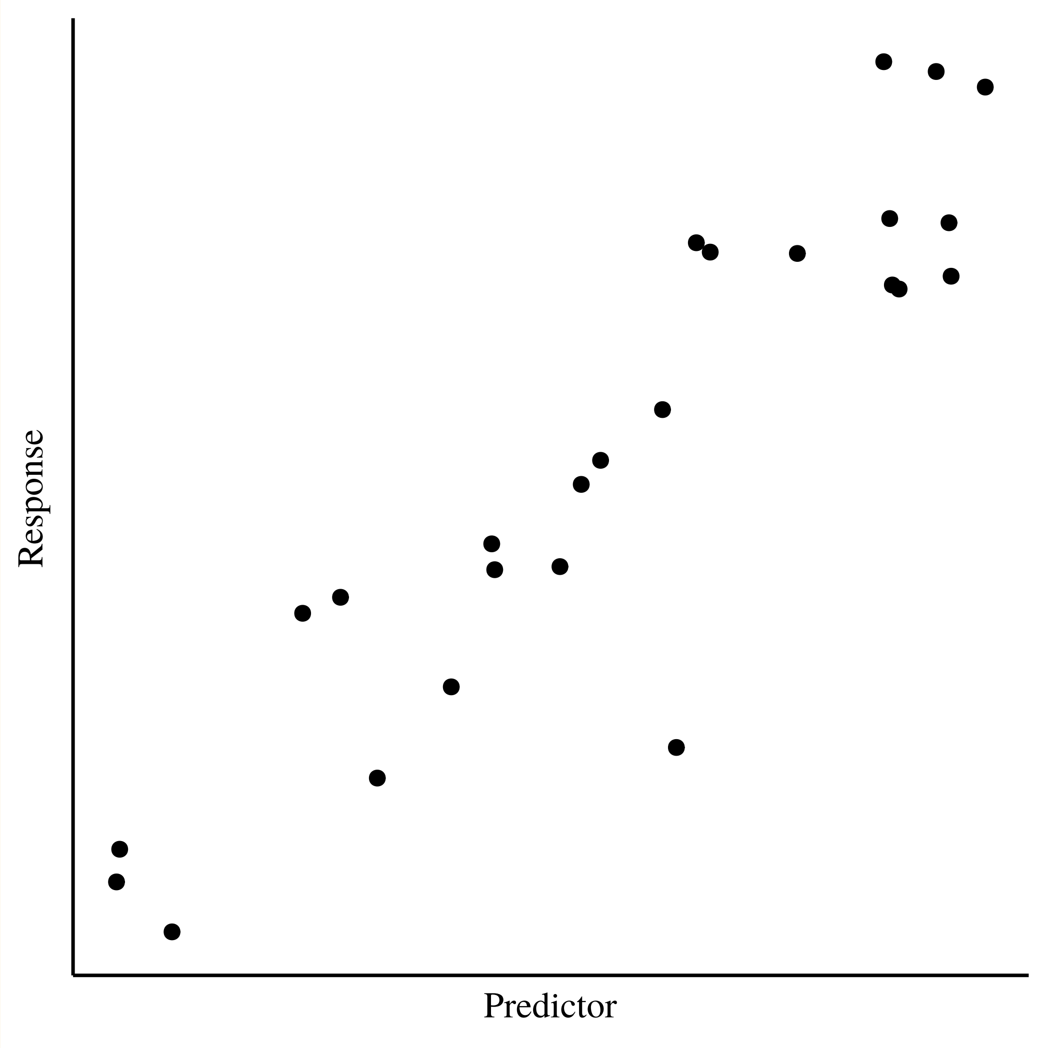
# Simple Linear Regression (SLR) model

Stat 230

April 01 2022

# Simple Linear Regression




Today:

- Introduce SLR model
- SLR model interpretation
- Parameter estimation
  - sampling distribution

# `cars` example

How does speed of cars relates to the distances taken to stop?

- `speed`: car speed in miles per hours (mph)
- `dist`: stopping distance in feet (ft)

```
str(cars)
```

```
'data.frame':	50 obs. of  2 variables:
 $ speed: num  4 4 7 7 8 9 10 10 10 11 ...
 $ dist : num  2 10 4 22 16 10 18 26 34 17 ...
```

# Statistical Modeling: EDA

For now: assume $y$ and $x$ are **quantitative** variables

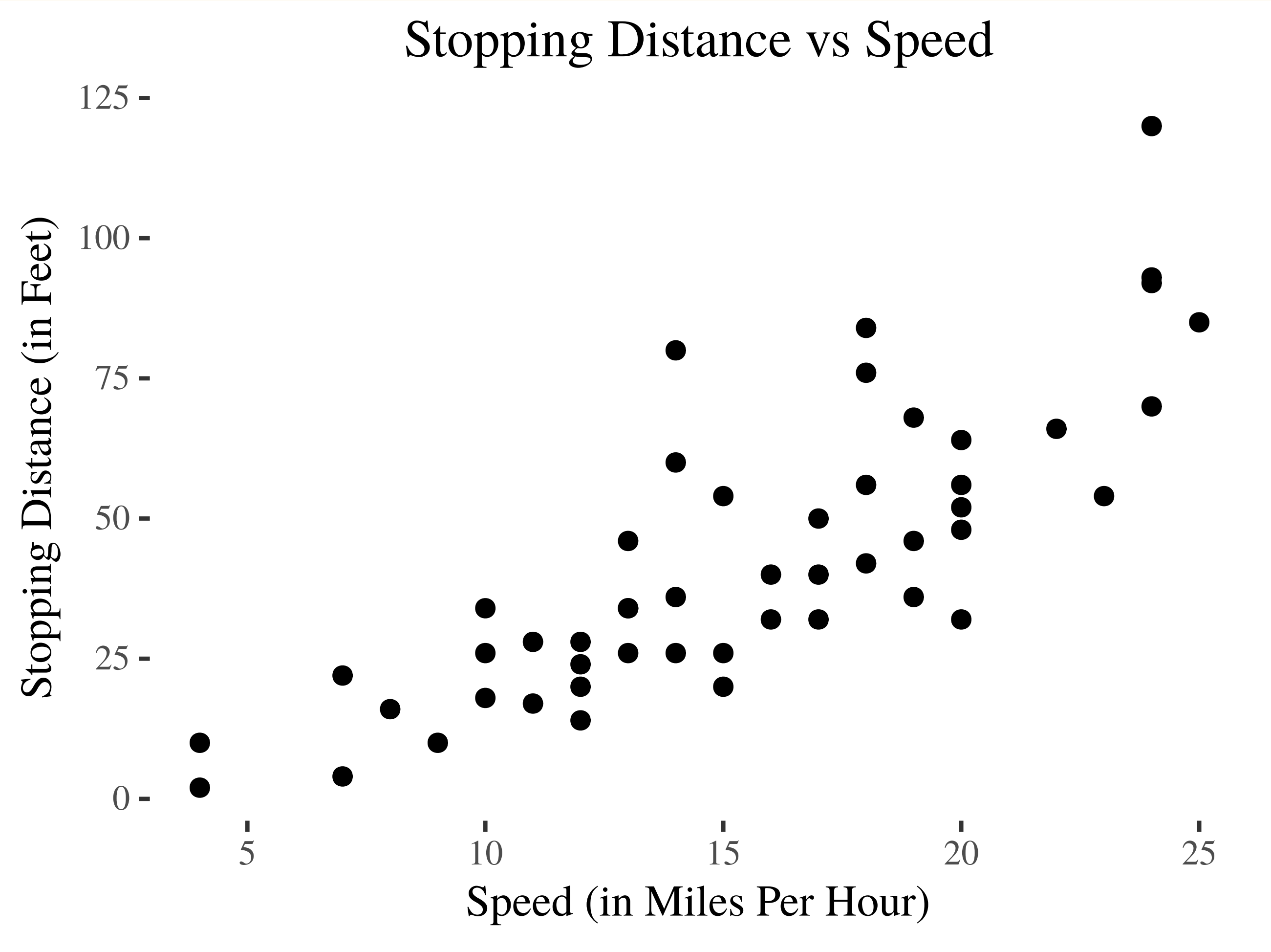
**univariate:** summary stats, boxplot/histogram

**bivariate:** scatterplot of $y$ against $x$

- relationship form
- direction
- relationship strength
- unusual cases

# Scatter plot

```
ggplot(cars, aes(x = speed, y = dist)) +
  geom_point() +
  labs(x='Speed (in Miles Per Hour)',
       y='Stopping Distance (in Feet)',
       title='Stopping Distance vs Speed') +
  theme(plot.title = element_text(hjust = 0.5)
```



# Statistical Modeling

$$\text{Response} = \text{Mean} + \text{Random error}$$

**Mean:** A population/theoretical mean value for the response

- Statistical models often focus on modeling this mean response as a function of other variables!

$$Y = \beta_0 + \beta_1 X + \epsilon$$

**Random error:** How much do individual responses vary from the mean

- Random errors are centered around 0 (why?)

# The Simple Linear Regression (SLR) Model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma)$$

**Linear Mean:** : The population mean value of $Y$ given a value of $x$ is

$$\mu_{y|x} = E(Y \mid x) = \beta_0 + \beta_1 x$$

- This mean value varies linearly with $(x)$

**Constant SD:** The population SD value of given a value of is

$$SD(Y \mid x) = \sigma$$

- The fact that this SD does not depend on the value of $x$ is called the contant variance, or homoscedastic, assumption

# The Simple Linear Regression (SLR) Model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma)$$

**Normality:** The population distribution of Y given a value of $x$ is

$$Y_i | x_i \sim N(\mu_{y|x} = \beta_0 + \beta_1 x_i, \sigma)$$

**Independence:** All responses given a value of $x$ occur independently of each other

# SLR model: quick example

$$\text{exam score}_i = 30 + 12\text{study hours}_i + \epsilon_i \qquad \epsilon_i \sim N(0, 3)$$

For all students who studied 5 hours, their scores are normally distributed with a mean of 90 and SD of 3.

$$\mu_{y|x=5} = 30 + 12 * 5 = 90$$

For all students who studied 1 hour, their scores are normally distributed with a mean of 42 and SD of $3$.

$$\mu_{y|x=1} = 30 + 12 * 1 = 42$$

# SLR model interpretation

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma)$$

There are a total of three parameters in the SLR model:

- the two mean parameters $\beta_0$ and $\beta_1$
- the SD parameter $\sigma$

# SLR model interpretation: Slope $\beta_1$

$\beta_1$ is the constant (additive) rate of change in the mean of $Y$ as $x$ varies.

- A 1 unit increase in $x$ is associated with a $\beta_1$ change in $Y$, on average.
- $\beta_1$ is the "effect" of $x$ on the response $Y$

$$\text{exam score}_i = 30 + 12\text{study hours}_i + \epsilon_i \quad \epsilon_i \sim N(0, 3)$$

Each additional hour studying is associated with a 12 points increase in exam score, on average.

# SLR model interpretion: Intercept $\beta_0$

$\beta_0$ is the mean value of $Y$ when $x = 0$

- Usually needed in the model but may not make sense to interpret unless we've observed x values around 0

$$\text{exam score}_i = 30 + 12\text{study hours}_i + \epsilon_i \qquad \epsilon_i \sim N(0, 3)$$

The mean exam score is 30 for all students who did not study (hours = 0).

- if we didn't see any cases with 0 study hours, then this would be an extrapolation!

# SLR model interpretation: $\sigma$

$\sigma$ is the standard deviation of $Y$ around $\mu_{y|x}$ given any predictor value $x$

$$\text{exam score}_i = 30 + 12\text{study hours}_i + \epsilon_i \qquad \epsilon_i \sim N(0, 3)$$

For any study hour value, individual exam scores vary around the model mean $30 + 12$ study hours with a SD of 3 points.

# SLR estimation

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma)$$

How do we estimate $\beta_0$, $\beta_1$, $\sigma$ given a sample of $n$ data points?

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

# Maximum likelihood estimation (MLE)

**Motivation:** To find the values of $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}$ that maximizes the likelihood (probability) of the observed data based on the SLR model

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

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

MLE yields the **same estimates** that you get from the "best fit" line found via "least squares" estimation

# Fitting a SLR in R

$$\text{lm}(y \sim x,\ \text{data} \ = \quad )$$

```
cars_lm <- lm(dist ~ speed, data = cars)
cars_lm
```

```

Call:
lm(formula = dist ~ speed, data = cars)

Coefficients:
(Intercept)        speed
    -17.579        3.932
```

# `cars` linear model estimates

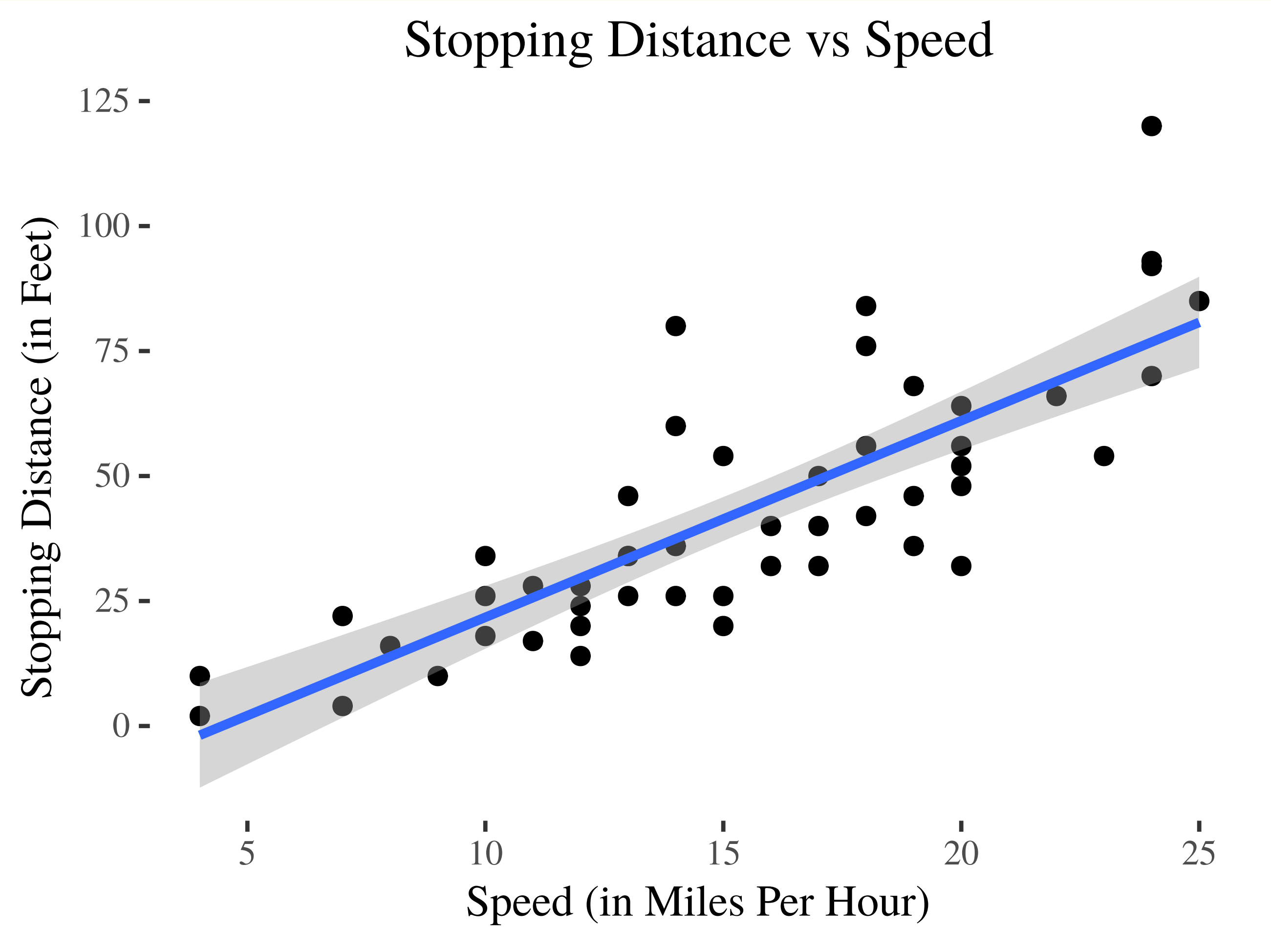
$$\hat{\beta}_0 = -17.579 \quad \hat{\beta}_1 = 3.932$$

A one miles per hour increase in speed is associated with an estimated 3.932 feet increase in mean stopping distance

At 0 speed, we estimate an average stopping of -17.579 feet

# Visualize the SLR line

```
ggplot(cars, aes(x = speed, y = dist)) +
  geom_point() +
  labs(x='Speed (in Miles Per Hour)',
       y='Stopping Distance (in Feet)',
       title='Stopping Distance vs Speed') +
  theme(plot.title = element_text(hjust = 0.5)
  geom_smooth(method = "lm", se = FALSE) +
  geom_smooth(method = "lm")
```



# Back to estimation theory

A one miles per hour increase in speed is associated with an estimated 3.932 feet increase in mean stopping distance

- How much uncertainty do we have in this estimated effect?

Need to understand the sampling distribution of our estimates

- Interested in the mean function parameters

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

# Sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

For a fixed set of predictor values $x_1, \ldots, x_n$ :

- our observed responses $y_1, \ldots, y_n$ will vary from sample to sample
- so our "best fit" line intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$ will vary from sample to sample

# Sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$

Probability theory tells us that the sampling distributions are normally distributed!!

$$\hat{\beta}_j \sim N\left(\beta_j, SD\left(\hat{\beta}_j\right)\right) \qquad \text{for } j = 0, 1$$

We need to estimate $SD\left(\hat{\beta}_j\right)$ with standard errors:

$$SE\left(\hat{\beta}_1\right) = \hat{\sigma}\sqrt{\frac{1}{(n-1)s_x^2}} \quad SE\left(\hat{\beta}_0\right) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}$$

# Sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$

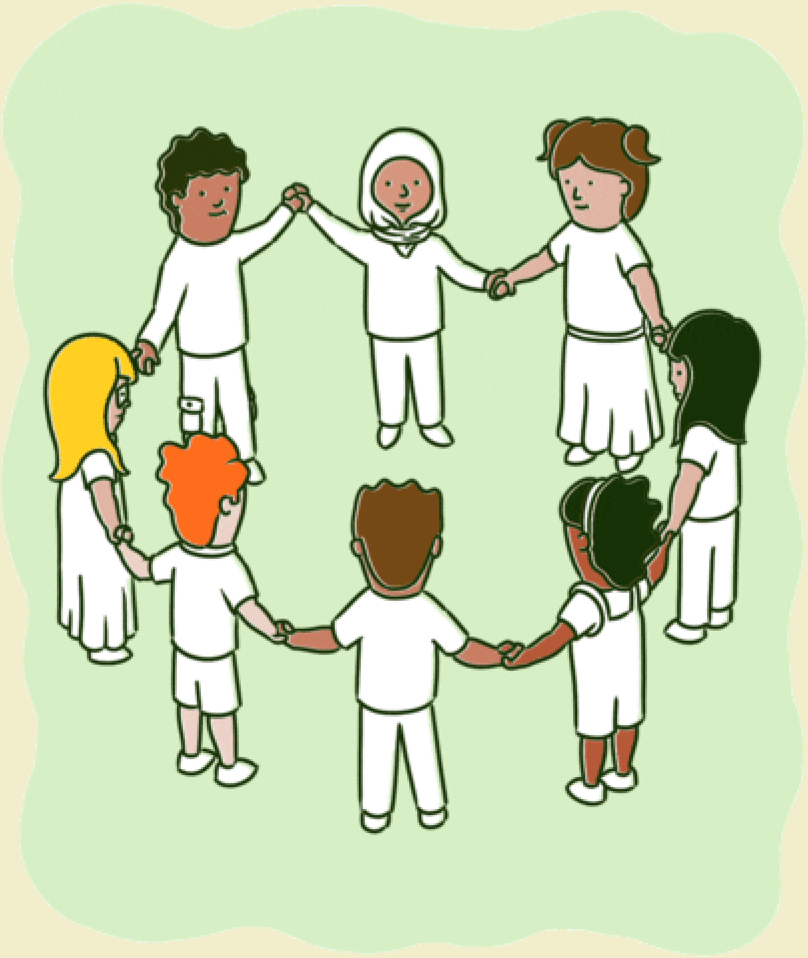
- Need advanced statistics for the proofs!!

Using simulations in R

- we can generate lots of different responses from the same set of $x$ values and see how the best fit slopes/intercepts vary!!

# Your Turn 1

05:00



- Get the in class activity file from moodle
- Use given functions to simulate many parameter estimates
- Infer the standard errors!