Lecture 1: Introduction

(Text Sections 6.1 and 6.3)

Prerequisite Material

- Concepts of estimation, sampling distributions, and hypothesis testing
- Experience with t-tests, ANOVA, linear regression, chi-squared tests of independence for 2-dimensional contingency tables
- Some knowledge of matrix algebra and calculus
- Basic familiarity with a statistical computing package (e.g. S-PLUS)

Overview of Course

Regression analysis is the study of the relationship between a response (outcome, dependent) variable and one or more predictor (explanatory, independent) variables (covariates). This analysis allows us to understand which variables influence the response, and to predict the response for other values of these variables.

Repeat observations of the response under identical conditions will usually differ. These differences are referred to as *noise* or *random error* in the model. Regression analysis gives insight into the nature and magnitude of the noise. In this way, we can quantify our uncertainty about the relationship between the covariates and the response, and about our predictions of future responses.

Linear regression (studied in Stat 302 and 350) assumes that the response variables are normally distributed with common variance. However, this is not necessarily the case. For example, the response could be a binary, count, categorical, or multinomial random variable. Stat 402/602 introduces the theory of generalized linear regression, which can be used in these cases.

Linear Regression

Example: Real estate. How does the selling price of a house in Burnaby depend on its age and its lot size?

- Let $Y_1, \ldots, Y_n$ be the selling prices of n independent houses. These are considered to be *random* quantities.
- Let x_{i2} be the age and let x_{i3} be the lot size of the i^{th} house, i.e. there are 2 predictor variables. We treat them as *fixed* quantities.

- Model:
$$Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$
where ϵ_i is the unobserved, random error or *residual* associated with observation Y_i, and β_1, β_2, and β_3 are unknown parameters.
- KEY ASSUMPTION: $\epsilon_1, \dots, \epsilon_n$ are independent, each with distribution $\mathrm{N}(0, \sigma^2)$.
- Consequently, $Y_1, \dots, Y_n$ are independent, and $Y_i \sim \mathrm{N}(\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}, \sigma^2)$.

Matrix Notation

- Let $\mathbf{Y}$ be the n-dimensional column vector $(Y_1, \dots, Y_n)$, and let $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)$.
- Let $\boldsymbol{\beta}$ be the p-dimensional column vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$.
- Let $\mathbf{X}$ be the $n \times p$-dimensional *design matrix*
$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & & \\ \vdots & \vdots & \ddots & \\ x_{n1} & x_{n2} & & x_{np} \end{bmatrix}$$
If $x_{i1} \equiv 1$ then β_1 is the *intercept* of the model.
- We can write the model as
$$\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Moments

In general, if a, b, and c are constants, and U and V are random variables,

$$\mathrm{E}[aU + bV + c] = a\mathrm{E}[U] + b\mathrm{E}[V] + c$$

and

$$\mathrm{Var}[aU + bV + c] = a^2\mathrm{Var}[U] + b^2\mathrm{Var}[V] + 2ab\mathrm{Cov}[U, V].$$

Under the linear regression model, the expectation (mean) of Y_i is

$$\begin{aligned} \mathrm{E}[Y_i] \equiv \mu_i &= \mathrm{E}[\sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i] \\ &= \sum_{j=1}^{p} \beta_j x_{ij} \end{aligned}$$

since $\mathrm{E}[\epsilon_i] = 0$. The variance of Y_i is

$$\begin{aligned} \mathrm{Var}[Y_i] &= \mathrm{Var}[\sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i] \\ &= \sigma^2. \end{aligned}$$

In matrix notation,

$$\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$$

and

$$\mathrm{Var}[\mathbf{Y}] = \sigma^2 \mathbf{I},$$

where $\mathbf{I}$ is the $n \times n$ identity matrix.

Linear Models

Definition: A model is *linear* if $\mathrm{E}[Y_i] \equiv \mu_i$ is linear in the unknown parameters.

Example 1: Linear regression model

For the model $Y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$,

$$\mu_i = \sum_{j=1}^{p} \beta_j x_{ij}.$$

This model is linear since μ_i is linear in the β_j's.

Example 2: Regression model with polynomial functions of the predictor variables

In the lumber industry, the selling price, Y_i, of tree i is related to its radius, x_i.

Question 1: Is the model

$$Y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \epsilon_i$$

linear?

Question 2: Is the model

$$Y_i = \beta_1 + e^{\beta_2 x_i} + \epsilon_i$$

linear?