STA13: Elementary Statistics

Lecture 22

Book Sections: 2.9, parts of Chapter 9

Dmitriy Izyumin

March 12 2018

Linear Relationships

- Suppose we have two random variables X and Y.
- Are they dependent?
- If so, can we use one to predict the other?
- We will briefly consider linear relationships only.

Sign of the Relationship

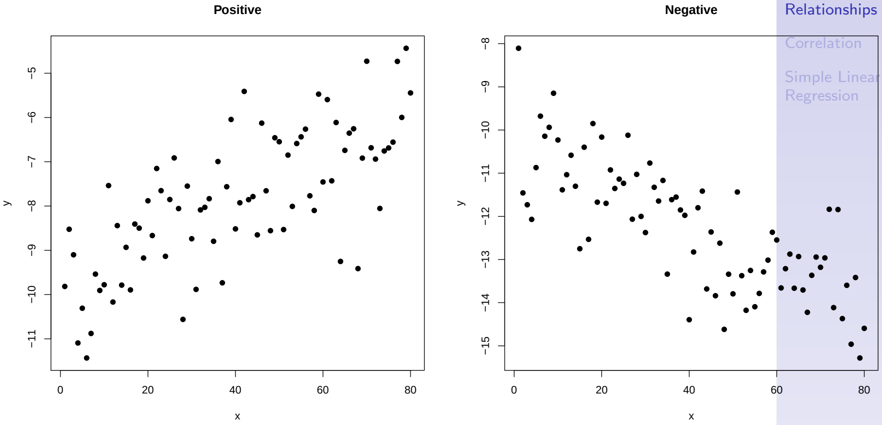
X and Y are positively linearly related if...

- If X increases, Y tends to also increase on average
- If X decreases, Y tends to also decrease on average

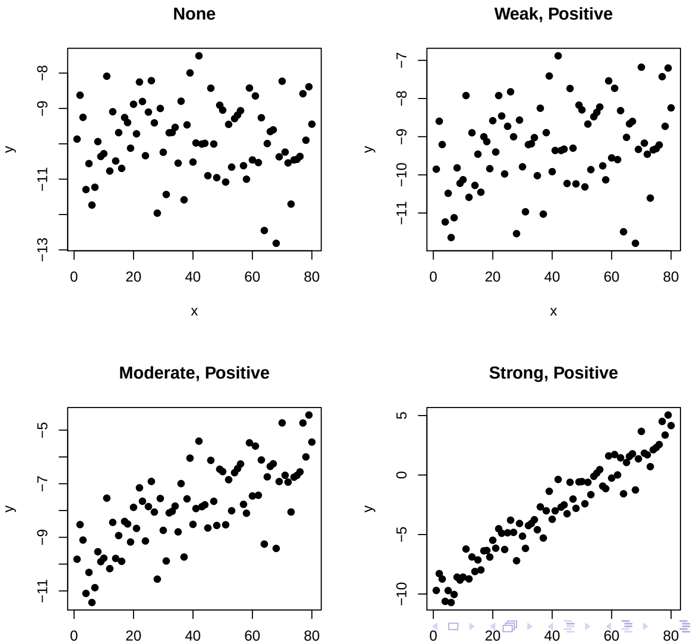
X and Y are negatively linearly related if...

- If X increases, Y tends to decrease on average
- If X decreases, Y tends to increase on average

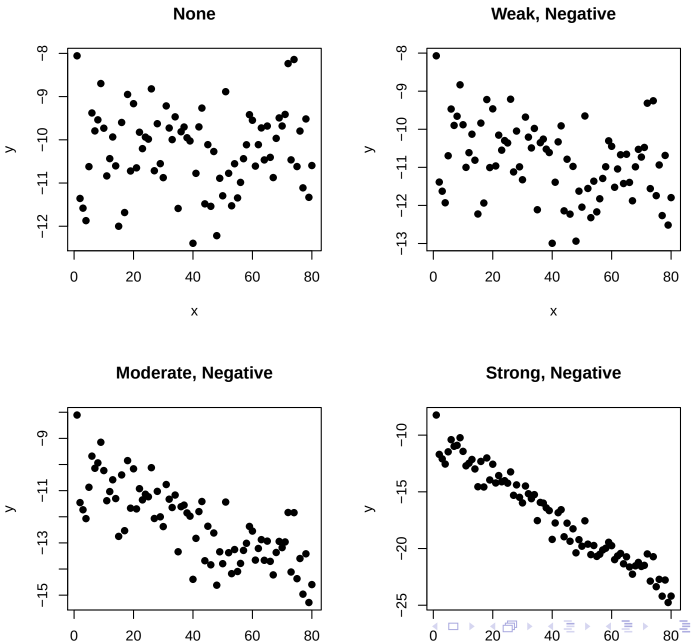
Sign of the Relationship



Strength of the Relationship



Strength of the Relationship



Coefficient of Correlation

The coefficient of correlation is a numerical measure of the strength of linear relationship between X and Y.

- Denoted by r.
- Computed from a sample of n pairs of (X, Y) values.
- Also known as Pearson's correlation coefficient.
- Also known as the sample correlation.

Coefficient of Correlation

- Value is between -1 and 1:

$$-1 \leq r \leq 1$$

- $r = 1$ means there is a perfectly linear positive relationship.
- $r = -1$ means there is a perfectly linear negative relationship.
- $r = 0$ means there is no linear relationship.
- $r = 0$ does not mean that there is no relationship between X and Y. It just means there is no *linear* relationship.

Coefficient of Correlation

- The farther $|r|$ is from 0, the stronger the linear relationship.
- $0 \leq |r| < 0.2$ no linear relationship, or a very weak linear relationship
- $0.2 \leq |r| < 0.4$ weak linear relationship
- $0.4 \leq |r| < 0.6$ moderate linear relationship
- $0.6 \leq |r| < 0.8$ strong linear relationship
- $0.80 \leq |r| < 1$ very strong linear relationship

Prediction

If two variables X and Y are related, a common goal is to predict Y given a value of X.

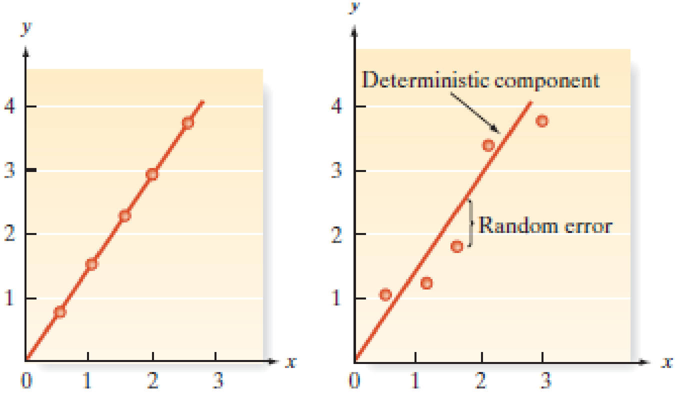
- Predict a person's height given their father's height.
- Predict precipitation given temperature.
- Can be extended to more variables.
  - Predict a person's life expectancy given their age, weight, blood pressure, etc.
  - Netflix: predict a person's rating of a movie given their customer information and viewing history.

Probabilistic Models

- Probabilistic models are used to model relationships that involve randomness.
- A probabilistic model has the form

  $Y =$ (deterministic part that depends on X)+(random error)
- Estimate model parameters from the sample. (Fit the model to the sample)
- Use the model to make inferences.

Probabilistic Relationship



a. Deterministic relationship: $y = 1.5x$

b. Probabilistic relationship: $y = 1.5x$ + Random error

Simple Linear Regression

- Very popular and influential type of probabilistic model.
- Idea has been around since the late 1800s.
- Extensive theory, and many variants exist.
- Covered in detail in Sta 108.
- We will just barely touch on some basics.

Simple Linear Regression

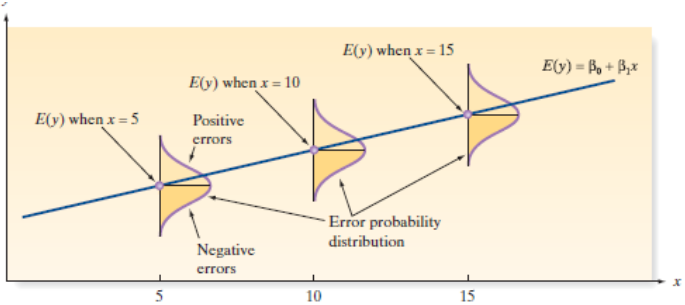
- Assume that Y depends *linearly* on X.

$$Y = \underbrace{\beta_0 + \beta_1 X}_{\text{depends on X linearly}} + \underbrace{\epsilon}_{\text{random}}$$

- On average, Y is linearly related to X. However there are random deviations from the relationship.

- Example: On average, people's weight is linearly related to their height. However, there are random deviations - people of the same height can have different weights.

Simple Linear Regression



Simple Linear Regression

- Assume that X and Y are linearly related.

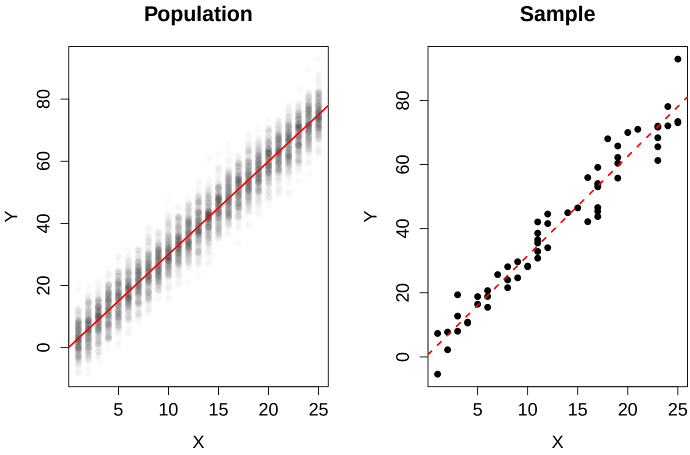
$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Assume that the ranfom errors (ϵ) are independent, and all distributed $N(0, \sigma^2)$.
- Use a sample to estimate $\beta_0, \beta_1, \sigma^2$.
- Check if the model seems to be a good fit.
- If it is a good fit, then we can use the model to make inferences.

Population and Sample

- Population: all possible (X, Y) pairs
- True Regression Line: $Y = \beta_0 + \beta_1 X$
- True line contains the expected values of Y at each value of X
- Sample: n pairs: $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$
- Estimated Regression Line: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

What We Assume and What We See



- Solid line is the true regression line $Y = \beta_0 + \beta_1 X$
- Dashed line is the estimated regression line $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

Inerpreting the Coefficients

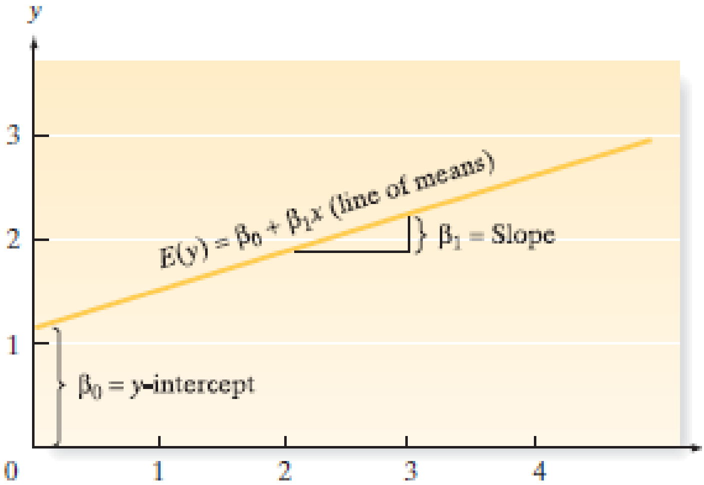
The estimated regression line is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

The slope $\hat{\beta}_1$ is the estimated average change in Y per unit increase in X. According to the estimated line, if X increases by 1, then Y increases by $\hat{\beta}_1$ on average.

The intercept $\hat{\beta}_0$ is the estimated average value of Y when $X = 0$. Keep in mind, that $X = 0$ may not be in the scope of the model, or may not make sense at all. That is alright. The quantity $\hat{\beta}_0$ is a feature of the model, but may not have a real world meaning.

Inerpreting the Coefficients



Fitted Values

- Start with a sample of pairs (X_i, Y_i), where $i = 1, \cdots, n$
- Fit a regression line $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
- Plug the sample values $X_1, \cdots, X_n$ into the equation to get the fitted values.
- Denoted by $\hat{Y}_1, \hat{Y}_2, \cdots, \hat{Y}_n$.
- Computed as $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ for $i = 1, \cdots, n$.

Sum of Squared Errors

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

- Sum of squared errors
- Squared differences between the observed values Y_i and the fitted values $\hat{Y}_i$.
- Serves as a measure of how well the estimated equation fits the sample.

How are the Coefficients Estimated?

- It is possible to fit many lines (make up values for slope and intercept).
- Each line would lead to different fitted values and a different SSE.
- Of all possible lines (or all values of slope and intercept), we use the one that leads to the smallest value of SSE.
- This is called least squares estimation.
- We will not be carrying out the estimation process in this class.

Predicting Y from X

- Check the range of X values in the sample. This is called the scope of the model.

- If X^* is in the scope, then we can predict the average value of Y at X^* as

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X^*.$$

- If X^* is not in the scope, then we cannot use the estimated line to predict the average value of Y at X^*.