Lesson 3 Multiple Linear Regression

Multiple Linear Regression

- In practice, we are often interested in the relationships between a set of explanatory or independent variables (x_1,x_2,…,x_k) and a dependent variable (y). Like simple linear regression, multiple linear regression allow us to quantify the relationship between our response variable and our explanatory variables as well as providing a tool for predicting the response of a new observation for a given set of values for x_1,x_2,…, and x_k.
- For example, we may be interested in understanding the relationship between blood pressure and other variables(such as age, weight, physical activity, and pulse rate).

Multiple Linear Regression

- The equation for the simple linear regression line is given by

$$y=\beta_0+\beta_1 x_1 +\beta_2 x_2 +\ldots + \beta_k x_k$$

where

- y is the response or dependent variable
- $x_1, x_2, \ldots, x_k$ are the explanatory or independent variables
- β_0 is the intercept (the value of y when all independent variables $x_1, x_2, \ldots, x_k$ are set to 0)
- β_i is the slope (the expected change in y for each one unit change in x_i, after adjusting for $x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k$)

Regression Equation

- The equation for the least-squares regression line for multiple linear regression is an extension of the equation for simple linear regression:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

  - where $\hat{y}$ (read "y hat") is the expected or predicted value of y for a given values of $x_1, x_2, \ldots, x_k$
  - $\hat{\beta}_0$ is the least-square estimate of β_0 (the intercept)
  - $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ are the least-squares estimates of $\beta_1, \beta_2, \ldots, \beta_k$, respectively

Estimation of the parameters by least squares

The method of least squares is applied in multiple linear regression to estimate the parameters $\beta_1,\beta_2,\ldots,\beta_k$, in the same way it was applied in simple linear regression. In least-squares regression, the estimates are selected as to minimize the sum of the squared differences between the observed response and the estimated response across all data points. That is, the estimates are selected in such a way that the sum of $(y_i - \hat{y}_i)^2$ is minimized:

In multiple regression, the least squares estimates of $\beta_1,\beta_2,\ldots,\beta_k$ do not have simple closed form equations. We will use R to estimates the parameters.

MLR Using R

Size (sqft)	#Rooms	House Price
1850	4	$229,500
2190	5	$273,300
2100	4	$247,000
1930	3	$195,100
2300	4	$261,000
1710	2	$179,700
1550	2	$168,500
1920	4	$234,400
1840	2	$168,800
1720	2	$180,400
1660	2	$156,200
2405	5	$288,350
1525	3	$186,750
2030	2	$202,100
2240	4	$256,800

MLR Using R

R code:

```
m<-lm(data$HousePrice~data$Size+data$Rooms)
summary(m)
```

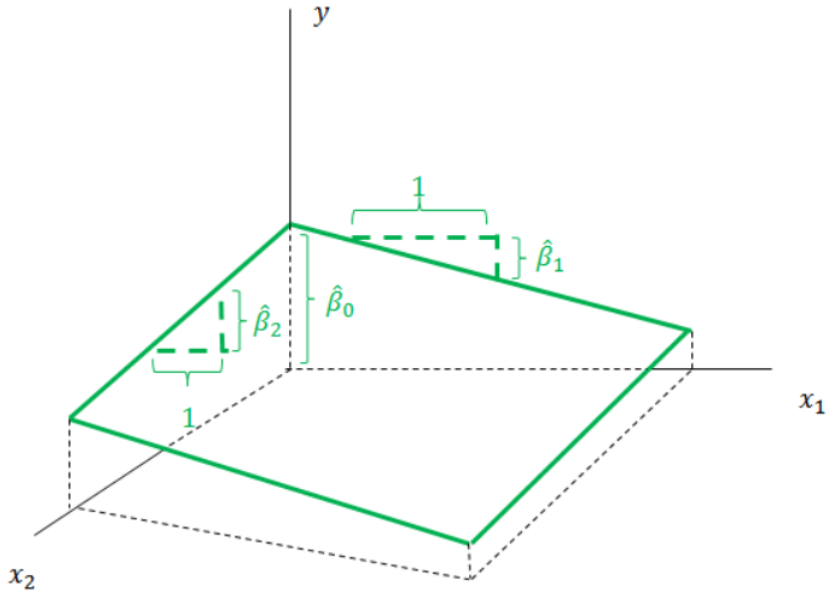
```
> summary(m)

Call:
lm(formula = data$HousePrice ~ data$Size + data$Rooms)

Residuals:
     Min       1Q   Median       3Q      Max
-15283.2   -894.8   1850.6   4798.8   8769.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12923.73   17568.20   0.736 0.476088
data$Size      65.61      12.03   5.454 0.000147 ***
data$Rooms  23613.14    2878.30   8.204  2.9e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8186 on 12 degrees of freedom
Multiple R-squared:  0.9688,    Adjusted R-squared:  0.9637
F-statistic: 186.6 on 2 and 12 DF,  p-value: 9.144e-10
```

Regression Equation:

House Price ($) = 65.6 1* Size (sqft) + 23613.14 * Rooms + 12923.73

Regression in multiple dimensional space

- In simple linear regression, the regression line corresponds to a line. In multiple linear regression, the regression line corresponds to a plane or a hyperplane (a k-dimensional plane in a k+1 dimensional space).



- In a three dimensional space, the beta estimates define the surface of the plane. $\hat{\beta}_0$ is the expected value of y when x_1 and x_2 are 0.
- $\hat{\beta}_1$ is the slope of the surface projected onto the x_1,y plane. $\hat{\beta}_2$ is the slope of the surface projected onto the x_2,y plane.

Interpretation

- Generally, the interpretation of a regression equation focuses on the slope parameters for the independent variables. The estimate of the slope parameter $\hat{\beta}_i$ in a multiple regression model gives the expected or average change in the dependent (response) variable (y) for a one unit increase in the independent variable (x_i) after controlling for the other independent variables. Or, the estimate can be thought of as the average increase in y for per unit increase in x_i holding all other independent variables constant.
- As in simple linear regression, the intercept's interpretation $\hat{\beta}_0$ is not meaningful unless values of the dependent variables near 0 are possible. The intercept in multiple linear regression can be interpreted as the expected or average value of the response when all independent variables are equal to 0.

Assessing the Fit of the Regression Line

- $\sum_{i=1}^{n}(y_i - \bar{y})^2$ (called the total sum of squares or Total SS) represents the sum of squares of the deviations of the individual sample points from the sample mean
- $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ (called the residual sum of squares or Res SS) represents the sum of squares of the residual components
- $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ (called the regression sum of squares or Reg SS) represents the sum of squares of the regression components
- That is, Total SS = Res SS + Reg SS.
- One of the measures that we use to assess the fit of the data is the coefficient of variation R^2. In the multiple linear regression setting, the coefficient of variation is often referred to as the multiple R-squared.
  - $R^2 = \frac{Reg\ SS}{Total\ SS}$

Compute Multiple R-squared By Hand

Note: fitted(m) extracts fitted values from the model

```
totalss <-sum((housedata$HousePrice -mean(housedata$HousePrice))^2)
regss <-sum((fitted(m) -mean(housedata$HousePrice))^2)
residss <-sum((housedata$HousePrice-fitted(m))^2)
rsquare <- regss/totalss
rsquare
```

Output:

```
rsquare
[1] 0.9688456
```

Using the equation for Predictions

- Same idea as simple linear regression.
- The equation of the least-squares regression line can be used to predict the expected value of the response variable for new values of the explanatory variables $x_1,x_2,\ldots,x_k$. This is done by substituting the new values of $x_1,x_2,\ldots,x_k$ into the regression equation and calculating the associated value of $\hat{y}$. The predicted value $\hat{y}$ for given values of $x_1,x_2,\ldots,x_k$ can be interpreted as the average value of the response for the given values of $x_1,x_2,\ldots,x_k$.

Advantages of Regression Modeling

- Easy to understand
- Provide simple equations
- Goodness of fit is measured by correlation coefficient (or R^2)
- Match and beat the predictive power of other modeling techniques
- Can include any number of variables
- Regression modeling tools are pervasive such as Excel

Disadvantages of Regression Models

- Can not cover for poor data quality issues
- Does not automatically take care of non-linearity
- Usually works only with numeric data