Computer Repair Data

Suresh Kumar Prajapati

2024-03-05

#To study the relationship between the length of a service call and the number of electronic components in the computer that must be repaired or replaced, a sample of records on service on service calls was taken. Import the data

```
setwd("C:\\Users\\Admin\\OneDrive\\Desktop\\santosh")
getwd()

## [1] "C:/Users/Admin/OneDrive/Desktop/santosh"

data=read.csv("comrepair.csv")
data

##    Minutes Units
## 1       23     1
## 2       29     2
## 3       49     3
## 4       64     4
## 5       74     4
## 6       87     5
## 7       96     6
## 8       97     6
## 9      109     7
## 10     119     8
## 11     149     9
## 12     145     9
## 13     154    10
## 14     166    10
```

#Alinear model $\acute{y}=\frac{\sum y_i}{n}, \acute{x}=\frac{\sum x_i}{n}$

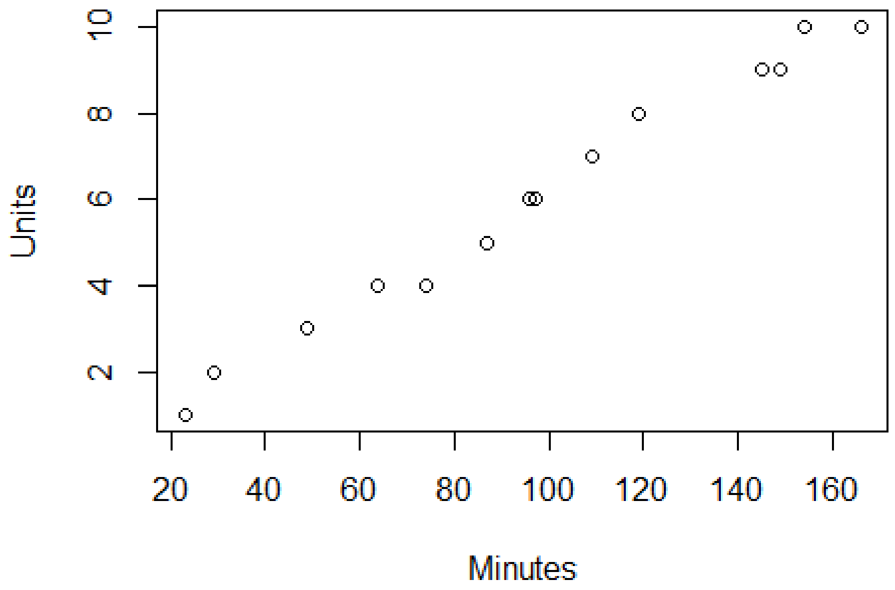
$$\text{Cov}(Y, X)=\frac{\sum(y_i-\acute{y})(x_i-\acute{x})}{n-1}$$

$$\text{Cor}(Y, X)=\frac{\sum(y_i-\acute{y})(x_i-\acute{x})}{\sqrt{\sum(y_i-\acute{y})^2\sum(x_i-\acute{x})^2}}$$

$\text{Cor}(Y, X)=\text{Cor}(X, Y)$ measures only pairwise relationships

Regression analysis is an attractive extension to correlation analysis because it postulates a model that can be used not only to measure the direction and the strength of a relationship between the response and predictor variables, but also to numerically describe that relationships.

```
M=data$Minutes
M

##  [1]  23  29  49  64  74  87  96  97 109 119 149 145 154 166

mean(M)

## [1] 97.21429

U=data$Units
U

##  [1]  1  2  3  4  4  5  6  6  7  8  9  9 10 10

mean(U)

## [1] 6

Cov=cov(M,U)
Cov

## [1] 136

plot(data)
```



#The simple linear regression Model $Y=\beta0+\beta1X +\epsilon$, where β0 and β1 are constants called the model regression coefficients or parameters, and ϵ is a random disturbance or error

```
              Y=βo+βiX +ϵi    i=1,2,3,4…..,n
```

Sometimes β0=intercepts β1=slope

Regression analysis differs in an important way from correlation analysis. The correlation coefficient is symmetric in the sense that Cor(Y, X) is the same as Cor(X, Y). The variables X and Y are of equal importance. In regression analysis the response variable Y is of primary importance. The Minutes=βo+βiUnits +ϵ

is assumed to represent the relationship between the length of service calls and the number of electronic components in the computer that must be repaired or replaced. See above figure.

Parameter estimation by least square method

$\text{errors}=\epsilon_i=Y_i-\beta_o-\beta_i X_i, i=1,2,3,\ldots\ldots n$

Vertical distance is the perpendicular (shortest)distance from each point to the line. The resultant line is called the orthogonal regression line.

Vertical distance is the perpendicular (shortest)distance from each point to the line. The resultant line is called the orthogonal regression line.

$S\left(\hat{\beta}_0,\hat{\beta}_1\right)=\epsilon_i^2=\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)^2$

The values of β0^ and β1^ that minimize S(β0 ,β1) are given by

$$\hat{\beta}_1=\frac{\sum\left(y_i-\acute{y}\right)\left(x_i-\acute{x}\right)}{\sum\left(x_i-\acute{x}\right)^2}$$

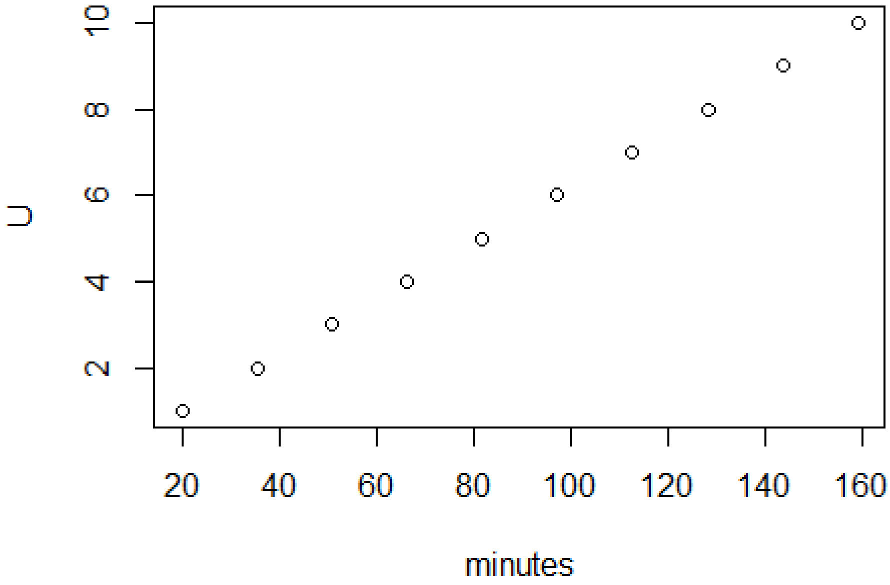
```
yi_ybar=(M-mean(M))
yi_ybar

##  [1] -74.2142857 -68.2142857 -48.2142857 -33.2142857 -23.2142857 -
10.2142857
##  [7]  -1.2142857  -0.2142857  11.7857143  21.7857143  51.7857143
47.7857143
## [13]  56.7857143  68.7857143

xi_xbar=(U-mean(U))
xi_xbar

##  [1] -5 -4 -3 -2 -2 -1  0  0  1  2  3  3  4  4

sqr=xi_xbar*xi_xbar
sqr

##  [1] 25 16  9  4  4  1  0  0  1  4  9  9 16 16
```

```
s=sum(sqr)
s

## [1] 114

pro=yi_ybar*xi_xbar
pro

##  [1] 371.07143 272.85714 144.64286  66.42857  46.42857  10.21429   0.00000
##  [8]   0.00000  11.78571  43.57143 155.35714 143.35714 227.14286 275.14286

S=sum(pro)
S

## [1] 1768

B1=S/s
B1

## [1] 15.50877
```

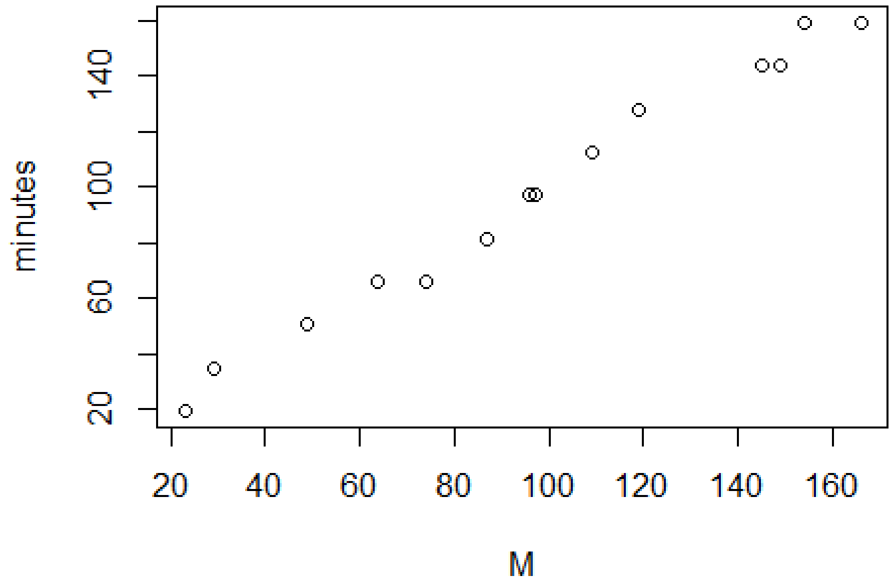
After plot the graph is linearity assumption is proof

$\widehat{\beta_o} = \acute{y} - \beta_1 \acute{x}$

```
B0=(mean(M)-B1*mean(U))
B0

## [1] 4.161654

minutes=4.161654+15.50877*U
minutes

##  [1]  19.67042  35.17919  50.68796  66.19673  66.19673  81.70550  97.21427
##  [8]  97.21427 112.72304 128.23181 143.74058 143.74058 159.24935 159.24935

plot(minutes,U)
```



```
minutes=4.161654+15.50877*U
minutes

##  [1]  19.67042  35.17919  50.68796  66.19673  66.19673  81.70550  97.21427
##  [8]  97.21427 112.72304 128.23181 143.74058 143.74058 159.24935 159.24935

resi=M-minutes
resi

##  [1]  3.329576 -6.179194 -1.687964 -2.196734  7.803266  5.294496 -1.214274
##  [8] -0.214274 -3.723044 -9.231814  5.259416  1.259416 -5.249354  6.750646

plot(M,minutes)
```



# The equation of the least squares regression line below.Thhe constant term represents the setup or startup time for each repair.

$$\hat{\beta}=\frac{cov(Y,X)}{var(X)}=\frac{cor(y,x)S_y}{S_x}$$

positive (negative) slope means positive (negative) correlation. To cheack linearty.2.5). If we observe a nonlinear pattern, we will have to take corrective action. For example, we may reexpress or transform the data before we continue the analysis,Data transformation

Using the properties of least squares estimators, one can develop statistical inference procedures (e.g., confidence interval estimation, tests of hypothesis, and goodness-of-fit test

#Test of Hypothesis

Hypothesis requires the following assumption. For every fixed value of X, the ϵ's are assumed to be independent random quantities normally distributed with mean zero and a common variance.

$$\mathrm{Var}\left(\hat{\beta}_0\right)=\sigma^2\left[\frac{1}{n}+\frac{\left(x-\acute{x}\right)^2}{\sum\left(x_i-\acute{x}\right)^2}\right)$$

$$\mathrm{Var}\left(\hat{\beta}_1\right)=\frac{\sigma^2}{\sum\left(x_i-\acute{x}\right)^2}$$

variances of $\widehat{\beta}_o$ and $\widehat{\beta}_1$ depend on the unknown parameter σ2 An unbiased estimate of σ2 is given by

${\sigma}^2=\frac{\sum e_i^2}{n-2}=\frac{\sum (y_i -\hat{y}_i)^2}}{n-2}=\frac{\text{SSE}}{n-2}$

It is equal to the number of observations minus the number of estimated regression coefficients.

```
sigma_2=sum(resi*resi)/(14-2)
sigma_2

## [1] 29.0707

sigma=sqrt(sigma_2)
sigma

## [1] 5.391725

seB0=sigma*sqrt(1/14+(mean(M)^2)/s)
seB0

## [1] 49.11254

seB1=sigma/sqrt(s)
seB1

## [1] 0.5049813
```

Null hypothesis H0: B1=0

Alternative Hypothesis H1:B1 !=0

$t1=\frac{\hat{\beta}_1}{s.e.\hat{\beta}_1}$, with n-2 degree of freedom

H0 is to be rejected at the significance level α if

|t1|≥ t(n-2,α/2), compare the p valuefor the t-Test with α with reject
H0 p(|t1|)≤α ;predictor variable X is a statistically significant predictor of the response variable Y.

T tab(12,0.05)=2.179

```
t1=B1/seB1
t1

## [1] 30.71158

#curve(,from=NULL,to=NULL,12,add=FALSE,type="I",xname="minutes",xlab=xname,ylab=NULL,xlim =NULL )
```

Since t cal=30.71158 > T tab=2.179 reject the null hypothesis. ## Testing NUll hypothesis ##Alternative Hypothesis

$H_0 : \beta_1 = \beta_{01}$ vs. $H_1 : \beta_1 \neq \beta_{01}$

T tab(12,0.05)=2.179 Using the Repair data ,let us suppose that the management expected the increase in service time for each additional unit to be repaired to be 12 minutes.Do the data support this conjector ?

```
T1=(B1-12)/seB1
T1

## [1] 6.94832
```

T1 cal=6.94832 > T tab=2.179 The result is highly significant,leading to the rejection the null hypothesis.The management's estimate of the increase in time for each additional component to be repaired in not supported by the data .Their estimate is too low. ## To testing Null hypothesis H0:B0=B0 ## Alternative Hypothesis H1:B0!=B0

```
t0=B0/seB0
t0

## [1] 0.0847371

datafit=lm(M~U, data = data)
datafit

## 
## Call:
## lm(formula = M ~ U, data = data)
## 
## Coefficients:
## (Intercept)            U  
##       4.162       15.509  

summary(datafit)

## 
## Call:
## lm(formula = M ~ U, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -9.2318 -3.3415 -0.7143  4.7769  7.8033 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    4.162      3.355    1.24    0.239    
## U             15.509      0.505   30.71 8.92e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
```

```
## Residual standard error: 5.392 on 12 degrees of freedom
## Multiple R-squared:  0.9874, Adjusted R-squared:  0.9864
## F-statistic: 943.2 on 1 and 12 DF,  p-value: 8.916e-13
```

A test using the correlation

##Testing $H_o:\beta_1=0$ ## Alternative Hypothesis $H_1\neq 0$

a test for determining whether the response and the predictor variables are linearly related

$$t_1=\frac{\mathrm{Cor}(Y,X)\sqrt{n-2}}{\sqrt{1-(\mathrm{Cor}(Y,X))^2}}$$

if Ho : p = 0 is rejected, it means that there is a statistically significant linear relationship between Y and X

```
cor_1=cor(M,U)
cor_1

## [1] 0.9936987

t_1=(cor_1*sqrt(14-2))/sqrt(1-(cor_1)^2)
t_1

## [1] 30.71158
```

#Confident intervals # (1-alpha)*100% confident interval for β_o β0^ ± t(n-2,α/2)×s.e.(β0^)

where t(n-2,α/2) is the (1 - α/2) percentile of a t distribution with n - 2 degrees of freedom. Similarly, limits of the (1 - α) x 100% confidence interval for**β**1 are given by

```
         β1^    ±   t(n-2,α/2)×s.e.(β1^)


conl=B0-(2.179*seB0)
conl

## [1] -102.8546

conu=B0+(2.179*seB0)
conu

## [1] 111.1779
```

#confident intervals # (1-alpha)*100% confident interval for B1

#confident intervals # (1-alpha)*100% confident interval for B1 This does not mean that a simultaneous (joint) confidence region for the two parameters is rectangular. Actually, the simultaneous confidence region is elliptical.

```
CONL=B1-(2.179*seB1)
CONL
```

```
## [1] 14.40842

CONU=B1+(2.179*seB1)
CONU

## [1] 16.60913
```

Prediction

#d. If Y4 denotes the predicted value Two type of predictions: The prediction of the value of the response variable Y which corresponds to any chosen value, X0, of the predictor variable.

2. The estimation of the mean response µ0, when X = x0. Predicted value y^0= β0^ + **β**1X^0

```
              s.e.(y^0) =σ^√(1/n+x/∑(x0-x̅)2/∑(xi-x̅)2)
```

the confidence limits for the predicted value with confidence coefficient (1 - α) are given by Y0^ ± t(n-2,α/2)×s.e.(Y0^)
for forcast use or prediction the second case, the mean response 110 is estimated by

```
        µ0= β0^ + �1µ^0
```

The standard error of this estimate

```
       s.e.( µ0^)=σ^√(1/n+(x0-x̅)2/∑(xi-x̅)2)
```

that the confidence limits for 110 with confidence coefficient (1 - a) are given by µ0^ ± t(n-2,α/2)×s.e.(µ0^)

for confident limits Note that the point estimate of µ0 is identical to the predicted response Y0. This can be seen by comparing. The standard error of µ0 is, however, smaller than the standard error of y0 can be seen by comparing. Intuitively, this makes sense. There is greater uncertainty (variability) in predicting one observation (the next observation) than in estimating the mean response when X = X0. The averaging that is implied in the mean response reduces the variability and uncertainty associated with the estimate.

```
y4=B0+(B1*4)
y4

## [1] 66.19674

sey4=sigma*sqrt(1+1/14+(4-mean(U))^2/s)
sey4

## [1] 5.671614
```

```
mu4=B0+(B1*4)
mu4

## [1] 66.19674

semu4=sigma*sqrt(1/14+(4-mean(U))^2/s)
semu4

## [1] 1.759688
```

#Measuring The Quality of Fit #Measuring The Quality of Fit we are interested not only in knowing whether a linear relationship exits, but also in measuring the quality of the fit of the model to the data.

1:-The larger the t (in absolute value) or the smaller the corresponding p-value, the stronger the linear relationship between Y and X. assumption of normality of the ϵ's.

2:-correlation coefficient Cor(Y, X), the closer the set of points to a straight line .

3:- Examine the scatter plot of Y versus Y^. measure the strength of the linear relationship

```
   Cor(Y, Y^) =Σ(Yi-Y—)(Y^i-Y^-)/√Σ(Yi-Y—)2Σ(Y^i-Y^-)2
```

n fact, the scatter plot of Y versus X and the scatter plot of Y versus Y^ are redundant because the patterns of points in the two graphs are identical in simple linear regression, the scatter plot of Y versus Y^ is redundant.But in multiple regression is not redundant.

```
                               Cor(Y, Y^) |Cor(Y, X)|  ;         -1≤Cor(Y,X)≤
1

                         Total sum of square deviation=SST= Σ(Yi-Y—)2  ;

                         Sum of square due to regression=SSR= Σ(Y^i-Y^-)2

                         Sum of square residuals(errors)=SSE=  Σ(Yi-Y^-)2

        SST = SSR + SSE    ; R2=SSR/SST=1-SSE/SST ;

                                             R2=  [Cor(Y,X)]2=[Cor(Y,Y^)]2

    in simple linear regression, R2 is equal to the square of the correlation
coefficient .
                                              0≤    R2 ≤ 1 since SSE≤SST .

              If R2 is near 1,this reason it is known as coefficient of
determination.


COR=cor(M,minutes)
COR
```

```
## [1] 0.9936987

ycapbar=mean(minutes)
ycapbar

## [1] 97.21427

fit=(minutes-ycapbar)
fit

##  [1] -77.54385 -62.03508 -46.52631 -31.01754 -31.01754 -15.50877   0.00000
##  [8]   0.00000  15.50877  31.01754  46.52631  46.52631  62.03508  62.03508

sm=sum(fit*fit)
sm

## [1] 27419.5

sy=sum(yi_ybar*yi_ybar)
sy

## [1] 27768.36

COR1=sum(yi_ybar*(fit))/sqrt((sm)*sy)
COR1

## [1] 0.9936987

SST=sum((M-mean(M))^2)
SST

## [1] 27768.36

SSR=sum((minutes-mean(M))^2)
SSR

## [1] 27419.5

SSE=sum((resi)^2)
SSE

## [1] 348.8484

SST=SSR+SSE
SST

## [1] 27768.35

Rsqr=SSR/SST
Rsqr

## [1] 0.9874372
```

```
rsqr=1-(SSE/SST)
rsqr

## [1] 0.9874372
```

#R^2 =0.987 indicates that nearly 99% of total variability in response variable(Minutes) is accounted for the predictor variable(Units).

```
Rr=COR^2
Rr

## [1] 0.9874372

r=cor_1*cor_1
r

## [1] 0.9874372
```