Packet 5: Linear Regression

Chap 6.5 A Simple Regression Problem

There is often interest in the relation between two variables. Simple linear regression is a way of evaluating the relationship between two continuous variables. One variable is regarded as the predictor variable, explanatory variable, or independent variable denoted by X. The other variable is regarded as the response variable, outcome variable, or dependent variable denoted by Y.

For example, we might we interested in investigating the relationship between:

- heights v.s. weights
- high school grade point averages v.s. college grade point averages
- outdoor temperature v.s. evaporation rate
- alcohol consumed v.s. blood alcohol content

Learning Objects:

- Learn theoretical background of regression models.
- Make statistical inferences using confidence intervals and prediction intervals.
- Learn how to interpret the analysis outcomes.

The Meaning of Regression

We need to distinguish between two different type of relationships, namely:

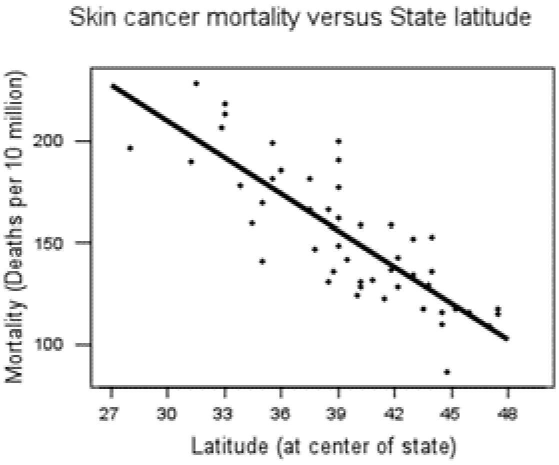
- Deterministic relationships:

  A deterministic (or functional) relationship is an exact relationship between the predictor X and the response Y. Take, for instance, the conversion relationship between temperature in degrees Celsius (C) and temperature in degrees Fahrenheit (F). We know the relationship is: F=1.8C+32.

- Statistical relationships:

  A statistical relationship is not an exact relationship but a trend between the predictor X and the response Y with uncertainty.

Researchers investigated the relationship between the latitude (in degrees) at the center of each of the 50 U.S. states and the mortality (in deaths per 10 million) due to skin cancer in each of the 50 U.S. states. As the latitude increases for the northern states, in which sun exposure is less prevalent and less intense, mortality due to skin cancer decreases, but not perfectly so.



In regression, instead of predicting the exact value of Y, we estimate $E(Y|X = x) = f(x)$ as a function of the predictor x which is known in advance.

- linear regression: $f(x) = \alpha + \beta x$.
- non-linear regression: $f(x) = \alpha e^{\beta x}$.
- classification and regression tree: $f(x) = \sum_k \beta_k 1_{\mathrm{X \in S_k}}$.

Suppose $(X_1, Y_1) \ldots (X_n, Y_n)$ are a random sample from a regression model

$$Y_i = f(X_i) + \epsilon_i,$$

where $f(x)$ is called a regression function , $\epsilon_i \sim N(0, \sigma^2)$, called a random error.

To study the relationship between X and Y, we start from a simple case, $f(x) = \alpha_1 + \beta x$.

$$Y_i = \alpha_1 + \beta X_i + \epsilon_i = (\alpha_1 + \beta \bar{X}) + \beta(X_i - \bar{X}) + \epsilon_i.$$

Let $\alpha = \alpha_1 + \beta \bar{X}$, we have

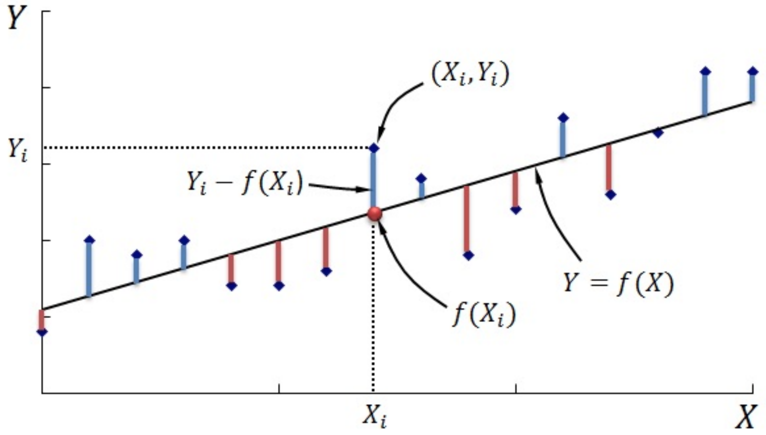
$$Y_i = \alpha + \beta(X_i - \bar{X}) + \epsilon_i.$$

Parameter interpretations:

Question: How to estimate α, β, and σ^2 (variance of ϵ_i)?

The method of least squares

It can be traced back to the late 1700's (Gauss, Legendre, Laplace, etc.). The basic idea is to choose a function of the predictors, $f(X)$, so that the sum of squared distances from the data to the function values is minimized.



Mathematically, finding a straight line, $y = \alpha + \beta_1(x - \bar{x})$, that minimizes the sum of squared "vertical distance" from the paired observations, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, to the values on the line, $\{(x_1, \alpha + \beta(x_1 - \bar{x})), \ldots, (x_n, \alpha + \beta(x_n - \bar{x}))\}$, is to find the values of (α, β) such that

Following the standard procedures for minimizing a function, we find the partial derivatives

Setting both to zero we get

From these, we solve for α and β, and treat them as estimates (i.e., putting on hats).

These are called the least square estimators for α and β_1. It is crucial to note that $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear functions of the y_i's. Later, we will treat dependent variables as r.v.s denoted by Y_i's (not y_i's).

Conventional notation:

If we use these notations, it is straightforward to show that

Similar procedure applies to more general situations with multiple predictors $x_1, x_2, \ldots, x_p$. We can try to find a function $f(x_1, x_2, \ldots, x_p)$, e.g., a linear function

$$f(x_1, x_2, \ldots, x_p) = \alpha + \beta_1(x_1 - \bar{x}_1) + \beta_2(x_2 - \bar{x}_2), \ldots \beta_p(x_p - \bar{x}_p).$$

such that the sum of "vertical" squared distances from the data to the function is minimized:

$$\sum_{i=1}^{n} \left(y_i - f(x_{1i}, x_{2i}, \ldots, x_{pi})\right)^2 .$$