Advanced Statistics DS2003 (BDS-4A) Lecture 21

Instructor: Dr. Syed Mohammad Irteza

Assistant Professor, Department of Computer Science, FAST

10 May, 2022

Previous Lecture

- Revision of Linear Regression
  - Sum of squares for x and y
  - Sum of products for x, y
- Least squares regression model
  - Residuals sum to zero, and the line always passes through $(\bar{x}, \bar{y})$

Today

- More in detail workings of multiple linear regression
  - Calculating the slopes for each independent variable (X_1, X_2, …, X_n)
  - Calculating the intercept (b_0)
- Preference for no multicollinearity
  - We don’t want correlation between different explanatory (independent) variables
  - We want to witness a correlation between explanatory variables with the response (dependent) variable

Recap of Linear Regression: 10/5/2022 (01)

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2 \Big\} \text{ sum of squared residuals.}$$

This is called the method of least squares:

SS_{xx} ("sum of squares" for X) $= \sum (X_i - \bar{X})^2$ | S_x^2 (sample variance) $= \frac{SS_{xx}}{n-1}$

SS_{YY} ("sum of squares" for Y) $= \sum (Y_i - \bar{Y})^2$ | S_Y^2 (sample variance) $= \frac{SS_{YY}}{n-1}$

SP_{XY} ("sum of products") $= \sum (X_i - \bar{X})(Y_i - \bar{Y})$ | sample covariance $\mathrm{Cov}(x,y) = \frac{SP_{XY}}{n-1}$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad ; \quad \hat{\beta}_1 = \frac{SP_{XY}}{SS_{XX}} = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}$$

For least squares regression:

→ The residuals sum to Ø ($\sum e_i = 0$)

→ The line passes thru $(\bar{x}, \bar{y})$.

Multiple Regression ⇒ a statistical measure that attempts to determine the strength of the relationship between 1 dependent variable (or response var, often denoted by Y), & a series of other changing variables (also known as explanatory or independent variables).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \boxed{u_i}$$

price, income, rand. variable

Y = quantity demanded of a commodity

X_1 = price of that commodity

X_2 = consumer income

$\Big\}$ $Y = f(X_1, X_2)\ldots$

→ we assume a linear relationship btw Y, X_1, X_2.

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\sum x_1 y = \sum X_1 Y - \frac{(\sum X_1)(\sum Y)}{N}$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\sum x_2 y = \sum X_2 Y - \frac{(\sum X_2)(\sum Y)}{N}$$

$$\sum x_1 x_2 = \sum X_1 X_2 - \frac{(\sum X_1)(\sum X_2)}{N}$$

Example: Multiple Linear Regression

10/

	Y (Quantity Demanded)	X_1 (Price)	X_2 (Income)	$X_1 \cdot Y$	$X_2 \cdot Y$	$X_1 \cdot X_2$	x_1^2	x_2^2
	100	5	1000	500	100,000	5,000	1	40,000
	75	7	600	525	45,000	4,200	1	40,000
	80	6	1200	480	96,000	7,200	0	160,000
	70	6	500	420	35,000	3,000	0	90,000
	50	8	300	400	15,000	2,400	4	250,000
	65	7	400	455	26,000	2,800	1	160,000
	90	5	1300	450	117,000	6,500	1	250,000
	100	4	1100	400	110,000	4,400	4	90,000
	110	3	1300	330	143,000	3,900	9	250,000
	60	9	300	540	18,000	2,700	9	250,000
um	800	60	8000	4500	705,000	42,100	30	1,580,000
ean	80	6	800	450	70,500	4,210	3	158,000

$b_0 = 111.8$

$b_1 = -7.18$

$b_2 = 0.014$

$$Y = b_0 + b_1 X_1 + b_2 X_2 = 111.8 - \underbrace{7.19 X_1}_{\text{price}} + \underbrace{0.014 X_2}_{\text{income.}}$$

It is important to understand how we can interpret the values for the intercept and the various slopes

Useful Links

- **Statistics 101: Multiple Linear Regression, The Very Basics**
  - https://www.youtube.com/watch?v=dQNpSa-bq4M

- **What is Multiple Regression | numerical explanation AND interpretation of Multiple regression (English and Urdu)**
  - https://www.youtube.com/watch?v=iCENCX60JpY

- Example:
  - http://faculty.cas.usf.edu/mbrannick/regression/Part3/Reg2.html

Sources

- openintro.org/os (Chapter 9, Section 9.1)