Contents

Introduction

A multiple linear regression model is a linear model where the target variable is depend on at least 2 regrresor variables.

Recall that a linear model is of the form: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

A multiple linear model is of the form:

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik}$

$= B_0 + \sum_{j=1}^{k} \beta_j X_{ij} + ei$

where:

Y_i is the observed random variable

β_0 is the intercept of the model

$\beta_j^{'s}$ are the the partial coefficients of the model

$X_i^{'s}$ are the observed non random variables

The fitted multiple linear model is of the from:

$\hat{Y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j X_{ij}$

$\hat{Y}_1$ estimate of Y_i

$\hat{\beta}_0$ is the estimate of β_0

$\hat{\beta_j^{'s}}$ are estimates of $\beta_j^{'s}$

For easier manual computations we employ the use of matrices to find estimates of Y_i $\beta_0, \beta_j^{'s}$ variances and standard errors, $SS_{reg}\ and\ SS_e$

Matrix representation of multiple linear model

$\underline{Y} = \underline{X\beta} + \epsilon$ is the matrix notation of a multiple linear model where ;

$$\underline{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

,

$$\underline{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

,

$$\underline{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$$

and,

$$\underline{X} = \begin{pmatrix} x_{11}\ x_{12} \ldots x_{1k} \\ x_{22}\ x_{22} \ldots x_{2k} \\ \vdots \\ x_{n1}\ x_{n2} \ldots x_{nk} \end{pmatrix}$$

The fitted line in matrix notation is given by

$\underline{\hat{Y}} = \underline{X}\underline{\hat{\beta}}$

where:

$\underline{\hat{\beta}} = (\underline{X^1}\ \underline{X})^{-1}.\underline{X^1}\ \underline{Y}$

where; $\underline{X^1}$ is the transpose of $\underline{X}$

The variance of $\underline{\hat{\beta}}$ is given by $var(\underline{\hat{\beta}}) = \sigma^2(\underline{X^1}.\underline{X})^{-1}$

The estimate of σ^2 is $\hat{\sigma^2} = \frac{1}{n-k}[\underline{Y^1}.\underline{Y} - \underline{\hat{\beta}^1}.\underline{X^1}.\underline{Y}]$, where k is the number of independent variables

The standard error of the coefficients $\underline{\hat{\beta}}$ is given by

$SE(\underline{\hat{\beta}}) = \sqrt{var(\underline{\hat{\beta}})}$

Test of Hypothesis

Test of hypothesis is done to get the following inferences

1. Overall test:

- Do the independent variables as well as the fitted model contribute significantly to the prediction of Y

2. Test for addition of a single variable

- We want to determine whether addition of a single independent variable of interest add significance to the prediction of Y

3. Test for addition of a group of variables

- We want to determine whether addition of a group of independent variables of interest contribute significantly to prediction of Y

An F-test is used to carry out hypothesis of a multiple linear model as it carries out an overall significance test for all the partial coeeficients

A t-test can be used but it makes the work tideous and more errors are bound to be made as significance test is carried out for each individual partial coefficient

The hypothesis is given by:

$H_0 : \beta_j = 0 \ for \ all \ j = 1, 2, \ldots, p-1 \ vs$
$H_1 : \beta_j \neq 0 \ for \ atleast \ 1 \ j = 1, 2, \ldots, p-1 \ where \ p = k+1$

The F- statistics is given by:

$F = \frac{SS_{reg}/p-1}{SS_e/n-p} = \frac{MSR}{MSE} \sim F(p-1, n-p)$, where;

$SS_{Reg} = \underline{\hat{\beta}}^1.\underline{X}^1.\underline{Y} - \frac{1}{n}[\sum_{i=1}^{n} Y_i]^2$

$SS_e = SS_T - SS_{Reg}$, where $SS_T = \underline{Y}^1.\underline{Y} - \frac{1}{n}[\sum_{i=1}^{n} Y_i]^2$

Confidence Interval

Here we find confidence interval for each of the predictors, the t-test is used.

$$t = \hat{\beta}_j \pm t_{\frac{\alpha}{2}(n-p)}.SE(\hat{\beta}_j), p = k+1$$

Examples

one

```
D <- data.frame(
x1=c(0.58, 0.86, 0.29, 0.20, 0.56, 0.28, 0.08, 0.41, 0.22,
0.35, 0.59, 0.22, 0.26, 0.12, 0.65, 0.70, 0.30, 0.70,
0.39, 0.72, 0.45, 0.81, 0.04, 0.20, 0.95),
x2=c(0.71, 0.13, 0.79, 0.20, 0.56, 0.92, 0.01, 0.60, 0.70,
0.73, 0.13, 0.96, 0.27, 0.21, 0.88, 0.30, 0.15, 0.09,
0.17, 0.25, 0.30, 0.32, 0.82, 0.98, 0.00),
y=c(1.45, 1.93, 0.81, 0.61, 1.55, 0.95, 0.45, 1.14, 0.74,
0.98, 1.41, 0.81, 0.89, 0.68, 1.39, 1.53, 0.91, 1.49,
1.38, 1.73, 1.11, 1.68, 0.66, 0.69, 1.98)
)
nrow(D)
```

```
## [1] 25
```

1. Calculate the parameter estimates ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and σ_2^2), in addition find the usual 95% confidence intervals for β_0 , β_1 , and β_2

```
#fit a multiple linear model
fit<-lm(y~x1+x2,data=D)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = D)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.15493 -0.07801 -0.02004  0.04999  0.30112
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.433547   0.065983   6.571 1.31e-06 ***
## x1          1.652993   0.095245  17.355 2.53e-14 ***
## x2          0.003945   0.074854   0.053    0.958
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1127 on 22 degrees of freedom
## Multiple R-squared:  0.9399, Adjusted R-squared:  0.9344
## F-statistic:   172 on 2 and 22 DF,  p-value: 3.699e-14
```

```r
#get confidence intervals
confint(fit)
```

```
##                  2.5 %    97.5 %
## (Intercept)  0.2967067 0.5703875
## x1           1.4554666 1.8505203
## x2          -0.1512924 0.1591822
```

HAPPY NEW MONTH

Have
A
Peachy,
Prosperous,
Yummy

New
Experience
With

Moments
Of
New
Triumphs
Happiness!