Correlation and Regression

October 17, 2013

Correlation

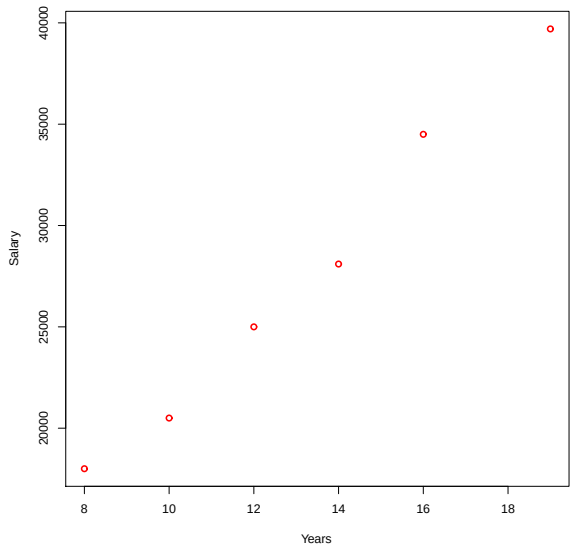
The tools used to explore the relationship between two continuous variable, such as height and weight, the concentration of an injected drug and heart rate, or the consumption level of some nutrient and weight gain etc., is the regression and correlation analysis. These tools can be used to find out if the outcome from one variable depends on the value of the other variable, which would mean a dependency from one variable on the other. Regression and correlation analysis can be used to describe the nature and strength of the relationship between two continuous variables. The first step in the investigation of the relationship between two continuous variables is a scatterplot. We create a scatterplot for the two variables and evaluate the quality of the relationship.

Example

Does the number of years invested in schooling pay of in the job market? Apparently so - the better educated you are, the more money you will earn. The data in the following table give the median annual income of full-time workers age 25 or older by the number of years of schooling completed. Let X: Years of schooling and Y: Salary(in Dollars)

X	8	10	12	14	16	19
Y	18,000	20,500	25,000	28,100	34,500	39,700

The scatteplot of the data is as below



Correlation Coefficient

The Pearson Product-Moment Correlation Coefficient (ρ), or correlation coefficient for short is a measure of the degree of linear relationship between two variables, usually labeled X and Y.
The correlation coefficient $\rho = \rho_{XY}$ is defined by

$$\rho = \frac{COV(X, Y)}{\sqrt{Var(X)Var(Y)}}.$$

where, $COV(X, Y) = E(X - \mu_X)(Y - \mu_Y) = E(XY) - \mu_X\mu_Y$ is called the covariance between X and Y.
Note that, $-1 \leq \rho \leq 1$.
In R use `cor(x,y)` to calculate the correlation coefficient and `cov(x,y)` to calculate the covariance between x and y.

Testing for Correlation Coefficient

Pearson Correlation Coefficient

```
cor.test(x,y)
```

Spearman Correlation Coefficient

```
cor.test(x,y,method="spearman")
```

Kendall (tau-b) Correlation Coefficient.

```
cor.test(x,y,method="kendall")
```

Pearson R Correlation Coefficient

Assumptions

- They must be approximately Gaussian distributed.
- There must be a significant linear relationship between them.
- They must be either interval or ratio measurements.
- There may not be any outliers.
- They must have similar variances.

The correlation coefficient $\rho = \rho_{XY}$ is defined by

$$\rho = \frac{COV(X, Y)}{\sqrt{Var(X)Var(Y)}}.$$

where, $COV(X, Y) = E(X - \mu_X)(Y - \mu_Y) = E(XY) - \mu_X\mu_Y$ is called the covariance between X and Y

Spearman Rank Correlation Coefficient

Assumptions

- They must be rank ordered.
- They are monotonically related.
- They need not be Gaussian distributed.
- They do not require the parameters of distribution.
- They do not require that the relationship between them being linear.
- They do not require to be measured on interval or ratio scale.

The Spearman rank correlation coefficient is Pearsons moment correlation formula applied to ordinals, X_i and $Y_i, i = 1, 2, \cdots, N$, with no ties in either X_i or Y_i.
We have

$$\rho = 1 - \frac{6 \sum_i d_i^2}{N(N^2 - 1)}$$

where, $d_i = R(X_i) - R(Y_i)$ is the difference in the ranks.

Kendall's Tau Coefficient

It is a non-parametric correlation coefficient like the Spearman correlation that can be used to find correlation between the variables X and Y. While Spearman rank correlation coefficient is Pearson correlation coefficient computed from ranked variables, the Kendall correlation rather represents the difference between the probabilities of the dependent variable Y increasing and decreasing with respect to X, and may not be necessary to rank order them. However, they are usually rank ordered to facilitate computation. If (U_i, V_i) and $(U_j, V_j), i, j \in N$ are two pairs of observations not necessarily rank ordered unlike the Spearman correlation coefficient, with no ties among them then if the pairs $(Ui - Uj)$ and $(Vi - Vj)$ are of the same sign for each i and j then these pairs are called `concordant` pairs c. If they have opposite signs then they are called `discordant` pairs d. The measure of correlation proposed by Kendall in case of no ties is

$$\tau = \frac{(N_c - N_d)}{N(N-1)/2}$$

where N_c and N_d are the number of concordant pairs and number of discordant pairs respectively.

Example

Twelve MBA graduates are studied to measure the strength of the relationship between their score on the GMAT which they took prior to entering graduate school, and their grade point average while they were in MBA program. Their GMAT scores(x) and GPA scores(y) are given below

GMAT	710	610	640	580	545	560	610	530	560	540	570	56
GPA	4.0	4.0	3.9	3.8	3.7	3.6	3.5	3.5	3.5	3.3	3.2	3.

```
x<-c(710,610,640,580,545,560,610,530,560,540,570,560)
y<-c(4.0, 4.0, 3.9, 3.8, 3.7, 3.6, 3.5, 3.5, 3.5, 3.3, 3.2, 3.2)
cor(x,y)
cor.test(x,y)
cor.test(x, y, method="s")
cor.test(x,y, method="k")
```

Regression Analysis

Regression analysis is a statistical technique for investigating and modeling the relationship between variables. Applications of regression are numerous and occur in almost every field, including engineering, the physical and chemical sciences, economics, management, life and biological sciences and the social science. It is the most widely used statistical technique.
We use regression analysis for explaining or modeling the relationship between a single variable y, called the response, output or dependent variable, and one or more predictor, input, independent or explanatory variables, $x_1, x_2, ..., x_p$.
When $p = 1$, it is called simple regression but when $p > 1$ it is called multiple regression or sometimes multivariate regression. When there is more than one y, then it is called multivariate multiple regression. The response must be a continuous variable but the explanatory variables can be continuous, discrete or categorical.
Regression analysis have several possible objectives including

- A general description of data structure
- Variable screening
- Prediction of future observations
- Assessment of the effect of, or relationship between, explanatory variables on the response

Simple Linear Regression Model

Consider a random sample of n observations of the form $(x_1, y_1), (x_2, y_2), ...(x_n, y_n)$ where x is independent variable and y is the dependent variable, both being scalars. The model that is applicable in the simplest regression structure is the simple linear regression model. Here the term simple implies a single regressor variable x and the term linear implies linear in the coefficients $\beta's$. The model is given by

$$y = \beta_0 + \beta_1 x + \epsilon$$

where β_0 and β_1 are the intercept and slope respectively, and ϵ is the model error.

Note that the variable x is often called the predictor or regressor variable and y as a response variable.

We assume that $\epsilon_i \sim N(0, \sigma^2)$ and we also assume that ϵ_i are uncorrelated from observation to observation. In addition, any error in the measurement of the x_i is assumed to be small compared to the range.

In a simple linear regression model we have

$$E(y_i) = \beta_0 + \beta_1 x_i$$

and the variance is σ^2.

Model Assumption

The standard analysis is based on the following assumptions about the regressor variable x and the random errors $\epsilon_i, i = 1, 2, ..., n$:

- The regressor variable is under the experimenter's control, who can set the values of $x_1, x_2, ..., x_n$. This means that x_i can be taken as constants, they are not random variables.
- $E(\epsilon_i) = 0, i = 1, 2, 3, ..., n$. This implies that $\mu_i = E(y_i) = \beta_0 + \beta_1 x_i, i = 1, 2, 3, \cdots, n$
- $Var(\epsilon_i) = \sigma^2$ is constant for all $i = 1, 2, \cdots, n$. This implies $Var(y_i) = \sigma^2$.
- Different errors ϵ_i and ϵ_j and hence the different responses y_i and y_j, are independent. This implies that $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$.

In summary, the regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ implies that the response y_i come from probability distributions whose means are $E(y_i) = \beta_0 + \beta_i x_i$ and whose variances are σ^2, the same for all levels of x. Furthermore, any two responses y_i and y_j are not correlated.

Objectives of the Analysis

Given a set of observations, we would like to answer the following questions:

- Can we establish a relationship between x and y?
- Can we predict y from x? To what extent can we predict y from x?
- Can we control y by using x?

In order to answer these questions within the simple regression framework we need to estimate the values of β_0, β_1 & σ^2 from the given data set. In particular, we are interested in β_1 as $\beta_1 = 0$ indicates the absence of linear association.

Parameter Estimation

Consider a simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, 2, 3, \cdots, n$$

The method of least squares is used more extensively than any other estimation procedure for estimating the regression parameters β_0 and β_1.
The method of least squares is designed to provide estimators b_0 and b_1 of β_0 and β_1, respectively and the fitted value

$$\widehat{y_i} = b_0 + b_1 x_i$$

of the response so that the residual sum of squares (RSS) or the error sum of squares (SSE)is minimized.
Note that

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = \sum_{i=1}^{n} [y_i - b_0 - b_1 x_i]^2$$

Hence b_0 and b_1 must satisfies the

$$\frac{\partial}{\partial b_0} \left[\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \right] = 0$$

$$\frac{\partial}{\partial b_1} \left[\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \right] = 0$$

Parameter Estimation

$$\begin{aligned}\sum_{i=1}^{n} y_i &= nb_0 + b_1 \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i y_i &= b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2\end{aligned}$$

Solving these normal equations simultaneously for b_0 and b_1

$$\begin{aligned}b_1 &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ b_0 &= \overline{y} - b_1\overline{x}\end{aligned}$$

Our estimators for intercept and slope are

$$b_0 = \overline{y} - b_1\overline{x}$$

$$b_1 = \frac{S_{xy}}{S_{xx}}$$

where $S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})y_i$ and $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$.

Example

Table below gives the measurements of systolic blood pressure(SBP) and age for a sample of 15 individuals older than age 40 years.

SBP(y)	164	220	133	146	162	144	166	152	140	145	135	150	170	122	120
Age(x)	65	63	47	54	60	44	59	64	51	49	57	56	63	41	43

For this data we have the summary statistic

$$\begin{aligned} n &= 15, \quad \sum_{i=1}^{n} x_i = 816, \\ \sum_{i=1}^{n} y_i &= 2269, \quad \sum_{i=1}^{n} x_i^2 = 45318, \\ \sum_{i=1}^{n} y_i^2 &= 351475, \quad \sum_{i=1}^{n} x_i y_i = 125445 \end{aligned}$$

Now,

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})y_i = \sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i = 2011.4$$

and

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 927.6$$

Hence,

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{2011.4}{927.6} = 2.1684$$

and

$$b_0 = \bar{y} - b_1\bar{x} = 151.267 - 2.1684 \times 54.4 = 33.306$$

Therefore, resulting least square line is

$$\hat{y} = 33.306 + 2.1684x$$

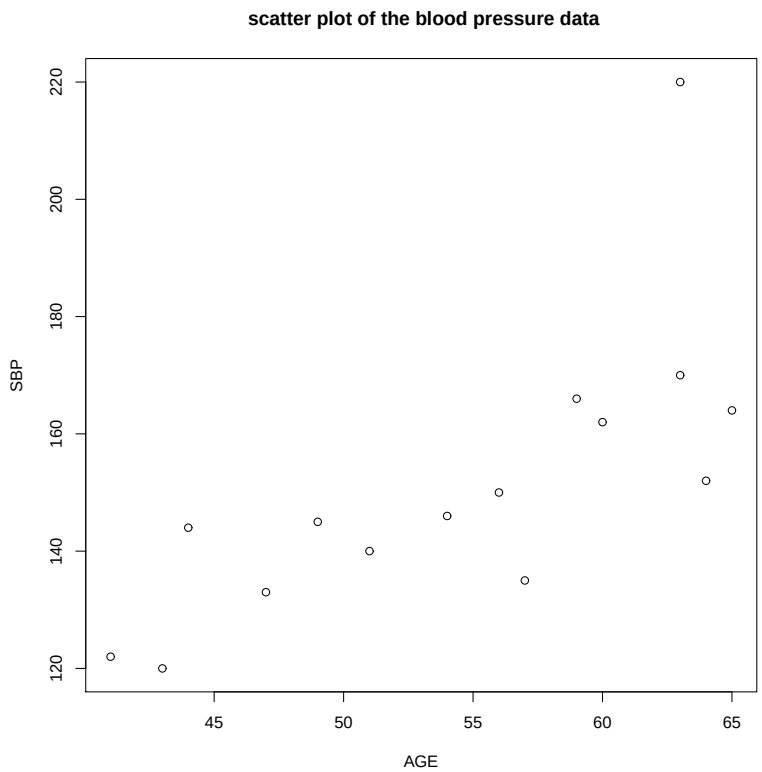
Example

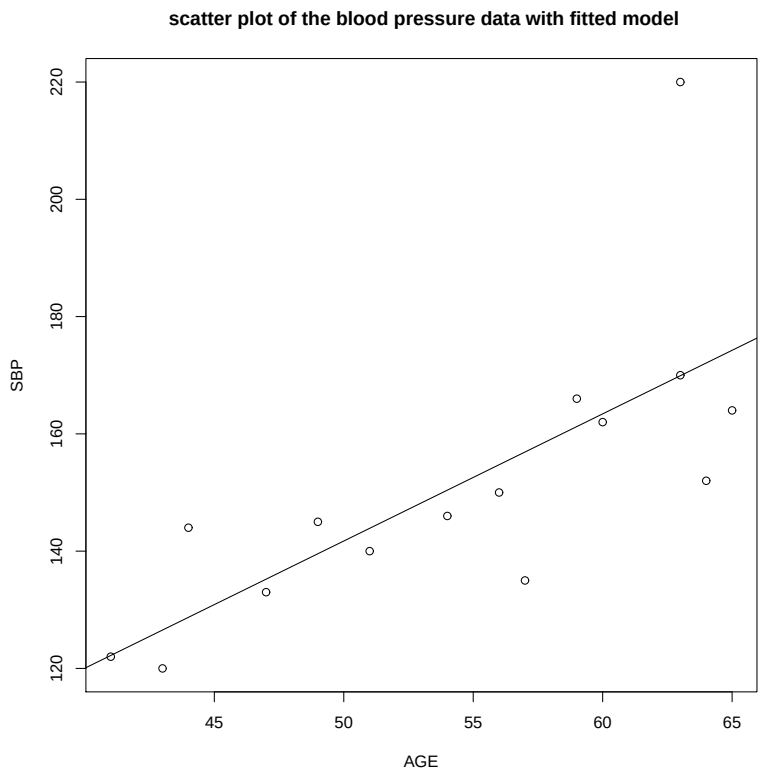
Using R to obtain the regression model

```
> x<-c( 65, 63, 47, 54, 60, 44, 59, 64, 51, 49, 57, 56, 63, 41, 43)
> y<-c(164,220,133,146,162,144,166,152,140,145,135,150,170,122,120)
> model=lm(y~x)
> model
Call:
lm(formula = y ~ x)
Coefficients:
(Intercept)            x
     33.306        2.168
```

Example

```
> x<-c( 65, 63, 47, 54, 60, 44, 59, 64, 51, 49, 57, 56, 63, 41, 43)
> y<-c(164,220,133,146,162,144,166,152,140,145,135,150,170,122,120)
> model=lm(y~x)
> plot(x,y,xlab="AGE",ylab="SBP",main="scatter plot of the blood pressu
> plot(x,y,xlab="AGE",ylab="SBP",main="scatter plot of the blood pressu
> abline(model)
```





Extractor Functions for the results of *lm*()

`summary()`	Returns summary information about the regression
`plot()`	makes diagnostic plots
`coef()`	returns the coefficients
`residuals()`	returns the residuals (can be abbreviated to resid())
`fitted()`	returns the fitted values
`confint()`	returns the confidence interval for the parameters
`deviance()`	returns RSS
`predict()`	performance predictions
`anova()`	finds various sums of squares
`AIC()`	is used for model selection

Properties of fitted regression line

The fitted regression line $\widehat{y} = b_0 + b_1 x$ has the following properties:

- The sum of the residuals is zero, i.e., $\sum_{i=1}^{n} e_i = 0$.
- The sum of the squared residuals, $\sum_{i=1}^{n} {e_i}^2$ is minimum.
- The sum of the observed values y_i equals the sum of the fitted values $\widehat{y_i}$,
  i.e., $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y_i}$
- The regression line always pases through the point $(\overline{x}, \overline{y})$.