# Welcome to STA 101!

Statistics is a confrontation with uncertainty.

Statistics confronts uncertainty by quantifying it.

## Data analysis

Transforming messy, incomplete, imperfect data into knowledge.

This knowledge usually takes the form of:

- pictures;
- a concise set of numerical summaries.

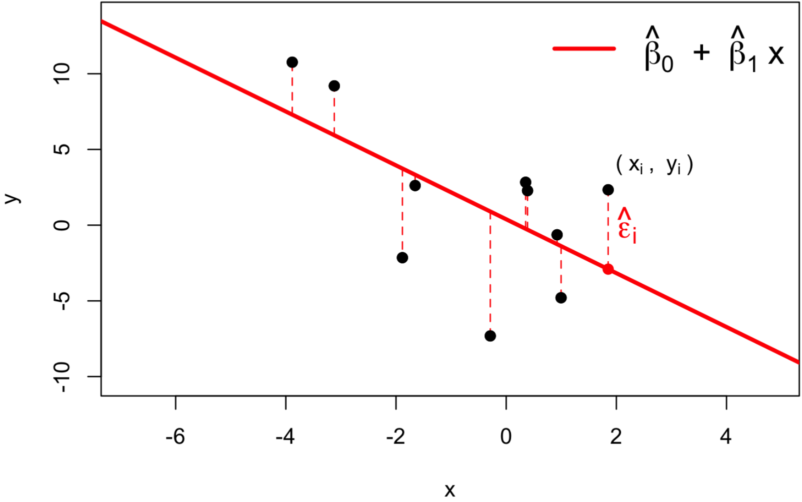
## Statistical inference

Quantifying our uncertainty about that knowledge:

- **Question**: Given data, what's our best guess at some quantity of interest?
- **Answer**: best-guess $\pm$ margin-of-error

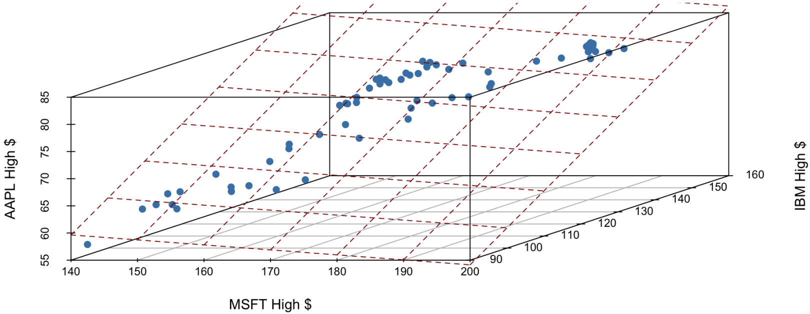
# Recap: linear regression

Numerical response and one numerical predictor:



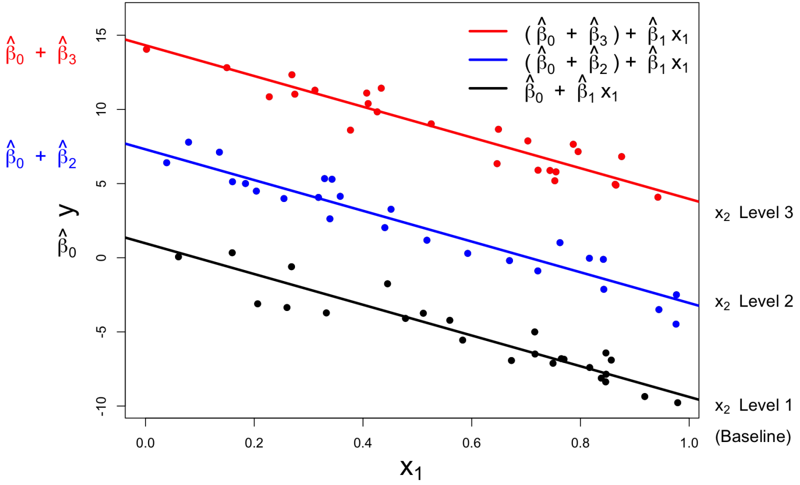
# Recap: linear regression

Numerical response and two numerical predictors:



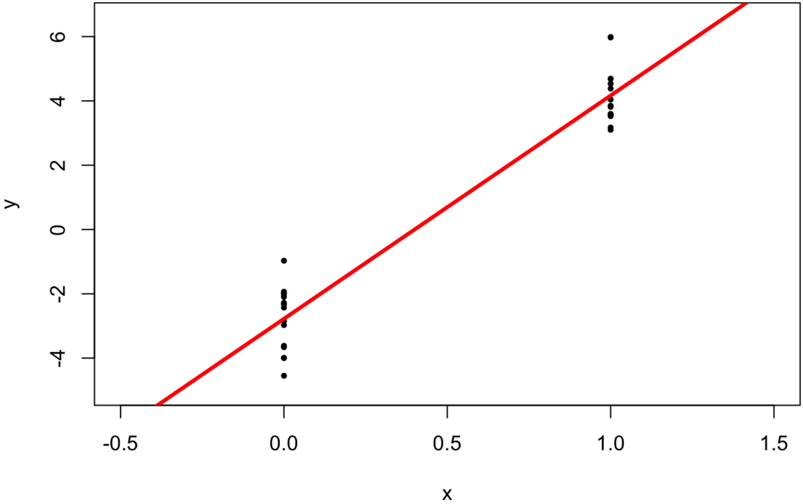
# Recap: linear regression

Numerical response, two predictors (numerical and categorical):

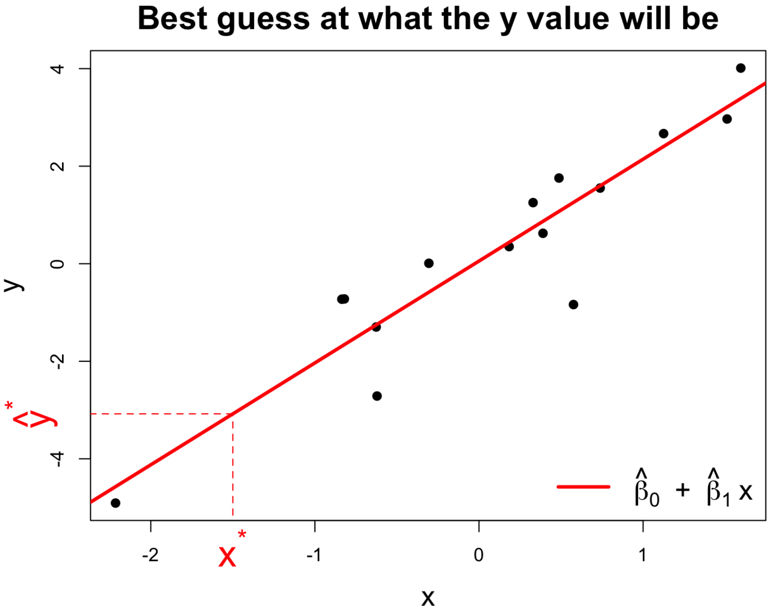


# Recap: linear regression

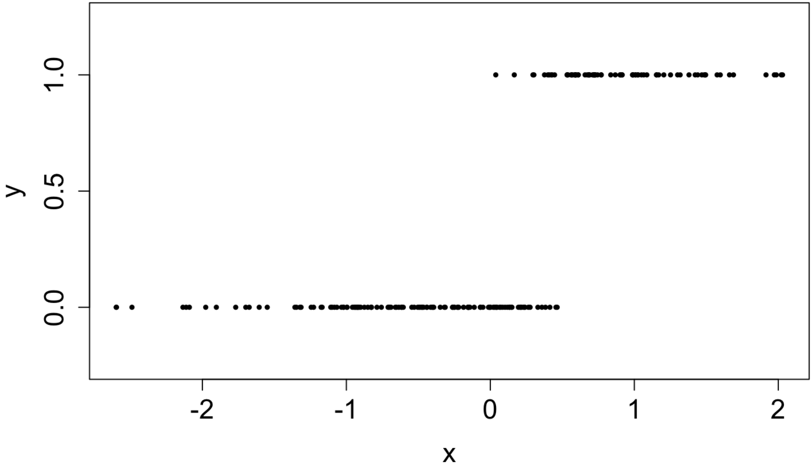
Numerical response and one categorical predictor (two levels):



# Primary use: out-of-sample prediction



# What if you have a categorical *response* with two levels?



# Examples

$$y_i = \begin{cases} 0 & \text{person } i \text{ repays loan} \\ 1 & \text{defaults.} \end{cases} \qquad x_i = \text{debt-to-income ratio}$$

# Examples

$$y_i = \begin{cases} 0 & \text{person } i \text{ repays loan} \\ 1 & \text{defaults.} \end{cases} \qquad x_i = \text{debt-to-income ratio}$$

$$y_i = \begin{cases} 0 & \text{email } i \text{ is spam} \\ 1 & \text{email } i \text{ is legit.} \end{cases} \qquad x_i = \text{occurrences of word ``money''}$$

# Examples

$$y_i = \begin{cases} 0 & \text{person } i \text{ repays loan} \\ 1 & \text{defaults.} \end{cases} \qquad x_i = \text{debt-to-income ratio}$$

$$y_i = \begin{cases} 0 & \text{email } i \text{ is spam} \\ 1 & \text{email } i \text{ is legit.} \end{cases} \qquad x_i = \text{occurrences of word ``money''}$$

$$y_i = \begin{cases} 0 & \text{person } i \text{ votes Green} \\ 1 & \text{votes Libertarian.} \end{cases} \qquad x_i = \text{income}$$

# Examples

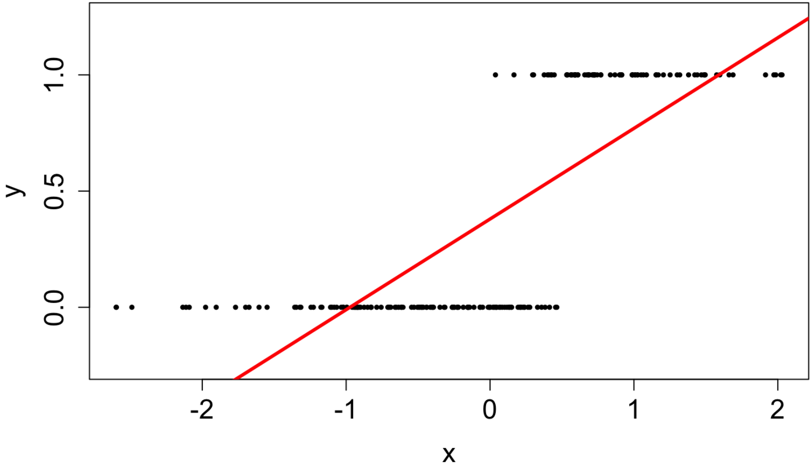
$$y_i = \begin{cases} 0 & \text{person } i \text{ repays loan} \\ 1 & \text{defaults.} \end{cases} \qquad x_i = \text{debt-to-income ratio}$$

$$y_i = \begin{cases} 0 & \text{email } i \text{ is spam} \\ 1 & \text{email } i \text{ is legit.} \end{cases} \qquad x_i = \text{occurrences of word ``money''}$$

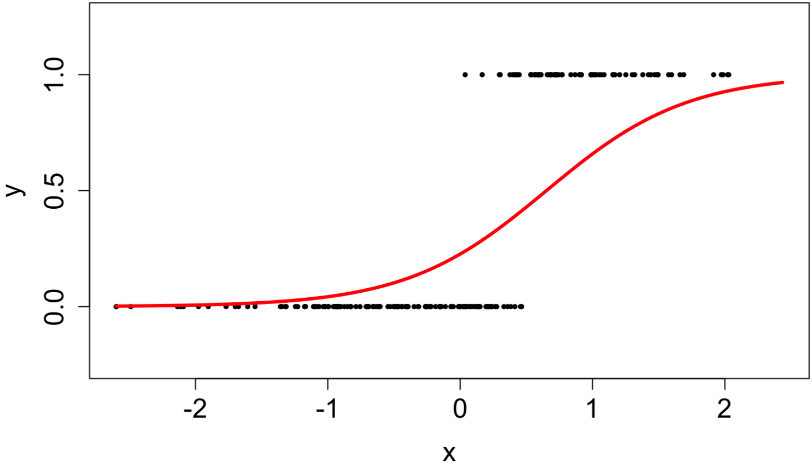
$$y_i = \begin{cases} 0 & \text{person } i \text{ votes Green} \\ 1 & \text{votes Libertarian.} \end{cases} \qquad x_i = \text{income}$$

$$y_i = \begin{cases} 0 & \text{Swiftie} \\ 1 & \text{Hater.} \end{cases} \qquad x_i = \text{times you shampoo hair per week}$$

# Straight line of best fit is a little silly here



# Instead: S-curve of best fit




Shaped like the letter “S.” Floor at 0, ceiling at 1. Neat-o!

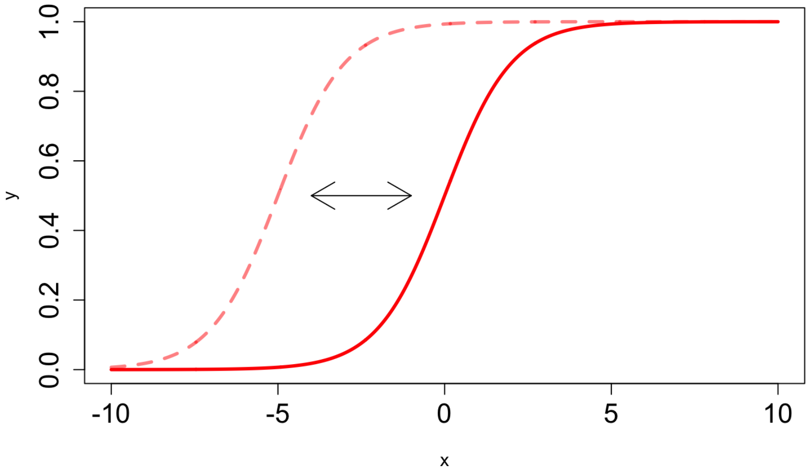
# The S-shaped curve is called the *logistic function*

Here you go:

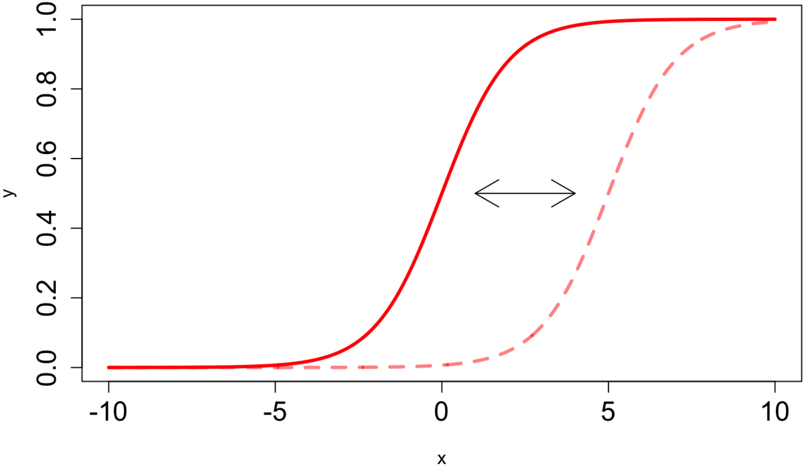
$$f(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}, \quad -\infty < x < \infty.$$

The parameters $\beta_0$ and $\beta_1$ control the shape of the curve.

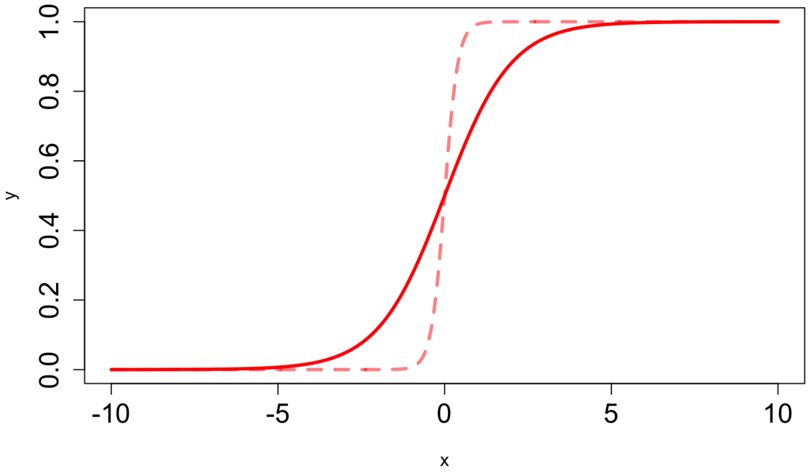
# Shifting it side to side



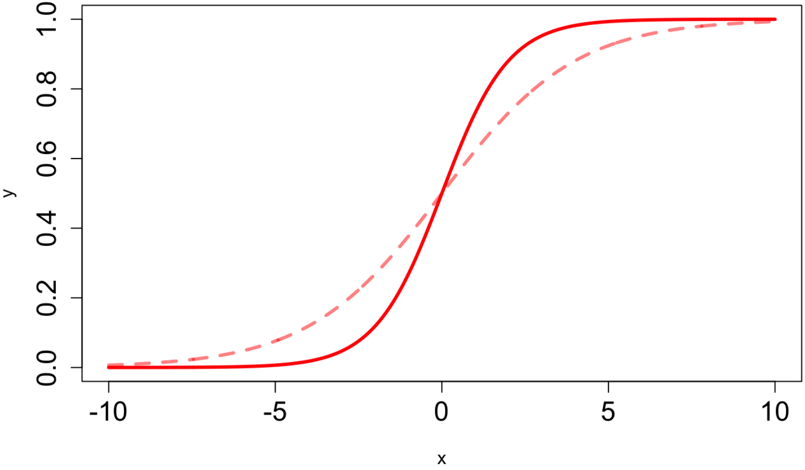
# Shifting it side to side



# Expanding or contracting it



# Expanding or contracting it



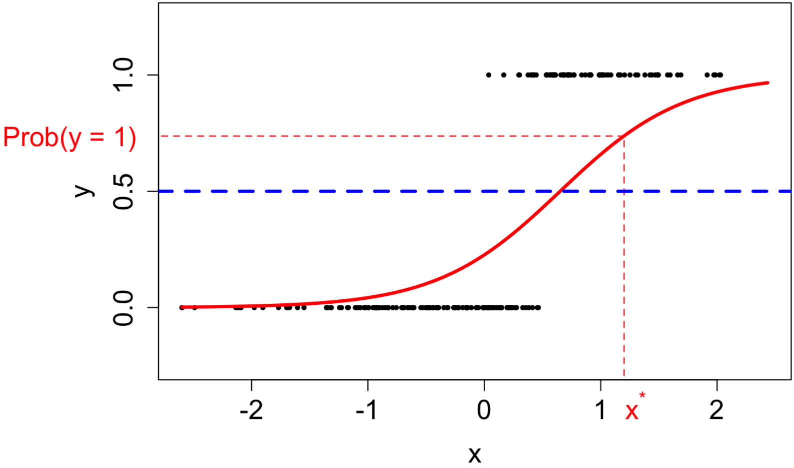
# Estimation

- Given data...

| $x$ | $y$ |
|---|---|
| 3.3 | 1 |
| 10.4 | 1 |
| -2.3 | 0 |
| 7.0 | 0 |
| 100.5 | 1 |
| $\vdots$ | $\vdots$ |

- Find estimates $\hat{\beta}_0, \hat{\beta}_1$ that give the "best fitting" S-curve;
- "Best" means *maximum likelihood* (don't worry about it);
- This is called *logistic regression*.

# Using the fitted model for prediction




- Points on the red line are $\text{Prob}(y = 1)$ at that $x$;
- If a new person $(x^\star)$ has probability above or below a chosen threshold, we predict they are $1/0$.

# “Multiple” logistic regression

There’s nothing special about a single predictor:

$$\mathrm{Prob}(y = 1) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p)}}.$$

# Decision boundary