Stat 154: Elementary Statistics

Jongyun Jung

Minnesota State University, Mankato

Ch 9: **Association between Two Variables**

jongyun.jung@mnsu.edu

April 12, 2019

Overview

1. Correlation Coefficient
   - Testing for the significance of the correlation coefficient
2. Simple Linear Regression
   - Testing for the significance of the regression line
   - $(1-\alpha) \times 100\%$ confidence interval

Correlation Coefficient

- The correlation coefficient measures the strength of the linear relationship between two variables. Let X and Y be two different measurements on an individual subject.
- The population correlation coefficient is measured as

$$\rho = \frac{\text{Covariance between} \quad X \quad \text{and} \quad Y}{\sqrt{(\text{Variance of} X)(\text{Variance of} Y)}}$$

- where $-1 \leq \rho \leq 1$.
- The covariance between X and Y is the mean of the product of the deviations from the respective means.
  - $\rho = 0$ indicates that there is no linear relationship between X and Y.
  - $\rho = 1$ indicates that there is a perfect positive linear relationship between X and Y.
  - $\rho = -1$ indicates that there is a perfect negative linear relationship between X and Y.
- Other values are interpreted about how strong the relationship is depending on how close the value is to -1 or +1.

Correlation Coefficient

- Let $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ be n pairs of measurements on X and Y. Then the sample correlation coefficient is computed as

$$\begin{aligned} r &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \\ &= \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sqrt{(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2)(\sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2)}} \end{aligned}$$

Correlation Coefficient

EXAMPLE 9.1 A college administers a student evaluation questionnaire for all its courses. For a random sample of 12 courses, the accompanying table and the student evaluation data file show both the average student ratings of the instructor (on a scale of 1 to 5) and the average expected grades of the students (on a scale from A = 4 to F = 0). Find the sample correlation coefficient between instructor ratings and expected grades.

Instructor rating:	2.8	3.7	4.4	3.6	4.7	3.5	4.1	3.2	4.9	4.2	3.8	3.3
Expected grade:	2.6	2.9	3.3	3.2	3.1	2.8	2.7	2.4	3.5	3.0	3.4	2.5

Solution: Here,

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{46.2}{12} = 3.85,\ \bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{35.4}{12} = 2.95,\ \sum_{i=1}^{n} X_i Y_i = 138.09,$$

$$\sum_{i=1}^{n} X_i^2 = 182.22, \text{ and } \sum_{i=1}^{n} Y_i^2 = 105.86.$$

Then,

$$r = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right)}} = \frac{138.09 - 12(3.85)(2.95)}{\sqrt{(182.22 - 12(3.85)^2)(105.86 - 12(2.95)^2)}}$$

$$= \frac{1.8}{\sqrt{(4.35)(1.43)}} = 0.7217,$$

there is a moderate positive correlation between instructor ratings and expected grades.

Figure: Ex 9-1

Testing for the significance of the correlation coefficient

$$Case1: \begin{array}{c} H_0: \rho \geq 0 \\ H_1: \rho < 0 \end{array} \quad Case2: \begin{array}{c} H_0: \rho \leq 0 \\ H_1: \rho > 0 \end{array} \quad Case3: \begin{array}{c} H_0: \rho = 0 \\ H_1: \rho \neq 0 \end{array}$$

- Case 1 is to test whether there is any significant negative correlation between the variables.
- Case 2 is to test whether there is any significant positive correlation between the variables.
- Case 3 is to test whether there is any significant correlation between the variables.
- The corresponding test statistic is

$$T = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \tag{1}$$

- will have a t-distribution with $n-2$ degrees of freedom.

Simple Linear Regression

- Let Y be the dependent variable and X be the independent variable. Then, the value of Y depends on the value of X. The relationship can be any nature or functional form of

$$Y = \beta_0 + \beta_1 X + \epsilon$$

, where β_0 is the intercept coefficient and β_1 is the slope coefficient. ϵ is the random fluctuation from the line that follows normal distribution with mean zero and variance σ^2.

- For a random sample of size n, the sum of squared errors can be written as

$$SSE = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Simple Linear Regression

- The SSE is minimized when β_1 is estimated as

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})Y_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}$$

- β_0 is estimated as $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$. $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are known as the least square estimates for β_0 and β_1, respectively.

Testing for the significance of the regression line

$$Case1: \begin{array}{l} H_0: \beta_1 \geq 0 \\ H_1: \beta_1 < 0 \end{array} \quad Case2: \begin{array}{l} H_0: \beta_1 \leq 0 \\ H_1: \beta_1 > 0 \end{array} \quad Case3: \begin{array}{l} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{array}$$

- Case 1 is to test whether there is any inverse linear relation between the variables.
- Case 2 is to test whether there is any direct linear relation between the variables.
- Case 3 is to test whether there is any significant linear relation between the variables.
- The corresponding test statistic is

$$T = \frac{\widehat{\beta_1} - 0}{\frac{s_e}{\sqrt{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}}} \tag{2}$$

- will have a t-distribution with $n-2$ degrees of freedom.

Testing for the significance of the regression line

- The corresponding test statistic is

$$T = \frac{\widehat{\beta}_1 - 0}{\frac{s_e}{\sqrt{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}}} \tag{3}$$

- will have a t-distribution with $n-2$ degrees of freedom, where **the standard error of estimate** s_e is

$$s_e = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{n-2}}$$

Testing for the significance of the regression line

- Similar tests involving β_0, the intercept coefficient, can also be obtained. The test statistic is

$$T = \frac{\widehat{\beta}_0 - 0}{s_e\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}}} \tag{4}$$

- which has t-distribution with $n-2$ degrees of freedom.

$(1-\alpha) \times 100\%$ confidence interval

- $(1-\alpha) \times 100\%$ confidence interval for β_1 can be computed as

$$\widehat{\beta}_1 \mp t_{\frac{\alpha}{2};n-2} \cdot \frac{s_e}{\sqrt{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}} \tag{5}$$

- $(1-\alpha) \times 100\%$ confidence interval for β_0 can be computed as

$$\widehat{\beta}_0 \mp t_{\frac{\alpha}{2};n-2} \cdot s_e \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}} \tag{6}$$

$(1-\alpha) \times 100\%$ confidence interval

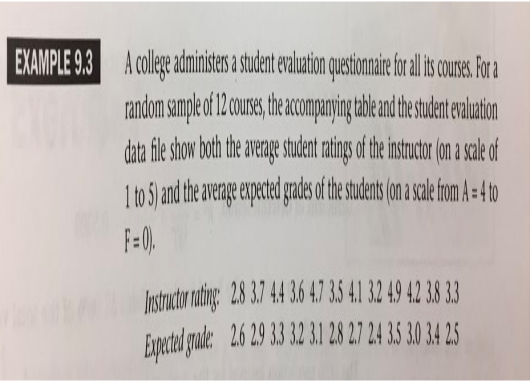
- $(1-\alpha) \times 100\%$ confidence interval for the mean response at $x = x_0, y(x_0) = \beta_0 + \beta_1 x_0$ can be computed as

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_0 \mp t_{\frac{\alpha}{2};n-2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{X})^2}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}} \tag{7}$$

- $(1-\alpha) \times 100\%$ prediction interval for the predicted response at $x = x_0, y(x_0) = \beta_0 + \beta_1 x_0$ can be computed as

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_0 \mp t_{\frac{\alpha}{2};n-2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{X})^2}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}} \tag{8}$$

Simple Linear Regression



EXAMPLE 9.3 A college administers a student evaluation questionnaire for all its courses. For a random sample of 12 courses, the accompanying table and the student evaluation data file show both the average student ratings of the instructor (on a scale of 1 to 5) and the average expected grades of the students (on a scale from A = 4 to F = 0).

Instructor rating:	2.8	3.7	4.4	3.6	4.7	3.5	4.1	3.2	4.9	4.2	3.8	3.3
Expected grade:	2.6	2.9	3.3	3.2	3.1	2.8	2.7	2.4	3.5	3.0	3.4	2.5

Figure: Ex 9-3

Simple Linear Regression

X	Y	X^2	Y^2	XY	$\hat{Y}$	$e = Y - \hat{Y}$
2.8	2.6	7.84	6.76	7.28	2.5155	0.0845
3.7	2.9	13.69	8.41	10.73	2.8880	0.0120
4.4	3.3	19.36	10.89	14.52	3.1776	0.1224
3.6	3.2	12.96	10.24	11.52	2.8466	0.3534
4.7	3.1	22.09	9.61	14.57	3.3018	-0.2018
3.5	2.8	12.25	7.84	9.80	2.8052	-0.0052
4.1	2.7	16.81	7.29	11.07	3.0535	-0.3535
3.2	2.4	10.24	5.76	7.68	2.6811	-0.2811
4.9	3.5	24.01	12.25	17.15	3.3845	0.1155
4.2	3.0	17.64	9.00	12.60	3.0949	-0.0949
3.8	3.4	14.44	11.56	12.92	2.9293	0.4707
3.3	2.5	10.89	6.25	8.25	2.7224	-0.2224
$\Sigma X = 46.2$	$\Sigma Y = 35.4$	$\Sigma X^2 = 182.22$	$\Sigma Y^2 = 105.86$	$\Sigma XY = 138.09$		$\Sigma e \approx 0$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} = \frac{138.09 - 12 \times \frac{46.2}{12} \times \frac{35.4}{12}}{182.22 - 12 \times \left(\frac{46.2}{12}\right)^2} = \frac{1.80}{4.35} = 0.4138,$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{35.4}{12} - 0.4138 \times \frac{46.2}{12} = 1.3569$$

Then the least square estimate of the regression line is $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = 1.3569 + 0.4138X$.

The computation for the estimate of the standard error,

$$s_e = \sqrt{\frac{\sum_{i=1}^{n} e^2}{n-2}} = \sqrt{\frac{0.6853}{12-2}} = 0.2618$$

Unexplained variation, SSE (sum of squared error) $\Sigma e^2 = 0.6853$.

Total variation, SST (sum of squares total)

$$\sum Y^2 - n\bar{Y}^2 = 105.86 - 12\left(\frac{35.4}{12}\right)^2 = 1.43$$

Figure: Ex 9-3

Simple Linear Regression

Explained variation, SSR (sum of squares for regression) = $SST - SSE = 1.43 - 0.6853 = 0.7447$

Coefficient of determination, $R^2 = \frac{SSR}{SST} = \frac{0.7447}{1.43} = 0.5208$

Thus, instructor rating in the fitted model explains 52.08% of the total variation in expected grade.

The 90% prediction interval for the predicted response at $x = 4$ can be computed as

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \mp t_{\alpha/2;n-2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{\left(x_0 - \overline{X}\right)^2}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}}$$

$$1.3569 + 0.4138(4) \mp 1.8125(0.2618)\sqrt{1 + \frac{1}{12} + \frac{(4 - 3.85)^2}{4.35}}$$

$$3.0121 \mp 0.4951 = (2.5170, 3.5072)$$

Figure: Ex 9-3

References

Mezbahur Rahman, Deepak Sanjel, Han Wu. Statistics Introduction, Revised Printing
KendallHunt