Mid Semester Practical Exam 1740256

**Jeevan**

11/09/2019

---------------------------------------------------------------------------------------------------------------------

**AIM:**

A random sample was collected on y, X1, X2, X3, and X4 from various cities of America to study about the health status of the population. The detailed description of the variables is as follows:  
y= death rate per 1000 residents  
X1 = doctor availability per 100,000 residents  
X2 = hospital availability per 100,000 residents  
X3 = annual per capita income in thousands of dollars  
X4 = population density people per square mile

**Data is given in separate excel file**

Use R package to analyze the above data with the following steps and draw your conclusions.

1. *Obtain the best subset model using backward elimination method to estimate the death rates.*
2. *Establish the linear relationship between death rate and various regressors.*
3. *Test the significance of the regression coefficients at 5% of level of significance.*
4. *Obtain the predicted death rates and check whether the sum of the observed values and expected values is approximately equal.*
5. *Obtain the adjusted coefficient of multiple determination.*

---------------------------------------------------------------------------------------------------------------------

**PROCEDURE:**

library(readxl)

## Warning: package 'readxl' was built under R version 3.5.2

data <- read\_excel("C:/Users/Jeevan/Desktop/Christ University/Statistics/Linear Regression/death rate-mid sem practical 2.xlsx")  
# View(data)  
attach(data)  
fullmodel = lm(Y~.,data = data)  
summary(fullmodel)

##   
## Call:  
## lm(formula = Y ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.6404 -0.7904 0.3053 0.9164 2.7906   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.2662552 2.0201467 6.072 1.95e-07 \*\*\*  
## X1 0.0073916 0.0069336 1.066 0.2917   
## X2 0.0005837 0.0007219 0.809 0.4228   
## X3 -0.3302302 0.2345518 -1.408 0.1656   
## X4 -0.0094629 0.0048868 -1.936 0.0587 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.601 on 48 degrees of freedom  
## Multiple R-squared: 0.1437, Adjusted R-squared: 0.07235   
## F-statistic: 2.014 on 4 and 48 DF, p-value: 0.1075

confint(be,level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) 8.2044779685 1.632803e+01  
## X1 -0.0065494163 2.133265e-02  
## X2 -0.0008677875 2.035219e-03  
## X3 -0.8018281444 1.413677e-01  
## X4 -0.0192884855 3.627158e-04

fit\_best = fitted.values(fullmodel)  
fit\_best

## 1 2 3 4 5 6 7   
## 8.972027 8.785975 9.768070 10.164920 8.401402 8.525122 8.486422   
## 8 9 10 11 12 13 14   
## 9.341207 8.771974 9.516408 7.406281 8.700588 9.185219 8.372307   
## 15 16 17 18 19 20 21   
## 9.533879 8.345707 8.922888 8.706957 8.770364 9.233007 10.204736   
## 22 23 24 25 26 27 28   
## 9.862588 10.031174 8.820786 9.094704 9.768036 9.333990 8.998103   
## 29 30 31 32 33 34 35   
## 10.268026 8.981212 8.969487 9.507752 9.545355 9.634755 9.361534   
## 36 37 38 39 40 41 42   
## 10.290094 8.540195 10.047817 8.755423 9.721895 9.500673 9.664476   
## 43 44 45 46 47 48 49   
## 9.240410 9.950695 9.883561 8.690609 9.790375 9.481179 10.501729   
## 50 51 52 53   
## 10.063890 10.009360 9.304791 9.469863

sum(Y)

## [1] 493.2

sum(fit)

## [1] 493.2

---------------------------------------------------------------------------------------------------------------------

**INTERPRETATION:**

The “readxl” library has to be imported in order to read the data from the given excel sheet provided for our analysis.

1. The best subset model using backward elimination method to eliminate the death rates has been obtained using RStudio.

step(fullmodel,direction = "backward")

## Start: AIC=54.65  
## Y ~ X1 + X2 + X3 + X4  
##   
## Df Sum of Sq RSS AIC  
## - X2 1 1.6763 124.75 53.369  
## - X1 1 2.9139 125.99 53.892  
## <none> 123.07 54.652  
## - X3 1 5.0825 128.16 54.797  
## - X4 1 9.6144 132.69 56.639  
##   
## Step: AIC=53.37  
## Y ~ X1 + X3 + X4  
##   
## Df Sum of Sq RSS AIC  
## <none> 124.75 53.369  
## - X1 1 5.1882 129.94 53.529  
## - X3 1 6.1544 130.91 53.921  
## - X4 1 8.3192 133.07 54.791

##   
## Call:  
## lm(formula = Y ~ X1 + X3 + X4, data = data)  
##   
## Coefficients:  
## (Intercept) X1 X3 X4   
## 12.565900 0.009284 -0.359140 -0.008580

1. The linear relationship between death rate and various regressors has been obtained.

The parameters for the multiple linear regression model are -

be = lm(Y~X1+X2+X3+X4,data = data)  
summary(be)

##   
## Call:  
## lm(formula = Y ~ X1 + X2 + X3 + X4, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.6404 -0.7904 0.3053 0.9164 2.7906   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.2662552 2.0201467 6.072 1.95e-07 \*\*\*  
## X1 0.0073916 0.0069336 1.066 0.2917   
## X2 0.0005837 0.0007219 0.809 0.4228   
## X3 -0.3302302 0.2345518 -1.408 0.1656   
## X4 -0.0094629 0.0048868 -1.936 0.0587 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.601 on 48 degrees of freedom  
## Multiple R-squared: 0.1437, Adjusted R-squared: 0.07235   
## F-statistic: 2.014 on 4 and 48 DF, p-value: 0.1075

1. The significance of the regression coefficients at 5% level of significance for the best model and their slope and interecept parameter is -confint(fullmodel,level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) 8.2044779685 1.632803e+01  
## X1 -0.0065494163 2.133265e-02  
## X2 -0.0008677875 2.035219e-03  
## X3 -0.8018281444 1.413677e-01  
## X4 -0.0192884855 3.627158e-04

1. The predicted death rates (fitted values of the model) are -

fit = fitted.values(be)  
fit

## 1 2 3 4 5 6 7   
## 8.972027 8.785975 9.768070 10.164920 8.401402 8.525122 8.486422   
## 8 9 10 11 12 13 14   
## 9.341207 8.771974 9.516408 7.406281 8.700588 9.185219 8.372307   
## 15 16 17 18 19 20 21   
## 9.533879 8.345707 8.922888 8.706957 8.770364 9.233007 10.204736   
## 22 23 24 25 26 27 28   
## 9.862588 10.031174 8.820786 9.094704 9.768036 9.333990 8.998103   
## 29 30 31 32 33 34 35   
## 10.268026 8.981212 8.969487 9.507752 9.545355 9.634755 9.361534   
## 36 37 38 39 40 41 42   
## 10.290094 8.540195 10.047817 8.755423 9.721895 9.500673 9.664476   
## 43 44 45 46 47 48 49   
## 9.240410 9.950695 9.883561 8.690609 9.790375 9.481179 10.501729   
## 50 51 52 53   
## 10.063890 10.009360 9.304791 9.469863

It is found that the sum of observed values of the best model are equal to the sum of expected values.

sum(fit\_best)

## [1] 493.2

sum(Y)

## [1] 493.2

1. The adjusted coefficient of multiple determination is

The adjusted coefficient of multiple determination is ***0.07234595*** and this shows that the model is a good fit to the data.

summary(be)$adj.r.squared

## [1] 0.07234595

---------------------------------------------------------------------------------------------------------------------