Simple Linear Regression

Legesse Kassa Debusho, UNISA, South Africa and Ziv Shkedy,
Hasselt University, Belgium

July 3, 2017

Introduction

- Regression analysis is a body of knowledge dealing with the formulation of mathematical models that depict relationships among variables.
- It is used 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, $X_1, X_2, \ldots, X_p$.
- When $p = 1$, it is called simple regression but when $p \geq 1$ it is called multiple regression (this will be discussed in Chapter 3).
- In simple regression a single predictor or independent variable is used for predicting the variable of interest, i.e. dependent variable.
- In this course we will consider parametric regression models and we will assume a relationship that is linear in the parameters.

Introduction

- Therefore, a simple linear regression (SLR) attempts to model the relationship between a single explanatory or predictor variable and a response variable by fitting an equation to observed data that is linear in the parameters.
- The parameters are called regression parameters or regression coefficients.
- Both the response and independent variables in SLR must be a continuous variables.
- Objectives of regression analyses can be
  - to predict a continuous dependent variable from one or a number of independent variables.
  - to check whether there is a relationship between a dependent or response variable and one or more than one independent variables.
  - to describe the data structure.

The simple linear regression model

- The model expresses the value of a response variable as a linear function of a predictor variable and an error term. The simple linear regression model can be stated as follows:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, ..., n \tag{1}$$

  where
  - y_i represents the ith value of the response variable Y,
  - x_i represents the ith value of the predictor variable X,
  - β_0 and β_1 are constants called regression coefficients or parameters.
  - ε_i is a random disturbance or error.

The simple linear regression model

- The coefficient β_1, called the slope, may be interpreted as the change in Y for a unit change in X.
- The coefficient β_0, called the constant coefficient or intercept, is the predicted value of Y when $X = 0$.
- Note that the random term ε_i does not contain any systematic information for determining y_i that is not already captured by x_i.

Matrix notation of simple linear regression model

- Model (1) implies that

$$\left\{\begin{array}{lcl} y_1 & = & \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ \vdots & & \\ y_i & = & \beta_0 + \beta_1 x_i + \varepsilon_i \\ \vdots & & \\ y_n & = & \beta_0 + \beta_1 x_n + \varepsilon_n \end{array}\right. \tag{2}$$

Matrix notation of simple linear regression model

- This can be written in matrix form as follows

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- or

$$\underset{(n \times 1)}{\mathbf{y}} = \underset{(n \times 2)}{\mathbf{X}} \ \underset{(2 \times 1)}{\boldsymbol{\beta}} + \underset{(n \times 1)}{\boldsymbol{\epsilon}} \tag{3}$$

Assumptions

- The errors have an expected value of zero, i.e. $E(\varepsilon_i)$ or $E(\boldsymbol{\epsilon}) = \mathbf{0}$. This means that on average the errors balance out.
- The independent variable is non-random. In an experiment, the values of the independent variable would be fixed by the experimenter and repeated samples could be drawn with the independent variable fixed at the same values in each sample.
- The disturbances are homoscedastic. This means that the variance of the disturbance is the same for each observation, i.e. $Var(\varepsilon_i) = \sigma^2$, for all i or $Var(\boldsymbol{\epsilon}) = \sigma^2\,\mathbf{I}_n$.
- The disturbances are not autocorrelated. This means disturbances associated with different observations are uncorrelated.
- For statistical inferences (confidence intervals and test of hypothesis), the distribution of ε_i is normal.

Assumptions

- In short, the above assumptions related to ε_i can be stated as $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ or $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, s^2\,\mathbf{I}_n)$.
- Note the following
  - From model (1), note that the observed value of Y in the ith observation is the sum of two components, namely the constant term $\beta_0 + \beta_1\, x_i$ and the random error term ε_i. Hence y_i is a random variable.
  - From the above assumptions it follow that
    - (i) $E(y_i) = E(\beta_0 + \beta_1\, x_i + \varepsilon_i) = \beta_0 + \beta_1\, x_i$.
    - (ii) $Var(y_i) = Var(\beta_0 + \beta_1\, x_i + \varepsilon_i) = Var(\varepsilon_i) = \sigma^2$.
    - (iii) $Cov(Y_i, Y_j) = 0$, i.e., the response y_i and y_j are uncorrelated.
    - (iv) Using matrix notation, $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\,\boldsymbol{\beta}, \sigma^2\,\mathbf{I}_n)$.

Least squares estimation

- The least squares criterion states that the estimator $\widehat{\beta}$ of β must be found in such a way that $\boldsymbol{\epsilon}'\boldsymbol{\epsilon}$, the sum of squares of the residuals, is a minimum. That is, to find the least squares estimates we have to minimise

$$\begin{aligned} Q(\boldsymbol{\beta}) &= \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = \sum_{i=1}^{n} \varepsilon_i^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}'\mathbf{y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

- In order to minimise $Q(\boldsymbol{\beta})$ we set $\left.\frac{dQ(\boldsymbol{\beta})}{d\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} = \mathbf{0}$. We find that

$$\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

and these are called the normal equations.

Least squares estimation

- Provided that $\mathbf{X}$ has full column rank or $\mathbf{X}'\mathbf{X}$ is non-singular, the least square estimates are given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}.$$

- Hence the least squares regression line is given by $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$.
- For simple linear regression the least squares estimates can also be given by

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

and

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

Example: Galapagos Islands Data

- A data frame with 30 Galapagos islands and 7 variables is available in R under the library named faraway:
  - Species: the number of plant species found on the island
  - Endemics: the number of endemic species
  - Area: the area of the island (km^2)
  - Elevation: the highest elevation of the island (m)
  - Nearest: the distance from the nearest island (km)
  - Scruz: the distance from Santa Cruz island (km)
  - Adjacent: the area of the adjacent island (square km)
- The data were presented by Johnson and Raven (1973) and also appear in Weisberg (1985). The original dataset contained several missing values which have been filled for convenience.

Example: Galapagos Islands Data

- The relationship between the number of plant species and several geographic variables is of interest.
- In this chapter we are interested in the relationship between the number of plant species and the area of the adjacent island (square km).
- As a preliminary analysis a scatter plot of the two variables are presented in Figure 2.1 below.

Example: Galapagos Islands Data

- Figure 2.1 shows that the relationship between the number of species found on the island and the highest elevation of the island (m) can be described by a linear model.

Simple Linear Regression in R

- Fitting a simple linear regression model in R is done using the lm() function. A general call of the function has the form of:

  `lm(response~predictor)`

  which corresponds to the model

  $$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \tag{4}$$

Example 1: Galapagos Islands Data

- Read the data into R as shown below.

```
library(faraway)
data(gala)
attach(gala)
```

- The library(faraway) makes the data used in this book available while data(gala) calls up this particular dataset.

Example 1: Galapagos Islands Data

- Then the model fitted using the lm() function, a linear regression model in which `Species` is the response and `Elevation` is a predictor can be fitted by

  ```
  SLR.fit.1 <- lm(Species~Elevation)
  summary(SLR.fit.1)
  ```

- The result of fitted model is stored in an object called `SLR.fit.1`".

Example 1: Galapagos Islands Data

- The function summary() can be used to see the estimated parameters.

```
Call:
lm(formula = Species ~ Elevation)
Residuals:
     Min        1Q    Median        3Q       Max
-218.319   -30.721   -14.690     4.634   259.180
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.33511   19.20529   0.590     0.56
Elevation    0.20079    0.03465   5.795 3.18e-06 ***
---
Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’
Residual standard error: 78.66 on 28 degrees of freedom
Multiple R-squared:  0.5454,  Adjusted R-squared:  0.5291
F-statistic: 33.59 on 1 and 28 DF,  p-value: 3.177e-06
```

Example 1: Galapagos Islands Data

- The estimates for the model intercept and the slope are 11.33511 and 0.20079. Therefore, the fitted or the least squares regression line is

$$\hat{y} = 11.335 + 0.201\,x$$

- The slope show that there is a positive relationship (consistent with the scatter plot), in other words, as the elevation of the island increase there will be an increase in the number of species of tortoise found on the island.

Example 1: Galapagos Islands Data

- The multiple R-squared or $R^2 = 0.5454$

```
Residual standard error: 78.66 on 28 degrees of freedom
Multiple R-squared:  0.5454,  Adjusted R-squared:  0.5291
F-statistic: 33.59 on 1 and 28 DF,  p-value: 3.177e-06
```

- The coefficient of determination: may be interpreted as the proportion of the total variability in the number of species of tortoise found on the island (i.e. the response variable) that is accounted for the elevation of the island (i.e. the predictor variable).
-

Example 1: Galapagos Islands Data

- Therefore, 54.54% of the total variability in the number of species of tortoise found on the island is accounted for by the elevation of the island.
- For simple linear regression, the correlation coefficient can be calculated using the multiple R-squared value from the above output as $r = \sqrt{0.5454} \approx 0.74$.
- This shows that there is a strong positive linear relationship between the two variables.

Example 2: The cars Data

- The data give the speed of cars and the distances taken to stop. Note that the data were recorded in the 1920s.
- Use

  ```
  >help(cars)
  ```

  to get more information about the data.

Example 2: The `cars` Data

- We would like to model the stopping distance as a function of the car's speed.
- We formulate the following model:

$$\text{dist}_i = \beta_0 + \beta_1 \text{speed}_i + \varepsilon_i.$$

- In R the model is fitted using

```
> fit.lm<-lm(cars$dist~cars$speed)
```

Example 2: The cars Data

- The function `summary()` can be used to see the estimated parameters.

```
> summary(fit.lm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
cars$speed    3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Properties of the least square estimator of β

- Recall that for the general linear model $\mathbf{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$ we assume that
- $E(\varepsilon) = \mathbf{0}, \quad Var(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I_n}$ or $\boldsymbol{\epsilon} \sim (\mathbf{0}, \sigma^2\mathbf{I}_n)$.
- Furthermore, do to these assumptions we have $\mathbf{y} \sim (\mathbf{X}\beta, \sigma^2\mathbf{I}_n)$.
- Thus, for $\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}$

  (i) $E(\widehat{\beta}) = E((\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta = \beta$ and
  (ii) $Var(\widehat{\beta}) = Var((\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{I_n}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}$.

Properties of the least square estimator of β

- For simple linear regression

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} & -\frac{\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ -\frac{\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} & \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{pmatrix}.$$

- Thus $Var(\widehat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ is a variance-covariance matrix with

$$Var(\widehat{\beta}_0) = \sigma^2 \sum_{i=1}^{n} x_i^2 / n \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right],$$
$$Var(\widehat{\beta}_1) = \sigma^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2, \text{ and}$$
$$Cov(\widehat{\beta}_0, \widehat{\beta}_1) = -\sigma^2 \sum_{i=1}^{n} x_i / n \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Estimating σ^2

- The variances of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ depend on the unknown parameter σ^2.
- In the output below, $\hat{\sigma}^2 = 78.66$.

```
. .
Residual standard error: 78.66 on 28 degrees of freedom
Multiple R-squared: 0.5454, Adjusted R-squared: 0.5291 F-sta
33.59 on 1 and 28 DF, p-value: 3.177e-06
```