Linear Regression

1 Introduction

It is often interesting to study the effect of a variable on a response. In ANOVA, the response is a continuous variable and the variables are discrete / categorical. What if the variables are also continuous? That is, how to examine the relationship between a continuous outcome and one or more continuous variables? For example, what is the relationship between kid's height and the parents'? Another example, the relationship between gas mileage and vehicle size.

To answer these questions, we introduce linear regression. Linear regression studies the association between a dependent (response) variable and several independent variables.

Dependent and independent variables

(from wiki) Dependent and independent variables refer to values that change in relationship to each other. The dependent variables are those that are observed to change in response to the independent variables. The independent variables are those that are deliberately manipulated to invoke a change in the dependent variables.

Synonyms of dependent variable: response variable, outcome variable

Synonyms of independent variable: predictor variable, explanatory variable

Linear regression can be use to

(1) Describe or quantify realtionship between variables

(2) Make inference

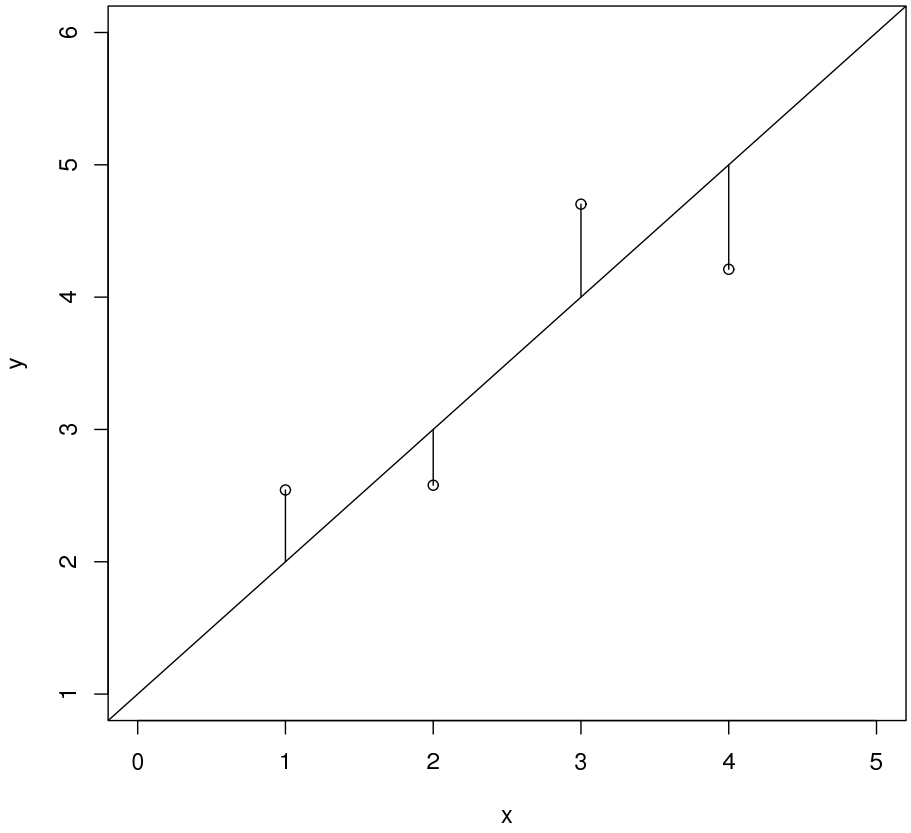
(3) Predict.

2 Least Squares

Consider an independent variable x and a dependent variable y and denote the observed data by (x_i, y_i), where $i = 1, 2, \cdots, n$. In the method of least squares, we fit a straight line to the points by choosing the slope and intercept that minimizes

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The least squares estimate (LSE) of (β_0, β_1) are chosen to minimize the sum of squared


y
x
1
2
3
4
5
6
0
1
2
3
4
5

vertical deviations, or predictor errors. Let's derive the LSE of β_0 and β_1. The pair $(\hat{\beta}_0, \hat{\beta}_1)$ that minimizes the sum of squares must satisfy the following equations:

$$\begin{aligned} \frac{\partial S}{\partial \beta_0} &= -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial S}{\partial \beta_1} &= -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \end{aligned}$$

We have

$$\begin{aligned} \sum_{i=1}^{n} y_i &= n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \ (n\bar{y} = n\hat{\beta}_0 + n\hat{\beta}_1\bar{x}) \\ \sum_{i=1}^{n} x_i y_i &= \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \end{aligned}$$

The first equation implies that

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Working on the two equations leads to

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{n\sum_{i=1}^{n} x_i y_i - n^2\bar{x}\bar{y}}{n\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The main results of LSE

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} \end{aligned}$$

Example. Consider the example of the overall ratings of 30 companies. The data can be found from https://stat.ethz.ch/R-manual/R-patched/library/datasets/html/attitude.html

The dependent variable (y) is overall rating and the independent variable (x) is score of handling of employee complaints. Using the data, we found

$$n = 30, \bar{x} = 66.6, \bar{y} = 64.63, \sum(x_i - \bar{x})(y_i - \bar{y}) = 3879.6, \sum(x_i - \bar{x})^2 = 5141.2$$

Using the formula we derived,

$$\begin{aligned}\hat{\beta}_1 &= \frac{3879.6}{5141.2} = 0.7546 \\ \hat{\beta}_0 &= 64.63 - 0.7546 \times 66.6 = 14.3763\end{aligned}$$

Try the following code in R

```
require(stats); require(graphics) #to access the data set
#the description of the data can be found from
#https://stat.ethz.ch/R-manual/R-patched/library/datasets/html/attitude.html

#sample size
n=dim(attitude)[1]

# check the relationship between complaints and rating. Do them look linear?
plot(rating~complaints, data=attitude)

#Use the least squares  we derived in class to find beta1.hat and beta0.hat
x=attitude[,2] #the x-variable
y=attitude[,1] #the y-variable

beta1.hat=sum( (x-mean(x)) * (y-mean(y)) ) /
sum( (x-mean(x))^2 )
beta1.hat

beta0.hat= mean(y) - beta1.hat * mean(x)
beta0.hat

#You can also find the LSE by fitting a lienar model using the R function "lm"
lm(rating~complaints, data=attitude)
```

The method of least squares provides an approximate relationship between the independent and dependent variables. But how do you know whether or not you can use linear regression to describe the relationship between variables?

- Is the fitted model reasonable and accurate enough to use? How much do we believe the fitted model?
- How can we test the hypothesis that $\beta_1 = 0$ (there is no linear relationship between x and y)? When $\beta_1 \neq 0$, we say there is regression effect.

3 Simple Linear Regression (regression with one predictor variable)

The goal of this section is to study how to fit a straight line to data.

3.1 Statistical Model

In the method of least squares, we didn't explicitly model noise. Consequently, we have not addressed the reliability of the slope and intercept in the presence of noise.

The standard, which is also the simplest statistical model is

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

Here ϵ_i are independent with mean zero and variance σ^2, i.e., $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$. In this course, we also assume that the errors are normally distributed, i.e., $\epsilon \overset{iid}{\sim} N(0, \sigma^2)$.

The above model contains two components: systematic and random components

Random components:

1. Normal error
2. constant variance
3. independent error

3.1.1 Interpretation of regression parameters

β_0 and β_1 are model parameters or regression coefficients. For the intercept, we have $\beta_0 = E(y|x = 0)$

For the slope, we have $\beta_1 = E(x|x = x' + 1) - E(y|x = x')$. This says that β_1 is the expected change in y that is associated with 1-unit increase in X. A graphical interpretation of the the regression parameters:

3.1.2 Estimation of Coefficients

Recall that the sum of squared difference between the observed y_i and its fitted value under the simple linear regression is

$$S = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=n}^{n} \epsilon_i^2$$

We showed that the LSE of the parameters are:

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{aligned}$$

Theorem 1 Under the assumptions of linear regression model, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1, respectively.

Proof: From the assumptions,

$$\begin{aligned} E(y_i) &= \beta_0 + \beta_1 x_i \\ E(\bar{y}) &= \frac{1}{n}\sum_{i=1}^{n} E[y_i] = \frac{1}{n}\sum(\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1 \bar{x} \end{aligned}$$

Thus,

$$\begin{aligned} E(\hat{\beta}_1) &= \frac{\sum(x_i - \bar{x})E[y_i - \bar{y}]}{\sum(x_i - \bar{x})^2} \\ &= \frac{\sum(x_i - \bar{x})[\beta_0 + \beta_1 x_i - \beta_0 - \beta_1 \bar{x}]}{\sum(x_i - \bar{x})^2} \\ &= \frac{\beta_1 \sum(x_i - \bar{x})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \\ &= \beta_1 \end{aligned}$$

$$\begin{aligned} E(\hat{\beta}_0) &= E(\bar{y} - \hat{\beta}_1 \bar{x}) \\ &= (\beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x}) \\ &= \beta_0 \end{aligned}$$

Theorem 2 Under the assumption of linear regression,