Multiple Regression

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

April 5, 2017

Table of contents

Introduction

- Regression analysis is a body of knowledge dealing with the formulation of mathematical models that depict relationships among variables.
  - These modeled relationships usually used for the purpose of prediction and statistical inference on the regression curve.
- Multiple regression is a method used to model the relationship between a dependent variable and two or more independent variables.
- In this course we will consider parametric regression models and we will assume a relationship that is linear in the parameters.

Introduction

- Therefore, multiple linear regression attempts to model the relationship between two or more explanatory or predictor variables 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.
- Multiple linear regression is mainly used
  - to check whether there is a relationship between a dependent or response variable and more than one independent variables, also called predictors or regressors.
  - to predict a continuous dependent variable from a number of independent variables.

Introduction

- The independent variables used in multiple linear regression can be either continuous or dichotomous.
- Independent categorical variables with more than two levels can also be used in multiple linear regression analyses, but each of the levels first must be converted into variables that have only two levels. This is called dummy coding.
- Although regression analysis is usually used with naturally-occurring variables rather than experimentally manipulated variables, it can also be used with experimentally manipulated variables.

Introduction

- One point to keep in mind with regression analysis is that causal relationships among the variables cannot be determined. Given that $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ with $\mathbf{x}$ a vector, and the components are called the predictors or regressors. Then the terminology is in such way that $\mathbf{x}$ "predicts" y, we cannot say that $\mathbf{x}$ "causes" y.
- The general premise of multiple regression is similar to that of simple linear regression. However, in multiple regression, we are interested in examining more than one predictor of our criterion variable.
  - Often this is done to determine whether the inclusion of additional predictor variables leads to improved prediction of the outcome variable.

The Multiple linear regression model

- The model expresses the value of a response variable as a linear function of two or more predictor variables and an error term. The multiple linear regression model has the following general form

$$y_i = \mu(x_{i1}, \ldots, x_{i,p-1}) + \varepsilon_i \quad i = 1, ..., n$$

  - $x_{i1}, \ldots, x_{i,p-1}$ are the particular (deterministic) values of the $(p-1)$ regressors
  - y_i is the corresponding continuous response random variable
  - $\mu(\cdot)$ the unknown mean response function
  - $\varepsilon_1, ..., \varepsilon_n$ independent (unobservable) random error terms with $E(\varepsilon_i) = 0$

The Multiple linear regression model

- To estimate $\mu(\cdot)$ we can use methods from nonparametric regression estimation theory or we can be more specific and assume a parametric functional form.
- We shall consider the linear model

$$\mu(x_1, \ldots, x_{p-1}) = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_j$$

- The parametric assumption simplifies the problem, since we only have to estimate a finite number of parameters, the **regression parameters** $\beta_0, \ldots, \beta_{p-1}$.

The Multiple linear regression model

- In other words, we define the general linear regression model as

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i,p-1} + \varepsilon_i.$$

- Thus, we have the equations

$$\begin{cases} y_1 &= \beta_0 + \beta_1 x_{11} + \ldots + \beta_{p-1} x_{1,p-1} + \varepsilon_1 \\ \vdots & \\ y_i &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i,p-1} + \varepsilon_i \\ \vdots & \\ y_n &= \beta_0 + \beta_1 x_{n1} + \ldots + \beta_{p-1} x_{n,p-1} + \varepsilon_n \end{cases} \quad (1)$$

The Multiple linear regression model

- This can be written in matrix terms as follows

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1,p-1} \\ 1 & x_{21} & \dots & x_{2,p-1} \\ \vdots & & & \\ 1 & x_{n1} & \dots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- or

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

The Multiple linear regression model

where

- **X** is the **design matrix,** also called the regression matrix
- **y** is a vector of responses
- β is a vector of parameters
- ϵ is a vector of independent random variables with expectation
- $E(\epsilon) = 0$ and variance covariance matrix $\mathcal{D}(\epsilon) = \sigma^2 \mathbf{I_n}$.
- The model

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_{p-1} x_{p-1} + \varepsilon$$

  is called the **general linear regression model**.
- This model describes a hyperplane in the (p -1) dimensional space of the explanatory or predictor variables x_j.

Least squares estimation

- Recall from Chapter 2 that the least squares estimator of $\boldsymbol{\beta}$ for the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is given by

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

  of provided that $\mathbf{X}'\mathbf{X}$ is non-singular.
- Similar to the simple linear regression model, a multiple linear regression model can also be fitted using the function lm() in R.

Example: Galapagos Islands Data

- Recall the Galapagos Islands Data example. Consider now the relationship between the number of plant species with the four geographic variables and the number of endemic species.
- The linear model then fitted using the following R code:

```
MLR.fit.1 <- lm(Species ~ Area + Elevation + Nearest
                          + Scruz + Adjacent, data=gala)
summary(MLR.fit.1)
```

Example: Galapagos Islands Data

```
Call:
lm(formula = Species ~ Area + Elevation + Nearest +
             Scruz + Adjacent, data = gala)
Residuals:
     Min       1Q   Median       3Q      Max
-111.679  -34.898   -7.862   33.460  182.584

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.068221  19.154198   0.369 0.715351
Area        -0.023938   0.022422  -1.068 0.296318
Elevation    0.319465   0.053663   5.953 3.82e-06 ***
```

Example: Galapagos Islands Data

```
Nearest      0.009144   1.054136   0.009 0.993151
Scruz       -0.240524   0.215402  -1.117 0.275208
Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-squared: 0.7658, Adjusted R-squared: 0.7171
F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

Example: Galapagos Islands Data

- The fitted or the least squares regression line is

  $\hat{y} = 7.068 - 0.024\, Area + 0.319\, Elevation + 0.009\, Nearest - 0.241\, Scruz +$

- $R^2 = 0.76.58$: 76.58% of the total variability in the number of species of tortoise found on the island is accounted for by the fitted regression line (or by six predictor variables).

The Multiple linear regression model

- The general linear regression model is not restricted to linear response surfaces. That is, any linear regression model that is linear in the parameters is a linear regression model regardless of the surface that it generates.
- For example, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ produces a flat surface, as shown in Figure 1, while the surface that the model $y = \beta_0 + \beta_1 x_1^2 + \beta_2 x_2^2 + \epsilon$ produces is not flat shown in Figure 2. See the R codes in the file "R codes for Multiple Regression I Chapter".

The Multiple linear regression model

The Multiple linear regression model

The Gauss-Markov Theorem

- **Estimable function**: A linear combination of the parameters $\psi = \mathbf{c}'\,\beta$ is estimable if and only if there exists a linear combination $\mathbf{a}'\,\mathbf{y}$ such that

$$E(\mathbf{a}'\,\mathbf{y}) = \mathbf{c}'\,\beta \qquad \forall \beta.$$

- **The Gauss-Markov Theorem**: Suppose $\mathbf{y} = \boldsymbol{X}'\beta + \epsilon$, where $\mathrm{E}(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2\mathbf{I}_n$. Let $\psi = \mathbf{c}'\,\beta$ be an estimable function, then in the class of all unbiased linear estimates of ψ, $\widehat{\psi} = \mathbf{c}'\widehat{\beta}$ has the minimum variance and is unique.

Proof

- Suppose $\mathbf{a}'\mathbf{y}$ is some unbiased estimate of $\mathbf{c}'\beta$ so that $E(\mathbf{a}'\mathbf{y}) = \mathbf{c}'\beta \quad \forall\ \beta$, since $E(\mathbf{y}) = \mathbf{X}\beta$ it follows that $\mathbf{a}'\mathbf{X}\beta = \mathbf{c}'\beta \quad \forall \beta$ which means that $\mathbf{a}'\mathbf{X} = \mathbf{c}'$.
- This implies that $\mathbf{c}$ must be in the range space of $\mathbf{X}'$ which in turn implies that $\mathbf{c}$ is also in the range space of $\mathbf{X}'\mathbf{X}$ which means there exists $\boldsymbol{\lambda}$ such that

$$\begin{aligned} \mathbf{c} &= \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\lambda} \\ \Rightarrow \mathbf{c}'\widehat{\beta} &= \boldsymbol{\lambda}'\boldsymbol{X}'\boldsymbol{X}\widehat{\beta} = \boldsymbol{\lambda}'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\mathbf{y}) = \boldsymbol{\lambda}'\boldsymbol{X}'\mathbf{y} \end{aligned}$$

- Now, let us compute the variance of $\mathbf{a}'\mathbf{y}$.

$$\begin{aligned} Var(\mathbf{a}'\mathbf{y}) &= Var(\mathbf{a}'\mathbf{y} + \mathbf{c}'\widehat{\beta} - \mathbf{c}'\widehat{\beta}) \\ &= Var(\mathbf{a}'\mathbf{y} - \boldsymbol{\lambda}'\mathbf{X}'\mathbf{y} + \mathbf{c}'\widehat{\beta}) \\ &= Var(\mathbf{a}'\mathbf{y} - \boldsymbol{\lambda}'\mathbf{X}'\mathbf{y}) + Var(\mathbf{c}'\widehat{\beta}) + 2\,Cov(\mathbf{a}'\mathbf{y} - \boldsymbol{\lambda}'\boldsymbol{X}'\mathbf{y}\boldsymbol{\lambda}'\boldsymbol{X}'\mathbf{y} \end{aligned}$$

Proof

- but

$$\begin{aligned} Cov(\mathbf{a}'\mathbf{y} - \lambda'\mathbf{X}'\mathbf{y}, \lambda'\mathbf{X}'\mathbf{y}) &= Cov((\mathbf{a}' - \lambda'\boldsymbol{X}')\mathbf{y}, \lambda'\boldsymbol{X}'\mathbf{y}) \\ &= (\mathbf{a}' - \lambda'\boldsymbol{X}')Var(\mathbf{y})(\lambda'\boldsymbol{X}')' \\ &= (\mathbf{a}' - \lambda'\boldsymbol{X}')\sigma^2\mathbf{I}\boldsymbol{X}\lambda \\ &= (\mathbf{a}'\mathbf{X} - \lambda\boldsymbol{X}'\boldsymbol{X})\sigma^2\lambda \\ &= (\mathbf{c}' - \mathbf{c}')\sigma^2\lambda = \mathbf{0} \end{aligned}$$

Proof

- So $Var(\mathbf{a}'\mathbf{y}) = Var(\mathbf{a}'\mathbf{y} - \boldsymbol{\lambda}'\boldsymbol{X}'\mathbf{y}) + Var(\mathbf{c}'\widehat{\boldsymbol{\beta}})$.
- Since $Var(\mathbf{a}'\mathbf{y} - \boldsymbol{\lambda}'\boldsymbol{X}'\mathbf{y}) \geq 0$, $Var(\mathbf{a}'\mathbf{y}) \geq Var(\mathbf{c}'\widehat{\boldsymbol{\beta}})$.
- Thus $\mathbf{c}'\widehat{\boldsymbol{\beta}}$ has minimum variance.

Proof

- $Var(\mathbf{a}'\mathbf{y}) = Var(\mathbf{c}'\widehat{\beta})$, if $Var(\mathbf{a}'\mathbf{y} - \boldsymbol{\lambda}'\boldsymbol{X}'\mathbf{y}) = \mathbf{0}$ which would require that $\mathbf{a}' - \boldsymbol{\lambda}'\boldsymbol{X}' = \mathbf{0}$ which means that $\mathbf{a}'\mathbf{y} = \boldsymbol{\lambda}'\boldsymbol{X}'\mathbf{y} = \mathbf{c}'\widehat{\beta}$.
- Therefore, equality occurs only if $\mathbf{a}'\mathbf{y} = \mathbf{c}'\widehat{\beta}$ revealing that the estimator is unique.

Implications

- The Gauss-Markov theorem shows that the least squares estimate $\widehat{\beta}$ is a good choice, but if the errors are correlated or have unequal variance, there will be better estimators.
- Even if the errors behave but are non-normal then non-linear or biased estimates may work better in some sense. So this theorem does not tell one to use least squares all the time, it just strongly suggests it unless there is some strong reason to do otherwise.

Implications

- Situations where estimators other than ordinary least squares should be considered are
  1. When the errors are correlated or have unequal variance, generalized least squares should be used (this will be discussed in Chapter 5).
  2. When the error distribution is long-tailed, then robust estimates might be used. Robust estimates are typically not linear in y (this is a topic of Chapter 10).
  3. When the predictors are highly correlated (collinear), then biased estimators such as ridge regression might be preferable (this will be discussed in Chapter 10 if time allows).

Estimation of σ^2

- The variance-covariance matrix $Var(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ contains the unknown parameter σ^2. In this section we give an **unbiased** estimator of σ^2. We start with some preliminaries from matrix theory.

Lemma 1: Let $\mathbf{A}$ be a symmetric $(n \times n)$ matrix $(\mathbf{A}' = \mathbf{A})$, then

(i) all eigenvalues $\lambda_1, \ldots, \lambda_n$ of $\mathbf{A}$ are real and

$$|\mathbf{A}| = \prod_{i=1}^{n} \lambda_i, \qquad tr(A) = \sum_{i=1}^{n} \lambda_i$$

Estimation of σ^2

(ii) the eigenvectors $u_1, \ldots, u_n$ are orthogonal

(iii) $\mathbf{A}$ can be expressed as a diagonal matrix

$$\mathbf{A} = \mathbf{UDU}^{-1}$$

where $\mathbf{U} = (u_1, \ldots, u_n)$ and $\mathbf{D} = diagonal(\lambda_1, \ldots, \lambda_n)$.

(iv) $rank(\mathbf{A})$ is the number of nonzero eigenvalues.

(v) Let $\mathbf{A}$ be an idempotent projection $(n \times n)$ matrix $(\mathbf{A^2} = \mathbf{A})$, then $\lambda_i = 1$ or 0, $i = 1, \ldots, n$.

(vi) Let $\mathbf{A}$ be an idempotent projection matrix, then rank $(\mathbf{A}) = tr(\mathbf{A})$.

Estimation of σ^2

- **Proof of (vi)**

$$tr(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i = \#\{i : \lambda_i = 1\} = \text{rank}(\mathbf{A})$$

- Let $\mathbf{X}$ denote the design matrix and assume that it has full rank. Define the $(n \times n)$ hat matrix, $\mathbf{H}$ given by

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

- Note that $\widehat{\theta} = \mathbf{X}\widehat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$. In the next lemma we show that $\mathbf{H}$ is an **orthogonal projection** (which corresponds with the fact that $\widehat{\theta}$ is chosen such that $\mathbf{y} - \widehat{\theta} \perp \mathbf{R}(\mathbf{X})$.

Lemma 2

Assume that the design matrix $\mathbf{X}$ has full rank p.

(i) $\mathbf{H}$ is symmetric and idempotent. The same holds for $\mathbf{I}_n - \mathbf{H}$.

(ii) $\text{rank}(\mathbf{I}_n - \mathbf{H}) = tr(\mathbf{I}_n - \mathbf{H}) = n - p$.

(iii) $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \mathbf{0}$.

Proof

(i) $\mathbf{H}' = \mathbf{H},\ \mathbf{I}_n - \mathbf{H})' = \mathbf{I}_n - \mathbf{H}$

(ii) $\mathbf{H}^2 = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{I}_p\mathbf{X}' = \mathbf{H}$

(iii) $(\mathbf{I}_n - \mathbf{H})^2 = \mathbf{I}_n^2 - 2\mathbf{H} + \mathbf{H}^2 = \mathbf{I}_n - \mathbf{H}$

(iv) $tr(\mathbf{H}) = tr(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = tr(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}) = tr(\mathbf{I}_p) = \mathbf{p}$

(v) $rank(\mathbf{I}_n - \mathbf{H}) = tr(\mathbf{I}_n - \mathbf{H}) = tr(\mathbf{I}_n) - tr(\mathbf{H})$

(vi) $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \mathbf{X} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{X} - \mathbf{X} = \mathbf{0}$

A quadratic form

Let $\mathbf{A}$ be a symmetric $(n \times n)$ matrix of constants and let $\mathbf{y}$ be an $(n \times 1)$ random vector. A quadratic form is defined as $\mathbf{y}'\mathbf{A}\mathbf{y} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}y_i y_j$, note that $a_{ij} = a_{ji}$.

Lemma 3 If $E(\mathbf{y}) = \boldsymbol{\theta}$ and $\sigma^2(\mathbf{y}) = \boldsymbol{\Sigma}$, then

$$E(\mathbf{y}'\mathbf{A}\mathbf{y}) = tr(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\theta}'\mathbf{A}\boldsymbol{\Sigma}$$

A quadratic form

Proof

$$
\begin{aligned}
E(\mathbf{y}'\mathbf{A}\mathbf{y}) &= E[tr(\mathbf{y}'\mathbf{A}\mathbf{y})] \\
&= E[tr(\mathbf{A}\mathbf{y}\mathbf{y}')] \\
&= tr[E(\mathbf{A}\mathbf{y}\mathbf{y}')] \\
&= tr[\mathbf{A}E(\mathbf{y}\mathbf{y}')] \\
&= tr[\mathbf{A}(\boldsymbol{\Sigma} + \boldsymbol{\theta}'\boldsymbol{\theta})], \quad \text{because } \boldsymbol{\Sigma} = E(\mathbf{y}\mathbf{y})' - \boldsymbol{\theta}\boldsymbol{\theta}' \\
&= tr(\mathbf{A}\boldsymbol{\Sigma}) + tr(\mathbf{A}\boldsymbol{\theta}\boldsymbol{\theta}') \\
&= tr(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\theta}'\mathbf{A}\boldsymbol{\theta}'
\end{aligned}
$$

Theorem: Unbiased estimator of σ^2

Assume (i) $\mathbf{X}$ has full rank, (ii) $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and (iii) $Var(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}_n$. An unbiased estimator of σ^2 is given by

$$\widehat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})}{n - p}$$

Proof

$$\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$$

Therefore

$$\begin{aligned}(n-p)MSE &= \mathbf{y}'(\mathbf{I}_n - \mathbf{H})'(\mathbf{I}_n - \mathbf{H})\mathbf{y} \\ &= \mathbf{y}'(\mathbf{I}_n - \mathbf{H})^2\mathbf{y} \\ &= \mathbf{y}'(\mathbf{I}_n - \mathbf{H})\mathbf{y}\end{aligned}$$

Proof

and

$$\begin{aligned} E((n-p)MSE) &= E(\mathbf{y}'(\mathbf{I}_n - \mathbf{H})\mathbf{y}) \\ &= tr[\mathbf{I}_n - \mathbf{H})\sigma^2(\mathbf{y})] + E(\mathbf{y})'\mathbf{I}_n - \mathbf{H})E(\mathbf{y}) \\ &= \sigma^2 tr\mathbf{I}_n - \mathbf{H}) + \beta'\mathbf{X}'(\mathbf{I}_n - \mathbf{H})\mathbf{X}\beta \\ &= \sigma^2(n-p) \end{aligned}$$

because $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \mathbf{0}$ by Lemma 2(iii)