Lecture 1 Stat 331

Regression Model: Quantify / Infer relationship between a response variable and a set of explanatory variables.

Response Variable: (Y)
- Variable of primary interest
- Understand how Y depends on other variables
- Dependent Variable, outcome Variable.

Explanatory Variables: $(X_1, X_2, \ldots, X_p)$
- Variables that are used to explain the response variable Y.
- Regression model process: determines which of $(X_1, \ldots, X_p)$ are associated w/ Y.
- Independent Variables, predictors, covariates, features.

Applications of Regression Modelling:

Area	Response Variable	Explanatory Variables
Public Health	Cognitive Function	Age, Sex, Education, Occupation, Lifestyle factors (smoking, drinking...)
	Lung Function (FEV)	Lifestyle factors, Age, Sex, Height
Economics	Crime Rate	Unemployment rate, average income, education region

In regression modeling, we try to explain Y in term of $(X_1, \ldots, X_p)$ through a function $f(\cdot)$, s.t. $Y = f(X_1, \ldots, X_p)$
- In linear regression, $f(\cdot)$ is a linear function: Y is a linear combination of $(X_1, \ldots, X_p)$
  - Y is linear in its parameters.

- Y is continuous variable
- $(X_1, \ldots, X_p)$: continuous, discrete, binary.

What does linear model look like?

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

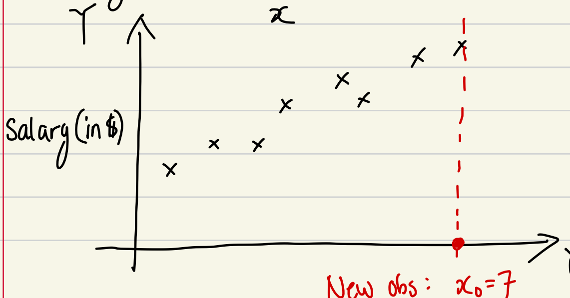
- Y: response variable
- $(x_1, \ldots, x_p)$: explanatory variables (fixed constants)
- $(\beta_0, \ldots, \beta_p)$: parameters
  - β_0 : intercept (average of response when $x_1 = x_2 = \cdots = x_p = 0$)
  - β_j : parametes that quantify the association between x_j and Y for $j = 1, \ldots, p$.
- ε : error term
  $\varepsilon \sim N(0, \sigma^2)$

$$\begin{aligned}
E(Y) &= E(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon) \\
&= E(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p) + \underbrace{E(\varepsilon)}_{=0} \\
&= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p
\end{aligned}$$

$$\begin{aligned}
Var(Y) &= Var(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon) \\
&= Var(\varepsilon) \\
&= \sigma^2
\end{aligned}$$

$$Y \sim N(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \ \sigma^2)$$

$\underbrace{\text{Salary}}_{Y}$ vs. $\underbrace{\text{Years of work experience}}_{x}$



1. Try to explain the relationship between salary and work experience.

2. Predict salary for a new observation.

Topics to Cover:

- Parameter estimation and inference
- Model interpretation
- Prediction
- Variable / Model selection
- Multicollinearity, outliers, influential points
- Detect violations against model assumptions
- Nonlinear regression.

Review of Statistical Concepts:

① $X \sim N(\mu, \sigma^2)$, $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-(x-\mu)^2}{2\sigma^2} \right\}$

② Expected Value : For a R.V X: $\xrightarrow{}$ pmf

$$E(X) = \begin{cases} \sum_x x \, P(X=x) \text{ , when } X \text{ is discrete} \\ \int_x x \underbrace{f(x)}_{\text{pdf.}} dx \text{ , when } X \text{ is cts.} \end{cases}$$

Property of Expected Value : Linearity

a) $E(X+Y) = E(X) + E(Y)$

b) $E(cX) = cE(X)$, c is constant.

Sample mean of X:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

② Variance : $Var(X) = E\left[(X - E(X))^2\right]$

$$= E(X^2) - [E(X)]^2$$

Properties : For R.Vs (X, Y):

a) $Var(X+Y) = Var(X) + Var(Y) + \underbrace{2Cov(X,Y)}_{=0 \text{ if } X,Y \text{ independent}}$

b) $Var(cX+a) = c^2 Var(X)$, a, c constants

Sample variance :

$$s_{xx} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

③ Covariance : $Cov(X,Y) = E\left[(X - E(X))(Y - E(Y))\right]$

$$= E(XY) - E(X)E(Y).$$

Properties :

a) $Cov(X,X) = Var(X)$

b) $Cov(aX+c, bY+d) = ab\,Cov(X,Y)$; a,b,c,d constants

Sample covariance :

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$