Regression Analysis

Regression Analysis

Introduction

Describing and assessing the significance of relationships between variables is very important in research. We will first learn how to do this in the case when the two variables are quantitative. Quantitative variables have numerical values that can be ordered according to those values.

Main idea

We wish to study the relationship between two quantitative variables.

- Generally one variable is the dependent variable, denoted by y. This variable measures the outcome of the study and is also called the *response variable*.

- The other variable is the independent variable, denoted by x and is also called the *explanatory variable*. It is the variable that is thought to explain the changes we see in the response/ dependent variable.
- The first step in examining the relationship is to use a graph - a scatterplot - to display the relationship. We will look for an overall pattern and see if there are any departures from this overall pattern.

- If a linear relationship appears to be reasonable from the scatterplot, we will take the next step of finding a model (an equation of a line) to summarize the relationship. The resulting equation may be used for predicting the response for various values of the explanatory variable. Let's begin with an example that we will carry throughout our discussions.

Graphing the Relationship

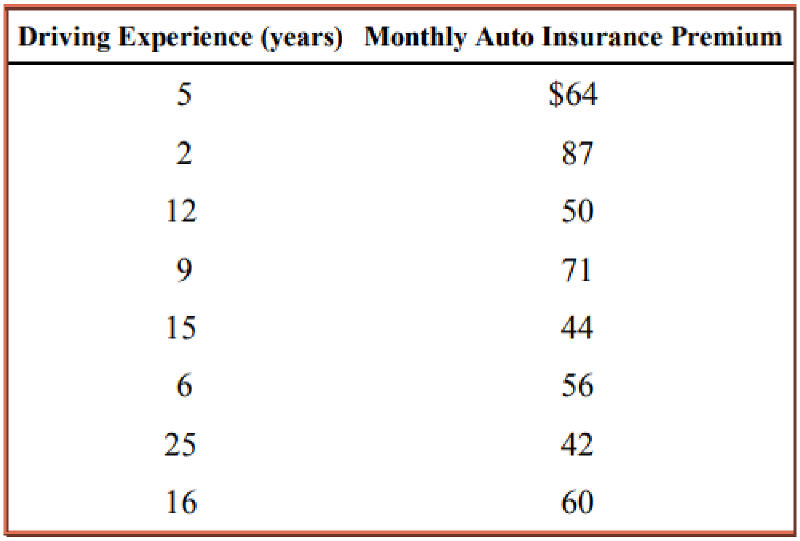
EXAMPLE

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.

A complete example of regression analysis.

PhotoDisc, Inc./Getty Images

a. Does the insurance premium depend on the driving experience or does the driving experience depend on the insurance premium?
b. Do you expect a positive or a negative relationship between these two variables?
c. Find the least squares regression line
d. Interpret the meaning of the values of a and b calculated in part c.
e. Plot the scatter diagram and the regression line. Interpret your scatter plot
f. Calculate coefficient of correlation r , interpret the value of r.
g. Predict the monthly auto insurance premium for a driver with 10 years of driving experience.



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

Here we have

Response (dependent) variable y = monthly auto insurance premium

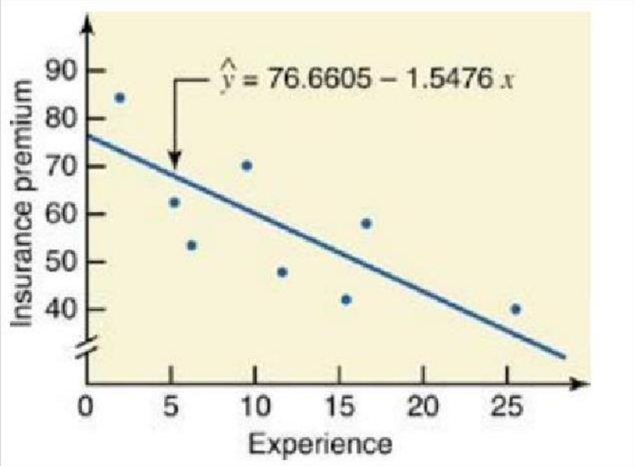
Explanatory (independent) variable x = driving experience

Step 1: Examine the data graphically with a scatterplot.

Solution (a and b)

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

Solution e (cont……)



Interpret the scatterplot in terms of …

- overall form (is the average pattern look like a straight line or is it curved?)
- direction of association (positive or negative)
- strength of association (how much do the points vary around the average pattern?)
- any deviations from the overall form?

Describing a Linear Relationship with a Regression

Line Regression analysis is the area of statistics used to examine the relationship between a quantitative response variable and one or more explanatory variables. A key element is the estimation of an equation that describes how, on average, the response/ dependent variable is related to the explanatory/ independent variables. A regression equation can also be used to make predictions. The simplest kind of relationship between two variables is a straight line, the analysis in this case is called linear regression.

Step 2: Finding the Regression Line/ least square line of regression

Solution c

Remember the equation of a line? In statistics we denote the regression line for a sample as:

$$y = a + bx$$

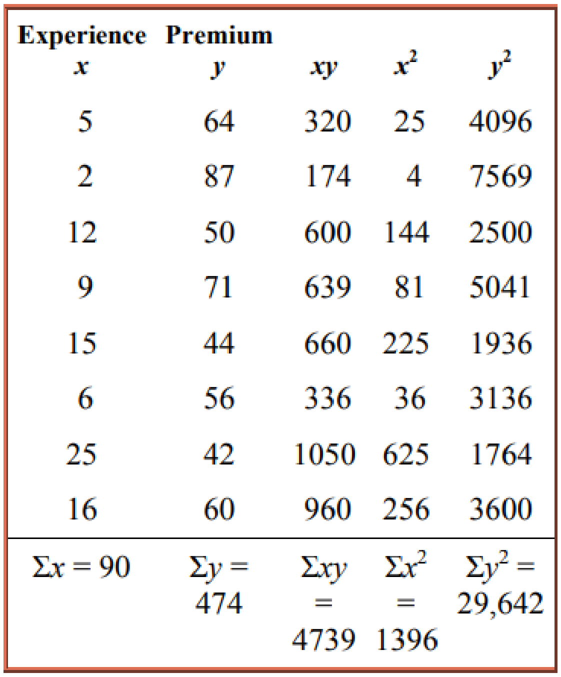
where:

a is the regression constant/ intercept of the regression line

b is the regression coefficient/ slope of the regression line.

$$b \quad = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2} \qquad a = \overline{Y} - b\overline{X}$$

,



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	Σxy = 4739	Σx^2 = 1396	$\Sigma y^2 =$ 29,642

Use the table to find out the values of a (regression constant) and b (regression coefficient)

The values of x and $y$is

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

Use the following formulas,

$$b \quad = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2}$$

$$a = \overline{Y} - b\overline{X}$$

$$b = -1.5476$$

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

$$a = 76.6605$$

$$y = 76.6605 - 1.5476\,x$$

Solution d

From the above regression equation

INTERPETATION SENTENCE FOR INTERCEPT:

y for zero x is ______________________________

- put your dependent variable instead of ‘y’ in above sentence
- put your independent variable instead of ‘x” in above sentence

In this example Intercept or a = 76.6605 gives the value of ŷ for x = 0; that is,

“the monthly autoinsurance premium for zero driving experience is $76.6”

OR

The monthly auto insurance premium for a driver with no driving experience is $76.6

INTERPETATION SENTENCE FOR SLOPE:

For one unit change x , y increased by / decreased by ________________________

Here b = −1.5476 indicates that,

“For one unit change in driving experience, monthly auto insurance premium decreased by $1.5”

OR

on average, for every extra year of driving experience, the monthly auto insurance premium decreases by $1.55.

Solution e (cont…)

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.

The meaning of intercept and slope (another Example)

Let's delve into the fascinating world of linear regression equations and unpack the meaning behind their intercept and slope. Imagine you're a researcher at a pharmaceutical company studying the effects of a new anti-anxiety medication. You conduct a clinical trial where participants are given different dosages (X) of the drug, and you measure their anxiety levels (Y) afterwards.

The linear regression equation you obtain from this experiment might look something like this:

$$y = a + bx$$

Here, Y represents the anxiety level, X represents the dosage of the medication, and a and b are the coefficients we're interested in.

The Intercept (a):

Think of the intercept (a) as the starting point of your anxiety level on the Y-axis. It tells you the predicted anxiety level when the dosage (X) is ZERO. In this medication example, the intercept might represent the average anxiety level of participants before taking any medication. It's important to consider whether this value makes sense in the context of your study. For example, if the intercept is a negative value, it would suggest that some participants had negative anxiety levels before the medication, which is unlikely. If the intercept is a positive value, it would suggest that some participants had positive anxiety levels before the medication.

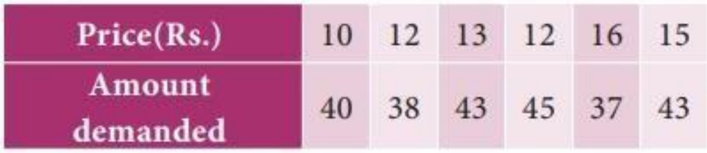
The Slope (b):

The slope (b) is where the real magic happens! It represents the rate of change in anxiety level with respect to the medication dosage. Here's how to interpret it:

- **Positive Slope (b > 0):** This indicates a positive relationship. As the dosage (X) increases, the predicted anxiety level (Y) also increases. In our example, this would suggest that higher medication dosages lead to higher anxiety levels, which wouldn't be the desired outcome!
- **Negative Slope (b < 0):** This indicates a negative relationship. As the dosage (X) increases, the predicted anxiety level (Y) decreases. This is the ideal scenario for our medication - higher dosages are associated with lower anxiety levels.
- **Zero Slope (b = 0):** This means there's no change in anxiety level regardless of the medication dosage. The medication has no effect in this case.

TRY IT

Calculate the regression equation of demanded amount *Y* on the price from the data given below. Estimate the likely demand when the price is Rs.20.



Price(Rs.)	10	12	13	12	16	15
Amount demanded	40	38	43	45	37	43

Steps to be followed

- Plot the scatter diagram, observe the pattern
- Interpret the pattern of scatter plot
- Find the values of intercept and slope
- Find the least square line of equation
- Interpret the slope and intercept
- Find the estimated value of demand when price is Rs.20