Question

In an experiment to measure the stiffness of a spring the length of the spring under different loads was measured as

Length (y) : 10 12 15 18 20 22 27 30 32 34

Weight(x): 3 5 6 9 10 12 15 20 22 28

Fit the simple linear regression model and compute the residuals

Test the heteroscedasticity at 5 % level of significance using Spearman rank correlation Test.

Multiple Regression Model

Multiple regression analysis is the study of how a dependent variable y is related to two or more independent variables. In the general case, we will use p to denote the number of independent variables.

Regression Model and Regression Equation

The concepts of a regression model and a regression equation introduced in the preceding chapter are applicable in the multiple regression case. The equation that describes how the dependent variable y is related to the independent variables $x_1, x_2, \ldots, x_p$ and an error term is called the **multiple regression model**. We begin with the assumption that the multiple regression model takes the following form.

MULTIPLE REGRESSION MODEL

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \qquad \textbf{(15.1)}$$

In the multiple regression model, $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the parameters and the error term ϵ (the Greek letter epsilon) is a random variable. A close examination of this model reveals that y is a linear function of $x_1, x_2, \ldots, x_p$ (the $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ part) plus the error term ϵ. The error term accounts for the variability in y that cannot be explained by the linear effect of the p independent variables.

ESTIMATED MULTIPLE REGRESSION EQUATION

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \qquad \textbf{(15.3)}$$

where

$b_0, b_1, b_2, \ldots, b_p$ are the estimates of $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$

$\hat{y}$ = estimated value of the dependent variable

Question

A shoe store developed the following estimated regression equation relating sales to inventory investment and advertising expenditures.

$$\hat{y} = 25 + 10x_1 + 8x_2$$

where

x_1 = inventory investment ($1000s)
x_2 = advertising expenditures ($1000s)
y = sales ($1000s)

a. Estimate sales resulting from a $15,000 investment in inventory and an advertising budget of $10,000.
b. Interpret b_1 and b_2 in this estimated regression equation.

Solution

a-

		($1000)	
Investment in inventory	= $150,000	$15	$= x_1$
advertising expenditures	= $10,000	$10	$= x_2$

So $\hat{y} = 25 + 10x_1 + 8x_2$

$\hat{y} = 25 + 10(15) + 8(10)$

$\hat{y} = 25 + 150 + 80 \Rightarrow$ **$1605** → Estimating Sales

Interpretation

$b_0 = 25$ indicate the average estimating Sales when there are no investment in inventory & adv. Exp.

$b_1 = 10$ indicate that 1 $ increase in investment in Inventory will increase the estimating Sales by 10 $ when effect of Adv. Exp is held constant

$b_2 = 8$ indicate that 1 $ increase in Adv. Exp will increase the estimating Sales by 8 $ when effect of investment in inventory is held constant.

Multiple Coefficient of Determination

In simple linear regression we showed that the total sum of squares can be partitioned into two components: the sum of squares due to regression and the sum of squares due to error. The same procedure applies to the sum of squares in multiple regression.

RELATIONSHIP AMONG SST, SSR, AND SSE

$$\text{SST} = \text{SSR} + \text{SSE} \tag{15.7}$$

where

$$\text{SST} = \text{total sum of squares} = \Sigma(y_i - \bar{y})^2$$
$$\text{SSR} = \text{sum of squares due to regression} = \Sigma(\hat{y}_i - \bar{y})^2$$
$$\text{SSE} = \text{sum of squares due to error} = \Sigma(y_i - \hat{y}_i)^2$$

MULTIPLE COEFFICIENT OF DETERMINATION

$$R^2 = \frac{\text{SSR}}{\text{SST}} \tag{15.8}$$

Many analysts prefer adjusting R^2 for the number of independent variables to avoid overestimating the impact of adding an independent variable on the amount of variability explained by the estimated regression equation. With n denoting the number of observations and p denoting the number of independent variables, the **adjusted multiple coefficient of determination** is computed as follows.

ADJUSTED MULTIPLE COEFFICIENT OF DETERMINATION

$$R_a^2 = 1 - (1 - R^2)\frac{n - 1}{n - p - 1} \tag{15.9}$$

Question

the following estimated regression equation relating sales to inventory investment and advertising expenditures was given.

$$\hat{y} = 25 + 10x_1 + 8x_2$$

The data used to develop the model came from a survey of 10 stores; for those data, SST = 16,000 and SSR = 12,000.

a. For the estimated regression equation given, compute R^2.
b. Compute R_a^2.
c. Does the model appear to explain a large amount of variability in the data? Explain.

Solution

Sol.. The given information is

$n = 10$

$SST = 16000$

$SSR = 12000$

As we know that

$$R^2 = \frac{SSR}{SST} = \frac{12000}{16000} = \frac{3}{4} = 0.75$$

Interpretation

$R^2 = 0.75$ indicate that 75% of the variation in Sales is explained by the ~~Inventory~~ investment in Inventory and Advt. Exp. so it indicate the model is good fit.

Adjusted R^2

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

$p = 2$ (no. of independent variables in the model)

$$= 1 - (1 - 0.75)\frac{10-1}{10-2-1}$$

$$= 1 - (0.25)\cdot\frac{9}{7}$$

$$= 1 - (0.25)(1.2857)$$

$$= 1 - 0.3214 \Rightarrow \boxed{0.68}$$

$R_a^2 = 0.68$ indicate that 68% of the variation in Sales is explained by x_1 and x_2 after adjusting the impact of adding independent variable on the amount of variability of y.

Testing for Significance

In this section we show how to conduct significance tests for a multiple regression relationship. The significance tests we used in simple linear regression were a t test and an F test. In simple linear regression, both tests provide the same conclusion; that is, if the null hypothesis is rejected, we conclude that $\beta_1 \neq 0$. In multiple regression, the t test and the F test have different purposes.

1. The F test is used to determine whether a significant relationship exists between the dependent variable and the set of all the independent variables; we will refer to the F test as the test for *overall significance.*
2. If the F test shows an overall significance, the t test is used to determine whether each of the individual independent variables is significant. A separate t test is conducted for each of the independent variables in the model; we refer to each of these t tests as a test for *individual significance.*

F TEST FOR OVERALL SIGNIFICANCE

H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$
H_a: One or more of the parameters is not equal to zero

TEST STATISTIC

$$F = \frac{\text{MSR}}{\text{MSE}} \qquad (15.14)$$

REJECTION RULE

p-value approach:	Reject H_0 if *p*-value $\leq \alpha$
Critical value approach:	Reject H_0 if $F \geq F_\alpha$

where F_α is based on an F distribution with p degrees of freedom in the numerator and $n - p - 1$ degrees of freedom in the denominator.

$$\text{MSR} = \frac{\text{SSR}}{p} \qquad (15.12)$$

$$\text{MSE} = \frac{\text{SSE}}{n - p - 1} \qquad (15.13)$$

Question

the following estimated regression equation relating sales to inventory investment and advertising expenditures was given.

$$\hat{y} = 25 + 10x_1 + 8x_2$$

The data used to develop the model came from a survey of 10 stores; for these data SST = 16,000 and SSR = 12,000.

a. Compute SSE, MSE, and MSR.
b. Use an *F* test and a .05 level of significance to determine whether there is a relationship among the variables.

Solution

Sol.. the given information is

$n=10$, $SST = 16000$, $SSR = 12000$, $P = 2$

SSE. As we know that

$SST = SSR + SSE$

$SSE = SST - SSR$

$SSE = 16000 - 12000$

$\boxed{SSE = 4000}$

$$MSE = \frac{\Sigma(y-\hat{y})^2}{n-p-1} = \frac{\Sigma e^2}{n-p-1} = \frac{SSE}{n-p-1}$$

$$MSE = \frac{4000}{10-2-1} = 571.4286$$

$$MSR = \frac{\Sigma(\hat{y}-\bar{y})^2}{p} = \frac{SSR}{p} = \frac{12000}{2} = 6000$$

$H_0: \beta_1 = \beta_2 = 0$

H_1: Not all β's are equal to zero

$\alpha = 0.05$

Test - Statistic

$$F = \frac{MSR}{MSE} = \frac{6000}{571.4286} = 10.4999$$

C.R $F \geq F_\alpha(p, n-p-1)$

$F \geq F_{0.05}(2, 7)$

$F \geq 4.74$

Conclusion Reject H_0, the investment in inventory and Adv. Exp have significant impact on Sales.

t Test

If the *F* test shows that the multiple regression relationship is significant, a *t* test can be conducted to determine the significance of each of the individual parameters. The *t* test for individual significance follows.

t TEST FOR INDIVIDUAL SIGNIFICANCE

For any parameter β_i

$$H_0: \beta_i = 0$$
$$H_a: \beta_i \neq 0$$

TEST STATISTIC

$$t = \frac{b_i}{s_{b_i}} \qquad (15.15)$$

REJECTION RULE

p-value approach:	Reject H_0 if *p*-value $\leq \alpha$
Critical value approach:	Reject H_0 if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$

where $t_{\alpha/2}$ is based on a *t* distribution with $n - p - 1$ degrees of freedom.

Question

the following estimated regression equation based on 10 observations was

$$\hat{y} = 29.1270 + .5906x_1 + .4980x_2$$

Here SST = 6724.125, SSR = 6216.375, $s_{b_1} = .0813$, and $s_{b_2} = .0567$.

a. Compute MSR and MSE.
b. Compute *F* and perform the appropriate *F* test. Use $\alpha = .05$.
c. Perform a *t* test for the significance of β_1. Use $\alpha = .05$.
d. Perform a *t* test for the significance of β_2. Use $\alpha = .05$.

Solution

Sol:-

The given information is

$$\hat{y} = 29.1270 + 0.5906x_1 + 0.4980x_2$$

$SST = 6724.125 \qquad n = 10$

$SSR = 6216.375 \qquad P = 2$

$s_{b_1} = 0.0813$

$s_{b_2} = 0.0567$

(a) $MSR = ?$

$$MSR = \frac{SSR}{P} = \frac{6216.375}{2} = 3108.1875$$

$$MSE = \frac{SSE}{n-P-1} = \frac{507.75}{10-2-1} = 72.5357 \quad [SSE = SST - SSR]$$

(b)

Overall Significance Test

$H_0: \beta_1 = \beta_2 = 0$

H_1: at least one of the β's are not equal to zero.

$\alpha = 0.05$

Test-Statistic

$$F = \frac{MSR}{MSE} = \frac{3108.1875}{72.5357} \Rightarrow 42.8504$$

C.R

$F \geq F_\alpha(P, n-P-1)$

$F \geq F_{0.05}(2, 7)$

$F \geq 4.74$

Conclusion

~~Accept~~ Reject H_0 so x_1 and x_2 have significant effect on y.

(c)

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

$\alpha = 0.05$

Test-Statistic

$$t = \frac{b_1}{s_{b_1}}$$

$$t = \frac{0.5906}{0.0813}$$

$t = 7.2644$

C.R

$|t| \geq t_{\alpha/2}(n-P-1)$

$|t| \geq t_{0.025}(7)$

$|t| \geq 2.365$

Conclusion

Reject H_0 so x_1 has significant effect on y.

(d)

$H_0: \beta_2 = 0$

$H_1: \beta_2 \neq 0$

$\alpha = 0.05$

Test-Statistic

$$t = \frac{b_2}{s_{b_2}}$$

$$t = \frac{0.4980}{0.0567}$$

$t = 8.7807$

C.R

$|t| \geq t_{\alpha/2}(n-P-1)$

$|t| \geq t_{0.025}(7)$

$|t| \geq 2.365$

Conclusion

Reject H_0 so x_2 has significant effect on y.

ASSUMPTIONS ABOUT THE ERROR TERM ϵ IN THE MULTIPLE REGRESSION MODEL $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$

1. The error term ϵ is a random variable with mean or expected value of zero; that is, $E(\epsilon) = 0$.
   Implication: For given values of $x_1, x_2, \ldots, x_p$, the expected, or average, value of y is given by

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p. \qquad \textbf{(15.11)}$$

   Equation (15.11) is the multiple regression equation we introduced in Section 15.1. In this equation, $E(y)$ represents the average of all possible values of y that might occur for the given values of $x_1, x_2, \ldots, x_p$.
2. The variance of ϵ is denoted by σ^2 and is the same for all values of the independent variables $x_1, x_2, \ldots, x_p$.
   Implication: The variance of y about the regression line equals σ^2 and is the same for all values of $x_1, x_2, \ldots, x_p$.
3. The values of ϵ are independent.
   Implication: The value of ϵ for a particular set of values for the independent variables is not related to the value of ϵ for any other set of values.
4. The error term ϵ is a normally distributed random variable reflecting the deviation between the y value and the expected value of y given by $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$.
   Implication: Because $\beta_0, \beta_1, \ldots, \beta_p$ are constants for the given values of $x_1, x_2, \ldots, x_p$, the dependent variable y is also a normally distributed random variable.

Multicollinearity

Question

A data analyst is want to predict the price of games app. The price of the games varies with respect to the rating of apps and number of downloading of the apps. Therefore, Data analyst consider the two independent variables; $x_1 =$ ratings of apps, $x_2 =$ number of downloading of the apps, for predicting the price of apps. He collects the data for 10 games apps. The multiple linear regression model is used and the estimated regression model is given as

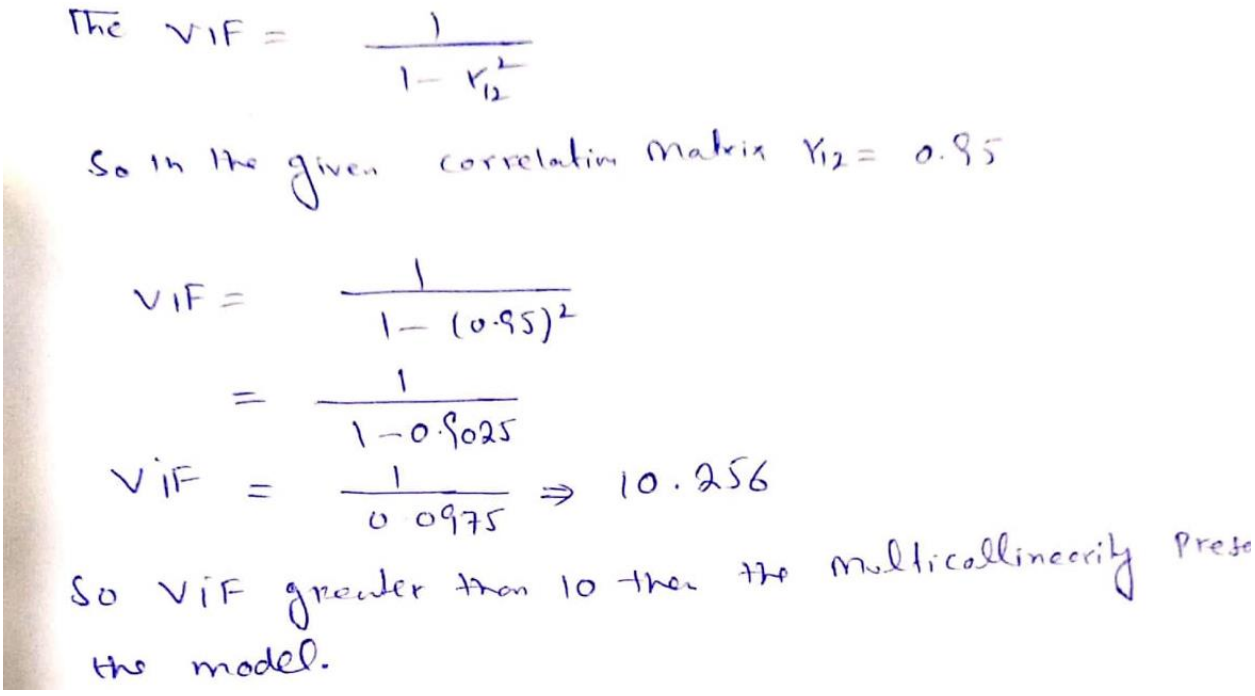
$$\hat{y} = 30 + 2.5x_1 + 0.2x_2$$

The correlation matrix between independent variables is given as

	x_1	x_2
x_1	1	0.95
x_2	0.95	1

Compute the variance inflation factor (VIF) and comment on the multicollinearity.

Solution



Question

The owner of Showtime Movie Theaters, Inc., would like to estimate weekly gross revenue as a function of advertising expenditures. Historical data for a sample of eight weeks follow.

Weekly Gross Revenue ($1000s)	Television Advertising ($1000s)	Newspaper Advertising ($1000s)
96	5.0	1.5
90	2.0	2.0
95	4.0	1.5
92	2.5	2.5
95	3.0	3.3
94	3.5	2.3
94	2.5	4.2
94	3.0	2.5

a. Develop an estimated regression equation with the amount of television advertising as the independent variable.
b. Develop an estimated regression equation with both television advertising and newspaper advertising as the independent variables.
c. Is the estimated regression equation coefficient for television advertising expenditures the same in part (a) and in part (b)? Interpret the coefficient in each case.

d. What is the estimate of the weekly gross revenue for a week when $3500 is spent on television advertising and $1800 is spent on newspaper advertising?

(ii)

a. Compute and interpret R^2 and R_a^2.
b. When television advertising was the only independent variable, $R^2 = .653$ and $R_a^2 = .595$. Do you prefer the multiple regression results? Explain.

c-

Use an F test and a .05 level of significance to determine whether there is a relationship among the variables.

d-

Use $\alpha = .05$ to test the significance of β_1. Should x_1 be dropped from the model?
Use $\alpha = .05$ to test the significance of β_2. Should x_2 be dropped from the model?

e- Test the autocorrelation at 5 % level of significance using Durbin-Watson Test

f-Test the Heteroscedasticity at 5% level of significance using Spearman Rank Correlation Test

g- check and test the normality of the residuals

h- Examine about the multicollinearity.

Solution

This question is solved using R-code as

###---Multiple Regression---###

```
rm(list=ls())
y=c(96,90,95,92,95,94,94,94)
x1=c(5,2,4,2.5,3,3.5,2.5,3)
x2=c(1.5,2,1.5,2.5,3.3,2.3,4.2,2.5)
dta=data.frame(y,x1,x2)
plot(dta)
reg1=lm(y~x1)
reg=lm(y~x1+x2)
summary(reg)
res=residuals(reg)
qqnorm(res)
qqline(res)
shapiro.test(res)  # test for normailty
plot(res)
library(lmtest)
dwtest(reg)   # test for autocorrelation
gqtest(reg)  # test for heteroscedasticity
ycp=fitted.values(reg)
plot(ycp,res^2)
cor(dta[,-1])   # for multicollinearity
```