University of Guelph Numeracy Project

Multiple Regression

TABLE OF CONTENTS

Multiple Regression 1
What is MULTIPLE REGRESSION? 1
Population Regression Equation 1
Population (Multiple) Regression Equation 1
Multiple Linear Regression Model 1
Multiple Linear Regression Model 1
Estimating Multiple Regression Parameters 2
Inference in Multiple Regression 2
Residuals 2
Estimating Variance (σ^2) 2
Confidence Intervals 3
Significance Tests 3
ANOVA F Test 3
ANOVA F Test 3
Glossary 5
References 6

Multiple Regression

What is MULTIPLE REGRESSION?

- Multiple regression contrasts with simple linear regression, in that while simple linear regression considers a single explanatory variable x to account for a response variable y, multiple regression considers two or more explanatory variables.

Population Regression Equation

Population (Multiple) Regression Equation

- Linear Regression Equation:

  $\mu_y = \beta_0 + \beta_1 x$

  Here, it is assumed that the explanatory variable x properly accounts for the response variable y.

- Population (Multiple) Regression Equation:

  $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$

  Here, it is assumed that each of the explanatory variables x_n contributes to a proper account for the response variable y. All explanatory variables are denoted by x_n, where n= 1, 2, 3,...,k.

Multiple Linear Regression Model

Multiple Linear Regression Model

- Refer back to the population regression equation:

  $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$

  The equation above operates under the assumption that exactly x explanatory variables can account for y, without considering random error. The equation below represents a modification to the population regression equation that accounts for random error:

$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + E_i$

In the above equation, E_i is a term used to represent the random error.

Estimating Multiple Regression Parameters

- In multiple regression, as in linear regression, least squares is used to find predictors of the intercept and slope of the regression line.

- Referring back to the multiple linear regression model:

  $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + E_i$

  let $b_0, b_1, b_2, ..., b_n$ be predictors of:

  $\beta_0, \beta_1, \beta_2, ..., \beta_n$

  resulting in the following prediction equation:

  $\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n$

Inference in Multiple Regression

Residuals

- A residual represents the difference between an observed and predicted response. The i^{th} residual is calculated by the following:

  e_i = observed response - predicted response
  $e_i = y_i - \hat{y}_i$
  $e_i = y_i - b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n$

Estimating Variance (σ^2)

- Variance (σ^2) is a measure of how the scores about the population regression equation vary. Similar to simple linear regression, we estimate variance by averaging the squared residuals. This estimation is represented by the following:

$$s^2 = \frac{\Sigma e_i^2}{n - p - 1} \quad \text{OR} \quad s^2 = \frac{\Sigma (y_i - \hat{y}_i)^2}{n - p - 1}$$

n - p - 1 represents the degrees of freedom of s^2.

Confidence Intervals

- Confidence intervals and significance tests for each regression coefficient, β_j, can be calculated in a similar fashion to those in simple linear regression.

  A confidence interval for β_j of level C is given by the following:

  $b_j \pm t^* SE_{bj}$

  Where t* value has n - p - 1 degrees of freedom

Significance Tests

- Testing the null hypothesis (H_o: $\beta_j = 0$) requires the calculation of the corresponding t statistic, given by the following:

$$t = \frac{b_j}{SE_{bj}}$$

ANOVA F Test

ANOVA F Test

- In simple linear regression, a F test is the same as the two-sided t-test of the null hypothesis, which asserts the slope of the regression line is equal to zero. Similarly, in multiple regression, the F test examines the following null hypothesis:

  H_o: $\beta_1 = \beta_2 = ... = \beta_n = 0$

  Where all slopes are simultaneously = 0

- The alternative hypothesis is as follows:

  H_a: at least one β_j is not = 0

  Note, if the resulting p-value is large, none of the explanatory variables will be helpful in predicting a response.

- The ANOVA table for multiple regression is as follows:

Source	Degrees of freedom	Sum of squares	Mean square	F
Model (regression)	p	$SSM = \Sigma (\hat{y}_i - \bar{y}_i)^2$	MSM= SSM/p	MSM/MSE
Error (residuals)	n-p-1	$SSE = \Sigma (y_i - \hat{y}_i)^2$	MSE= SSE/n-p-1	
Total	n-1	$SST = \Sigma (y_i - \bar{y}_i)^2$	SST/n-1	

- In multiple regression, similar to simple linear regression, both the sum of squares and their degrees of freedom can be summed as follows:

  SST = SSM + SSE

  DFT = DFM + DFE

  As before, the estimate of the variance is given by the MSE in the ANOVA table

  s^2 = MSE

- The F statistic for measuring the null hypothesis against the alternative hypothesis is calculated by dividing MSM by MSE (MSM/MSE).

  Large F values support the rejection of H_0. When H_0 is supported, F has an F (p, n-p-1) distribution.

Glossary

Multiple Regression:	A method which uses two or more explanatory variables to account for a single response variable.
Population Regression Equation:	The multiple regression equation, using population parameters, under the assumption that the mean response can be expressed as a linear function of the explanatory variables.
Residuals:	Is the difference between the observed and predicted response.
ANOVA Table:	Is a summary table in an analysis of variance for a multiple linear regression which gives the degrees of freedom, sum of squares and mean squares for the model, error and total sources of variation.
F-Statistic:	Is the ratio MSM/MSE, used to test the null hypothesis H_o: $\beta_1 = \beta_2 = ... = \beta_n = 0$.

References

McCabe, George P. & Moore, David S. *Introduction to the Practice of Statistics*, Fourth Edition. 2003. W.H. Freeman and Company, New York.