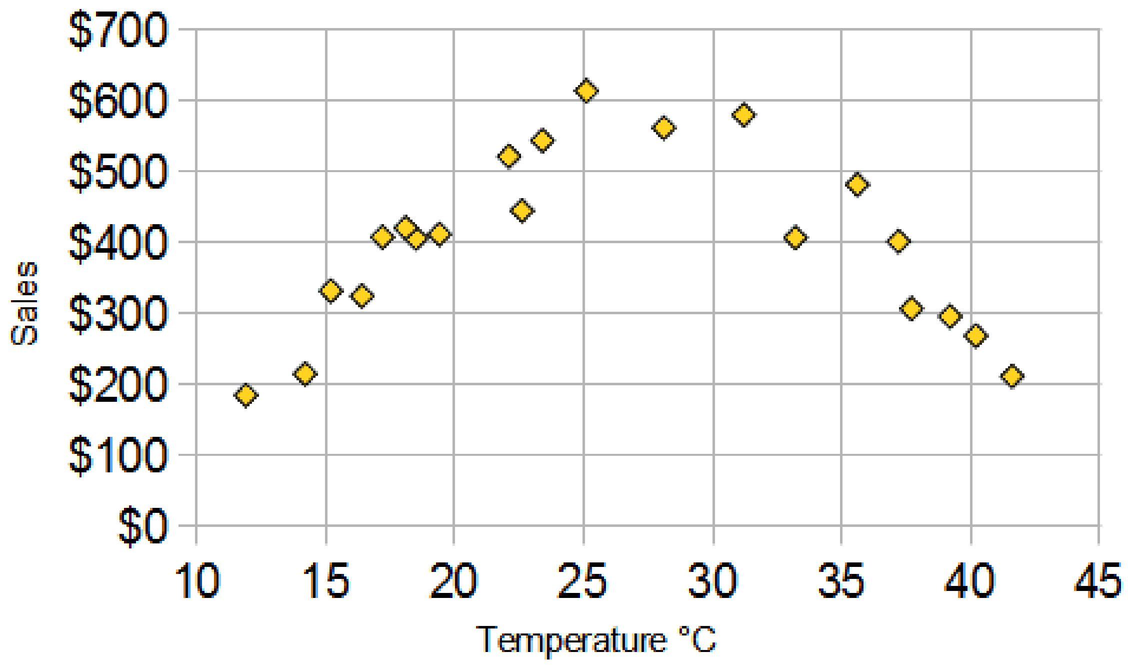
Avoid trying to apply the regression line to predict values far from those that were used to create the line.

- i.e., do not extrapolate too far

Basic regression caution # 2

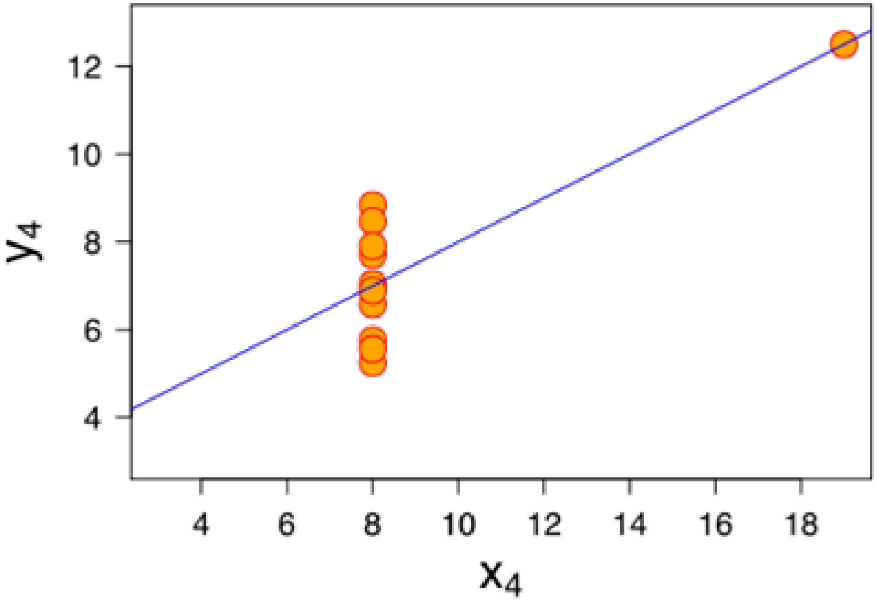
Plot the data! Linear regression is only appropriate when there is a linear trend in the data.

- We will discuss a set of checks on the appropriateness of using linear models soon



Basic regression caution #3

Be aware of outliers and high leverage points. They can have a large effect on the regression line.



Outlier: big $| y_i - \bar{y} |$

Leverage: big $| x_i - \bar{x} |$

Influential point: big outlier and leverage

There are statistics that quantify/describe influential points

- We will discuss these soon as well

Let’s try simple linear regression in R…

Faculty salaries

- Predict faculty salaries based on the size of a university's endowment

Inference for simple linear regression

Warning: there is a minefield of poor/misleading terminology out there

- So be careful when reading material related to inference on regression models

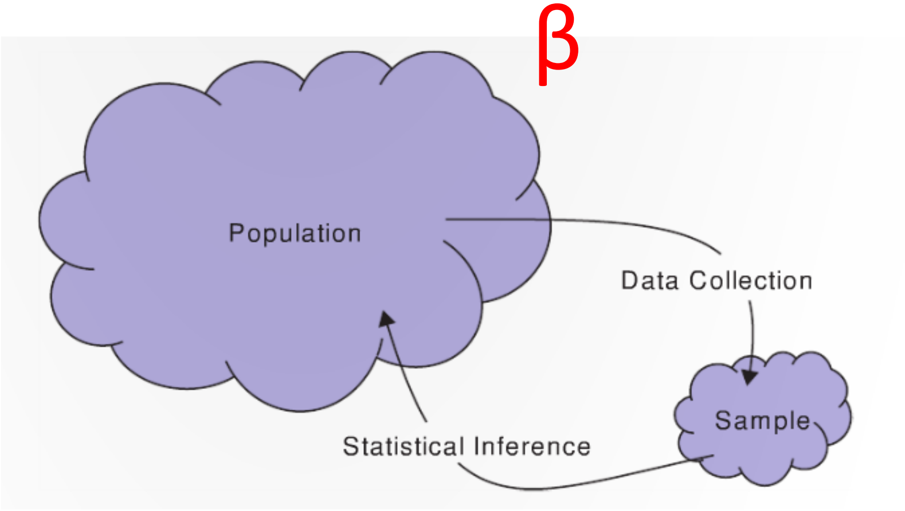
I will try to help you navigate this…

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***

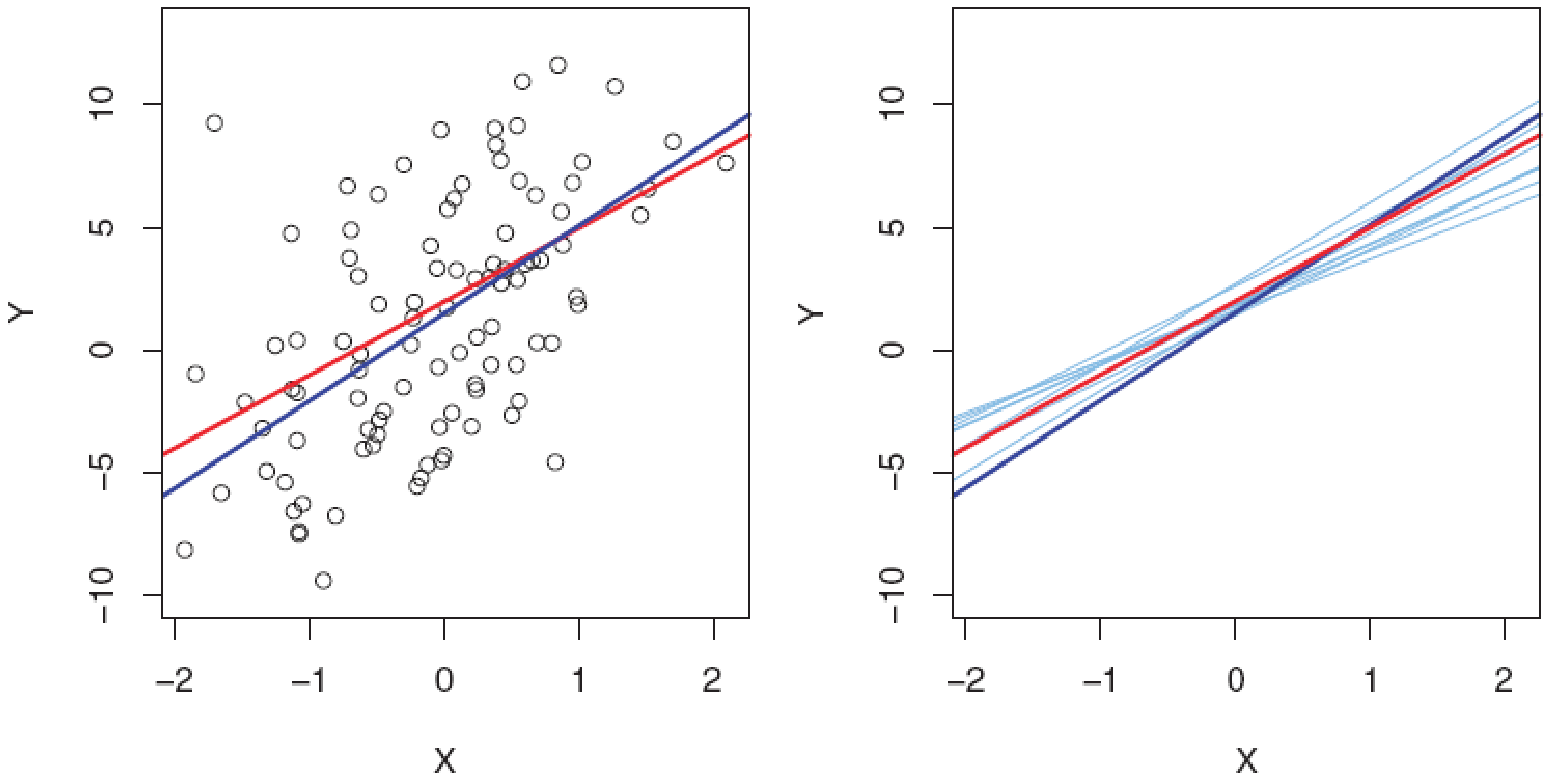
β



b $\hat{\beta}$

Population: β

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



Linear regression underlying model

Intercept   Slope   }   *Parameters*

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

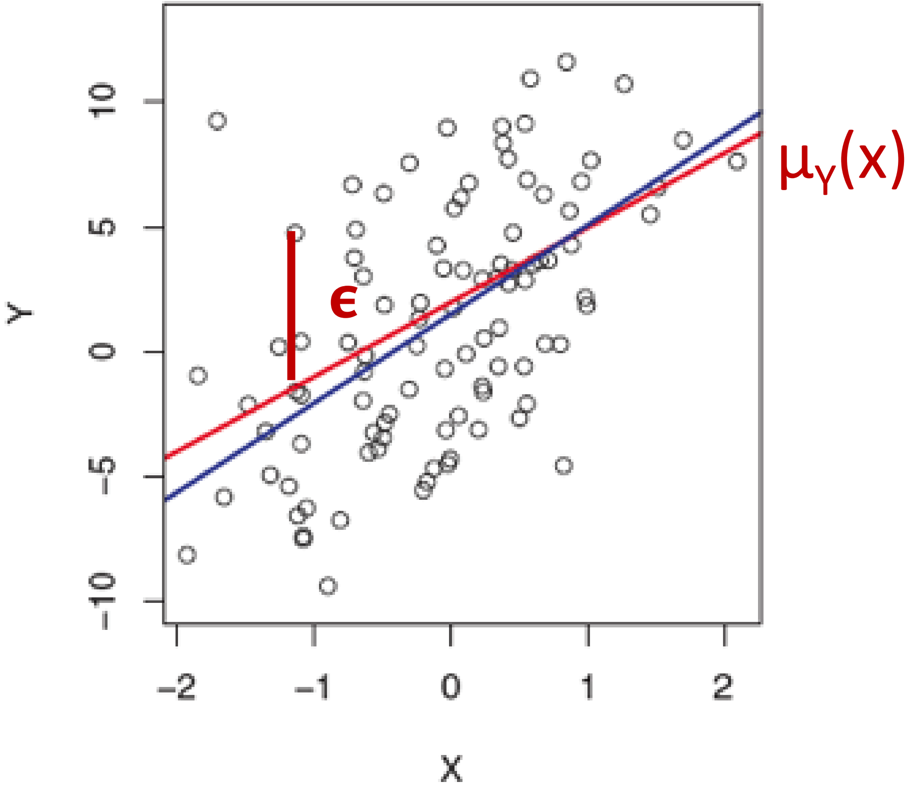
Error

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

$$= \mu_Y(x) + \epsilon$$

Errors ϵ_i are the difference between the **true regression line** $\mu_Y(x_i)$ and observed data points Y_i

- $\epsilon_i = Y_i - \mu_Y(x_i)$



Linear regression underlying model

Intercept   Slope   }   *Parameters*

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

Error

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

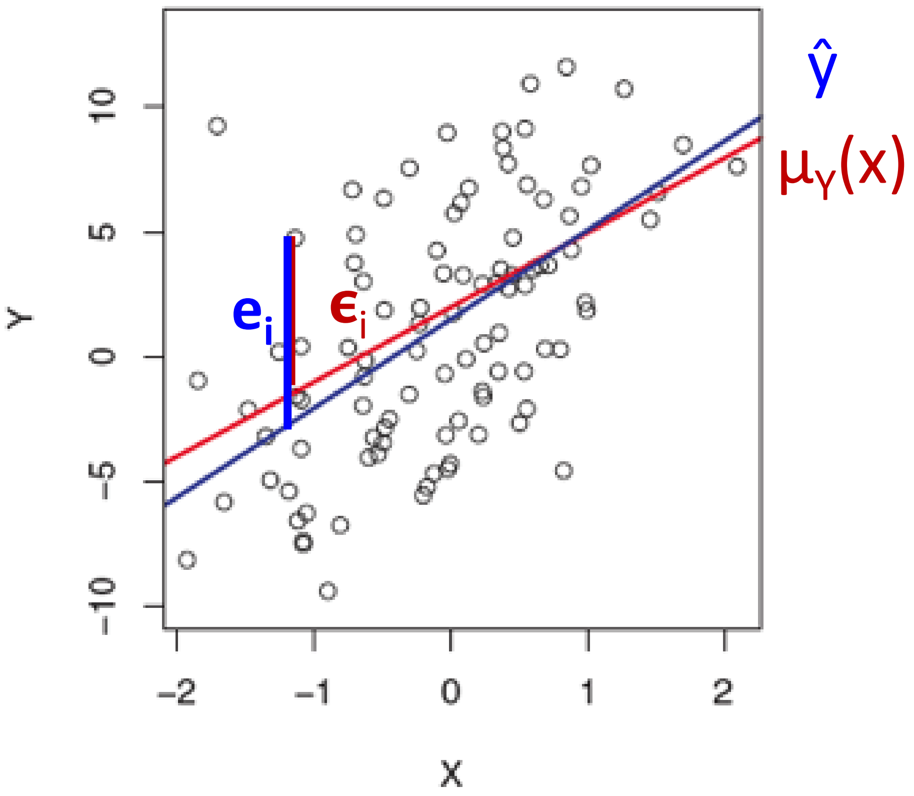
Estimated regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Errors ϵ_i are the difference between the **true regression line** $\mu_Y(x_i)$ and observed data points Y_i

- $\epsilon_i = Y_i - \mu_Y(x_i)$

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

- $e_i = Y_i - \hat{y}_i$



Linear regression underlying model

Intercept Slope } *Parameters*

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

Error

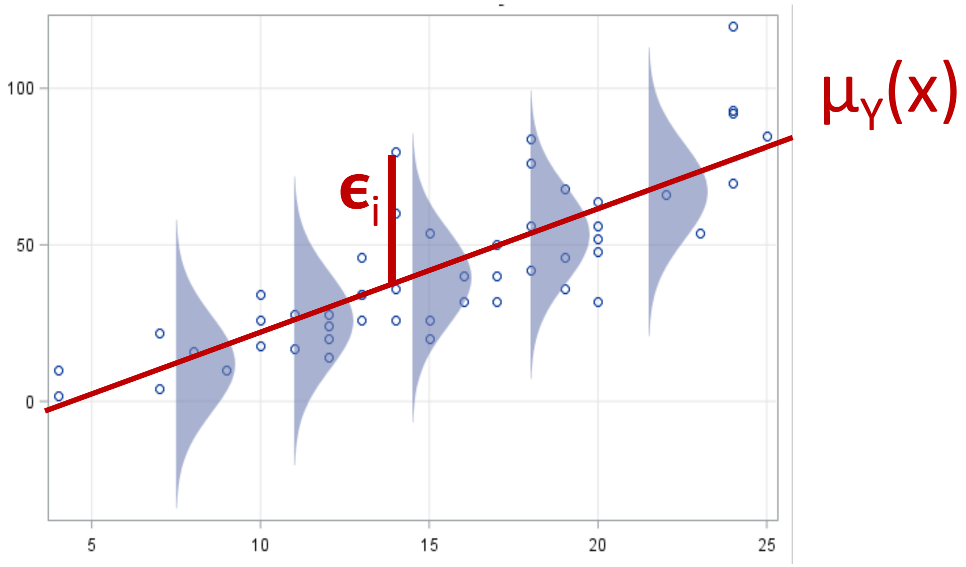
Observed data point: $$Y = \beta_0 + \beta_1 x + \epsilon \qquad \epsilon \sim N(0, \sigma_\epsilon)$$

Errors ϵ_i are the difference between the **true regression line** $\mu_Y(x_i)$ and observed data points Y_i

- $\epsilon_i = Y_i - \mu_Y(x_i)$

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

- This is needed for inference using parametric methods
  - e.g., to use t-distributions and F-distributions



Recap: Errors vs. residuals

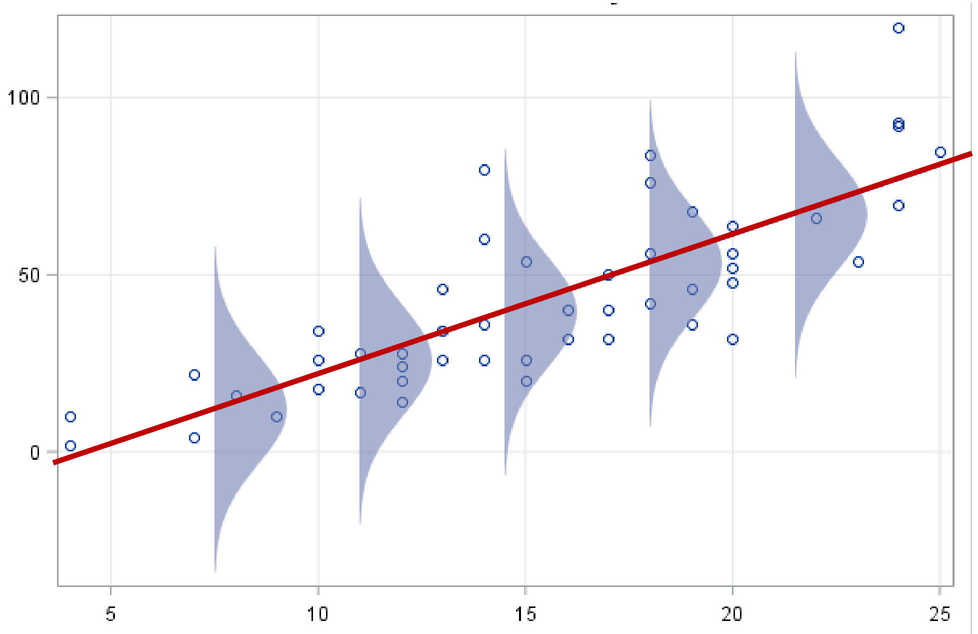
The data: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ $\epsilon_i \sim N(0, \sigma_\epsilon)$

"True" model: $\mu_Y(x_i) = \beta_0 + \beta_1 x_i$

- Errors: $\epsilon_i = Y_i - \mu_Y(x_i)$

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

- Residuals: $e_i = Y_i - \hat{y}_i$



Standard deviation of the errors: σ_ε

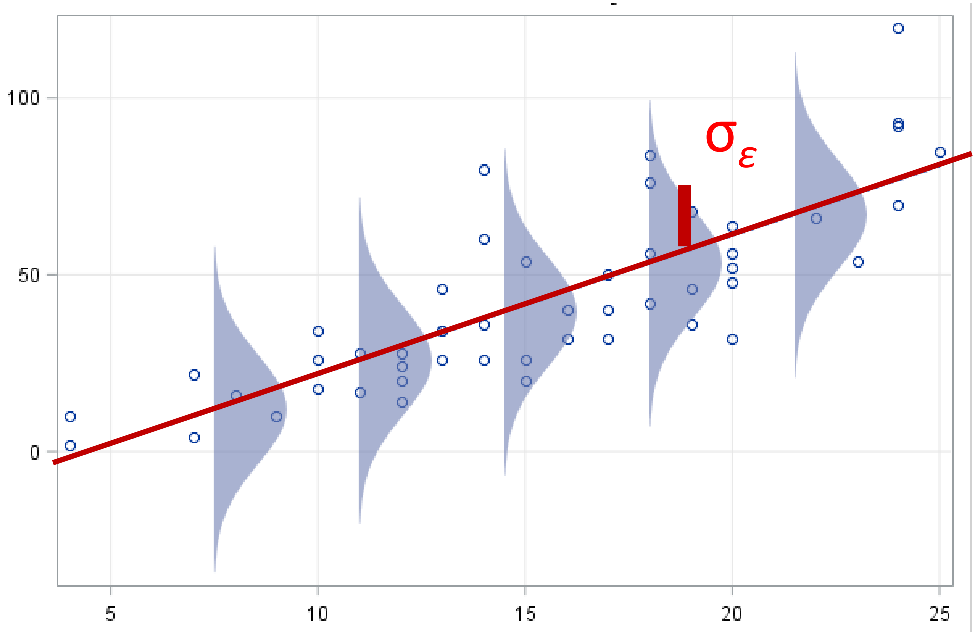
The standard deviation of the errors is denoted $\boldsymbol{\sigma_\varepsilon}$

We can use the **standard deviation of residuals** $\hat{\sigma}_\varepsilon$ as an estimate standard deviation of the errors σ_ε

- $\hat{\sigma}_\varepsilon$ often called the ~~"residual standard error”~~
- $\hat{\sigma}_\varepsilon$ we will call it the "residual standard deviation”

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

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

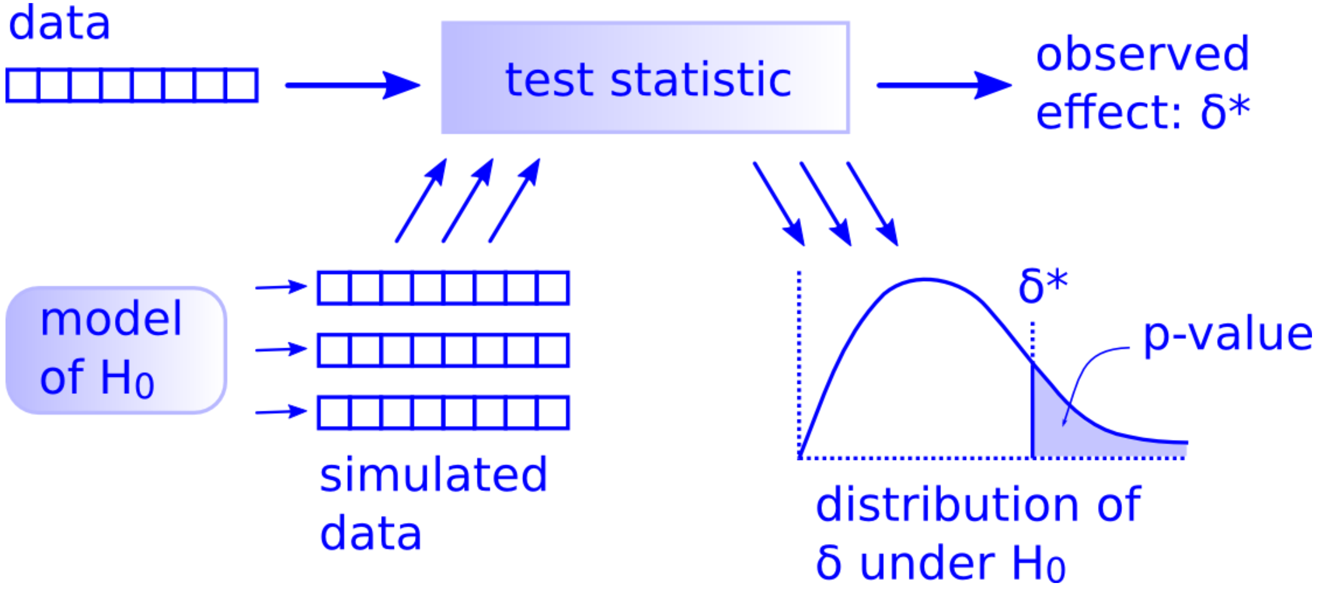


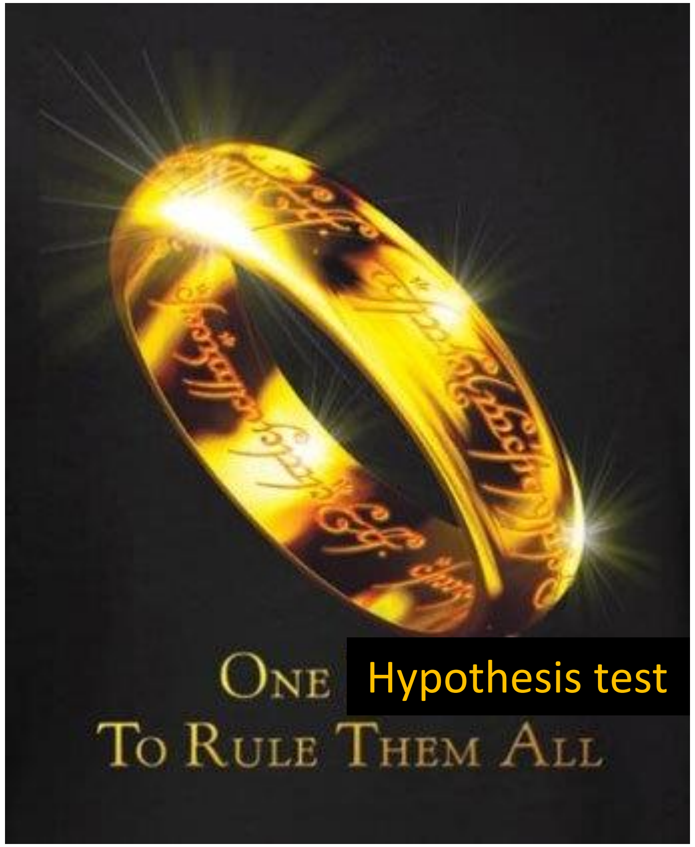
Let's quickly try this in R…

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: $t = \frac{\hat{\beta}_1 - 0}{\hat{SE}_{\hat{\beta_1}}}$

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

$$\hat{SE}_{\hat{\beta}_1} = \frac{\hat{\sigma}_\epsilon}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad \hat{SE}_{\hat{\beta}_0} = \hat{\sigma}_\epsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

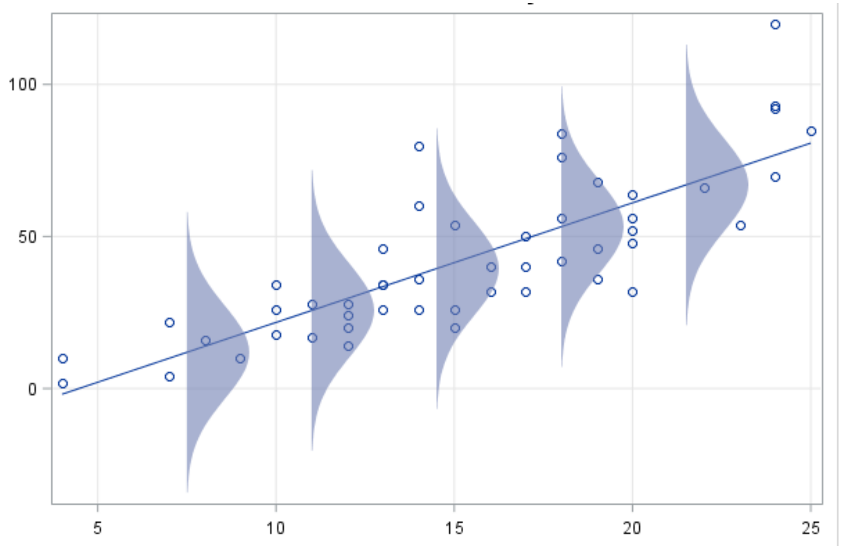
Inference using parametric methods

When using parametric methods, we make the following (LINE) assumptions:

- **Linearity**: A line can describe the relationship between x and y
- **Independence**: each data point is independent from the other points
- **Normality**: errors are normally distributed
- **Equal variance (homoscedasticity)**: constant variance of errors over the whole range of x values

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma_\epsilon)$$



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

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 $\mu_Y(x)$

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 a quantile value from a t-distribution with n-2 degrees of freedom obtain to get a desired confidence level

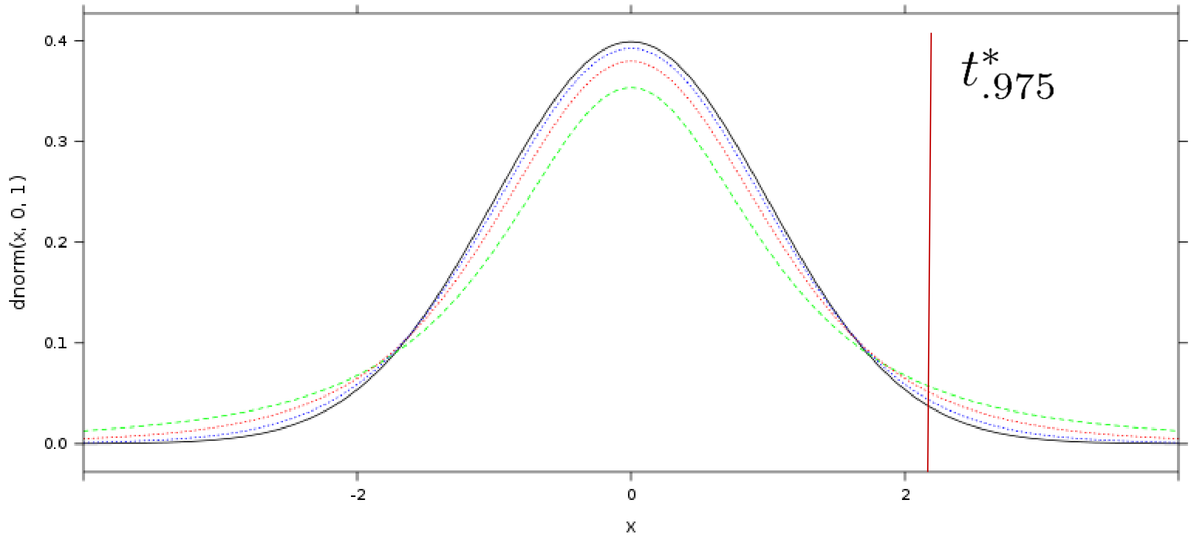
qt(.975, df)

N(0, 1)

df = 2

df = 5

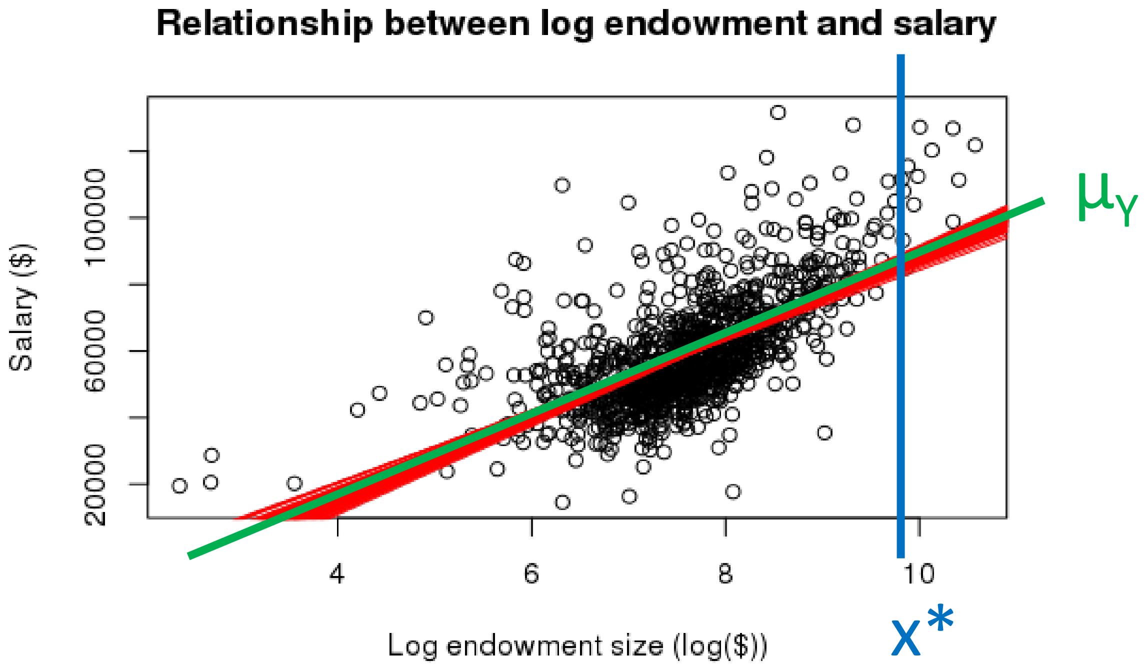
df = 15



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

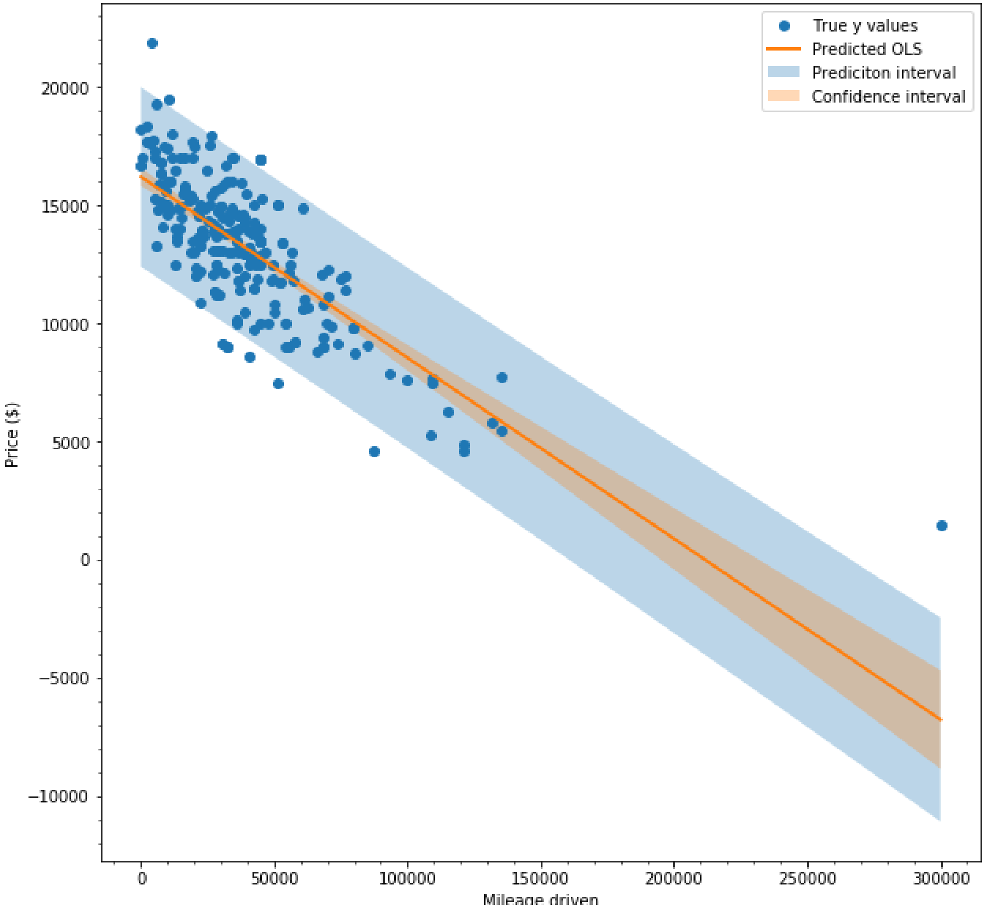
Prediction intervals

Confidence intervals give us a measure of uncertain about 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:

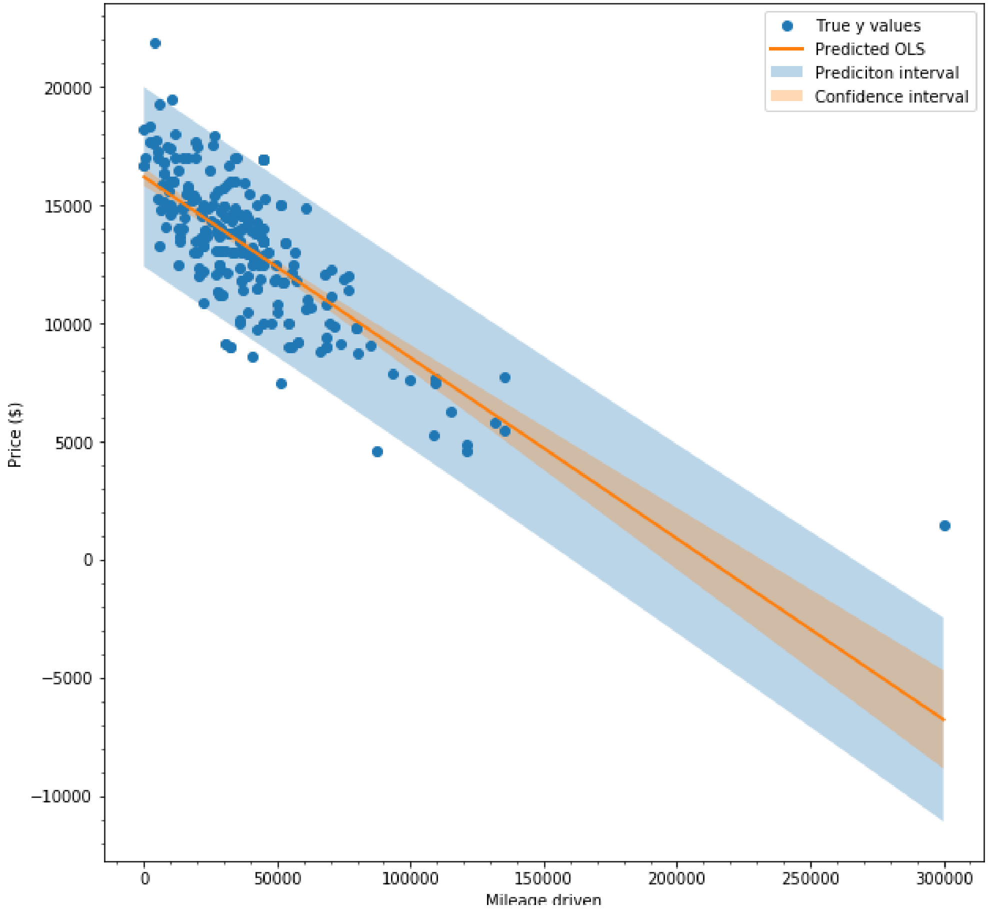
$$\hat{y} \;\pm\; t^* \cdot SE_{\hat{y}}$$

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*

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

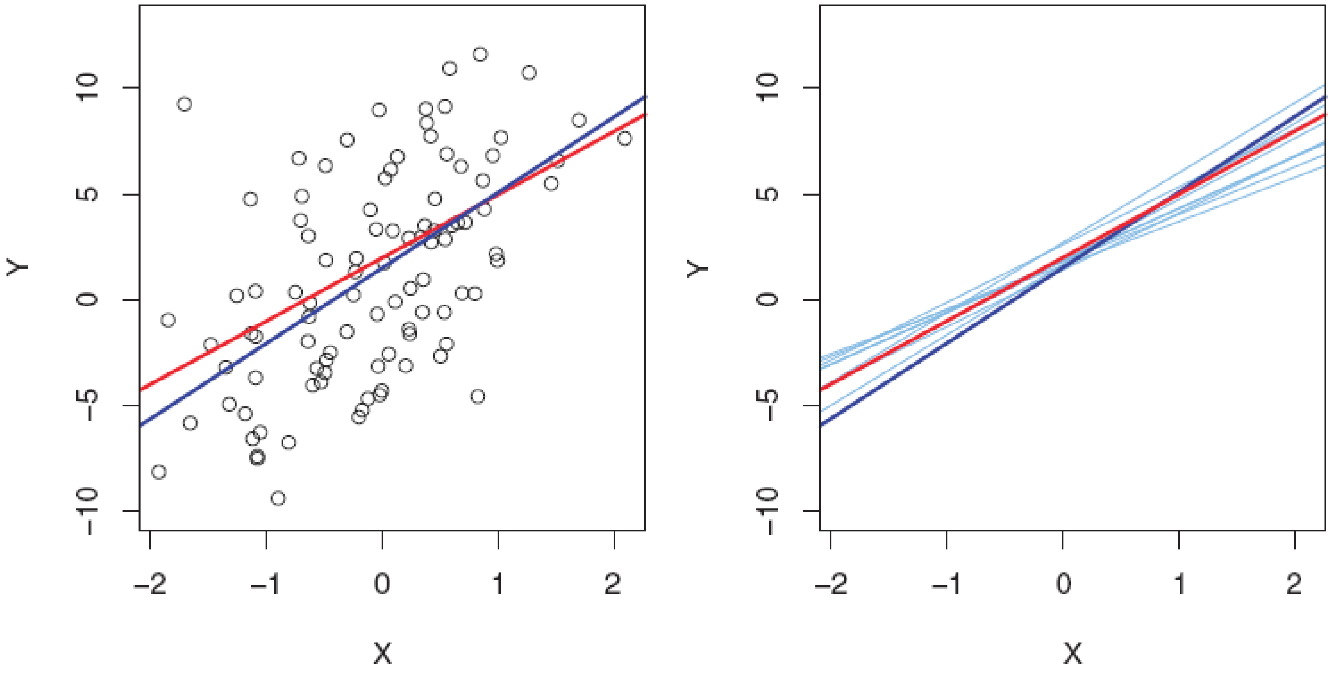
3. Prediction interval y

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

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 confidence and prediction intervals for simple linear regression in R…