8. REGRESSION ANALYSIS

Correlation, simple linear regression, multiple linear regression, and other regression models

Correlation

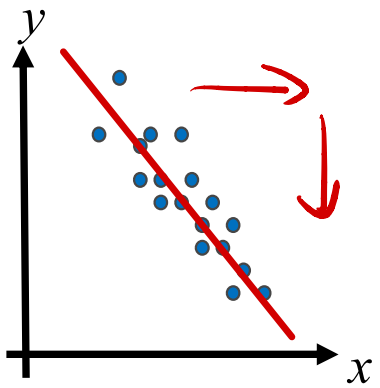
Correlation

- We are interested in looking at the relationship between two variables.

- The data can be represented by ordered pairs (x, y)
  - x is the independent (or explanatory) variable
  - y is the dependent (or response) variable

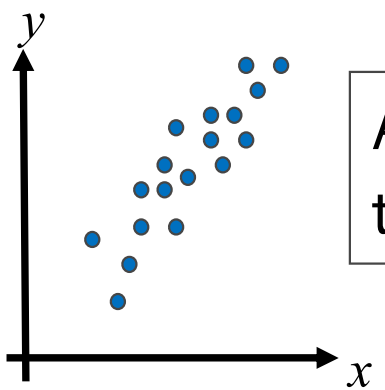
- We can draw a scatter plot to visually inspect the relationship

Types of correlation



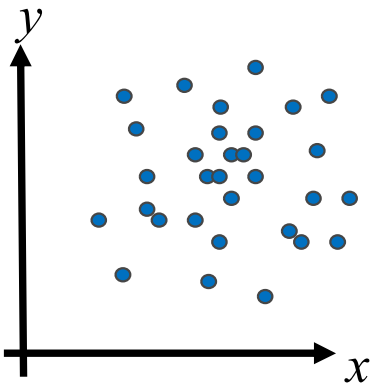
As x increases, y tends to decrease.

Negative Linear Correlation

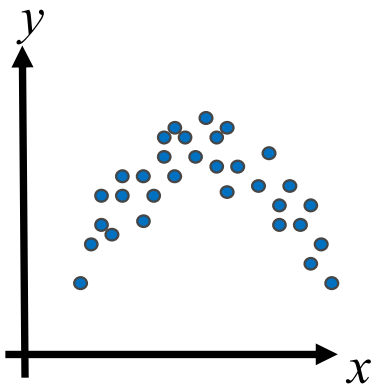


As x increases, y tends to increase.

Positive Linear Correlation



No Correlation



Nonlinear Correlation

Correlation coefficient

- A measure of the strength and the direction of a linear relationship between two variables.
- The symbol r represents the sample correlation coefficient.
- A formula for r is

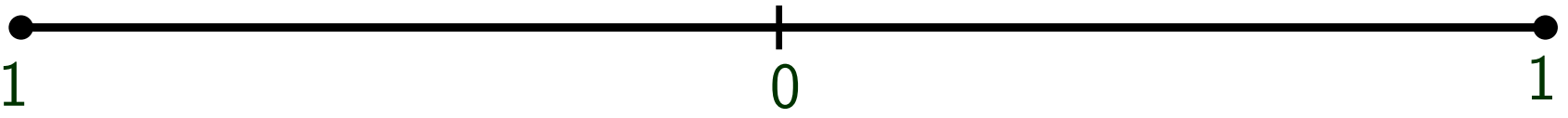
$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

- This is also called the Pearson correlation coefficient.

- The population correlation coefficient is represented by ρ (rho).

Correlation coefficient

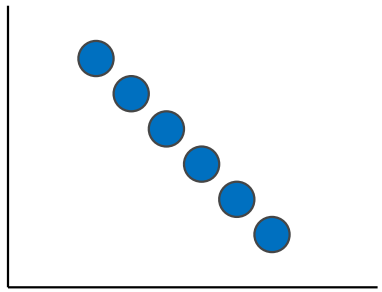
- The range of the correlation coefficient is -1 to 1.



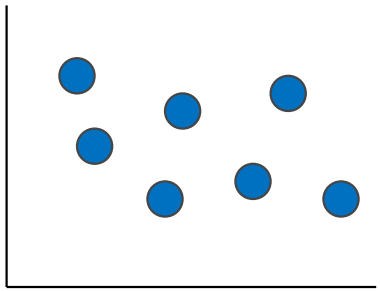
If $r = -1$ there is a perfect negative correlation

If r is close to 0 there is no linear correlation

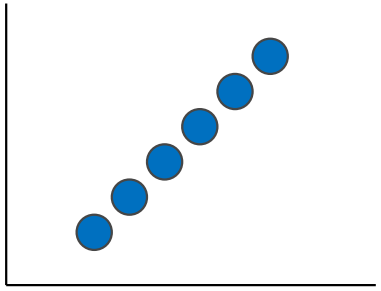
If $r = 1$ there is a perfect positive correlation



-1

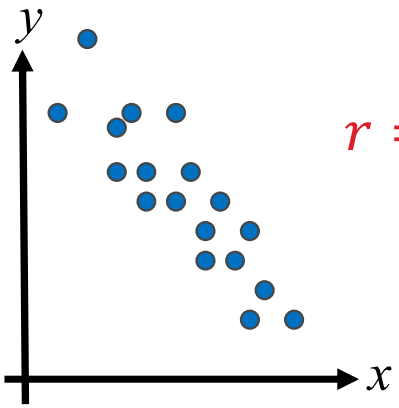


0



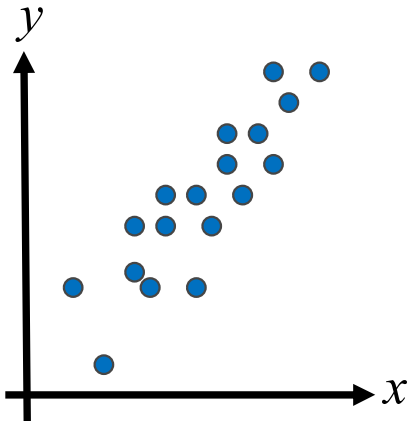
1

Correlation coefficient



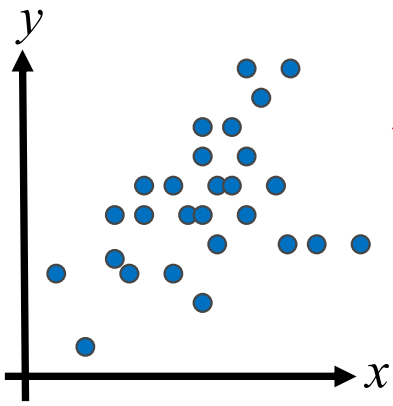
$r = -0.91$

Strong negative correlation



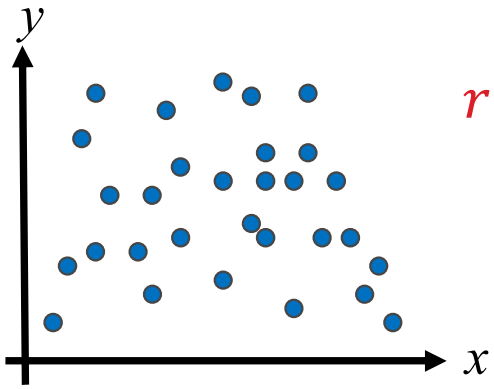
$r = 0.88$

Strong positive correlation



$r = 0.42$

Weak positive correlation



$r = 0.07$

Weak Correlation

Introduction to regression analysis

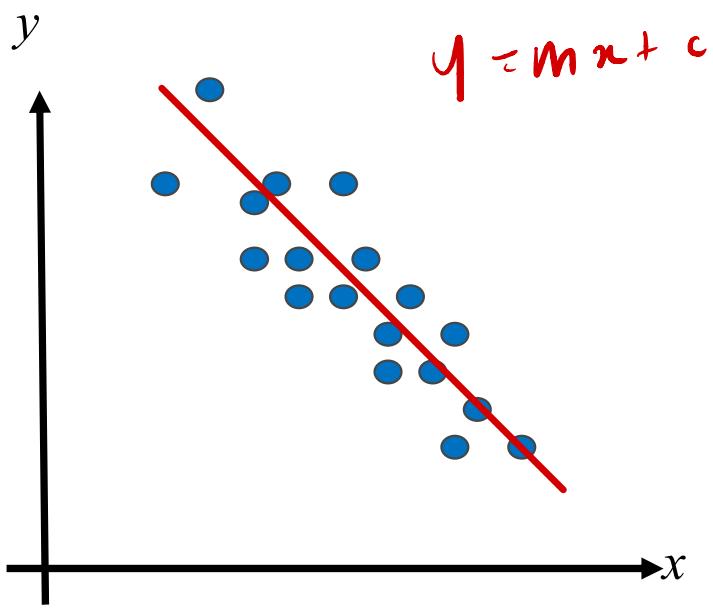
Introduction to regression analysis

- Regression analysis is a statistical technique that is useful for studying relationship between variables.
- For examples:
  - Relationship between expenses and monthly income of one family.
  - Relationship between air pollution rates and the number of vehicles on the road.
  - Relationship between age and salary.
- In regression analysis, we study the
  - relationship between 2 or more variables
  - predict the value of the variable of interests.

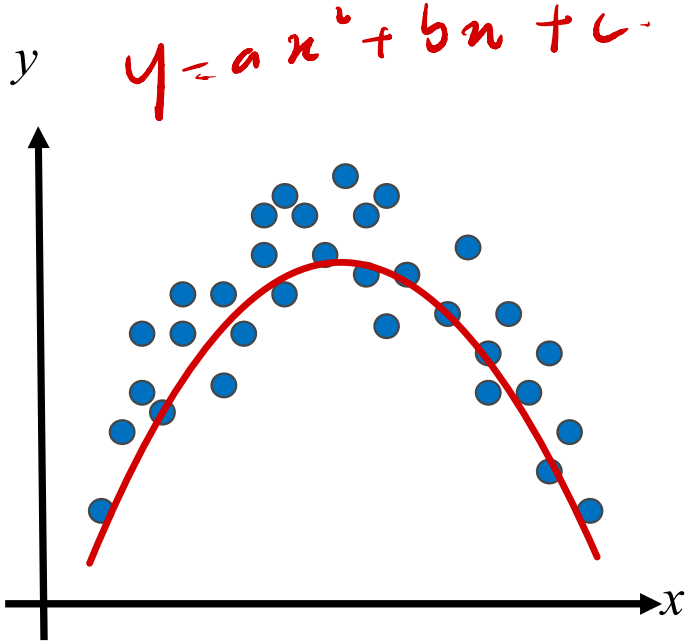
What about linear regression?

- The most basic type of regression, is the linear regression.
- For linear regression, we assume that there is an underlying linear relationship between the dependent variable y and the independent variable x.
- Other types of regression:
  - Polynomial regression
  - Poisson regression
  - Binomial regression
  - Logistic regression

Linear regression



- Has a linear relationship
- Can apply linear regression



- Does not have a linear relationship
- Cannot apply linear regression
- Maybe use polynomial regression

Basic steps in regression analysis

1. Plot the variables and look at the relationship between them.
2. Does the relationship fit with the model assumption?
3. Predict the parameters in the model.
4. Check for the suitability of the model build based on the data collected. Should the model be modified or accepted?
5. Prediction from the model

Examples of regression models

Simple linear regression

- Model and assumption:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

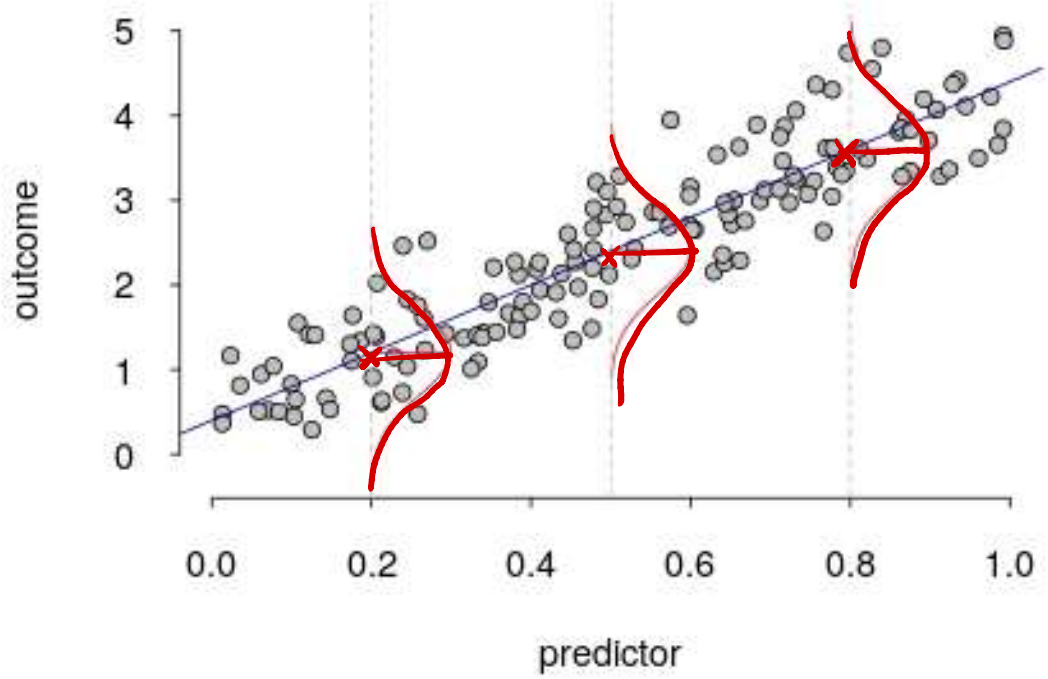
where $\varepsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$, $i = 1,2,\ldots,n$

- Used for:
  - Modelling linear relationship between two variables, x and y
  - The variable y is a continuous variable

Simple linear regression

- From the model, we can write down

$$y_i|x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$



Example (oxygen purity)

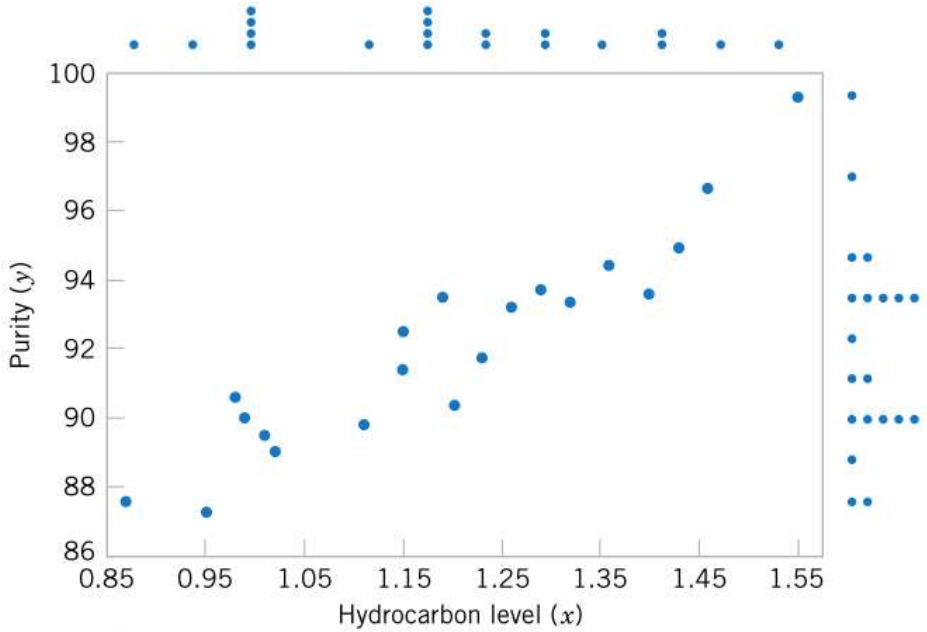


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

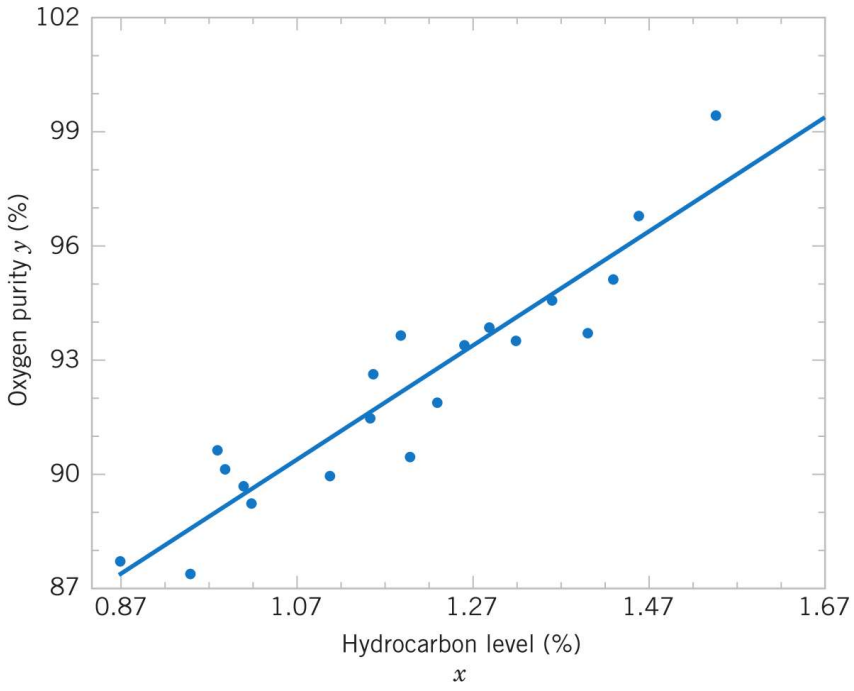


Figure 11-4 Scatter plot of oxygen purity y versus hydrocarbon level x and regression model $\hat{y} = 74.283 + 14.947x$.

Multiple linear regression

- Model and assumption:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

$$= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i$$

where $\varepsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$, $i = 1,2, \ldots, n$

- Used for:
  - Similar to simple linear regression, but now with more than one predictor/regressor variable
  - The variable y is a continuous variable

Example (wire pull strength)

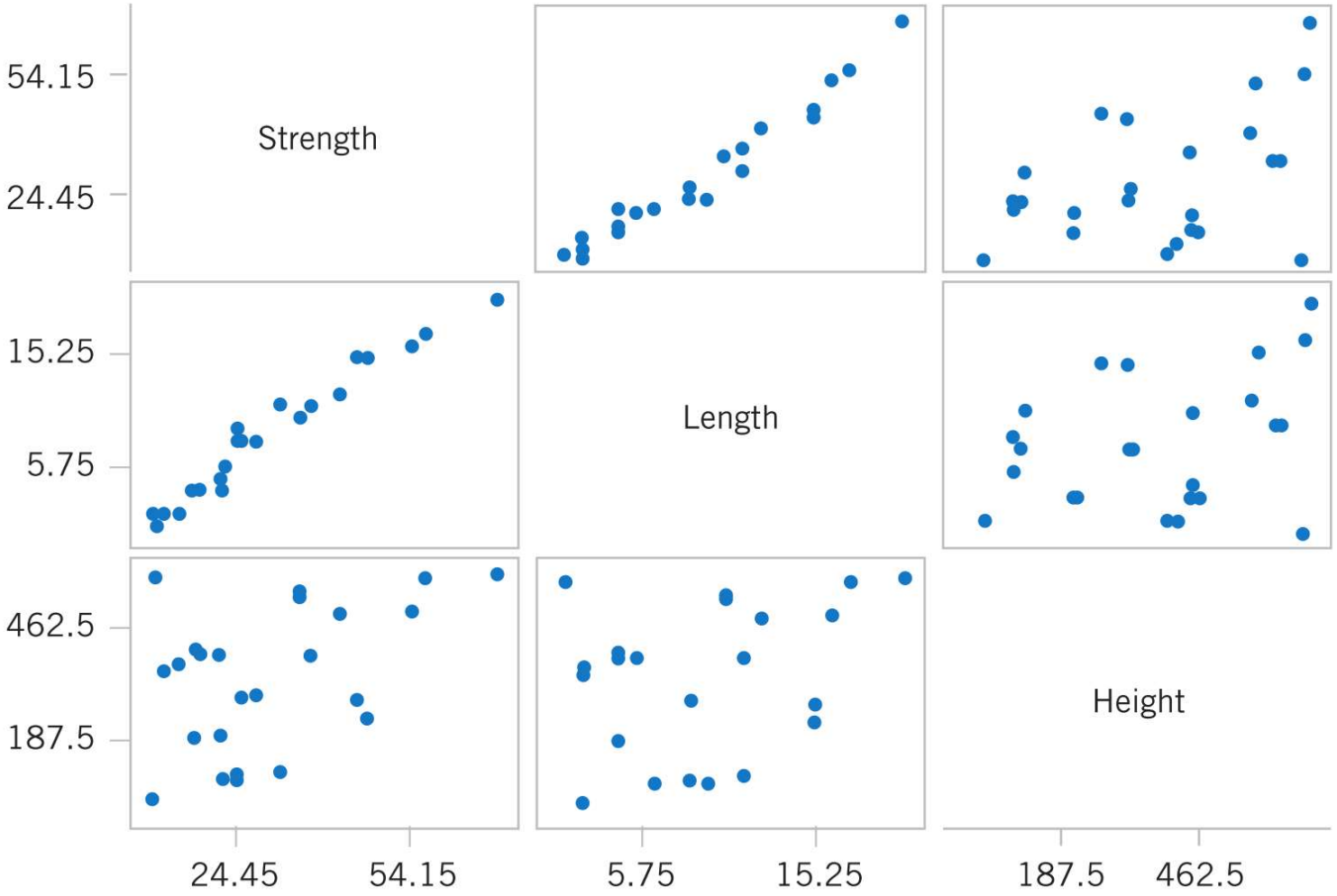


Figure 12-4 Matrix of scatter plots (from Minitab) for the wire bond pull strength data in Table 12-2.

Plot indicates strong linear relationship between strength and wire length.

Example (wire pull strength)

- The model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

y: wire pull strength

x_1: wire length

x_2: die height

ε: random error term

- The estimated parameters:

$$\hat{\beta}_0 = 2.26379, \qquad \hat{\beta}_1 = 2.74427, \qquad \hat{\beta}_2 = 0.01253$$

- Therefore, the fitted regression is

$$\hat{y} = 2.26379 + 2.74427x_1 + 0.01253x_2$$

Polynomial regression

- Model and assumption:

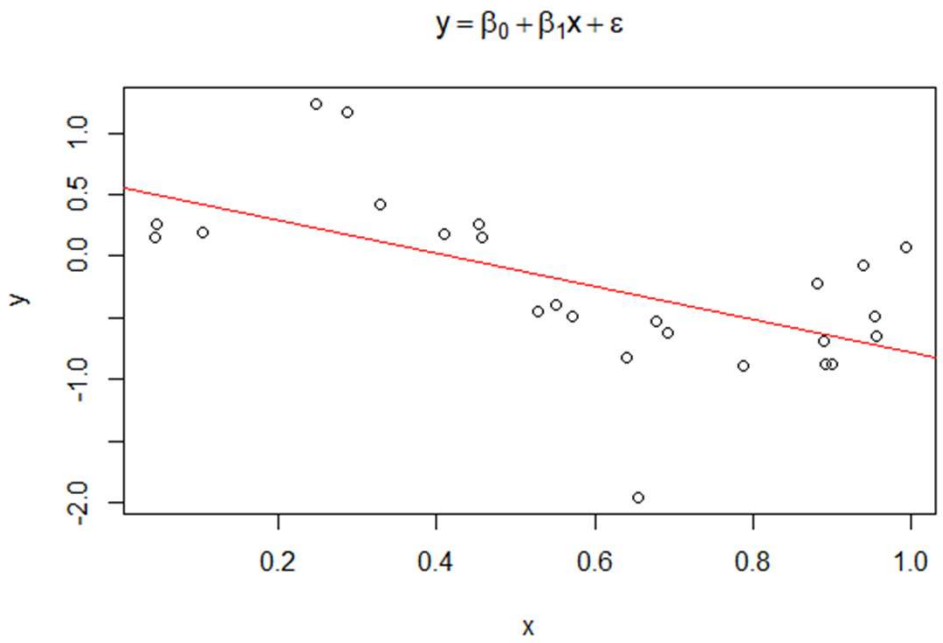
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \varepsilon_i$$

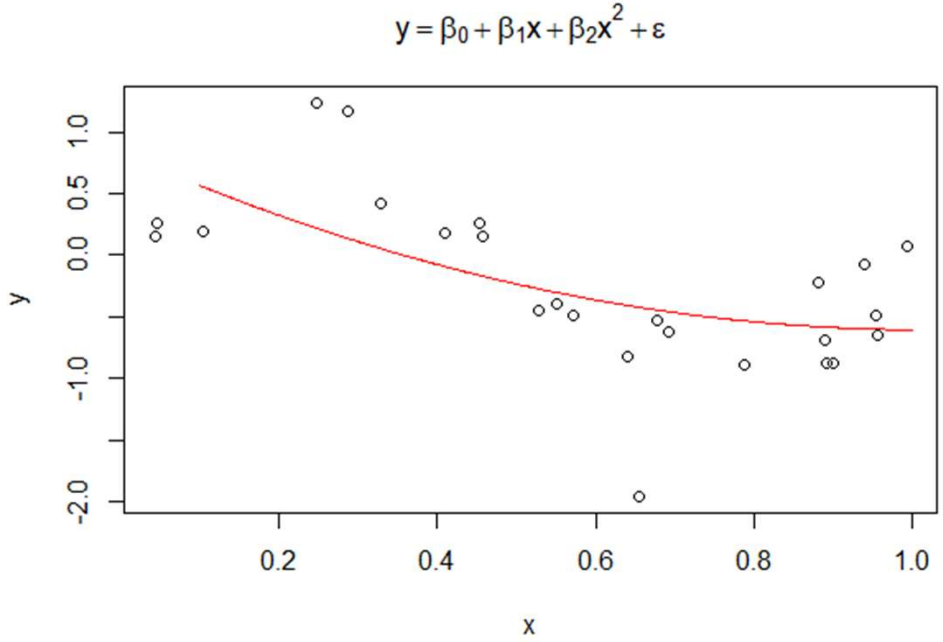
$$= \beta_0 + \sum_{j=1}^{k} \beta_j x_i^j + \varepsilon_i$$

where $\varepsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$, $i = 1,2, \ldots, n$

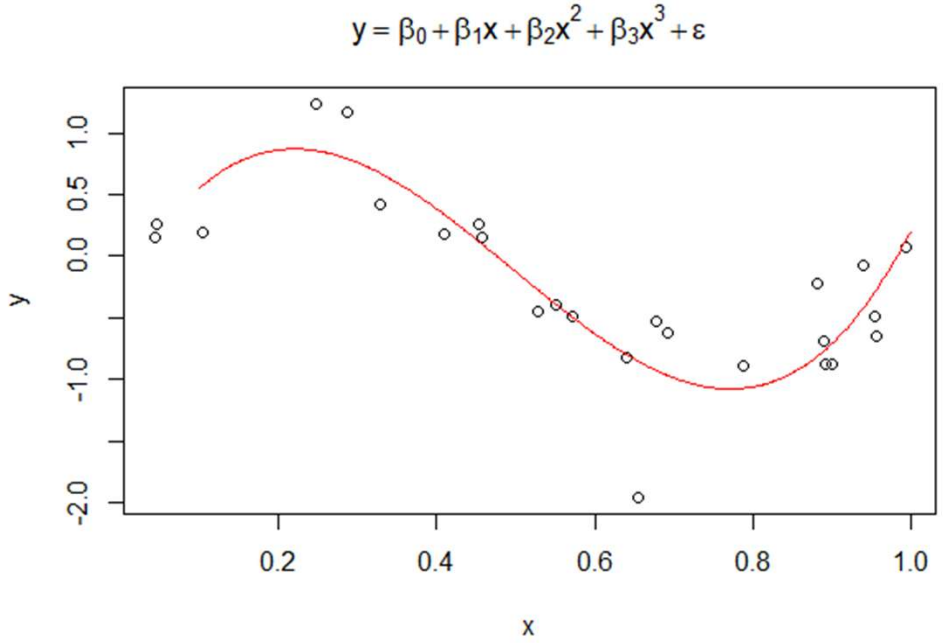
- Used for:
  - Modelling polynomial relationship between two variables, x and y
  - The variable y is a continuous variable
  - Works similar to multiple linear regression but uses polynomial of x as regressors

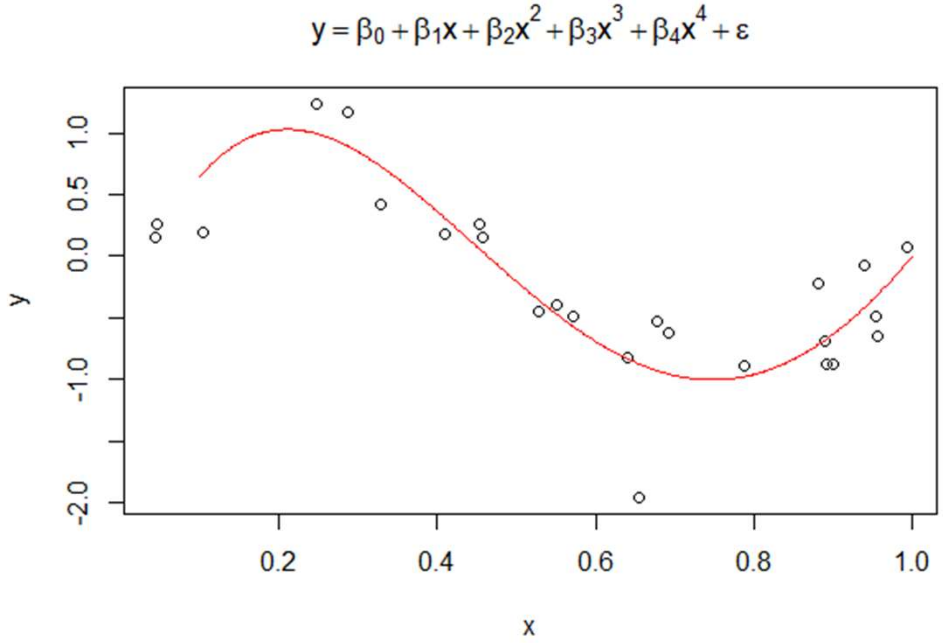
Example (polynomial)





Example (polynomial)





Poisson regression

- Model and assumption:

$$y_i \sim \text{Poi}(\mu_i)$$

where

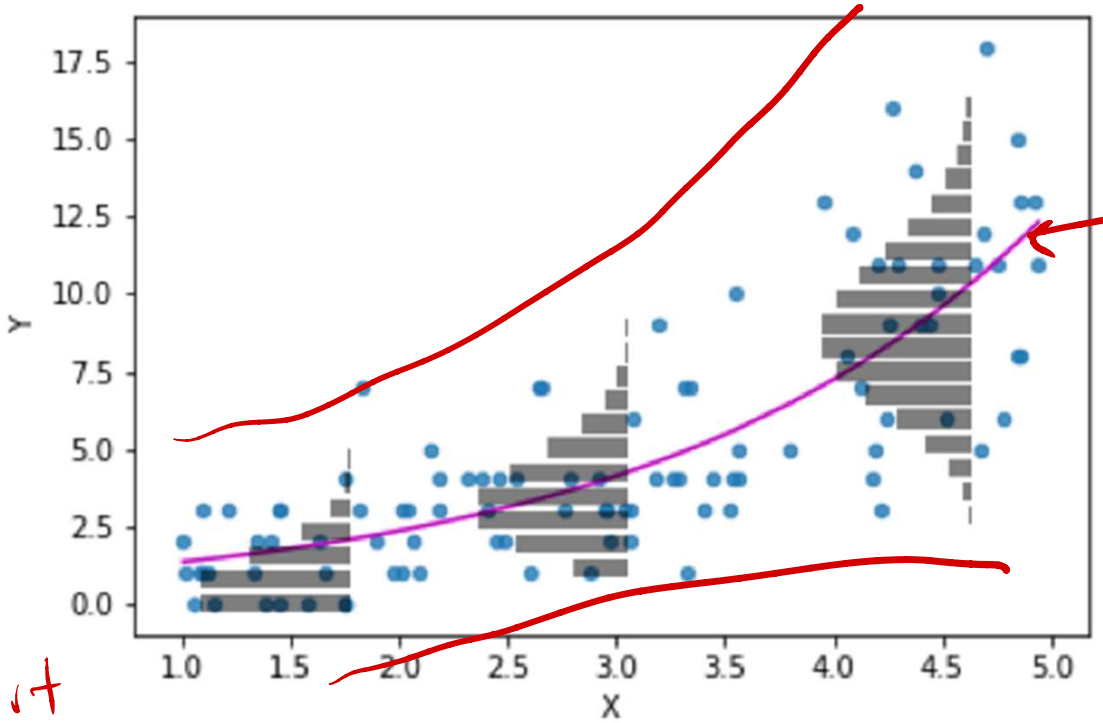
$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

$$= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

- Used for:
  - Modelling count or rate data

Poisson regression

$y_i \sim Pois(\mu_i)$



μ_i

$\mu \uparrow$,

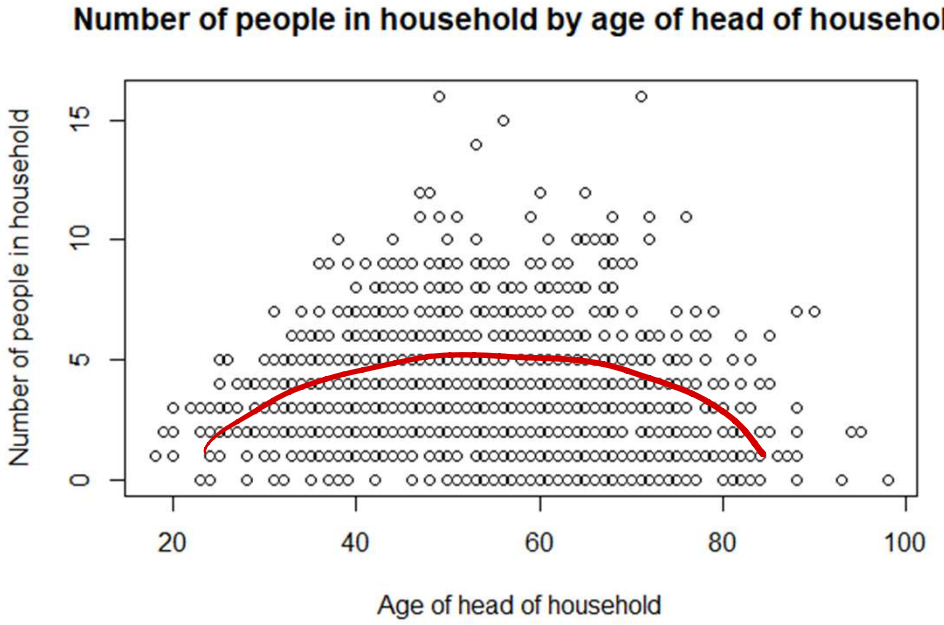
var ↑

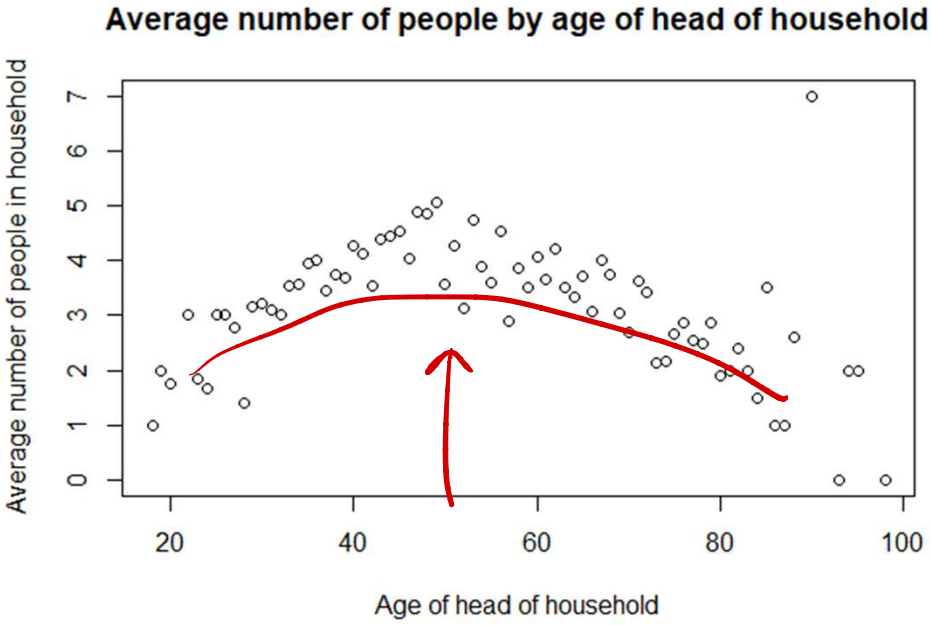
Poisson dist

Mean = var

Example (household size)

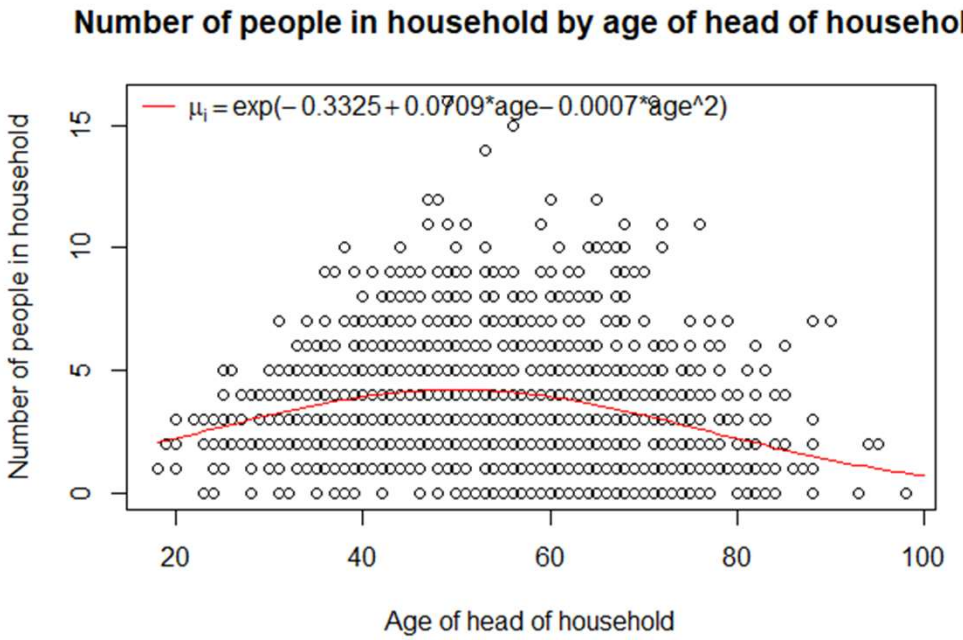
- A survey was conducted by The Philippine Statistics Authority to study family's income and expenditure. The data also include the household size and the age of the head of household. We are interested in studying how the age of head of household relates to the number of people in the household, using the 2015 data and 1500 households.

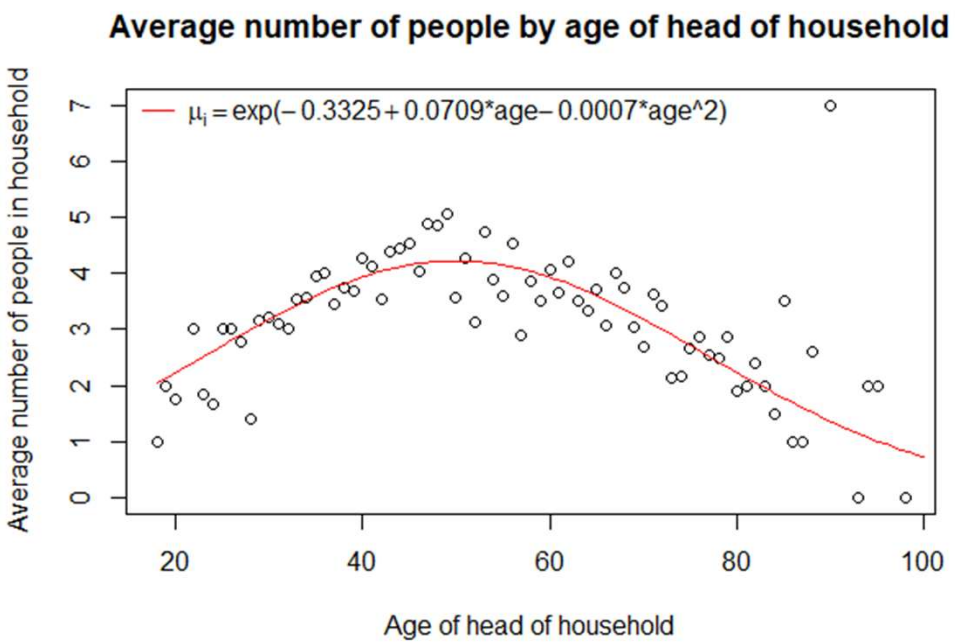




Example (household size)

- Poisson regression:





Logistic regression

- Model and assumption:

$$y_i \sim \text{Bernoulli}(p_i)$$

where

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

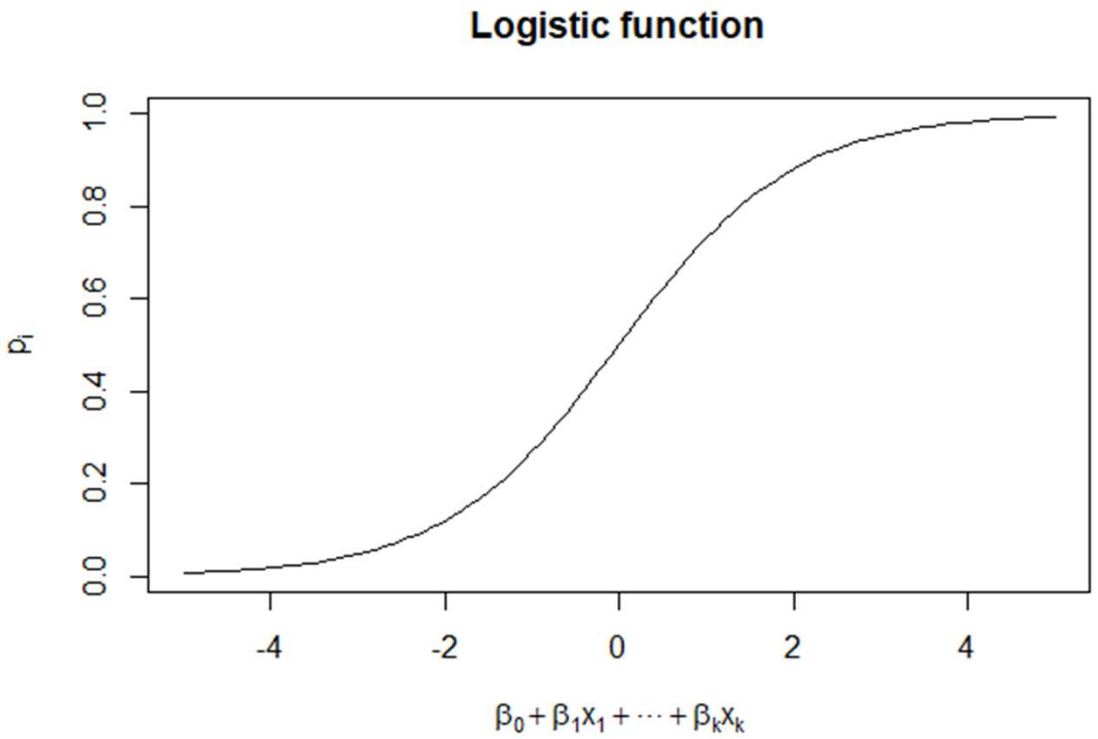
$$= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

- Used for:
  - Modelling binary data (0 or 1)
  - Classifying data into two categories

Logistic regression

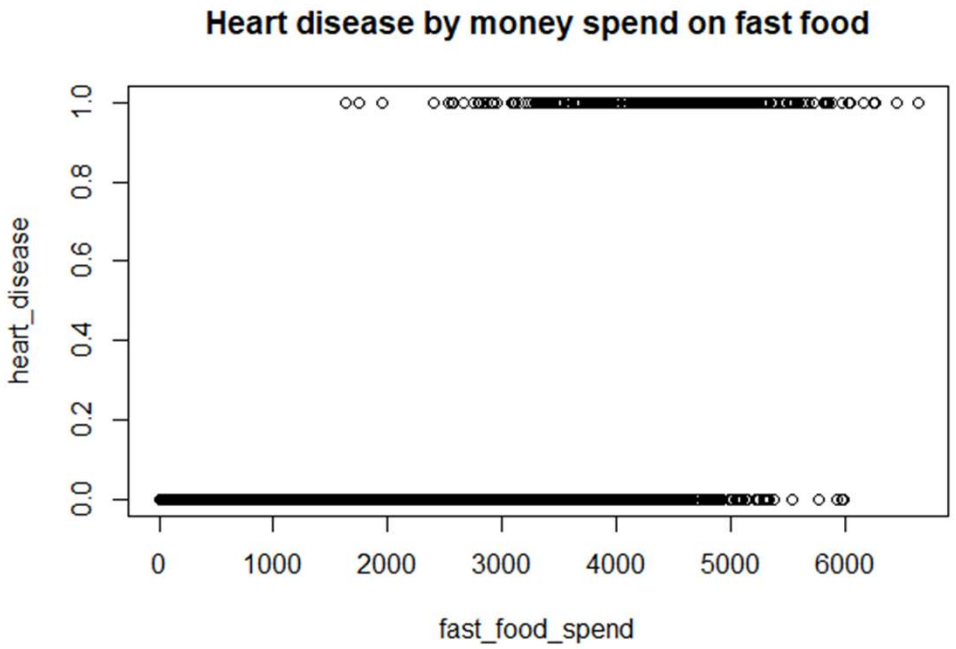
- By inversing the function,

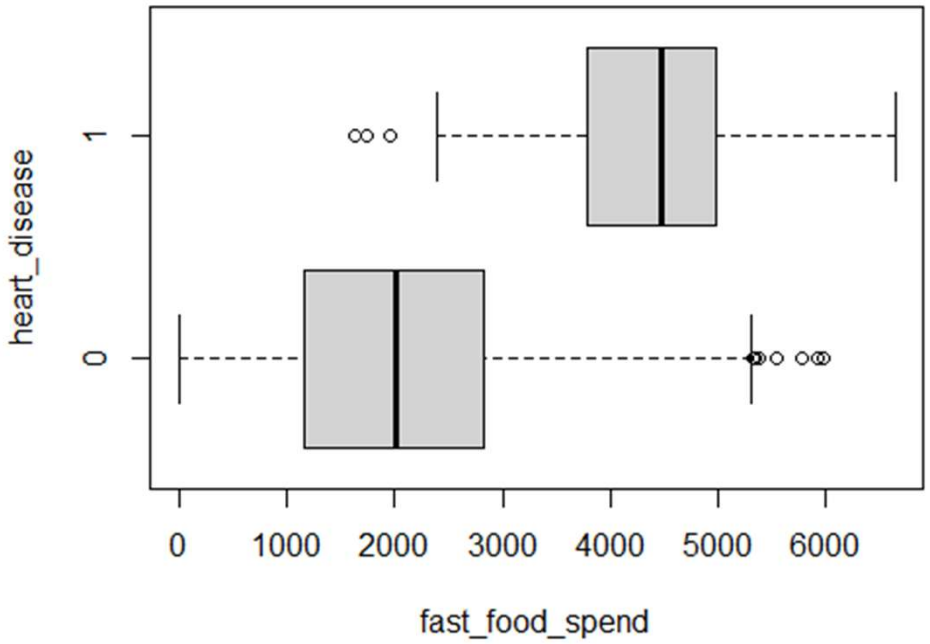
$$p_i = \text{logistic}\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)}$$



Example (heart disease)

- Heart disease is often associated with fast food. You are given a set of data that shows how much a person spend on fast food annually, and whether or not the person has a heart disease.

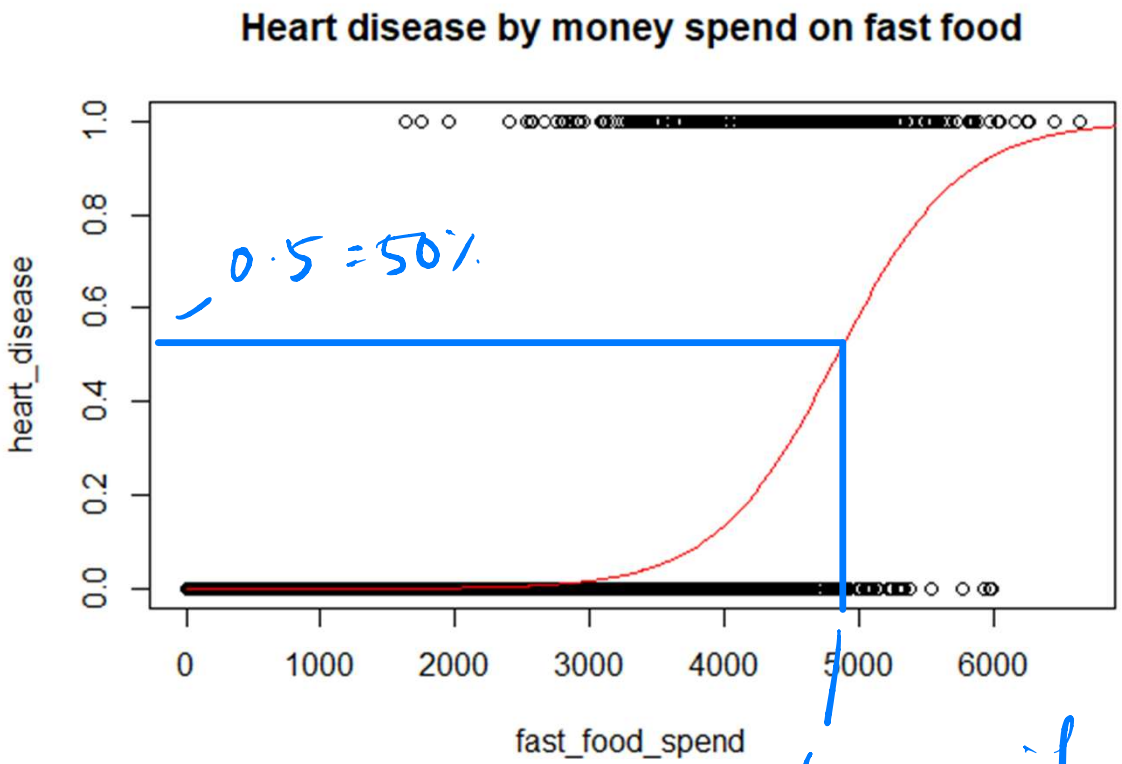




Example (heart disease)

- Using logistic regression:

$$\text{Prob}(\text{heart\_disease}) = \text{logistic}(-10.6513 + 0.0022 \times \text{fast\_food\_spend})$$



Summary

- Correlation

- Some examples of regression models:
  - Simple linear regression
  - Multiple linear regression
  - Polynomial regression
  - Poisson regression
  - Logistic regression