# Chapter 11: Linear regression

Part 2: Simple linear regression

https://dzwang91.github.io/stat324/

UNIVERSITY OF WISCONSIN–MADISON

# Reference

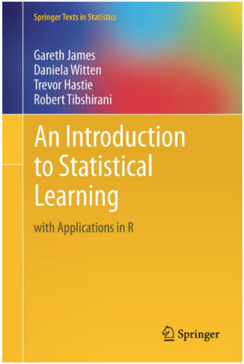



Figure: `http://www-bcf.usc.edu/~gareth/ISL/`

- Reading for today's lecture: Section 3.1

# Outline

1. Pearson correlation coefficient
2. Simple linear regression
3. Estimating the coefficients
4. Assessing the accuracy of the coefficient estimates
5. Assessing the accuracy of the model

# Motivation

A natural question: for two random variables $X$ and $Y$, how can we measure their association?

# Pearson correlation coefficient

- For two random variables $X$ and $Y$, Pearson correlation coefficient is defined as

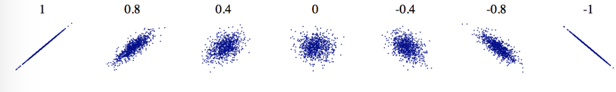
$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y},$$

  where
  - $cov(X,Y) = \mathbb{E}(X - \mu_X)(Y - \mu_Y)$, and $\mu_X$ and $\mu_Y$ are expectations of $X$ and $Y$.
  - $\sigma_X = \sqrt{Var(X)}$: standard deviation of $X$
  - $\sigma_Y = \sqrt{Var(Y)}$: standard deviation of $Y$
- Range of $\rho_{X,Y}$:

$$-1 \leq \rho_{X,Y} \leq 1$$

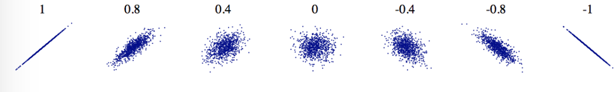
# Pearson correlation coefficient continued

- Visualization of $\rho_{X,Y}$:



# Pearson correlation coefficient continued

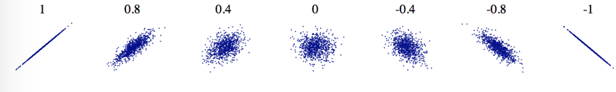
- Visualization of $\rho_{X,Y}$:




- When $\rho_{X,Y} = 1$, scatter is perfect straight line sloping up
- When $\rho_{X,Y} = -1$, scatter is perfect straight line sloping down
- When $\rho_{X,Y} = 0$, there is no linear association, then we call $X$ and $Y$ are uncorrelated.

# Pearson correlation coefficient continued

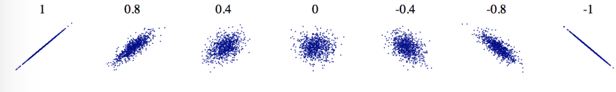
- Visualization of $\rho_{X,Y}$:




- When $\rho_{X,Y} = 1$, scatter is perfect straight line sloping up
- When $\rho_{X,Y} = -1$, scatter is perfect straight line sloping down
- When $\rho_{X,Y} = 0$, there is no linear association, then we call $X$ and $Y$ are uncorrelated.
- Conclusion: Pearson correlation coefficient measures the linear association

# Pearson correlation coefficient continued

- Visualization of $\rho_{X,Y}$:




- When $\rho_{X,Y} = 1$, scatter is perfect straight line sloping up
- When $\rho_{X,Y} = -1$, scatter is perfect straight line sloping down
- When $\rho_{X,Y} = 0$, there is no linear association, then we call $X$ and $Y$ are uncorrelated.
- Conclusion: Pearson correlation coefficient measures the linear association
- When $\rho_{X,Y} > 0$, we say $X$ and $Y$ have a positive linear association
- When $\rho_{X,Y} < 0$, we say $X$ and $Y$ have a negative linear association

# Sample Pearson correlation coefficient

- Given $n$ pairs of data $(x_1, y_1), ..., (x_n, y_n)$, sample Pearson correlation coefficient is defined as

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

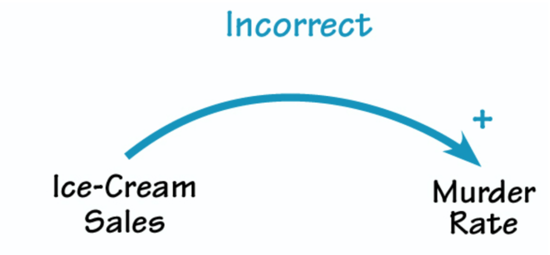
# Correlation is not causation[1]

- A famous example: Ice cream sales is correlated with homicides in New York,

[1]Reading: Why correlation does not imply causation?

# Correlation is not causation[1]

- A famous example: Ice cream sales is correlated with homicides in New York, but ice cream is not causing the death of people.




[1] Reading: Why correlation does not imply causation?

# Correlation is not causation continued

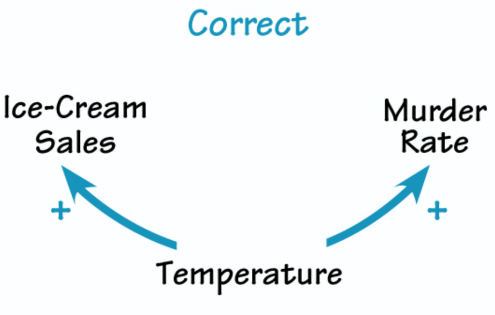
- Why are ice cream sales and homicides correlated?

# Correlation is not causation continued

- Why are ice cream sales and homicides correlated?
- There are some hidden factors which cause both of ice cream sales and homicides.

# Correlation is not causation continued

- Why are ice cream sales and homicides correlated?
- There are some hidden factors which cause both of ice cream sales and homicides.



# Outline

1. Pearson correlation coefficient
2. Simple linear regression
3. Estimating the coefficients
4. Assessing the accuracy of the coefficient estimates
5. Assessing the accuracy of the model

# Example

- Sir Francis Galton (1822-1911) was interested in how children resemble their parents. One simple measure of this is height.
- Galton measured the heights of father son pairs (in inches) at maturity.
- In the actual study, 1078 pairs were measured. For convenience, we will use a small subsample of $n = 14$ pairs:

# Example continued

| Family | Father's Height | Son's Height |
|---|---|---|
| 1 | 71.3 | 68.9 |
| 2 | 65.5 | 67.5 |
| 3 | 65.9 | 65.4 |
| 4 | 68.6 | 68.2 |
| 5 | 71.4 | 71.5 |
| 6 | 68.4 | 67.6 |
| 7 | 65.0 | 65.0 |
| 8 | 66.3 | 67.0 |
| 9 | 68.0 | 65.3 |
| 10 | 67.3 | 65.5 |
| 11 | 67.0 | 69.8 |
| 12 | 69.3 | 70.9 |
| 13 | 70.1 | 68.9 |
| 14 | 66.9 | 70.2 |

- Goal: predict sons' height from father's height.

# Example continued

- Which variable is input variable?

# Example continued

- Which variable is input variable? Father's height

# Example continued

- Which variable is input variable? Father's height
- Which variable is output variable?

# Example continued

- Which variable is input variable? Father's height
- Which variable is output variable? Son's height

# Example continued

- Which variable is input variable? Father's height
- Which variable is output variable? Son's height
- A simple linear regression:

$$\text{Son's height} = \beta_0 + \beta_1 * \text{ Father's height} + \text{ Random error}$$

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

# Example continued

- Which variable is input variable? Father's height
- Which variable is output variable? Son's height
- A simple linear regression:

$$\text{Son's height} = \beta_0 + \beta_1 * \text{ Father's height} + \text{Random error}$$
$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Denote the height of son $i$ by $y_i$, the height of father $i$ by $x_i$, and the random error by $\epsilon_i$, so that the model becomes:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

# Example continued

- Which variable is input variable? Father's height
- Which variable is output variable? Son's height
- A simple linear regression:

$$\text{Son's height} = \beta_0 + \beta_1 * \text{ Father's height} + \text{ Random error}$$
$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Denote the height of son $i$ by $y_i$, the height of father $i$ by $x_i$, and the random error by $\epsilon_i$, so that the model becomes:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

- Why do we add a random error term?

# Example continued

- Which variable is input variable? Father's height
- Which variable is output variable? Son's height
- A simple linear regression:

$$\text{Son's height} = \beta_0 + \beta_1 * \text{ Father's height} + \text{ Random error}$$

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

- Denote the height of son $i$ by $y_i$, the height of father $i$ by $x_i$, and the random error by $\epsilon_i$, so that the model becomes:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

- Why do we add a random error term? The random error term picks up sources of variation in an individual son's height that are not due to his father's height (mother's genetics, environmental factors, etc.) and which cause the points to be "off line."

# Intercept and slope

- $\beta_0$ is the intercept. It is the expected value of $Y$ when $X = 0$.
- $\beta_1$ is the slope. It is the average increase of $Y$ associated with a one-unit increase in $X$.
- Our goal: estimate the values of $\beta_0$ and $\beta_1$ from data. (what is the available (training) data?)

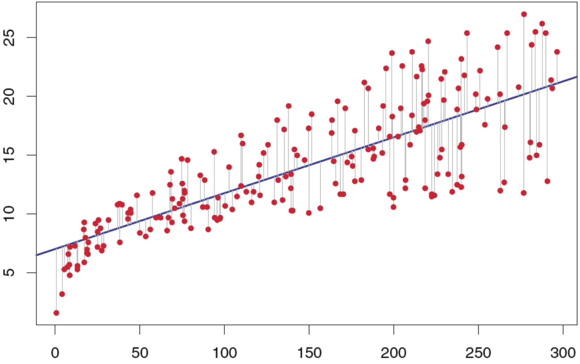
# Outline


1. Pearson correlation coefficient
2. Simple linear regression
3. Estimating the coefficients
4. Assessing the accuracy of the coefficient estimates
5. Assessing the accuracy of the model

- We assume a model

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

- Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be $n$ observation pairs, each of which consists of a measurement of $X$ and a measurement of $Y$.



# Residual sum of squares

- Suppose $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimates, then the estimated (fitted) value for given $x_i$ is:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

- The i-th residual: the difference between $\hat{y}_i$ and the observed $y_i$.

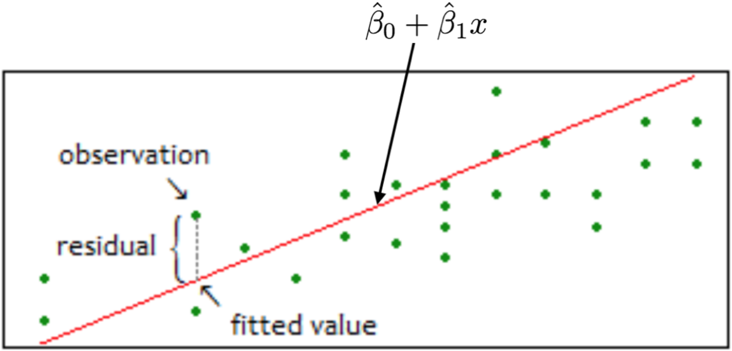
$$e_i = y_i - \hat{y}_i.$$

- Residual sum of squares (RSS):

$$\text{RSS} = e_1^2 + e_2^2 + ... + e_n^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2.$$

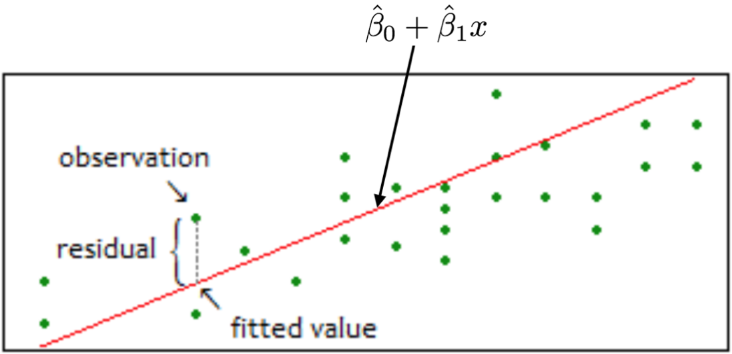
- RSS measures how well the line fits the data.

# Ordinary least squares




- How can we decide the line?

# Ordinary least squares




- How can we decide the line?
- Ordinary least squares (OLS):

$$\text{minimize } \text{RSS}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

# OLS estimator

- OLS estimator:

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

  where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
- We call $\hat{\beta}_0 + \hat{\beta}_1 x$ the least squares/best fit/regression line.
- The residual sum of squares for the least squares line is also called the sum of squared errors (SSE). SSE is the smallest possible residual sum of squares in the universe of all possible lines.

# OLS estimator

- OLS estimator:

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

  where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$.

- We call $\hat{\beta}_0 + \hat{\beta}_1 x$ the least squares/best fit/regression line.
- The residual sum of squares for the least squares line is also called the sum of squared errors (SSE). SSE is the smallest possible residual sum of squares in the universe of all possible lines.
- Exercise: calculate $\hat{\beta}_0$ and $\hat{\beta}_1$ for the father and son data. ($\hat{\beta}_1 = 0.65$ and $\hat{\beta}_0 = 23.64$)

# Back to Pearson correlation

- OLS estimator of slope:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2} = \frac{S_{xy}}{S_x S_y}\frac{S_x S_y}{S_x^2} = r_{xy}\frac{S_y}{S_x}$$

# Back to Pearson correlation

- OLS estimator of slope:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2} = \frac{S_{xy}}{S_x S_y}\frac{S_x S_y}{S_x^2} = r_{xy}\frac{S_y}{S_x}$$

- We have

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = (\bar{y} - \hat{\beta}_1\bar{x}) + \hat{\beta}_1 x_i,$$

# Back to Pearson correlation

- OLS estimator of slope:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2} = \frac{S_{xy}}{S_x S_y} \frac{S_x S_y}{S_x^2} = r_{xy} \frac{S_y}{S_x}$$

- We have

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x_i,$$

$$\hat{y}_i - \bar{y} = \hat{\beta}_1 (x_i - \bar{x}) = r_{xy} \frac{S_y}{S_x} (x_i - \bar{x}),$$

# Back to Pearson correlation

- OLS estimator of slope:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2} = \frac{S_{xy}}{S_x S_y}\frac{S_x S_y}{S_x^2} = r_{xy}\frac{S_y}{S_x}$$

- We have

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = (\bar{y} - \hat{\beta}_1\bar{x}) + \hat{\beta}_1 x_i,$$

$$\hat{y}_i - \bar{y} = \hat{\beta}_1(x_i - \bar{x}) = r_{xy}\frac{S_y}{S_x}(x_i - \bar{x}),$$

$$\frac{\hat{y}_i - \bar{y}}{S_y} = \hat{\beta}_1(x_i - \bar{x}) = r_{xy}\frac{(x_i - \bar{x})}{S_x}$$

# Back to Pearson correlation

- OLS estimator of slope:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2} = \frac{S_{xy}}{S_x S_y} \frac{S_x S_y}{S_x^2} = r_{xy} \frac{S_y}{S_x}$$

- We have

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x_i,$$

$$\hat{y}_i - \bar{y} = \hat{\beta}_1 (x_i - \bar{x}) = r_{xy} \frac{S_y}{S_x} (x_i - \bar{x}),$$

$$\frac{\hat{y}_i - \bar{y}}{S_y} = \hat{\beta}_1 (x_i - \bar{x}) = r_{xy} \frac{(x_i - \bar{x})}{S_x}$$

- Conclusion: $r_{xy}$ is the slope of the regression line for standardized data points.

# Outline

1. Pearson correlation coefficient
2. Simple linear regression
3. Estimating the coefficients
4. Assessing the accuracy of the coefficient estimates
5. Assessing the accuracy of the model

# Unbiasedness

- The OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased:

$$\mathbb{E}(\hat{\beta}_0) = \beta_0,\ \mathbb{E}(\hat{\beta}_1) = \beta_1$$

# Unbiasedness

- The OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased:

$$\mathbb{E}(\hat{\beta}_0) = \beta_0, \ \mathbb{E}(\hat{\beta}_1) = \beta_1$$

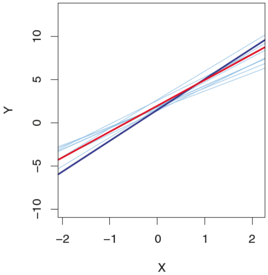



Figure: The simple linear regression is $Y = 2 + 3X + \epsilon$. The red line is the true regression function $2 + 3X$. The light blue lines are least squares lines for different sample. On average, the least squares lines are close to the true regression function.

# Standard error of OLS estimator [2]

- Standard error of $\hat{\beta}_0$:

$$SE(\hat{\beta}_0) = \sqrt{Var(\hat{\beta}_0)} = \sqrt{\sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}]}$$

  where $\sigma^2 = Var(\epsilon)$.

- Standard error of $\hat{\beta}_1$:

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

[2] Proofs are not required. See the extra slides in our course website.

# Standard error of OLS estimator [2]

- Standard error of $\hat{\beta}_0$:

$$SE(\hat{\beta}_0) = \sqrt{Var(\hat{\beta}_0)} = \sqrt{\sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}]}$$

  where $\sigma^2 = Var(\epsilon)$.

- Standard error of $\hat{\beta}_1$:

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

- When $\sigma$ is unknown, we need to estimate the standard error. Replace $\sigma^2$ by its estimator $\frac{SSE}{n-2}$ where $SSE = \sum_{i=1}^{n} e_i^2$

[2] Proofs are not required. See the extra slides in our course website.

# Confidence interval

- If
  1. The linear model is correct.
  2. The observations are independent.
  3. The variance around the true regression line is constant for all values of $x$.
  4. The random error around the true line is normal.

  Assumptions 2-4 are equivalent to $\epsilon_i \sim N(0, \sigma^2)$ with an unknown $\sigma$ and $\epsilon_i$ are i.i.d.

# Confidence interval

- If
  1. The linear model is correct.
  2. The observations are independent.
  3. The variance around the true regression line is constant for all values of $x$.
  4. The random error around the true line is normal.

  Assumptions 2-4 are equivalent to $\epsilon_i \sim N(0, \sigma^2)$ with an unknown $\sigma$ and $\epsilon_i$ are i.i.d.

  Then

$$\frac{\hat{\beta}_1 - \beta_1}{\widehat{SE(\hat{\beta}_1)}} \sim t_{n-2}$$

  where $\widehat{SE(\hat{\beta}_1)}$ is the estimated standard error of $\hat{\beta}_1$.

# Confidence interval

- If
  1. The linear model is correct.
  2. The observations are independent.
  3. The variance around the true regression line is constant for all values of $x$.
  4. The random error around the true line is normal.

  Assumptions 2-4 are equivalent to $\epsilon_i \sim N(0, \sigma^2)$ with an unknown $\sigma$ and $\epsilon_i$ are i.i.d.

  Then

$$\frac{\hat{\beta}_1 - \beta_1}{\widehat{SE(\hat{\beta}_1)}} \sim t_{n-2}$$

  where $\widehat{SE(\hat{\beta}_1)}$ is the estimated standard error of $\hat{\beta}_1$.
- Therefore,

$$\mathbb{P}(-t_{n-2,\alpha/2} \leq \frac{\hat{\beta}_1 - \beta_1}{\widehat{SE(\hat{\beta}_1)}} \leq t_{n-2,\alpha/2}) = 1 - \alpha$$

# Confidence interval continued [3]

- $100(1-\alpha)\%$ confidence interval of $\beta_1$ is

$$[\hat{\beta}_1 - t_{n-2,\alpha/2}\widehat{SE(\hat{\beta}_1)}, \hat{\beta}_1 + t_{n-2,\alpha/2}\widehat{SE(\hat{\beta}_1)}]$$

[3] The CI formulas here are slightly different with those in Section 3.1 of ISLR.

# Confidence interval continued [3]

- $100(1-\alpha)\%$ confidence interval of $\beta_1$ is

$$[\hat{\beta}_1 - t_{n-2,\alpha/2}\widehat{SE(\hat{\beta}_1)}, \hat{\beta}_1 + t_{n-2,\alpha/2}\widehat{SE(\hat{\beta}_1)}]$$

- Similarly, the $100(1-\alpha)\%$ confidence interval of $\beta_0$ is

$$[\hat{\beta}_0 - t_{n-2,\alpha/2}\widehat{SE(\hat{\beta}_0)}, \hat{\beta}_0 + t_{n-2,\alpha/2}\widehat{SE(\hat{\beta}_0)}]$$

[3] The CI formulas here are slightly different with those in Section 3.1 of ISLR.

# Hypothesis testing

- The most common hypothesis test in the simple linear regression is

$$H_0 : \text{There is no relationship between } X \text{ and } Y$$

vs.

$$H_A : \text{There is some relationship between } X \text{ and } Y$$

# Hypothesis testing

- The most common hypothesis test in the simple linear regression is

  $$H_0 : \text{There is no relationship between } X \text{ and } Y$$

  vs.

  $$H_A : \text{There is some relationship between } X \text{ and } Y$$

  Mathematically,

  $$H_0 : \beta_1 = 0 \text{ vs. } H_A : \beta_1 \neq 0,$$

  since if $\beta_1 = 0$, then the model reduces to $Y = \beta_0 + \epsilon$, and then $X$ is not associated with $Y$.

# Hypothesis testing

- The most common hypothesis test in the simple linear regression is
$$H_0 : \text{There is no relationship between } X \text{ and } Y$$
vs.
$$H_A : \text{There is some relationship between } X \text{ and } Y$$
Mathematically,
$$H_0 : \beta_1 = 0 \text{ vs. } H_A : \beta_1 \neq 0,$$
since if $\beta_1 = 0$, then the model reduces to $Y = \beta_0 + \epsilon$, and then $X$ is not associated with $Y$.
- Test statistic:
$$t = \frac{\hat{\beta}_1}{\widehat{SE(\hat{\beta}_1)}}$$

# Hypothesis testing

- The most common hypothesis test in the simple linear regression is
$$H_0 : \text{There is no relationship between } X \text{ and } Y$$
vs.
$$H_A : \text{There is some relationship between } X \text{ and } Y$$
Mathematically,
$$H_0 : \beta_1 = 0 \text{ vs. } H_A : \beta_1 \neq 0,$$
since if $\beta_1 = 0$, then the model reduces to $Y = \beta_0 + \epsilon$, and then $X$ is not associated with $Y$.
- Test statistic:
$$t = \frac{\hat{\beta}_1}{\widehat{SE(\hat{\beta}_1)}}$$
- Under same assumptions with confidence interval, if $H_0$ is true, then $t$ has a t distribution with $n - 2$ degrees of freedom.

# Hypothesis testing

- The most common hypothesis test in the simple linear regression is

  $$H_0 : \text{There is no relationship between } X \text{ and } Y$$

  vs.

  $$H_A : \text{There is some relationship between } X \text{ and } Y$$

  Mathematically,

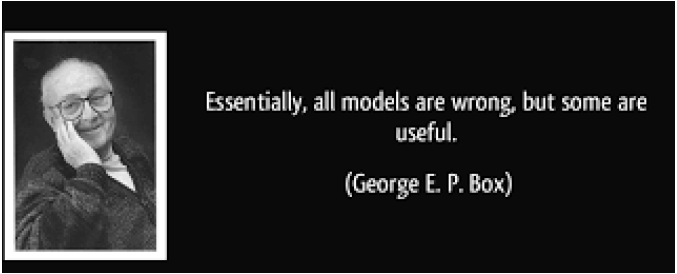
  $$H_0 : \beta_1 = 0 \text{ vs. } H_A : \beta_1 \neq 0,$$

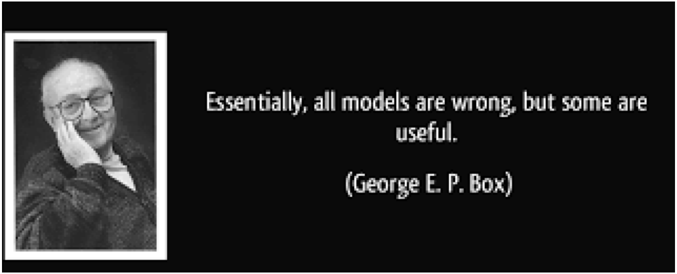
  since if $\beta_1 = 0$, then the model reduces to $Y = \beta_0 + \epsilon$, and then $X$ is not associated with $Y$.
- Test statistic:

  $$t = \frac{\hat{\beta}_1}{\widehat{SE}(\hat{\beta}_1)}$$
- Under same assumptions with confidence interval, if $H_0$ is true, then $t$ has a t distribution with $n - 2$ degrees of freedom.
- Exercise: For the father and son data, $\hat{\sigma} = 1.78$, so $\widehat{SE(\hat{\beta}_1)} = 0.24$, and $t_{obs} = 2.70$. Comparing this to a $t_{12}$, the p-value is 0.0193. So we would reject at the 5% level, and conclude that father's height is related to son's height.

# Outline


1. Pearson correlation coefficient
2. Simple linear regression
3. Estimating the coefficients
4. Assessing the accuracy of the coefficient estimates
5. Assessing the accuracy of the model


Essentially, all models are wrong, but some are useful.
(George E. P. Box)




How good does the linear model fit the data?

# R squared

- Total sum of squares:

$$SSTot = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Regression sum of squares:

$$SSReg = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- Residual sum of squares/ sum of squares error:

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

# R squared

- Total sum of squares:

$$SSTot = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

- Regression sum of squares:

$$SSReg = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

- Residual sum of squares/ sum of squares error:

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

- Sum of squares law:

$$\text{SSTot} = \text{SSReg} + \text{SSE}$$

# R squared continued

- R squared is defined as

$$R^2 = \frac{SSTot - SSE}{SSTot} = \frac{SSReg}{SSTot}.$$

# R squared continued

- R squared is defined as

$$R^2 = \frac{SSTot - SSE}{SSTot} = \frac{SSReg}{SSTot}.$$

- It's interpreted as the fraction of total sum of squares (variability) that is explained by the regression line.

# R squared continued

- R squared is defined as

$$R^2 = \frac{SSTot - SSE}{SSTot} = \frac{SSReg}{SSTot}.$$

- It's interpreted as the fraction of total sum of squares (variability) that is explained by the regression line.
- Exercise: for the father and son data, $R^2 = 0.38$. So we can say that about 38% of the variability in sons' heights can be explained by fathers' heights.

# What's the next?

We'll discuss how to use R to run linear regression in next lecture.