14.5: Inferential Statistics for b and r

Learning Objectives

- State the assumptions that inferential statistics in regression are based upon
- Identify heteroscedasticity in a scatter plot
- Compute the standard error of a slope
- Test a slope for significance
- Construct a confidence interval on a slope
- Test a correlation for significance

This section shows how to conduct significance tests and compute confidence intervals for the regression slope and Pearson's correlation. As you will see, if the regression slope is significantly different from zero, then the correlation coefficient is also significantly different from zero.

Assumptions

Although no assumptions were needed to determine the best-fitting straight line, assumptions are made in the calculation of inferential statistics. Naturally, these assumptions refer to the population, not the sample.

1. **Linearity**: The relationship between the two variables is linear.
2. **Homoscedasticity**: The variance around the regression line is the same for all values of X. A clear violation of this assumption is shown in Figure 14.5.1. Notice that the predictions for students with high high-school GPAs are very good, whereas the predictions for students with low high-school GPAs are not very good. In other words, the points for students with high high-school GPAs are close to the regression line, whereas the points for low high-school GPA students are not.

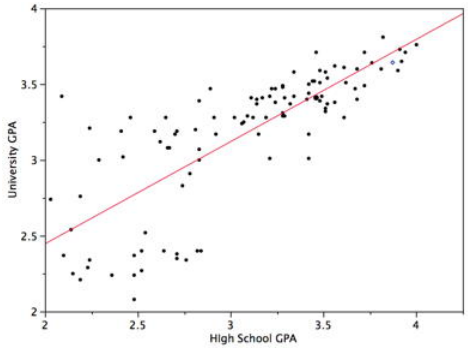

Figure 14.5.1: University GPA as a function of High School GPA

3. The errors of prediction are distributed normally. This means that the deviations from the regression line are normally distributed. It does not mean that X or Y is normally distributed.

Significance Test for the Slope (b)

Recall the general formula for a t test:

$$t = \frac{\text{statistic-hypothesized value}}{\text{estimated standard error of the statistic}} \tag{14.5.1}$$

As applied here, the statistic is the sample value of the slope (b) and the hypothesized value is 0. The number of degrees of freedom for this test is:

$$df = N - 2 \tag{14.5.2}$$

where N is the number of pairs of scores.

The estimated standard error of b is computed using the following formula:

$$s_b = \frac{s_{est}}{\sqrt{SSX}} \tag{14.5.3}$$

where s_b is the estimated standard error of b, s_{est} is the standard error of the estimate, and SSX is the sum of squared deviations of X from the mean of X. SSX is calculated as

$$SSX = \sum (X - M_X)^2 \tag{14.5.4}$$

where M_x is the mean of X. As shown previously, the standard error of the estimate can be calculated as

$$s_{est} = \sqrt{\frac{(1 - r^2)SSY}{N - 2}} \tag{14.5.5}$$

These formulas are illustrated with the data shown in Table 14.5.1. These data are reproduced from the introductory section. The column X has the values of the predictor variable and the column Y has the values of the criterion variable. The third column, x, contains the differences between the values of column X and the mean of X. The fourth column, x_2, is the square of the x column. The fifth column, y, contains the differences between the values of column Y and the mean of Y. The last column, y_2, is simply square of the y column.

Table 14.5.1: Example data

	X	Y	x	x2	y	y2
	1.00	1.00	-2.00	4	-1.06	1.1236
	2.00	2.00	-1.00	1	-0.06	0.0036
	3.00	1.30	0.00	0	-0.76	0.5776
	4.00	3.75	1.00	1	1.69	2.8561
	5.00	2.25	2.00	4	0.19	0.0361
Sum	15.00	10.30	0.00	10.00	0.00	4.5970

The computation of the standard error of the estimate (s_{est}) for these data is shown in the section on the standard error of the estimate. It is equal to 0.964.

$$s_{est} = 0.964 \tag{14.5.6}$$

SSX is the sum of squared deviations from the mean of X. It is, therefore, equal to the sum of the x^2 column and is equal to 10.

$$SSX = 10.00 \tag{14.5.7}$$

We now have all the information to compute the standard error of b:

$$s_b = \frac{0.964}{\sqrt{10}} = 0.305 \tag{14.5.8}$$

As shown previously, the slope (b) is 0.425. Therefore,

$$t = \frac{0.425}{0.305} = 1.39 \tag{14.5.9}$$

$$df = N - 2 = 5 - 2 = 3 \tag{14.5.10}$$

The p value for a two-tailed t test is 0.26. Therefore, the slope is not significantly different from 0.

Confidence Interval for the Slope

The method for computing a confidence interval for the population slope is very similar to methods for computing other confidence intervals. For the 95% confidence interval, the formula is:

$$\text{lower limit} : b - (t_{0.95})(s_b) \tag{14.5.11}$$

$$\text{upper limit} : b + (t_{0.95})(s_b) \tag{14.5.12}$$

where $t_{0.95}$ is the value of t to use for the 95% confidence interval.

The values of t to be used in a confidence interval can be looked up in a table of the t distribution. A small version of such a table is shown in Table $14.5.2$ The first column, df, stands for degrees of freedom.

Table $14.5.2$: Abbreviated t table

df	0.95	0.99
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
8	2.306	3.355
10	2.228	3.169
20	2.086	2.845
50	2.009	2.678
100	1.984	2.626

You can also use the "inverse t distribution" calculator to find the t values to use in a confidence interval.

Applying these formulas to the example data,

$$\text{lower limit} : 0.425 - (3.182)(0.305) = -0.55 \tag{14.5.13}$$

$$\text{upper limit} : 0.425 + (3.182)(0.305) = 1.40 \tag{14.5.14}$$

Significance Test for the Correlation

The formula for a significance test of Pearson's correlation is shown below:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{14.5.15}$$

where N is the number of pairs of scores. For the example data,

$$t = \frac{0.627\sqrt{5-2}}{\sqrt{1-0.627^2}} = 1.39 \tag{14.5.16}$$

Notice that this is the same t value obtained in the t test of b. As in that test, the degrees of freedom is $N - 2 = 5 - 2 = 3$.