Regression Analysis: A Complete Example

This section works out an example that includes all the topics we have discussed so far in this chapter.

A complete example of regression analysis.

PhotoDisc, Inc./Getty Images

A random sample of eight drivers insured with a company and having similar auto insurance policies was selected. The following table lists their driving experiences (in years) and monthly auto insurance premiums.

Driving Experience (years)	Monthly Auto Insurance Premium
5	$64
2	87
12	50
9	71
15	44
6	56
25	42
16	60

a. Does the insurance premium depend on the driving experience or does the driving experience depend on the insurance premium? Do you expect a positive or a negative relationship between these two variables?

b. Compute SS_{xx}, SS_{yy}, and SS_{xy}.

c. Find the least squares regression line by choosing appropriate dependent and independent variables based on your answer in part a.

d. Interpret the meaning of the values of a and b calculated in part c.

e. Plot the scatter diagram and the regression line.

f. Calculate r and r^2 and explain what they mean.

g. Predict the monthly auto insurance premium for a driver with 10 years of driving experience.

h. Compute the standard deviation of errors.

i. Construct a 90% confidence interval for B.

j. Test at the 5% significance level whether B is negative.

k. Using $\alpha = .05$, test whether ρ is different from zero.

Solution

a. Based on theory and intuition, we expect the insurance premium to depend on driving experience. Consequently, the insurance premium is a dependent variable and driving experience is an independent variable in the regression model. A new driver is considered a high risk by the insurance companies, and he or she has to pay a higher premium for auto insurance. On average, the insurance premium is expected to decrease with an increase in the years of driving experience. Therefore, we expect a negative relationship between these two variables. In other words, both the population correlation coefficient ρ and the population regression slope B are expected to be negative.

b. Table 13.5 shows the calculation of Σx, Σy, Σxy, Σx^2, and Σy^2.

Table 13.5

Experience x	Premium y	xy	x^2	y^2
5	64	320	25	4096
2	87	174	4	7569
12	50	600	144	2500
9	71	639	81	5041
15	44	660	225	1936
6	56	336	36	3136
25	42	1050	625	1764
16	60	960	256	3600
$\Sigma x = 90$	$\Sigma y = 474$	$\Sigma xy = 4739$	$\Sigma x^2 = 1396$	$\Sigma y^2 = 29{,}642$

The values of x and y are

$$\overline{x} = \Sigma x / n = 90 / 8 = 11.25$$
$$\overline{y} = \Sigma y / n = 474 / 8 = 59.25$$

The values of SS_{xy}, SS_{xx}, and SS_{yy} are computed as follows:

$$SS_{xy} = \Sigma xy = \frac{(\Sigma x)(\Sigma y)}{n} = 4739 - \frac{(90)(474)}{8} = -593.5000$$

$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 1396 - \frac{(90)^2}{8} = 383.5000$$

$$SS_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 29{,}642 - \frac{(474)^2}{8} = 1557.5000$$

c. To find the regression line, we calculate a and b as follows:

$$b = \frac{SS_{xy}}{SS_{xx}} = \frac{-593.5000}{383.5000} = -1.5476$$
$$a = \overline{y} - b\,\overline{x} = 59.25 - (-1.5476)(11.25) = 76.6605$$

Thus, our estimated regression line $\hat{y} = a + bx$ is

$$\hat{y} = 76.6605 - 1.5476x$$

d. The value of $a = 76.6605$ gives the value of $\hat{y}$ for $x = 0$; that is, it gives the monthly auto insurance premium for a driver with no driving experience. However, as mentioned earlier in this chapter, we should not attach much importance to this statement because the sample contains drivers with only two or more years of experience. The value of b gives the change in $\hat{y}$ due to a change of one unit in x. Thus, $b = -1.5476$ indicates that, on average, for every extra year of driving experience, the monthly auto insurance premium decreases by \$1.55. Note that when b is negative, y decreases as x increases.

e. Figure 13.21 shows the scatter diagram and the regression line for the data on eight auto drivers. Note that the regression line slopes downward from left to right. This result is consistent with the negative relationship we anticipated between driving experience and insurance premium.

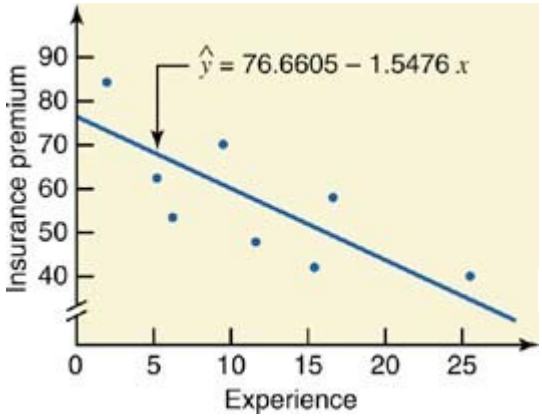

Figure 13.21 Scatter diagram and the regression line.

f. The values of r and r^2 are computed as follows:

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} = \frac{-593.5000}{\sqrt{(383.5000)(1557.5000)}} = -.77$$

$$r^2 = \frac{b\,SS_{xy}}{SS_{yy}} = \frac{(-1.5476)(-593.5000)}{1557.5000} = .59$$

The value of $r = -.77$ indicates that the driving experience and the monthly auto insurance premium are negatively related. The (linear) relationship is strong but not very strong. The value of $r^2 = .59$ states that 59% of the total variation in insurance premiums is explained by years of driving experience and 41% is not. The low value of r^2 indicates that there may be many other important variables that contribute to the determination of auto insurance premiums. For example, the premium is expected to depend on the driving record of a driver and the type and age of the car.

g. Using the estimated regression line, we find the predicted value of y for $x = 10$ is

$$\hat{y} = 76.6605 - 1.5476x = 76.6605 - 1.5476(10) = \mathbf{\$\,61.18}$$

Thus, we expect the monthly auto insurance premium of a driver with 10 years of driving experience to be \$61.18.

h. The standard deviation of errors is

$$s_e = \sqrt{\frac{SS_{yy} - b\,SS_{xy}}{n-2}} = \sqrt{\frac{1557.5000 - (-1.5476)(-593.5000)}{8-2}} = \mathbf{10.3199}$$

i. To construct a 90% confidence interval for B, first we calculate the standard deviation of b:

$$s_b = \frac{s_e}{\sqrt{SS_{xx}}} = \frac{10.3199}{\sqrt{383.5000}} = .5270$$

For a 90% confidence level, the area in each tail of the t distribution is

$$\alpha / 2 = (1 - .90) / 2 = .05$$

The degrees of freedom are

$$df = n - 2 = 8 - 2 = 6$$

From the t distribution table, the t value for .05 area in the right tail of the t distribution and 6 df is 1.943. The 90% confidence interval for B is

$$\begin{aligned} b \pm ts_b &= -1.5476 \pm 1.943(.5270) \\ &= -1.5476 \pm 1.0240 = \mathbf{-2.57 \text{ to } -.52} \end{aligned}$$

Thus, we can state with 90% confidence that B lies in the interval −2.57 to −.52. That is, on average, the monthly auto insurance premium of a driver decreases by an amount between $.52 and $2.57 for every extra year of driving experience.

j. We perform the following five steps to test the hypothesis about B.
Step 1. *State the null and alternative hypotheses*.
The null and alternative hypotheses are written as follows:

$$H_0: B = 0 \quad (B \text{ is not negative})$$
$$H_1: B < 0 \quad (B \text{ is negative})$$

Note that the null hypothesis can also be written as H_0: $B \geq 0$.
Step 2. *Select the distribution to use*.
Because σ_ε is not known, we use the t distribution to make the hypothesis test.
Step 3. *Determine the rejection and nonrejection regions*.
The significance level is .05. The < sign in the alternative hypothesis indicates that it is a left-tailed test.

$$\text{Area in the left tail of the } t \text{ distribution} = \alpha = .05$$
$$df = n - 2 = 8 - 2 = 6$$

From the t distribution table, the critical value of t for .05 area in the left tail of the t distribution and 6 df is −1.943, as shown in Figure 13.22.

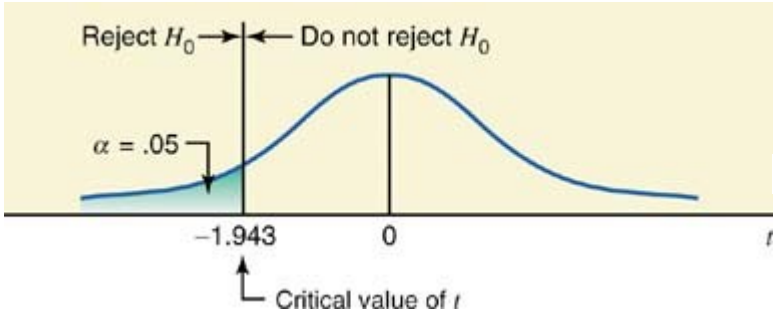

Figure 13.22

Step 4. *Calculate the value of the test statistic*.
The value of the test statistic t for b is calculated as follows:

$$t = \frac{b - B}{s_b} = \frac{-1.5476 - 0}{.5270} = -2.937$$

(From H_0)

Step 5. *Make a decision*.

The value of the test statistic $t = -2.937$ falls in the rejection region. Hence, we reject the null hypothesis and conclude that B is negative. That is, the monthly auto insurance premium decreases with an increase in years of driving experience.

Using the *p*-Value to Make a Decision

We can find the range for the *p*-value from the *t* distribution table (Table V of Appendix C) and make a decision by comparing that *p*-value with the significance level. For this example, $df = 6$ and the observed value of t is −2.937. From Table V (the *t* distribution table) in the row of $df = 6$, 2.937 is between 2.447 and 3.143. The corresponding areas in the right tail of the *t* distribution are .025 and .01. But our test is left-tailed and the observed value of t is negative. Thus, $t = -2.937$ lies between −2.447 and −3.143. The corresponding areas in the left tail of the *t* distribution are .025 and .01. Therefore the range of the *p*-value is

$$.01 < p\text{-value} < .025$$

Thus, we can state that for any α equal to or greater than .025 (the upper limit of the *p*-value range), we will reject the null hypothesis. For our example, $\alpha = .05$, which is greater than the upper limit of the *p*-value of .025. As a result, we reject the null hypothesis.

Note that if we use technology to find this *p*-value, we will obtain a *p*-value of .013. Then we can reject the null hypothesis for any $\alpha > .013$.

k. We perform the following five steps to test the hypothesis about the linear correlation coefficient ρ.

Step 1. *State the null and alternative hypotheses*.

The null and alternative hypotheses are:

$$H_0: \rho = 0 \quad \text{(The linear correlation coefficient is zero)}$$
$$H_1: \rho \neq 0 \quad \text{(The linear correlation coefficient is different from zero)}$$

Step 2. *Select the distribution to use*.

Assuming that variables x and y are normally distributed, we will use the *t* distribution to perform this test about the linear correlation coefficient.

Step 3. *Determine the rejection and nonrejection regions*.

The significance level is 5%. From the alternative hypothesis we know that the test is two-tailed. Hence,

$$\text{Area in each tail of the } t \text{ distribution} = .05 / 2 = .025$$
$$df = n - 2 = 8 - 2 = 6$$

From the *t* distribution table, Table V of Appendix C, the critical values of t are −2.447 and 2.447. The rejection and nonrejection regions for this test are shown in Figure 13.23.

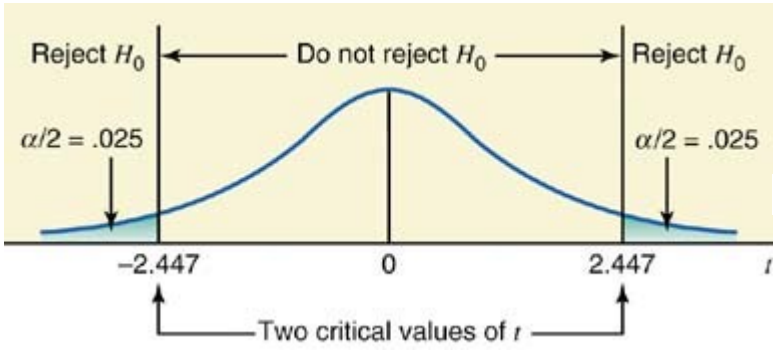

Figure 13.23

Step 4. *Calculate the value of the test statistic*.

The value of the test statistic t for r is calculated as follows:

$$t = r\sqrt{\frac{n-2}{1-r^2}} = (-.77)\sqrt{\frac{8-2}{1-(-.77)^2}} = -2.956$$

Step 5. *Make a decision*.
The value of the test statistic $t = -2.956$ falls in the rejection region. Hence, we reject the null hypothesis and conclude that the linear correlation coefficient between driving experience and auto insurance premium is different from zero.

Using the *p*-Value to Make a Decision

We can find the range for the *p*-value from the *t* distribution table and make a decision by comparing that *p*-value with the significance level. For this example, $df = 6$ and the observed value of *t* is −2.956. From Table V (the *t* distribution table) in the row of $df = 6$, $t = 2.956$ is between 2.447 and 3.143. The corresponding areas in the right tail of the *t* distribution curve are .025 and.01. Since the test is two tailed, the range of the *p*-value is

$$2(.01) < p\text{-value} < 2(.025) \quad \textbf{or} \quad .02 < p\text{-value} < .05$$

Thus, we can state that for any α equal to or greater than .05 (the upper limit of the *p*-value range), we will reject the null hypothesis. For our example, $\alpha = .05$, which is equal to the upper limit of the *p*-value. As a result, we reject the null hypothesis.

```
Welcome to Minitab, press F1 for help.
```

Regression Analysis: Premium y versus Experience x

-

```
The regression equation is
Premium y = 76.7 - 1.55 Experience x

Predictor        Coef  SE Coef      T      P
Constant       76.660    6.961  11.01  0.000
Experience x  -1.5476   0.5270  -2.94  0.026


S = 10.3199   R-Sq = 59.0%   R-Sq(adj) = 52.1%


Analysis of Variance

Source          DF      SS     MS     F      P
Regression       1   918.5  918.5  8.62  0.026
Residual Error   6   639.0  106.5
Total            7  1557.5


Predicted Values for New Observations

New
Obs    Fit  SE Fit      95% CI          95% PI
  1  61.18    3.71  (52.11, 70.26)  (34.35, 88.02)


Values of Predictors for New Observations

New
Obs  Experience x
  1          10.0
```

The following table gives information on ages and cholesterol levels for a random sample of 10 men.

Age	58	69	43	39	63	52	47	31	74	36
Cholesterol level	189	235	193	177	154	191	213	165	198	181

a. Taking age as an independent variable and cholesterol level as a dependent variable, compute SS_{xx}, SS_{yy}, and SS_{xy}.
b. Find the regression of cholesterol level on age.
c. Briefly explain the meaning of the values of a and b calculated in part b.
d. Calculate r and r^2 and explain what they mean.
e. Plot the scatter diagram and the regression line.
f. Predict the cholesterol level of a 60-year-old man.
g. Compute the standard deviation of errors.
h. Construct a 95% confidence interval for B.
i. Test at the 5% significance level if B is positive.
j. Using $\alpha = .025$, can you conclude that the linear correlation coefficient is positive?

Answer:

a. $SS_{xx} = 1895.6000$; $SS_{yy} = 4798.4000$; $SS_{xy} = 1231.8000$
b. $\hat{y} = 156.3302 + .6498x$
d. $r = .41$; $r^2 = .17$
f. 195.3182
g. $s_e = 22.3550$
h. −.53 to 1.83
i. H_0: $B = 0$; H_1: $B > 0$; critical value: $t = 1.860$; test statistic: $t = 1.265$; do not reject H_0
j. H_0: $\rho = 0$; H_1: $\rho > 0$; critical value: $t = 2.306$; test statistic: $t = 1.271$; do not reject H_0