Lecture 2 (STAT 153/248)

Simple Linear Regression

$y \longleftrightarrow x \rightarrow$ Simple Regression

$y \longleftrightarrow \underbrace{(x_1, \ldots, x_p)}_{\text{Multiple Regression}}$

Eg: y : Height of an Adult

x_1 : Height of parent of adult

x_2 : Gender

Linear Regression

Model: $\boxed{y = \beta_0 + \beta_1 x + \text{error}}$

Response Variable

Covariate

Parameters (Coefficients)

β_0 : value of y when $x = 0$

β_1 : change in y when x changes by 1 unit

Data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n}$$

Goal: To estimate β_0 & β_1 from the data $(x_1, y_1) \ldots (x_n, y_n)$.

Time Series: $\quad y_1 \ldots y_n$

$t_1 \qquad t_n$

Two ways of using Regression for Time Series:

(a) Regression using time (Time as Covariate) : $\boxed{x_i = i}$

y_i : observation for y at time i

(b) Lagged values of y as Covariate } Auto Regression

$x_i = y_{i-1}$

Eg: CPI (Consumer Price Index)

used to measure inflation.

y_t : CPI_t

y_t over t

or y_t over y_{t-1}

$$\boxed{CPI_t = \beta_0 + \beta_1 t + \varepsilon_t}$$ ✓

β_0, β_1

β_0 : CPI_0

β_1 : numerical increase in CPI from one month to the next

$$\boxed{y_t = \log CPI_t}$$

$$\boxed{y_t = \beta_0 + \beta_1 t + \varepsilon_t}$$

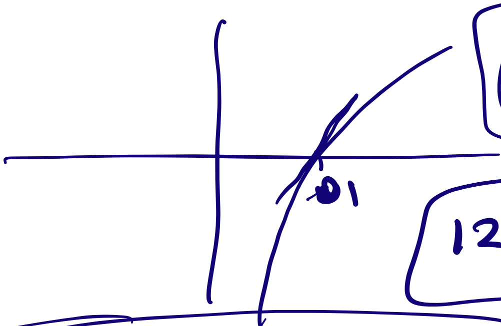
β_1 : change in log CPI month-to-month

$$\beta_1 = \log CPI_t - \log CPI_{t-1}$$

$$= \log \boxed{\frac{CPI_t}{CPI_{t-1}}} = \frac{CPI_t}{CPI_{t-1}} - 1$$

$$= \frac{CPI_t - CPI_{t-1}}{CPI_{t-1}}$$

$\boxed{\log x \approx x - 1 \text{ when } x \text{ is close to } 1}$



$\boxed{100\beta_1}$ = monthly inflation percent

$\boxed{12 \times 100 \times \beta_1} \rightarrow$ annual inflation rate

$\boxed{\text{statsmodels in Python}}$

OLS

Alternate ways of measuring Inflation

① average of $\frac{CPI_t - CPI_{t-1}}{CPI_{t-1}} \times 100$ over all t.

② $\dfrac{CPI_T - CPI_1}{\#\ \text{intervening months}}$

3 → geometric mean of

1 → 3.6 %

3.58 to 3.62

AutoRegression

$$\log CPI_t = \beta_0 + \beta_1 \log CPI_{t-1} + \text{error}$$

β_0 → 0.0039

β_1 = 0.9998

$$\log CPI_t = 0.0039 + \log CPI_{t-1}$$

$$\log \frac{CPI_t}{CPI_{t-1}} = 0.0039 + \text{error}$$

$$100 \times \frac{CPI_t - CPI_{t-1}}{CPI_{t-1}} = \boxed{0.39} + \text{error}$$

basically computing an average of monthly inflation rates

$\boxed{100 \times \beta_0}$

Linear Regression Math

How are the coefficient estimates $\hat{\beta}_0, \hat{\beta}_1$? (β_0, β_1)

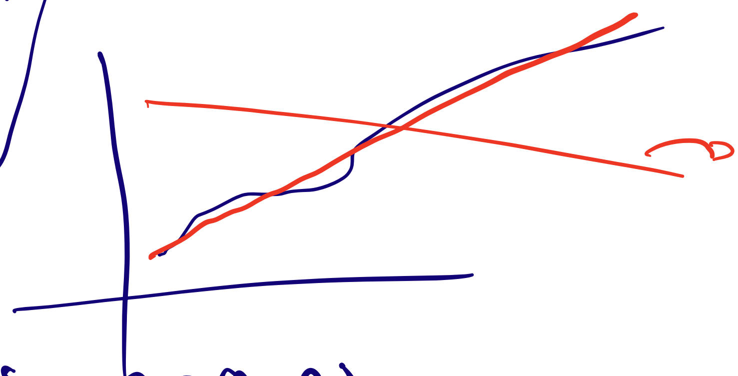
LEAST SQUARES (sm.OLS)

$(x_1, y_1), \ldots, (x_n, y_n)$

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

Find β_0 & β_1 which minimize $S(\beta_0, \beta_1)$

$$\boxed{\begin{array}{l} S(0.0039, 1) \\ S(1, 5) \\ S(-1, 3) \end{array}}$$



How to solve this minimization?

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

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

Ans: $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad \bar{x} = \frac{x_1 + \cdots + x_n}{n}, \quad \bar{y} = \frac{y_1 + \cdots + y_n}{n}$$

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

Maximum Likelihood Estimation

Assumption: $\varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} \boxed{N(0, \sigma^2)}$

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

σ : scale of the deviations from the model equation parameters: $\beta_0, \beta_1, \boxed{\sigma}$

MLE of β_0, β_1 under this Regression Model with Gaussian Errors = Least Squares Estimates of β_0, β_1

Why: Just Calculate the MLE.

Likelihood function = joint density of data with fixed values of the parameters

$$y_i = \boxed{\beta_0 + \beta_1 x_i} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

$$y_i \overset{\text{independent}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$$

Assume $x_i = i$

$y_1 \dots y_n$

Likelihood: $\prod_{i=1}^{n}$ [normal density with mean $\beta_0 + \beta_1 x_i$ & variance σ^2]

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

θ_1, θ_2

MLE: maximize over β_0, β_1, σ

Check MLEs of β_0, β_1 = Least squares estimates of β_0, β_1