GITAM Hyderabad

MATH2561: Unit 5 Notes and Formula list

Dr. Harsha K V

Correlation

In statistics, correlation usually refers to the degree to which a pair of variables are linearly related. Consider a data set $(x_1, y_1), \cdots, (x_n, y_n)$. Define

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2 \\ S_{yy} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2 \\ S_{xy} &= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n(\bar{x}\bar{y}). \end{aligned}$$

Karl Pearson's correlation coefficient is defined as

$$\begin{aligned} r &= \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} && (1) \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} && (2) \\ &= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\left(\sqrt{\sum x_i^2 - n(\bar{x})^2}\right)\left(\sqrt{\sum y_i^2 - n(\bar{y})^2}\right)}. && (3) \end{aligned}$$

- Pearson's correlation coefficient r always lie between $-1 \leq r \leq 1$.
- $r = 1$, we say that data is perfect positively correlated
- $r = -1$, data is perfect negatively correlated
- $r = 0$, data is uncorrelated.

Another type of correlation coefficient is the **Spearman's rank correlation coefficient** defined as follows:

- First assign rank (our convention: highest value gets rank 1) to each of the values of x_i. Then repeat the same for y_i.
- Assume that each x_i and y_i get unique rank (no ranks are repeated), then Spearman's rank correlation coefficient is defined as

  $$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

  d_i is the difference in ranks of x_i and y_i

- If any rank is repeated for any of the values of x or y, then add a factor of $\frac{1}{12}(m^3-m)$ to the above formula for each repeated observation, where m is the number of times that observation is repeated.
$$\rho = 1 - \frac{6}{n(n^2-1)}\left[\sum d_i^2 + \frac{1}{12}(m^3-m)\right].$$
- Spearman's rank correlation coefficient ρ always lie between $-1 \leq \rho \leq 1$.

For example, consider the following data. Let us see the procedure to rank a data. If we rank x values, 50 gets rank 1, 40 gets rank 2, and 20 gets rank 3. Here, 10 is repeating twice, and which corresponds to rank 4 and rank 5. Then assign rank $\frac{4+5}{2} = 4.5$ to the value 10. Then next value is 9 which gets rank 6 (we will not assign rank 4 or 5 to any values). Next 5 is repeating 3 times, which corresponds to rank 7, 8, 9. Then assign $\frac{7+8+9}{3} = 8$ to the value 5.

x	5	10	40	5	50	20	10	9	5
Rank	8	4.5	2	8	1	3	4.5	6	8

Here 10 is repeating twice. Hence we take $m_1 = 2$, and add a factor $\frac{1}{12}(m_1^3 - m_1) = \frac{1}{12}(2^3-2) = 0.5$ in the formula. Now 5 is repeating 3 times. Hence take $m_2 = 3$, add $\frac{1}{12}(m_2^3 - m_2) = \frac{1}{12}(3^3-3) = 2$ also in the formula.

Pearson's correlation assesses linear relationships, while Spearman's rank correlation coefficient assesses how well the relationship between two variables can be described using a monotonic function (may or may not be linear).

Regression and Least Square method

Regression means estimating the relationship between two variables. This helps us to predict the value of the dependent variable as a function of the independent variables. Consider a data set $(x_1, y_1), \cdots, (x_n, y_n)$. Suppose, we would like to express this data as $y = g(x)$ for some function g. Here we take x_i as independent variable and y_i are the dependent variable.

For example, we want to fit the data using a linear function of the form $y = a + bx$. Here, $g(x) = a + bx$, where a and b are unknown constants. This is known as **linear curve fitting**. Our goal is to find the value of the unknowns a and b which "best" describe the given data (x_i, y_i). But what is best fit means here? The least square method provides some criteria for determining the best fit. The least square method gives the optimal parameters a and b by minimizing the sum of the squared error (residual) as follows:

- Here observed value is y_i and the predicted value by our model is $g(x_i) = a + bx_i$
- Define residual as $r_i = y_i - g(x_i) = y_i - (a + bx_i)$
- Sum of the squared residual is
$$L = \sum_{i=1}^{n}(y_i - (a + bx_i))^2$$
- To get the best fit line, we find the optimal a and b which minimize L as given by
$$\min_{a,b} \sum_{i=1}^{n}(y_i - (a + bx_i))^2$$

- Note that our goal is to find a and b in terms of the given data (x_i, y_i) so that the error L is minimum!

- For minimizing L, we find the first partial derivatives of L with respect to a and b, then equate to zero, which gives

$$\frac{\partial L}{\partial a} = 0 \;\Rightarrow\; \sum y_i = an + b\sum x_i \tag{4}$$
$$\frac{\partial L}{\partial b} = 0 \;\Rightarrow\; \sum x_i y_i = a\sum x_i + b\sum x_i^2. \tag{5}$$

These equations are known as **Normal equations**.

By solving the above two equations, we obtain the constants a and b as follows:

$$\begin{aligned} b &= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n(\bar{x})^2} \\ a &= \bar{y} - b\bar{x}, \quad \bar{x} = \frac{\sum x_i}{n}, \bar{y} = \frac{\sum y_i}{n}. \end{aligned}$$

Hence for a given data (x_i, y_i), we obtain a linear function $y = a + bx$ which best fits the data. In fact this is the Regression line of y on x. We denote the constant b as b_{yx}, and is known as the **regression coefficient of y on x**. The regression line of y on x is of the form

$$y = a + b_{yx}x, \quad \text{where} \tag{6}$$
$$b_{yx} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n(\bar{x})^2} = \frac{S_{xy}}{S_{xx}} \tag{7}$$
$$a = \bar{y} - b_{yx}\bar{x}. \tag{8}$$

Similarly, the least square method can be applied to find function of the form $x = c + dy$, where y is the independent variable and x is the dependent variable. This would lead us to obtain the **regression line of x on y** as,

$$x = c + b_{xy}y, \quad \text{where} \tag{9}$$
$$b_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum y_i^2 - n(\bar{y})^2} = \frac{S_{xy}}{S_{yy}} \tag{10}$$
$$c = \bar{x} - b_{xy}\bar{y}. \tag{11}$$

Here the constant b_{xy} is known as the **regression coefficient of x on y**.
Note that Equations (6) and (9) can be written in the form

$$\begin{aligned} y - \bar{y} &= b_{yx}(x - \bar{x}) \\ x - \bar{x} &= b_{xy}(y - \bar{y}). \end{aligned}$$

These two lines of regression intersect at the point $(\bar{x}, \bar{y})$, and both lines have the slope b_{yx} and b_{xy} respectively. Using (1), (7), and (10), the regression coefficients can be written in terms of Pearson's correlation coefficient r as

$$\begin{aligned} b_{yx} &= \frac{S_{xy}}{S_{xx}} = \frac{r\sqrt{S_{yy}}}{\sqrt{S_{xx}}} \\ b_{xy} &= \frac{S_{xy}}{S_{yy}} = \frac{r\sqrt{S_{xx}}}{\sqrt{S_{yy}}}. \end{aligned}$$

This gives

$$r^2 = b_{yx} \times b_{xy} \quad \Rightarrow \quad r = \pm\sqrt{b_{yx} \times b_{xy}}.$$

Note that b_{xy} or b_{yx} will have the same sign as r.

Curve fitting using Least square method

The procedures that we did for linear curve fitting, can be done for other functions as well: For example,

- Quadratic function $g(x) = a + bx + cx^2$, where a, b, c are unknowns
- Exponential functions, $g(x) = ae^{bx}$ or $g(x) = ab^x$, where a, b are unknowns

Then using least square method, we can find the unknown values in each case by minimizing the function $L = \sum_{i=1}^{n}(y_i - g(x_i))^2$. In each case, we obtain the normal equations as follows:

- **Quadratic curve fitting of the form** $y = a + bx + cx^2$: Solve a, b, c using

$$\begin{aligned} \sum y_i &= an + b\sum x_i + c\sum x_i^2 \\ \sum x_i y_i &= a\sum x_i + b\sum x_i^2 + c\sum x_i^3 \\ \sum x_i^2 y_i &= a\sum x_i 2 + b\sum x_i^3 + c\sum x_i^4. \end{aligned}$$

- **Exponential curve fitting**
  - Form $y = ae^{bx}$: This can be converted to the linear curve fitting form as (here ln **is the natural logarithm with base** e)

$$y = ae^{bx} \quad \Rightarrow \quad \ln y = \ln a + bx$$

    Now define new variables as $Y = \ln y, A = \ln a, B = b, X = x$. We then have the linear form $Y = A + BX$. Then A, B can be solved using linear curve fitting method, given by

$$\begin{aligned} \sum Y_i &= nA + B\sum X_i \\ \sum X_i Y_i &= A\sum X_i + B\sum X_i^2. \end{aligned}$$

  - Form $y = ab^x$: This can be converted to the linear curve fitting form as

$$y = ab^x \quad \Rightarrow \quad \ln y = \ln a + x\ln b$$

    Define new variables as $Y = \ln y, A = \ln a, B = \ln b, X = x$. We then have the linear form $Y = A + BX$. Then A, B can be solved using linear curve fitting method, given by

$$\begin{aligned} \sum Y_i &= nA + B\sum X_i \\ \sum X_i Y_i &= A\sum X_i + B\sum X_i^2. \end{aligned}$$

    For the data set (x_i, y_i), form a table for $\ln y_i$ to do the calculations.

Consider a data set $(x_1, y_1), \cdots, (x_n, y_n)$.

	Concept	Formula
1	Pearson's correlation coefficient	$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$ $$r = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\left(\sqrt{\sum x_i^2 - n(\bar{x})^2}\right)\left(\sqrt{\sum y_i^2 - n(\bar{y})^2}\right)}$$
2	Spearman's rank correlation coefficient – No repetitions – With repetitions of rank (if any single observation is repeating m_1 times) – For each of the repeated observations with m_i no of times of repetition	$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$ d_i is the difference in ranks of x_i and y_i $$\rho = 1 - \frac{6}{n(n^2-1)}\left[\sum d_i^2 + \frac{1}{12}(m_1^3 - m_1)\right]$$ add a factor $\frac{1}{12}(m_i^3 - m_i)$ to the above formula
4	Regression line of y on x is of the form $y = a + b_{yx}x$,	$$b_{yx} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n(\bar{x})^2}$$ $$a = \bar{y} - b_{yx}\bar{x}.$$
5	Regression line of x on y is of the form $x = c + b_{xy}y$	$$b_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum y_i^2 - n(\bar{y})^2}$$ $$c = \bar{x} - b_{xy}\bar{y}.$$
5	Two regression lines can be written as $$\begin{aligned} y - \bar{y} &= b_{yx}(x - \bar{x}) \\ x - \bar{x} &= b_{xy}(y - \bar{y}) \end{aligned}$$ $(\bar{x}, \bar{y})$ is the intersection point	– b_{yx}: Regression coefficient of y on x – b_{xy}: Regression coefficient of x on y

	Concept	Formula
1	Linear curve fitting normal equations $y = a + bx$	Solve a, b from $$\begin{aligned}\sum y_i &= an + b\sum x_i\\ \sum x_i y_i &= a\sum x_i + b\sum x_i^2.\end{aligned}$$
2	Quadratic curve fitting normal equations $Y = a + bx + cx^2$	Solve a, b, c from $$\begin{aligned}\sum y_i = an + b\sum x_i + c\sum x_i^2\\ \sum x_i y_i = a\sum x_i + b\sum x_i^2 + c\sum x_i^3\\ \sum x_i^2 y_i = a\sum x_i^2 + b\sum x_i^3 + c\sum x_i^4.\end{aligned}$$
3	Exponential curve fitting – Form $y = ae^{bx}$: Define $Y = \ln y, A = \ln a, B = b, X = x$. – Form $y = ab^x$: Define $Y = \ln y, A = \ln a, B = \ln b, X = x$.	Solve A, B from $$\begin{aligned}\sum Y_i &= nA + B\sum X_i\\ \sum X_i Y_i &= A\sum X_i + B\sum X_i^2.\end{aligned}$$