Lecture 2 Stat 331

Review : Linear model $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$

Y: response variable

$(x_1, \cdots, x_p)$: explanatory variables (fixed constants)

ε : error term ; $\varepsilon \sim N(0, \sigma^2)$

Simple Linear Regression: Study relationship between a response variable and a single explanatory variable.

What does our data look like?

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

Salary dataset :

i	x_i	y_i
1	1.1	39343
⋮	⋮	⋮
n	10.5	121872

Explorative Analysis

① Scatterplot : helps us visualize what our data looks like

② Can consider correlation between two variables :

(sample) correlation coefficient

$$\underset{\text{(sample) correlation coefficient}}{r} = \frac{\overbrace{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}^{S_{xy}}}{\sqrt{\underbrace{\sum_{i=1}^{n}(x_i-\bar{x})^2}_{S_{xx}}\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{S_{yy}}}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \left(= \frac{\overset{\text{sample covariance of } x, y}{s_{xy}}}{\sqrt{\underbrace{s_{xx}}_{\text{sample var of } x}\ \underbrace{s_{yy}}_{\text{sample var of } y.}}} \right)$$

What does r tell us?

- The strength and direction of the linear relationship
- $r \in [-1, 1]$ (can be shown via Cauchy-Schwartz)
- $0 < r \leq 1$: positive linear relationship ($r \approx 1$: strong, positive)
- $-1 \leq r < 0$: negative " " ($r \approx -1$: strong, negative)
- $r = 0$: no linear relationship
- r is not sufficient to make predictions of Y given x. Need linear regression!

Suppose we observe $\{(x_i, y_i) : i = 1, \ldots, n\}$, consider a simple linear model for each observation :

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

Y_i: response variable of observation i

x_i: explanatory variable of observation i

β_0: intercept parameter ; β_1: slope parameter

ε_i: error for observation i

$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ (iid: independent and identically distributed).

$\Downarrow$

$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$: independent but not identically distributed.

$E(Y_i) = \beta_0 + \beta_1 x_i$; $Var(Y_i) = \sigma^2$

Interpret β_0, β_1:

β_0: average response when $x = 0$

β_1: average change in response for every unit increase in x.

How to estimate β_0 and β_1? Least-squares estimation

Method aims to estimate β_0 and β_1 by minimizing:

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

$$\arg\min_{\beta_0, \beta_1} S(\beta_0, \beta_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

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

Solve for a system of equations

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

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

$n\bar{y} = n\beta_0 + n\beta_1 \bar{x} \Rightarrow \beta_0 = \bar{y} - \beta_1 \bar{x}$

(2) $\sum_{i=1}^{n} (y_i - \bar{y} + \beta_1 \bar{x} - \beta_1 x_i) x_i = 0$

$$\sum_{i=1}^{n}(y_i - \bar{y}) x_i - \beta_1 \sum_{i=1}^{n}(x_i - \bar{x}) x_i = 0 \Rightarrow \beta_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}) x_i}{\sum_{i=1}^{n}(x_i - \bar{x}) x_i}$$

$$\text{(from below)} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

Aside: Note that $\sum_{i=1}^{n}(y_i-\bar{y})x_i = \sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})$

Pf: $RHS = \sum_{i=1}^{n}(y_i-\bar{y})x_i - \sum_{i=1}^{n}(y_i-\bar{y})\bar{x}$

$= \sum_{i=1}^{n}(y_i-\bar{y})x_i - \cancel{(\sum_{i=1}^{n} y_i)\bar{x}} + \cancel{n\bar{y}\bar{x}}$

$= LHS$ ∎

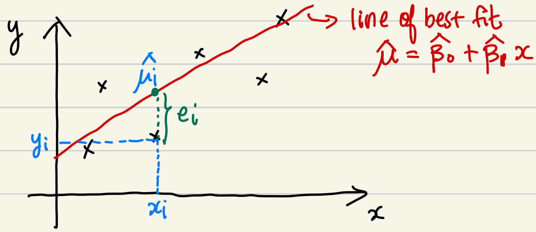
Similarly, show that $\sum_{i=1}^{n}(x_i-\bar{x})x_i = \sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})$

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

$\hat{\beta}_1 = S_{xy}/S_{xx}$

Note: Parameter estimates are denoted with hat-symbol $(\hat{\beta}_0, \hat{\beta}_1)$

- Call $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ fitted value corresponding to x_i in our dataset.
- Call $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ predicted value of response for a new observation x_0.



- Line of best fit: line "closest" to our data points.
- e_i: residual of the i^{th} observation

  $e_i = y_i - \hat{\mu}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

Example: Salary data

y_i: salary for i^{th} subject

x_i = work experience (in yrs) for i^{th} subject.

Q: Given $\bar{x} = 5.3133$, $\bar{y} = 76003$, $S_{xy} = 2207083$, $S_{xx} = 233.5547$.

Estimate β_0 and β_1.

A: $\hat{\beta}_1 = S_{xy}/S_{xx} = 2207083/233.5547 = 9449.96$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 76003 - 9449.96(5.3133) = 25792.20$

$\hat{\beta}_0$: average salary for an individual w/ no work experience is \$25 792.20

$\hat{\beta}_1$: for each additional year of work experience, on average salary increases by \$9450.

By the least-square estimation procedure:

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

This implies :

(1) $\sum_{i=1}^{n} e_i = 0 \Rightarrow \bar{e} = 0$

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

(3) $\sum_{i=1}^{n} e_i \hat{\mu}_i = 0$

Pf : $\sum_{i=1}^{n} e_i (\hat{\beta}_0 + \hat{\beta}_1 x_i) = \left(\cancelto{0}{\sum_{i=1}^{n} e_i}\right)\hat{\beta}_0 + \left(\cancelto{0}{\sum_{i=1}^{n} e_i x_i}\right)\hat{\beta}_1 = 0 \quad \blacksquare$

Variance σ^2 and its estimation

σ^2 is variability in random errors $\Rightarrow$ variability in responses

$(Var(\varepsilon_i) = Var(Y_i) = \sigma^2)$

$$\varepsilon_i = Y_i - \beta_0 - \beta_1 x_i$$
$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Naturally, we will use e_i to estimate σ^2. Specifically,

$$\hat{\sigma}^2 = \sum_{i=1}^{n} (e_i - \bar{e})^2 / n-2 .$$ (looks like a sample variance for e_i)

- $df = n - (\#$ of parameters in LS estimation , $\beta_0, \beta_1)$
- $\hat{\sigma}^2$ is an unbiased estimator of σ^2 (shown later)
- Lost two df from estimation of β_0 and β_1.

Let $\sum_{i=1}^{n} e_i^2 = SS(Res)$, then

$$\hat{\sigma}^2 = SS(Res)/n-2 .$$

Pf : $\sum_{i=1}^{n} (e_i - \bar{e})^2 = \sum_{i=1}^{n} (e_i^2 - 2\bar{e}e_i + \bar{e}^2)$

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