Mathematical Statistics II

Ch6.5 A Simple Regression Problem

Jungsoon Choi

jungsoonchoi@hanyang.ac.kr

Table of Contents

Introduction

Introduction

Introduction

Regression Analysis - Analysis of functional relationship between the response (or dependent) variable y and explanatory (or independent) variables $x_1, x_2, \ldots, x_k$ given the data.

- Simple Linear regression: $E(y_i) = \beta_0 + \beta_1 x_i$
- Multiple Linear regression model:

  $E(y_i) = \beta_0 + \beta_1 x_{1i} + \ldots \beta_k x_{ki}$

Main Purpose of Regression Analysis:

- Screening important independent variable X_j's
- Forecast y for a given set of x

Regression model

1) Linear regression model
2) Logistic regression model
3) Poisson regression model
⋮

Simple Linear Regression Model

Model

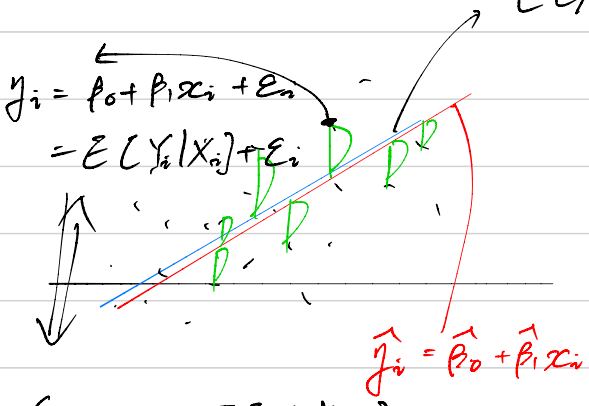
Data: (y_i, x_i), $i = 1, \ldots, n$

Model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

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

where β_0, β_1 are the regression coefficients and ϵ_i is the error (iid).



$\varepsilon_i = y_i - E(X_i|X_i)$

(x_i, y_i) $i = 1, \cdots, n$

Simple Linear Regression Model.

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ $i = 1, \cdots, n$

fixed & unknown

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

$\hat{\beta}_0$, $\hat{\beta}_1$: we don't know true one because of different dataset

→ we don't need distribution assumption

4). LSE

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

$$\sum y_i = \sum (\beta_0 + \beta_1 x_i) = \beta_0 n + \beta_1 \sum x_i$$

$$\sum x_i y_i = \sum x_i (\beta_0 + \beta_1 x_i) = \beta_0 \sum x_i + \beta_1 \sum x_i^2$$

$\sum (x_i - \bar{x})(y_i - \bar{y})$ = $\sum (x_i - \bar{x}) y_i$

$$\hat{\beta}_{1,LSE} = \frac{\sum (x_i - \bar{x}) y_i}{\sum x_i^2 - \bar{x} \sum x_i} = \frac{\sum (x_i - \bar{x}) y_i}{\sum x_i (x_i - \bar{x})} = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}$$

$$= \frac{S_{XY}}{S_{XX}}$$

$\frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x}) x_i}$

$\bar{y} = \beta_0 + \beta_1 \bar{x}$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i$

$= \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$

Estimation of the regression coefficients

Estimation of the regression coefficients

Least Square Estimation (LSE)

To minimize the distance between y_i and $E(y_i)$, we use the method of least squares,

$$\min_{\beta_0,\beta_1} Q(\beta_0,\beta_1) = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} \epsilon_i^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$\hat{\beta}_0, \hat{\beta}_1$ satisfying

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

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

Estimation of the regression coefficients

- Normal Equation

$$\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2$$

- LSE: $\hat{\beta}_0, \hat{\beta}_1$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$

Estimation of the regression coefficients - MLE

The likelihood function is given by

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right\} \right]$$

$$= \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \exp\left\{ -\frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right\}$$

$$\log L(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}$$

$$(\hat{\beta}_0, \hat{\beta}_1) = \max\left[\log L(\beta_0, \beta_1, \sigma^2)\right]$$

$$\Leftrightarrow (\hat{\beta}_0, \hat{\beta}_1) = \min\left[\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \right]$$

MLE of σ^2:

$$\frac{d \log L(\beta_0, \beta_1, \sigma^2)}{d\sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}{2(\sigma^2)^2} = 0$$

$$\hat{\sigma^2} = \frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}$$

$$= \frac{\sum_{i=1}^n e_i^2}{n}$$

where $e_i = y_i - \hat{y}_i$ is residual.

$y_i | x_i \overset{iid}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$ $i = 1, \dots, n$

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f(y_i)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{ - \frac{\sum (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right\}$$

$$(\hat{\beta}_0, \hat{\beta}_1) = \max \left[\log L(\beta_0, \beta_1, \sigma^2) \right]$$

$$\Leftrightarrow (\hat{\beta}_0, \hat{\beta}_1) = \min \left[\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right]$$

LSE of β_0, β_1 = MLE of β_0, β_1

$= \sum (x_i - \bar{x})(y_i - \bar{y})$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$: LSE, MLE, ME of β_1 $\sim N(\beta_1, \frac{\sigma^2}{\sum (x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}})$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$: LSE, MLE of β_0 $\sim N$

$\hat{\sigma}^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n}$: MLE of σ^2

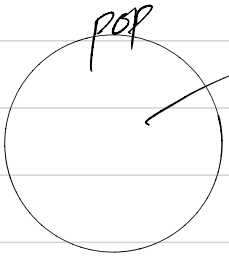
$y_i \sim N$

$\bar{y} = \sum \frac{y_i}{n} \sim N$

$N(\beta_0, \sigma^2 (\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}))$

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

$\Rightarrow \bar{y} = \frac{\sum y_i}{n} \sim N(\beta_0 + \beta_1 \bar{x}, \frac{\sigma^2}{n})$



$X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

what is dist^n of $\bar{X}$?

at this time

" of $\hat{\beta}_0, \hat{\beta}_1$

$$E[\hat{\beta}_1] = E\left[\frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}\right]$$

$$= \frac{\sum (x_i - \bar{x}) E[y_i]}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x}) \cdot (\beta_0 + \beta_1 x_i)}{\sum (x_i - \bar{x})^2}$$

$$= \frac{\sum (x_i - \bar{x}) \bar{x}}{\sum (x_i - \bar{x})^2} \beta_1 + \beta_1 \qquad \sum (x_i - \bar{x})(x_i - \bar{x})$$

$$E[\hat{\beta}_0] = E[\bar{y} - \hat{\beta}_1 \bar{x}] = E[\bar{y}] - E[\hat{\beta}_1 \bar{x}]$$

$$= E[\bar{y}] - \bar{x} E[\hat{\beta}_1] = \beta_0 + \beta_1 \bar{x} - \bar{x} \cdot \beta_1 = \beta_0$$

$$Var[\hat{\beta}_0] = Var(\bar{y}) + Var(\hat{\beta}_1 \bar{x}) - 2Cov(\bar{y}, \hat{\beta}_1 \bar{x})$$

$$= \frac{\sigma^2}{n} + \bar{x}^2 \times \frac{\sigma^2}{S_{xx}} \qquad Cov\left[\frac{\sum_{i=1}^{n} y_i}{n}, \frac{(x_i - \bar{x}) \bar{x} y_i}{\sum (x_i - \bar{x})^2 S_{xx}}\right]$$

$$\frac{\frac{1}{n} \sum (x_i - \bar{x}) \bar{x} \, Var(y_i)}{S_{xx}} = \frac{\bar{x} \sigma^2 \sum (x_i - \bar{x})}{n S_{xx}} = 0$$

If Y_i is indep.

$$Cov\left(\sum_{i=1}^{n} a_i Y_i, \sum_{i=1}^{n} b_i Y_i\right) = \sum_{i=1}^{n} a_i b_i Var(Y_i)$$

$$= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

Distribution of $\hat{\beta}_1$

Distribution of $\hat{\beta}_1$

Since $y_i = \beta_0 + \beta_1 x_i + e_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$,

$$E(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \bar{x}) E(y_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$= \beta_0 \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} + \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x}) x_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1$$

$$Var(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 Var(y_i)}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right)$$

where $S_{XX} = \sum_{i=1}^n (x_i - \bar{x})^2$.

Distribution of $\hat{\beta}_0$

Distribution of $\hat{\beta}_0$

$$E(\hat{\beta}_0) = E(\bar{y}) - \bar{x}E(\hat{\beta}_1) = (\beta_0 + \beta_1\bar{x}) - \beta_1\bar{x} = \beta_0$$

$$\begin{aligned} Var(\hat{\beta}_0) &= Var(\bar{y}) + \bar{x}^2 Var(\hat{\beta}_1) - 2Cov(\bar{y}, \hat{\beta}_1\bar{x}) \\ &= \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{S_{XX}} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right) \end{aligned}$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)\right)$$

Estimation of the variance

Estimation of the variance

Let $e_i = y_i - \hat{y}_i$ be the residual. The unbiased estimator of σ^2 is

$$\text{MSE} = \hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

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

Confidence Interval for the regression coefficients

C.I for β_0 and β_1

- C.I for β_1

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1) \rightarrow T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \sim t(n-2)$$

$$\Rightarrow \hat{\beta}_1 \pm t_{\alpha/2}(n-2)\sqrt{\hat{\sigma}^2/S_{xx}}$$

- C.I for β_0

$$T = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}\sqrt{(1/n + \overline{x}^2/S_{xx})}} \sim t(n-2)$$

$$\Rightarrow \hat{\beta}_0 \pm t_{\alpha/2}(n-2) \times \hat{\sigma}\sqrt{(1/n + \overline{x}^2/S_{xx})}$$

Example

Example 6.5-1

We have the data (x: pre-test score and y: final score).

$$x: 70,\ 74,\ 72,\ 68,\ 58,\ 54,\ 82,\ 64,\ 80,\ 61$$

$$y: 77,\ 94,\ 88,\ 80,\ 71,\ 76,\ 88,\ 80,\ 90,\ 69$$

- Find the estimated regression model.
- Compute the unbiased estimator of σ^2.
- Find the 95% C.I for β_0 and β_1.

x	y	x^2	xy	y^2	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$
70	77	4,900	5,390	5,929	82.561566	–5.561566	30.931016
74	94	5,476	6,956	8,836	85.529956	8.470044	71.741645
72	88	5,184	6,336	7,744	84.045761	3.954239	15.636006
68	80	4,624	5,440	6,400	81.077371	–1.077371	1.160728
58	71	3,364	4,118	5,041	73.656395	–2.656395	7.056434
54	76	2,916	4,104	5,776	70.688004	5.311996	28.217302
82	88	6,724	7,216	7,744	91.466737	–3.466737	12.018265
64	80	4,096	5,120	6,400	78.108980	1.891020	3.575957
80	90	6,400	7,200	8,100	89.982542	0.017458	0.000305
61	69	3,721	4,209	4,761	75.882687	–6.882687	47.371380
683	813	47,405	56,089	66,731		0.000001	217.709038