Evaluating Linear Regression

Nouman Riaz

October 18, 2016

The objective is to calculate multiple statistical measures for Simple and Linear Regression models to evaluate performance.

```r
require(MASS)
```

```
## Loading required package: MASS
```

```r
ads <- read.csv("Advertising.csv", stringsAsFactors=FALSE) #Reading Dataset
```

Part 1) Caclulate R-squared of Simple Linear Regression model for Sales vs input variables.

We know the formula for R-squared:

$$R^2 = 1 - \frac{TSS}{RSS}$$

and we also have the information of TSS & RSS calculation as:

$$TSS = \sum_{i=1}^{n} (y_i - y^-)^2$$

$$RSS = \sum_{i=1}^{n} (y_i - y_{est})^2$$

```r
y = ads$Sales #As 'sales' is the response variable for all models
TSS = sum((y - mean(y))*(y - mean(y))) #TSS for response variable
cat(TSS)
```

```
## 5417.149
```

a) Sales ~ TV

```r
#as we know the beta values already from the previous

beta_0_tv <- 7.03
beta_1_tv <- 0.0475

x <- ads$TV
y_est_tv <- beta_0_tv+(beta_1_tv*x) #Cacluating Y-estimated from TV.

RSS_tv <- sum((y-y_est_tv)*(y-y_est_tv)) #Caclulating RSS error

R_sqr_tv <- 1 - (RSS_tv/TSS) #Caclulating R-squared error for Sales~TV
cat(R_sqr_tv)
```

```
## 0.6118723
```

b) Sales ~ Radio

```
beta_0_radio <- 9.312
beta_1_radio <- 0.202

x <- ads$Radio

y_est_radio <- beta_0_radio+(beta_1_radio*x) #Cacluating Y-estimated from Radio.

RSS_radio <- sum((y-y_est_radio)*(y-y_est_radio)) #Caclulating RSS error

R_sqr_radio <- 1 - (RSS_radio/TSS) #Caclulating R-squared error for Sales~Radio
cat(R_sqr_radio)
```

```
## 0.3320259
```

c) Sales ~ Newspaper

```
beta_0_np <- 12.351
beta_1_np <- 0.055

x <- ads$Newspaper

y_est_np <- beta_0_np+(beta_1_np*x) #Cacluating Y-estimated from Newspaper.

RSS_np <- sum((y-y_est_np)*(y-y_est_np)) #Caclulating RSS error

R_sqr_np <- 1 - (RSS_np/TSS) #Caclulating R-squared error for Sales~Newspaper
cat(R_sqr_np)
```

```
## 0.05211583
```

Part 2) Caclulate P-values of Simple Linear Regression model for Sales vs input variables.

To find p-values, first we need to caclulate Standard error of beta values:

$$S.E(\beta_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{x^{-2}}{\sum_{i=1}^{n}(x_i - x^-)^2}\right]$$

$$S.E(\beta_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - x^-)^2}$$

and the t-statistic:

$$t - stat(\beta) = \frac{\beta}{S.E(\beta)}$$

a) Sales ~ TV

```
n <- length(ads$TV)
df <- n-2
sigma_sqr_tv <- RSS_tv/(n-2) #calculating sigma square

x <- ads$TV
x_bar <- mean(ads$TV)
```

```r
se_b0_tv <- sqrt((sigma_sqr_tv)*((1/n)+((x_bar*x_bar)/sum((x-x_bar)*(x-x_bar))))) #Calculating Standard
cat(se_b0_tv)
```

```
## 0.4578445
```

```r
t_stat_b0 <- beta_0_tv/se_b0_tv #The T-score of Beta.o
cat(t_stat_b0)
```

```
## 15.35456
```

```r
p_value_b0 <- 2*pt(t_stat_b0,df,lower.tail = FALSE) #Calculating P-value for Beta.o
cat(p_value_b0)
```

```
## 1.464053e-35
```

```r
se_b1_tv <- sqrt(sigma_sqr_tv/sum((x-x_bar)*(x-x_bar)))
cat(se_b1_tv)
```

```
## 0.002690617
```

```r
t_stat_b1 <- beta_1_tv/se_b1_tv
cat(t_stat_b1)
```

```
## 17.65395
```

```r
p_value_b1 <- 2*pt(t_stat_b1,df,lower.tail = FALSE) #Calculating P-value for Beta.1
cat(p_value_b1)
```

```
## 1.612191e-42
```

b) Sales ~ Radio

```r
n <- length(ads$TV)
df <- n-2
sigma_sqr_tv <- RSS_tv/(n-2) #calculating sigma square

x <- ads$Radio
x_bar <- mean(ads$Radio)

se_b0_tv <- sqrt((sigma_sqr_tv)*((1/n)+((x_bar*x_bar)/sum((x-x_bar)*(x-x_bar))))) #Calculating Standard
cat(se_b0_tv)
```

```
## 0.4290829
```

```r
t_stat_b0 <- beta_0_tv/se_b0_tv #The T-score of Beta.o
cat(t_stat_b0)
```

```
## 16.38378
```

```r
p_value_b0 <- 2*pt(t_stat_b0,df,lower.tail = FALSE) #Calculating P-value for Beta.o
cat(p_value_b0)
```

```
## 1.075494e-38
```

```r
se_b1_tv <- sqrt(sigma_sqr_tv/sum((x-x_bar)*(x-x_bar)))
cat(se_b1_tv)
```

```
## 0.01555895
```

```r
t_stat_b1 <- beta_1_tv/se_b1_tv
cat(t_stat_b1)
```

```
## 3.052904
```

```r
p_value_b1 <- 2*pt(t_stat_b1,df,lower.tail = FALSE) #Calculating P-value for Beta.1
cat(p_value_b1)
```

```
## 0.002577663
```

c) Sales ~ Newspaper

```r
n <- length(ads$TV)
df <- n-2
sigma_sqr_tv <- RSS_tv/(n-2) #calculating sigma square

x <- ads$Newspaper
x_bar <- mean(ads$Newspaper)

se_b0_tv <- sqrt((sigma_sqr_tv)*((1/n)+((x_bar*x_bar)/sum((x-x_bar)*(x-x_bar))))) #Calculating Standard
cat(se_b0_tv)
```

```
## 0.3976455
```

```r
t_stat_b0 <- beta_0_tv/se_b0_tv #The T-score of Beta.o
cat(t_stat_b0)
```

```
## 17.67906
```

```r
p_value_b0 <- 2*pt(t_stat_b0,df,lower.tail = FALSE) #Calculating P-value for Beta.o
cat(p_value_b0)
```

```
## 1.356373e-42
```

```r
se_b1_tv <- sqrt(sigma_sqr_tv/sum((x-x_bar)*(x-x_bar)))
cat(se_b1_tv)
```

```
## 0.01060677
```

```r
t_stat_b1 <- beta_1_tv/se_b1_tv
cat(t_stat_b1)
```

```
## 4.478272
```

```r
p_value_b1 <- 2*pt(t_stat_b1,df,lower.tail = FALSE) #Calculating P-value for Beta.1
cat(p_value_b1)
```

```
## 1.269891e-05
```

Part 3) Caclulate R-squared of Multiple Linear Regression model for Sales vs all input variables.

```r
#As we know the beta values of MLR from previous

X <- matrix(c(seq(from=1,to=1,length.out = nrow(ads)),ads$TV,ads$Radio,ads$Newspaper), nrow=nrow(ads), n

BetaV <- c(2.938889,0.04576465,0.18853,0.001037493)
```

```r
y_est <- X%*%BetaV

RSS <- sum((y-y_est)*(y-y_est)) #Caclulating RSS error

R_sqr <- 1 - (RSS/TSS) #Caclulating R-squared error for MLR
cat(R_sqr)
```

```
## 0.8969872
```

Part 4) Caclulate p-values of Multiple Linear Regression model for Sales vs all input variables.

```r
n = length(ads$Sales)
df=n-2
p = ncol(X)
sigma_sqr <- RSS/(n-p-1)
var_covar <- sigma_sqr*ginv((t(X)%*%X))          # variance covariance matrix
se_v <- sqrt(diag(var_covar))
cat(se_v) #Standard Error Matrix
```

```
## 0.3130466 0.001399988 0.008642663 0.005892438
```

```r
t_values <- BetaV/se_v
cat(t_values)
```

```
## 9.388023 32.68932 21.81388 0.176072
```

```r
p_values <- 2*pt(t_values,df,lower.tail = FALSE) #Calculating p-values
cat(p_values)
```

```
## 1.490564e-17 9.963258e-82 1.479456e-54 0.8604174
```

Part 5) Introducing interaction terms

```r
#a) TV*Radio as interaction term

X <- matrix(c(seq(from=1,to=1,length.out = nrow(ads)),ads$TV,ads$Radio,ads$Newspaper,ads$TV*ads$Radio),

#Finding beta vector for this
Xt <- t(X) #Taking transpose of input matrix

a <- Xt %*% X #Multiplaying input matrix with its transpose
b <- Xt %*% y #Multiplaying input matrix with output variable
inv <- ginv(a) #Taking inverse
BetaV <- inv %*% b #Caculating Beta vector
y_est <- X%*%BetaV

RSS <- sum((y-y_est)*(y-y_est)) #Caclulating RSS error

R_sqr2 <- 1 - (RSS/TSS) #Caclulating R-squared error for MLR
cat(R_sqr2)
```

```
## 0.8516222
```

```r
#b) Radio*Newspaper

X <- matrix(c(seq(from=1,to=1,length.out = nrow(ads)),ads$TV,ads$Radio,ads$Newspaper,ads$Radio*ads$News

#Finding beta vector for this
Xt <- t(X) #Taking transpose of input matrix

a <- Xt %*% X #Multiplaying input matrix with its transpose
b <- Xt %*% y #Multiplaying input matrix with output variable
inv <- ginv(a) #Taking inverse
BetaV <- inv %*% b #Caculating Beta vector
y_est <- X%*%BetaV

RSS <- sum((y-y_est)*(y-y_est)) #Caclulating RSS error

R_sqr3 <- 1 - (RSS/TSS) #Caclulating R-squared error for MLR
cat(R_sqr3)
```

```
## 0.8974526
```

```r
#c) Tv*Radio but no newspaper

X <- matrix(c(seq(from=1,to=1,length.out = nrow(ads)),ads$TV,ads$Radio,ads$TV*ads$Radio), nrow=nrow(ads

#Finding beta vector for this
Xt <- t(X) #Taking transpose of input matrix

a <- Xt %*% X #Multiplaying input matrix with its transpose
b <- Xt %*% y #Multiplaying input matrix with output variable
inv <- ginv(a) #Taking inverse
BetaV <- inv %*% b #Caculating Beta vector
y_est <- X%*%BetaV

RSS <- sum((y-y_est)*(y-y_est)) #Caclulating RSS error

R_sqr4 <- 1 - (RSS/TSS) #Caclulating R-squared error for MLR
cat(R_sqr4)
```

```
## 0.8461827
```

```r
cat(sprintf("R2 = %s\nR2_2 = %s\nR2_3 = %s\nR2_4 = %s\n",R_sqr,R_sqr2,R_sqr3,R_sqr4))
```

```
## R2 = 0.896987222260582
## R2_2 = 0.851622239746509
## R2_3 = 0.897452605460989
## R2_4 = 0.846182687323899
```

So the better R-square we have, is achieved by introducing “Radio*Newspaper" interaction term.