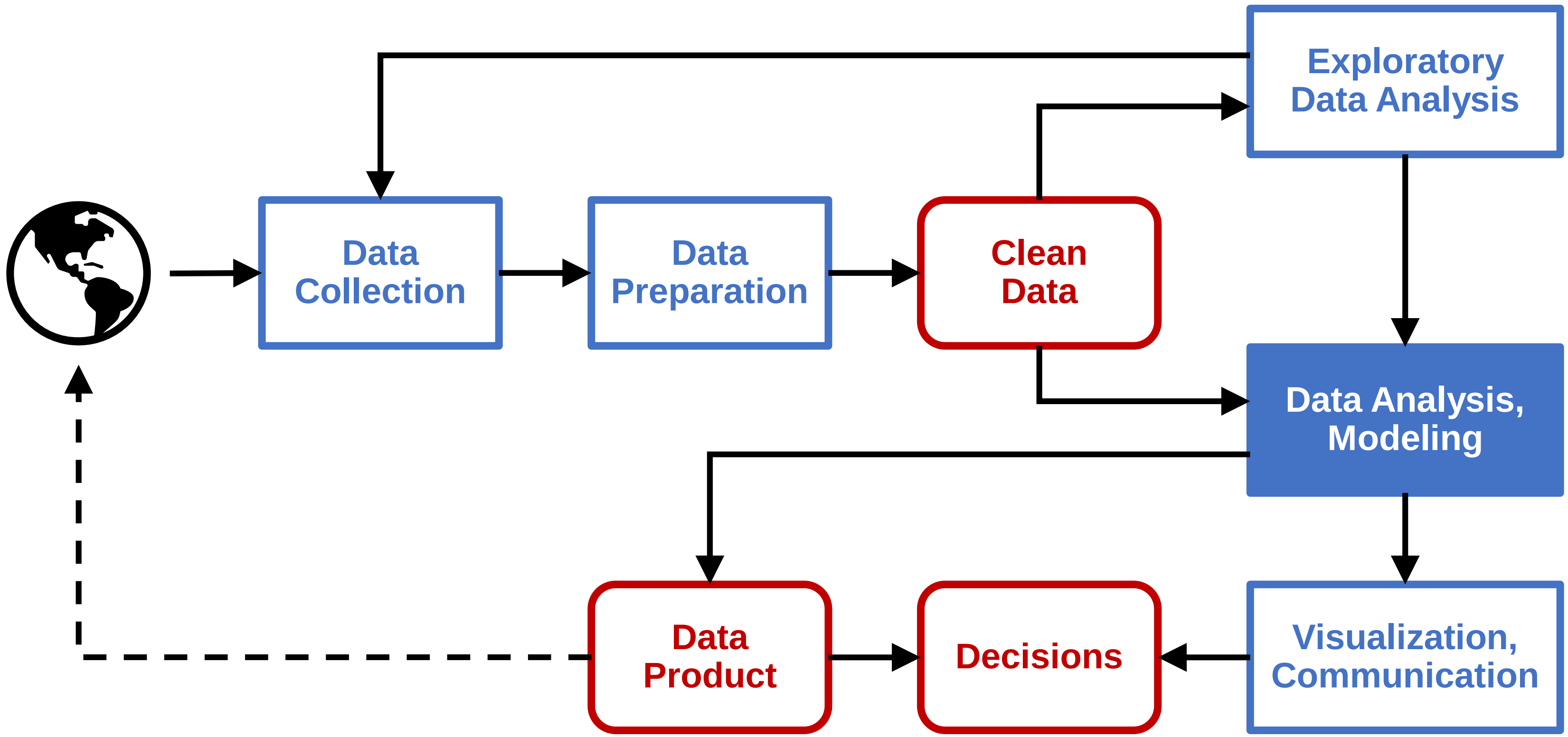
Regression:
Linear Regression

CS 418. Introduction to Data Science

Review
The Data Science Process

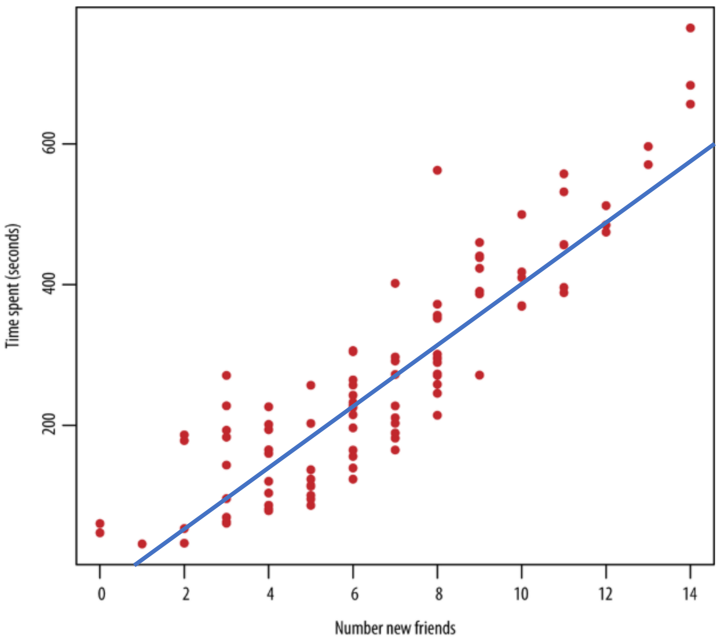
- The goal of the **data science process** is to **extract knowledge or insights from data**.



Adapted from: Cathy O'Neil and Rachel Schutt, *Doing Data Science* (2013)

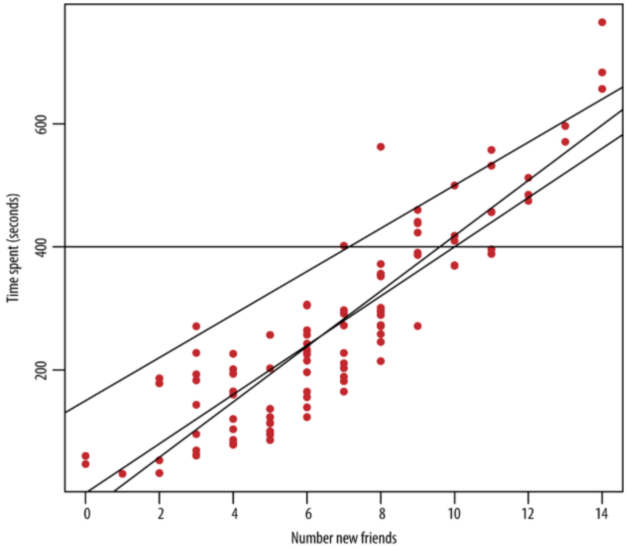
Linear Regression
Simple Linear Regression (I)

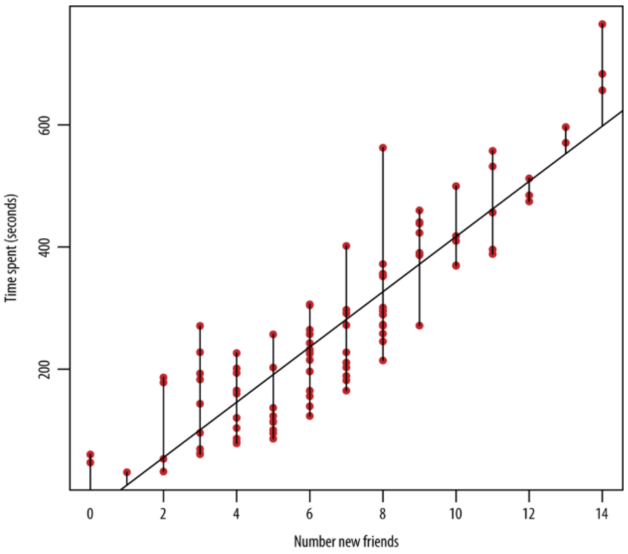
- **Simple linear regression** models a **linear relationship** between two quantitative variables.
- We can use this model to **predict** future outcomes of one of the variables or to **describe** the relationship between the variables.
- The variable being modeled or predicted is called the **response variable** (or **dependent variable** or **outcome** or **output**).
- The variable used to predict the response is called the **predictor variable** (or **independent variable** or **covariate** or **input**).



Linear Regression
Simple Linear Regression (II)

- The **simple linear regression** model is $\boldsymbol{Y = \beta_0 + \beta_1 X + \varepsilon}$ where β_0 and β_1 are the **regression parameters** and ε is the **regression error**.
- The **regression parameters** are typically unknown.
- We must estimate the **regression parameters** by finding the "best fitting" regression line for the observations.





Linear Regression
Least Squares Method

- The **least squares method** estimates the regression parameters by **minimizing the sum of the squared errors**.
- The **sum of squared errors** measures how far the observations are from the regression line.
- The **sum of squared errors** is given by the sum of the differences between the observed Y values and the values obtained from the linear regression model.

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

- For **simple linear regression**, the **estimated regression parameters** are given by:

$$\widehat{\beta_1} = r_{XY}\frac{s_Y}{s_X}$$

$$\widehat{\beta_0} = \bar{Y} - \beta_1\bar{X}$$

Linear Regression
Assessing Regression Parameters

- If $\boldsymbol{\beta_1 = 0}$, then no linear relationship exists between the response variable and the predictor variable.
- Given the **estimated** parameter $\widehat{\boldsymbol{\beta_1}}$, we can perform a $\boldsymbol{t}$**-test** to determine whether $\boldsymbol{\beta_1 = 0}$.
  - Set the **null and alternative hypotheses**: $\boldsymbol{H_0: \beta_1 = 0}$ and $\boldsymbol{H_a: \beta_1 \neq 0}$.
  - Compute the $\boldsymbol{t}$**-test statistic**, which is $\widehat{\boldsymbol{\beta_1}}$ divided by the standard error of $\widehat{\boldsymbol{\beta_1}}$.
  - Compute the **degrees of freedom**: $\boldsymbol{df = n - p}$, where $\boldsymbol{p}$ is the number of **regression parameters**.
  - Compute the corresponding $\boldsymbol{p}$**-value**.
  - Make a decision given a previously selected **significance level** $\boldsymbol{\alpha}$.
    - If $\boldsymbol{p}$**-value** $< \boldsymbol{\alpha}$, **reject** the null hypothesis.
    - If $\boldsymbol{p}$**-value** $\geq \boldsymbol{\alpha}$, there is **no sufficient evidence to reject** the null hypothesis.

Linear Regression
Coefficient of Determination

- The **coefficient of determination**, denoted by R^2, measures the **proportion of variance** of the response variable that is **explained** by the model.

$$R^2 = 1 - \frac{SSE}{SSTO} = \frac{SSR}{SSTO} = (r_{XY})^2$$

where:

$SSE = \sum_i \left(Y_i - \widehat{Y_i}\right)^2$ is the **residual sum of squares**.

$SSR = \sum_i \left(\widehat{Y_i} - \overline{Y}\right)^2$ is the **regression sum of squares**.

$SSTO = \sum_i (Y_i - \overline{Y})^2 = SSR + SSE$ is the **total sum of squares**.

r_{XY} is the **linear correlation coefficient**.

- The value of R^2 is **between 0 and 1**. The higher the value, the better the fit of the model.

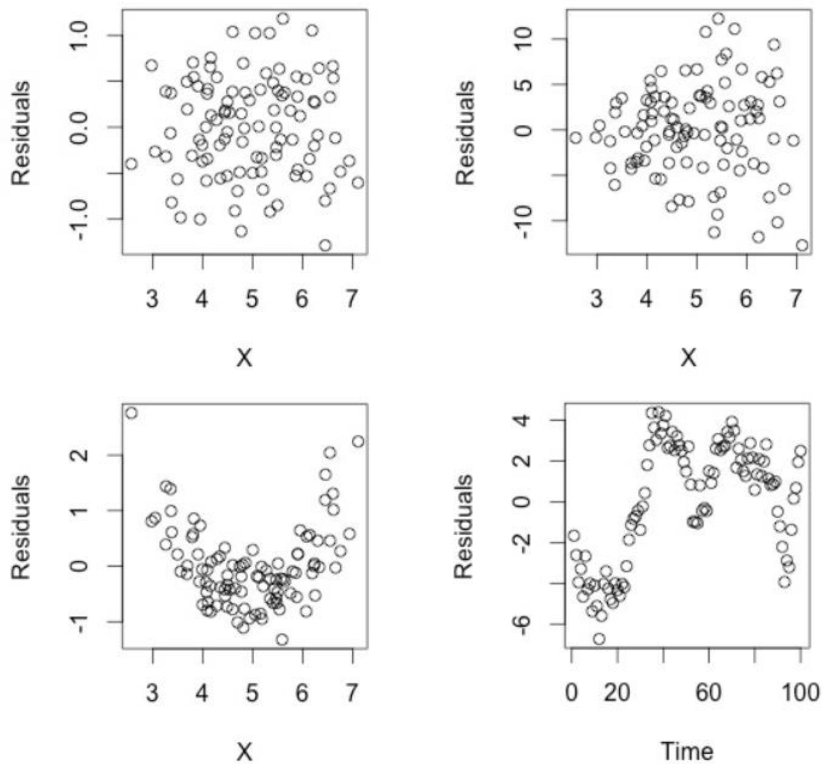
Linear Regression Assumptions

- There are four main **assumptions** that justify the use of a **linear regression** model:
  - **Linearity**. There is a **linear relationship** between the response variable and the predictor variable.
  - **Normality**. Errors are **normally distributed** with a **mean of 0**.
  - **Homoscedasticity**. Errors have **constant variance**.
  - **Independence**. Errors are **independent** of each other.
  - **How to diagnose?**
    - Analyze **scatterplots of residuals**.
    - Assessing whether a plot supports an **assumption** is subjective and requires a reasonably large sample size (at least 30 observations).
    - Ideally, all **assumptions** should hold for a model to be valid. However, **simple linear regression** models are reasonably robust to mild violations of the **assumptions**.

Linear Regression
Exercise 8.1

THINK
PAIR
SHARE

Analyze each of the following scatterplots of residuals.
Do the assumptions of linear regression hold for each plot?
Why or why not?



References

- Daniel Chen. *Pandas for Everyone* (2018).
- Joel Grus. *Data Science from Scratch* (2015).
- Cathy O’Neil and Rachel Schutt. *Doing Data Science* (2013).