STAT 331: Applied Linear Regression

Professor Kun Liang
LaTeXer Iris Jiang

Spring 2020

Contents

1. Introduction to Regression

1.1 What is regression

Definition 1.1 — Regression analysis.
Regression analysis is a statistical methodology that models the functional relationship between a response variable y and one or more explanatory variables $x_1, x_2, \ldots, x_p$.

A typical regression model is:

$$y = f(x_1, x_2, \ldots, x_p) + \epsilon$$

- y: dependent variable or response variable
- $x_1, x_2, \ldots, x_p$: covariates, explanatory variables, independent variables, or predictors
- ϵ: random error term

Regression models can be used to:

- Identify important predictors
- Estimate regression coefficients
- Estimate the response for given values of predictors
- Predict of future values of response

In STAT 331, we focus on the simplest form of regression: linear models

$$\begin{aligned} y &= f(x_1, x_2, \ldots, x_p) + \epsilon \\ &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon \end{aligned}$$

where the β's are the regression parameters(coefficients).
Linear in the parameter (not predictor). Linear model is the basic building block of more complicated models

R We refer to the model as linear in the parameters β's ($\frac{\partial f}{\partial \beta_i}$ do not depend on the parameters)

Are the following models linear?

(1) $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$
(2) $f(x) = \beta_0 + \beta_1 e^{\beta_2 x}$
(3) $f(x_1, x_2) = \beta_0 + \beta_1 x_1 x_2$

(1) This is a linear model. The predictor is x, this is not a linear model on the predictor but we define the linear model as to parameter, $\beta_0, \beta_1, \beta_2$ in this case.
(2) This is not a linear model. If taking derivative to β_1, the result involves β_2.
(3) This is a linear model.

1.2 Why linear model?

- Linear model is easy to implement and interpret
- All functions can be approximated locally by a linear function
- The simplest starting model to fit

1.3 Sample vs. population

Definition 1.2 — Sample.
A **sample** is the collection of units (people, animals, cities, whatever you study) that is actually measure or surveyed.

Definition 1.3 — Population.
The **population** is the large group of unites we are interested in, from which the sample was selected.

R We assume the data we have a representative sample (random sample) from a larger population

2. Simple Linear Regression (SLR)

2.1 Population model

$$y = \beta_0 + \beta_1 x + \epsilon$$

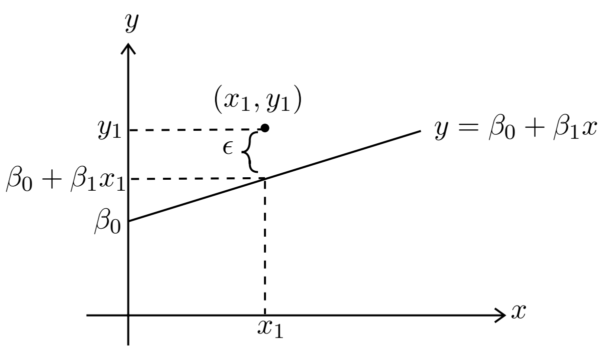
- y: response
- β_0, β_1: regression Coefficients
- x: predictor
- ϵ: random error
- $\beta_0 + \beta_1 x$: systematic (deterministic) part

Observed sample: suppose we have n pairs of observations (x_i, y_i), $i = 1, \ldots, n$. Then

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

- x_i: fixed and known (for this course)
- β: fixed and unknown
- ϵ_i: random and unknown
- y_i: random and known

“known" means we can observe them



- β_0: intercept
- β_1: slope

2.2 Assumptions

(1) $E(\epsilon_i) = 0$
(2) $\epsilon_1, \dots, \epsilon_n$ are statistically independent
(3) Constant variance: $\text{Var}(\epsilon_i) = \sigma^2 \implies \text{Var}(y_i) = \sigma^2$
The randomness of y_i comes from ϵ_i
(4) ϵ_i is normally distributed. $\epsilon_i \sim N(0, \sigma^2)$ and $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

Note 2.1 Assumption 1 to 3 are called Gauss-Markov assumptions.

Assumption 4 is stronger than all 3 assumptions combined.
1 to 3 are useful is you want to point estimate β.
4 is useful for further results.
There is no guarantee that the assumptions are correct. We will talk about model diagnostic and checking these assumptions.

2.3 Least Square Estimation (LSE)

2.3.1 Task

Given the sample observation (x_i, y_i), $i = 1, \dots, n$, estimate (β_0, β_1) as $(\hat{\beta}_0, \hat{\beta}_1)$ such that the values of

$$r_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x = y_i - \hat{y}_i$$

are “small".

- r_i: residual
- y_i: fitted value

We define discrepancy function

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} r_i^2$$

The reason we use square: the least square method provides an elegant solution; ϵ's follow normal distribution the Least Squared Estimation has the equivalence with Maximum Likelihood estimation.

2.3.2 Goal and Derivation

Minimize $S(\beta_0, \beta_1)$, i.e. solve $\begin{cases} \frac{\partial S(\beta_0,\beta_1)}{\partial \beta_0} = 0 \\ \frac{\partial S(\beta_0,\beta_1)}{\partial \beta_1} = 0 \end{cases}$.

$$\begin{cases} \frac{\partial S(\beta_0,\beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-1) = 0 \\ \frac{\partial S(\beta_0,\beta_1)}{\partial \beta_1} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0 \end{cases}$$

This is called Normal Equation

$$\implies \begin{cases} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 & \dots (1) \\ \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)(x_i) = 0 & \dots (2) \end{cases}$$

$$(1) \implies n\hat{\beta}_0 = \sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x_i \text{ or } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \text{ where } \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i \text{ and } \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$$

$$(2) \implies \sum_{i=1}^n x_i y_i - \hat{\beta}_0 \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0$$

we can replace $\hat{\beta}_0$ from previous results.

$$\implies \sum_{i=1}^n x_i y_i - \bar{y}\sum_{i=1}^n x_i - \hat{\beta}_1\bar{x}\sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0$$

$$\implies \hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \text{ where } S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) \text{ and } S_{xx} = \sum (x_i - \bar{x})^2$$

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ is the fractions of the sample covariance between x and y over the sampel variance of x

Thus, the solution is

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \\ \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \end{cases}$$

S_{xx} purely comes from x so is considered fixed, however S_{xy} is the joint quantity between x and y, so it is considered random. Both are functions of y.

2.3.3 The properties of $\hat{\beta}_0$ and $\hat{\beta}_1$

Property 2.1 — The properties of $\hat{\beta}_0$ and $\hat{\beta}_1$.

(1) LSEs are unbiased, i.e. $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$

(2) $\mathrm{Var}\left(\hat{\beta}_0\right) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}})$

$\mathrm{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{s_{xx}}$

(3) $\mathrm{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\frac{\sigma^2\bar{x}}{s_{xx}}$

Proof. Express $\hat{\beta}_1$ as a linear combination of y_i's.

$$\begin{aligned} \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{S_{xx}} \\ & \quad (\text{Since } \sum_{i=1}^n (x_i - \bar{x})\bar{y} = \bar{y}\sum_{i=1}^n (x_i - \bar{x}) = 0) \\ &= \frac{\sum (x_i - \bar{x}) y_i}{S_{xx}} \\ &= \sum \frac{x_i - \bar{x}}{S_{xx}} y_i \\ & \quad (\frac{x_i - \bar{x}}{S_{xx}} \text{ can be seen as some constant } c_i) \end{aligned}$$

The constant c_i's have a few properties.

(i) $\sum c_i = \sum \frac{x_i - \bar{x}}{S_{xx}} = 0$

(ii) $\sum c_i x_i = \frac{1}{S_{xx}} \sum x_i (x_i - \bar{x}) = \frac{1}{S_{xx}} \left(\sum x_i^2 - \bar{x}\sum x_i\right) = 1$

(iii) $\sum c_i^2 = \sum \frac{(x_i-\bar{x})^2}{S_{xx}} = \frac{1}{S_{xx}}$

$$E(\hat{\beta}_1) = E\left(\sum c_i y_i\right) = \sum c_i E(y_i)$$

Recall $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, and $\beta_0 + \beta_1 x_i$ is a fix constant; ϵ_i is the only random part, and we have $E(\epsilon_i) = 0$

$$\begin{aligned}
E(\hat{\beta}_1) &= \sum c_i E(y_i) = \beta_0 \underbrace{\sum c_i}_{=0} + \beta_1 \underbrace{\sum c_i x_i}_{=1} = \beta_1 \\
\mathrm{Var}\left(\hat{\beta}_1\right) &= \mathrm{Var}\left(\sum c_i y_i\right) \text{ (y_i's are independent)} \\
&= \sum c_i^2 \mathrm{Var}(y_i) = \sigma^2 \sum c_i^2 \\
&= \frac{\sigma^2}{S_{xx}} \\
E(\hat{\beta}_0) &= E(\bar{y} - \bar{x}\hat{\beta}_1) = \frac{1}{n}\sum E(y_i) - \bar{x}E(\hat{\beta}_1) \\
&= \frac{1}{n}\sum(\beta_0 + \beta_1 x_i) - \bar{x}\beta_1 \\
&= \beta_0 + \beta_1 \frac{1}{n}\sum x_i - \beta_1 \bar{x} \\
&= \beta_0
\end{aligned}$$

Note that

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 = \sum(\frac{1}{n}y_i - \bar{x}c_i y_i) = \sum(\frac{1}{n} - c_i\bar{x})y_i$$

$\frac{1}{n} - c_i\bar{x}$ is constant and let's call it k_i.

$$\begin{aligned}
\mathrm{Var}\left(\hat{\beta}_0\right) &= \mathrm{Var}\left(\sum k_i y_i\right) = \sum k_i^2 \mathrm{Var}(y_i) \\
&= \sigma^2 \sum k_i^2 = \sigma^2 \sum(\frac{1}{n} - c_i\bar{x})^2 \\
&= \sigma^2 \sum(\frac{1}{n^2} - \frac{2}{n}c_i\bar{x} + c_i^2\bar{x}^2) \\
&\quad \text{(Recall } \sum c_i = 0 \text{ and } \sum c_i^2 = \frac{1}{S_{xx}}\text{)} \\
&= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) \\
\mathrm{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) &= \mathrm{Cov}\left(\sum_{i=1}^n k_i y_i, \sum_{j=1}^n c_j y_j\right) = \sum_{i=1}^n \sum_{j=1}^n k_i c_i \mathrm{Cov}(y_i, y_j) \\
&\quad \left(\text{Note } \mathrm{Cov}(y_i, y_j) = \begin{cases} \sigma^2 & i = j \\ 0 & i \neq j \end{cases}\right) \\
&= \sigma^2 \sum_i k_i c_i = \sigma^s \sum(\frac{1}{n} - c_i\bar{x})c_i \\
&= \sigma^2\left(\frac{1}{n}\sum c_i - \bar{x}\sum c_i^2\right) \\
&= -\frac{\sigma^2\bar{x}}{S_{xx}}
\end{aligned}$$

■

2.3.4 Properties of the residual r_i

Recall

$$r_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1$$

Property 2.2 — Properties of the residual r_i.

Under LS fit

(1) $\sum_{i=1}^{n} r_i = 0$

(It comes from $\frac{\partial S(\beta_0,\beta_1)}{\partial \beta_0} = 0$)

(2) $\sum_{i=1}^{n} r_i x_i = 0$

(It comes from $\frac{\partial S(\beta_0,\beta_1)}{\partial \beta_1} = 0$)

(3) $r_i \hat{y}_i = 0$

(It comes from $\sum r_i(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \hat{\beta}_0 \sum r_i + \hat{\beta}_1 \sum r_i x_i = 0$)

(4) The point $(\bar{x}, \bar{y})$ is always on the fitted regression line

(It comes from $\hat{\beta}_0 + \hat{\beta}_1 bx = \bar{y}$)

The first property shows the residual vector and the 1 vector are perpendicular since the inner product equals to 0.
The second property shows the residual vector and the $\vec{x}$ are perpendicular since the inner product equals to 0
The third property shows the residual vector and the fitted value $\vec{y}$ are perpendicular since the inner product equals to 0