بسم الله الرحمن الرحيم=

STAT 332

Regression Analysis

Prepared by

Abdulrahman Alfaifi

King Saud University

Department of Statistics and Operation Research

alfaifi@ksu.edu.sa

@AlfaifiStat

A Alfaifi

ملاحظة : ليس بالضرورة ان تكون المذكرة شاملة للمقرر

9 Jan 2023

Proofs:

Least Square estimators of β_0 and β_1

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \Rightarrow \quad \varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

$$Q = \sum_{i=1}^{n} {\varepsilon_i}^2 = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0$$

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0$$

$$\sum_{i=1}^{n} Y_i - nb_0 - b_1 \sum_{i=1}^{n} X_i = 0 \cdots \boxed{1}$$

$$\frac{\partial Q}{\partial \beta_1} = 0 \Rightarrow -2X_i \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) \quad = 0$$

$$\Rightarrow X_i \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) \quad = 0$$

$$\sum_{i=1}^{n} Y_i X_i - b_0 \sum_{i=1}^{n} X_i - b_1 \sum_{i=1}^{n} {X_i}^2 = 0 \cdots \boxed{2}$$

<u>To find an estimator for $\boldsymbol{\beta_1}$</u>

$$\sum_{i=1}^{n} Y_i - nb_0 - b_1 \sum_{i=1}^{n} X_i = 0 \cdots\cdots \boxed{1}$$

$$\Rightarrow -\left(\sum_{i=1}^{n} Y_i\right)\left(\sum_{i=1}^{n} X_i\right) + nb_0\left(\sum_{i=1}^{n} X_i\right) + b_1\left(\sum_{i=1}^{n} X_i\right)^2 = 0$$

$$\sum_{i=1}^{n} Y_i X_i - b_0 \sum_{i=1}^{n} X_i - b_1 \sum_{i=1}^{n} {X_i}^2 = 0 \cdots\cdots \boxed{2}$$

$$\Rightarrow n \sum_{i=1}^{n} Y_i X_i - nb_0 \sum_{i=1}^{n} X_i - nb_1 \sum_{i=1}^{n} {X_i}^2 = 0$$

<u>By Adding</u>

$$n \sum_{i=1}^{n} Y_i X_i - \left(\sum_{i=1}^{n} Y_i\right)\left(\sum_{i=1}^{n} X_i\right) + b_1\left(\sum_{i=1}^{n} {X_i}^2 - n\sum_{i=1}^{n} {X_i}^2\right) = 0$$

$$b_1\left(\sum_{i=1}^{n} {X_i}^2 - n\sum_{i=1}^{n} {X_i}^2\right) = -n \sum_{i=1}^{n} Y_i X_i + \left(\sum_{i=1}^{n} Y_i\right)\left(\sum_{i=1}^{n} X_i\right)$$

$$\boxed{b_1 = \frac{-n\sum_{i=1}^{n} Y_i X_i + \left(\sum_{i=1}^{n} Y_i\right)\left(\sum_{i=1}^{n} X_i\right)}{\sum_{i=1}^{n} {X_i}^2 - n\sum_{i=1}^{n} {X_i}^2} = \frac{n\sum_{i=1}^{n} Y_i X_i - \left(\sum_{i=1}^{n} Y_i\right)\left(\sum_{i=1}^{n} X_i\right)}{n\sum_{i=1}^{n} {X_i}^2 - \sum_{i=1}^{n} {X_i}^2}}$$

<u>To find an estimator for $\boldsymbol{\beta_0}$ $(\boldsymbol{from\ 1})$</u>

$$\sum_{i=1}^{n} Y_i - nb_0 - b_1 \sum_{i=1}^{n} X_i = 0$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n} nb_0 - b_1 \frac{1}{n}\sum_{i=1}^{n} X_i = 0$$

$$\Rightarrow \qquad \bar{Y} - b_0 - b_1 \bar{X} = 0$$

$$\boxed{b_0 = \bar{Y} - b_1 \bar{X}}$$

Least Square estimators of β_1 when $\beta_0 = 0$

$$Y_i = \beta_1 X_i + \varepsilon_i \Rightarrow \varepsilon_i = Y_i - \beta_1 X_i$$

$$Q = \sum_{i=1}^{n} \varepsilon_i{}^2 = \sum_{i=1}^{n} (Y_i - \beta_1 X_i)^2$$

$$\frac{\partial Q}{\partial \beta_1} = 0$$

$$\Rightarrow -2\sum_{i=1}^{n} X_i\left(Y_i - \hat{\beta}_1 X_i\right) = 0$$

$$\Rightarrow \quad \sum_{i=1}^{n} X_i\left(Y_i - \hat{\beta}_1 X_i\right) = 0$$

$$\Rightarrow \quad \sum_{i=1}^{n} X_i Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i{}^2 = 0$$

$$b_1 = \hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i{}^2}$$

Unbiasedness of β_1 : $(E(b_1) = \beta_1)$

$$b_1 = \frac{n\sum_{i=1}^{n} Y_i X_i - (\sum_{i=1}^{n} Y_i)(\sum_{i=1}^{n} X_i)}{n\sum_{i=1}^{n} X_i{}^2 - \sum_{i=1}^{n} X_i{}^2}$$

$$= \frac{\sum_{i=1}^{n} Y_i X_i - \bar{X}(\sum_{i=1}^{n} Y_i)}{\sum_{i=1}^{n} X_i{}^2 - \frac{1}{n}(n^2\bar{X}^2)}$$

$$= \frac{\sum_{i=1}^{n}(Y_i X_i - \bar{X} Y_i)}{\sum_{i=1}^{n} X_i{}^2 - (n\bar{X}^2)}$$

$$= \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$= \sum_{i=1}^{n} k_i Y_i \qquad ; k_i = \frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$i. \ \sum_{i=1}^{n} k_i = \frac{\sum_{i=1}^{n} X_i - n\bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{n\bar{X} - n\bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = 0$$

$$ii. \ \sum_{i=1}^{n} k_i{}^2 = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2} = \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{1}{S_{xx}}$$

$$iii. \ \sum_{i=1}^{n} X_i k_i = \sum_{i=1}^{n} X_i\left(\frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right) = \frac{\sum_{i=1}^{n} X_i{}^2 - (n\bar{X}^2)}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} X_i{}^2 - (n\bar{X}^2)}{\sum_{i=1}^{n} X_i{}^2 - (n\bar{X}^2)} = 1$$

$$E(b_1) = E(\sum_{i=1}^{n} k_i Y_i) = \sum_{i=1}^{n} E(k_i Y_i)$$

$$= \sum_{i=1}^{n} k_i E(Y_i) = \sum_{i=1}^{n} k_i(\beta_0 + \beta_1 X_i)$$

$$= \beta_0 \sum_{i=1}^{n} k_i + \beta_1 \sum_{i=1}^{n} X_i k_i$$

$$= \beta_0(0) + \beta_1(1) = \beta_1$$

$Var(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

$Var(b_1) = Var(\sum_{i=1}^{n} k_i Y_i)$

$= \sum_{i=1}^{n} Var(k_i Y_i)$

$= \sum_{i=1}^{n} {k_i}^2 Var(Y_i)$

$= \sum_{i=1}^{n} {k_i}^2 \sigma^2$

$= \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

Unbiasedness of β_0 *:*$(E(b_0) = \beta_0)$

$\beta_0 = \bar{Y} - b_1\bar{X}$

$= \left(\frac{\sum_{i=1}^{n} Y_i}{n}\right) - (\sum_{i=1}^{n} k_i Y_i)\bar{X}$

$= \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}k_i\right)Y_i$

$= \sum_{i=1}^{n} a_i Y_i \qquad ; a_i = \frac{1}{n} - \bar{X}k_i$

i. $\sum_{i=1}^{n} a_i = \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}k_i\right) = \sum_{i=1}^{n}\left(\frac{1}{n}\right) - \sum_{i=1}^{n}(\bar{X}k_i) = 1 - \bar{X}\sum_{i=1}^{n}(k_i) = 1 - 0 = 1$

ii. $\sum_{i=1}^{n} {a_i}^2 = \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}k_i\right)^2 = \sum_{i=1}^{n}\left(\frac{1}{n^2} - \frac{2\bar{X}k_i}{n} + \bar{X}^2{k_i}^2\right)$

$= \sum_{i=1}^{n}\left(\frac{1}{n^2}\right) - \sum_{i=1}^{n}\left(\frac{2\bar{X}k_i}{n}\right) + \sum_{i=1}^{n}\left(\bar{X}^2{k_i}^2\right)$

$= \frac{1}{n} - 0 + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

iii. $\sum_{i=1}^{n} a_i X_i = \sum_{i=1}^{n}\left(\frac{X_i}{n} - \bar{X}k_i X_i\right) = \bar{X} - \bar{X}\sum_{i=1}^{n}(k_i X_i) = \bar{X} - \bar{X}(1) = 0$

$E(b_0) = E(\sum_{i=1}^{n} a_i Y_i) = \sum_{i=1}^{n} a_i\, E(Y_i)$

$= \sum_{i=1}^{n} a_i\,(\beta_0 + \beta_1 X_i)$

$= \sum_{i=1}^{n} \beta_0 a_i + \beta_1 \sum_{i=1}^{n} a_i\, X_i$

$= \beta_0 \sum_{i=1}^{n} a_i + \beta_1 \sum_{i=1}^{n} a_i\, X_i$

$= \beta_0(1) + \beta_1(0) = \beta_0$

$Var(b_0) = MSE \times \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$

$Var(b_0) = Var(\sum_{i=1}^{n} a_i Y_i)$

$= \sum_{i=1}^{n} Var(a_i Y_i)$

$= \sum_{i=1}^{n} a_i{}^2 Var(Y_i)$

$= \sigma^2 \sum_{i=1}^{n} a_i{}^2$

$= \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)$

$= \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right)$

The sum of the weighted residuals is zero when the residual in the ith trial is weighted by the level of the predictor variable in the ith trial

$$\sum_{i=1}^{n} e_i X_i = 0$$

$\sum_{i=1}^{n} e_i X_i = \sum_{i=1}^{n}\big(Y_i - \hat{Y}_i\big)X_i$

$= \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)X_i$

$= \sum_{i=1}^{n}(Y_i - (\bar{Y} - b_1\bar{X}) - b_1 X_i)X_i$

$= \sum_{i=1}^{n}(Y_i - \bar{Y} - b_1\bar{X} - b_1 X_i)X_i$

$= \sum_{i=1}^{n}\big(Y_i - \bar{Y} - b_1(X_i - \bar{X})\big)X_i$

$= \sum_{i=1}^{n}\left(X_i Y_i - \bar{Y}X_i - b_1\big(X_i{}^2 - X_i\bar{X}\big)\right)$

$= \sum_{i=1}^{n}(X_i Y_i) - \bar{Y}\sum_{i=1}^{n}(X_i) - b_1\sum_{i=1}^{n}\big(X_i{}^2 - X_i\bar{X}\big)$

$= \sum_{i=1}^{n}(X_i Y_i) - \left(\frac{\sum_{i=1}^{n} Y_i}{n}\right)\sum_{i=1}^{n}(X_i) - b_1\big(\sum_{i=1}^{n}\big(X_i{}^2\big) - \bar{X}\sum_{i=1}^{n}(X_i)\big)$

$= \sum_{i=1}^{n}(X_i Y_i) - \left(\frac{1}{n}\right)\sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i - b_1\left(\sum_{i=1}^{n}\big(X_i{}^2\big) - \left(\frac{1}{n}\right)\sum_{i=1}^{n} X_i{}^2\right)$

$= \sum_{i=1}^{n}(X_i Y_i) - \left(\frac{1}{n}\right)\sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i - \left(\frac{\sum_{i=1}^{n} Y_i X_i - \left(\frac{1}{n}\right)\sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i{}^2 - \left(\frac{1}{n}\right)\sum_{i=1}^{n} X_i{}^2}\right)\left(\sum_{i=1}^{n} X_i{}^2 - \left(\frac{1}{n}\right)\sum_{i=1}^{n} X_i{}^2\right) = 0$

The sum of the weighted residuals is zero when the residual in the ith trial is weighted by the level of the response variable in the ith trial

$$\sum_{i=1}^{n} e_i \hat{Y}_i = 0$$

$\sum_{i=1}^{n} e_i \hat{Y}_i = \sum_{i=1}^{n} e_i(b_0 + b_1 X_i)$

$= \sum_{i=1}^{n}(b_0 e_i + b_1 X_i e_i)$

$= \sum_{i=1}^{n}(b_0 e_i) + \sum_{i=1}^{n}(b_1 X_i e_i)$

$= b_0\sum_{i=1}^{n}(e_i) + b_1\sum_{i=1}^{n}(X_i e_i)$

$= b_0(0) + b_1(0) = 0$

خط الانحدار يمر بالنقطة $(\bar{X}, \bar{Y})$:

$$\hat{Y} = b_0 + b_1 X$$
$$= \bar{Y} - b_1\bar{X} + b_1 X$$
$$= \bar{Y} - b_1(\bar{X} - X)$$
$$= \bar{Y} - b_1(\bar{X} - \bar{X}) \qquad (When\ X = \bar{X})$$
$$= \bar{Y} - 0$$
$$= \bar{Y}$$

وبالتالي خط الانحدار يمر بالنقطة $(\bar{X}, \bar{Y})$

The sum of residuals is equal to zero ($\sum_{i=1}^{n} \boldsymbol{e_i} = \mathbf{0}$)

$$e_i = Y_i - \hat{Y}_i = Y_i - b_0 - b_1 X_i$$

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)$$
$$= \sum_{i=1}^{n}(Y_i - (\bar{Y} - b_1\bar{X}) - b_1 X_i)$$
$$= \sum_{i=1}^{n} Y_i - n\bar{Y} + nb_1\bar{X} - b_1 \sum_{i=1}^{n} X_i$$
$$= n\bar{Y} - n\bar{Y} + nb_1\bar{X} - nb_1\bar{X} = 0$$

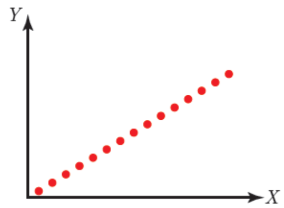
		$Variance$	$C.I$	$T.S$
b_1	$= \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$ $= \frac{\sum x_i y_i - \bar{x}\sum y_i}{\sum (x_i - \bar{x})^2}$ $= \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}$ $= \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$ $\boxed{\hat{\beta}_{2\times1} = (X'X)^{-1}X'Y}$	$\frac{MSE}{S_{xx}}$ $\frac{MSE}{\sum x_i^2 - n\bar{x}^2}$ $\boxed{Var(\hat{\beta}) = MSE(X'X)^{-1}}$	$b_1 \pm t_{(1-\alpha/_2, n-2)}\ S(b_1)$ $\boxed{b_i \pm t_{(1-\alpha/_2, n-p)}\ S(b_i)}$	$\frac{b_1 - \beta_1^{(0)}}{S(b_1)}$
b_o	$= \bar{y} - b_1\bar{x}$	$MSE \times \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$	$b_o \pm t_{(1-\alpha/_2, n-2)}\ S(b_o)$	$\frac{b_0 - \beta_0^{(0)}}{S(b_0)}$
Y_h	$\hat{y}_h = b_o + b_1 x_h$ $\boxed{\hat{Y}_h = X'_h\ b}$ $\boxed{X'_h = \{1 \quad X_h\}}$	$MSE \times \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right)$ $\boxed{\begin{gathered} = MSE(X'_h(X'X)^{-1}X_h) \\ = X'_h\ S^2\{b\}\ X_h \end{gathered}}$	$\hat{y}_h \pm t_{(1-\alpha/_2, n-2)}\ S(\hat{y}_h)$ $\boxed{\hat{y}_h \pm t_{(1-\alpha/_2, n-p)}\ S(\hat{y}_h)}$	
$Y_{h(new)}$		$MSE \times \left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right)$ $= MSE + S^2(\hat{y}_h)$	$\hat{y}_h \pm t_{(1-\alpha/_2, n-2)}\ S(\hat{y}_{h(new)})$ $\boxed{\hat{y}_h \pm t_{(1-\alpha/_2, n-p)}\ S(\hat{y}_{h(new)})}$	

ANOVA:

مصدر الخطأ	SS	df	MS	F^*
خطأ الانحدار $Regression$	$SSR = \Sigma(\hat{y}_i - \bar{y})^2 = b_1^2 S_{xx}$ $\boxed{\begin{gathered} = b'X'Y - \left(\frac{1}{n}\right)Y'JY \\ = Y'\left(H - \left(\frac{1}{n}\right)J\right)Y \end{gathered}}$ $\boxed{H = x(x'x)^{-1}x'}$	1 $\boxed{p-1}$	$MSR = \frac{SSR}{1}$ $\boxed{MSR = \frac{SSR}{p-1}}$	$\frac{MSR}{MSE}$
خطأ عشوائي $Error$	$SSE = \Sigma(y_i - \hat{y}_i)^2$ $= \Sigma y_i^2 - b_o \Sigma y_i - b_1 \Sigma x_i y_i$ $= S_{yy} - \frac{S_{xy}^2}{S_{xx}}$ $= \left[\Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}\right] - \frac{\left(\Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n}\right)^2}{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}}$ $\boxed{\begin{gathered} = e'e = Y'Y - b'X'Y \\ = Y'(I - H)Y \end{gathered}}$	$n-2$ $\boxed{n-p}$	$MSE = \frac{SSE}{n-2}$ $\boxed{MSE = \frac{SSE}{n-p}}$	
خطأ كلي $Total\ error$	$SST = \Sigma(y_i - \bar{y})^2 = S_{yy}$ $\boxed{\begin{gathered} = Y'Y - \left(\frac{1}{n}\right)Y'JY \\ = Y'\left(I - \left(\frac{1}{n}\right)J\right)Y \end{gathered}}$ $SST = SSR + SSE$	$n-1$		

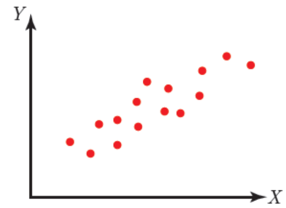
$Coefficient\ of\ Determination = R^2 = \frac{SSR}{SST}$, $Coefficient\ of\ Correlation = r_{xy} = \pm\sqrt{R^2} = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}} = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\sqrt{\Sigma x_i^2 - n\bar{x}^2}\sqrt{\Sigma y_i^2 - n\bar{y}^2}}$

- ***Correlation:***



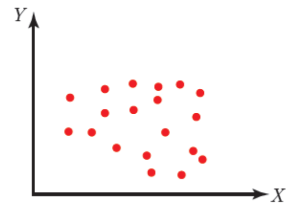
Maximum positive correlation

($\mathbf{r}$ = 1.0)



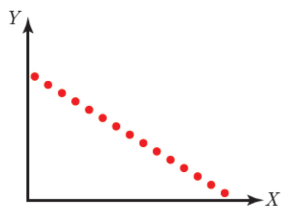
Strong positive correlation

($\mathbf{r}$ = 0.80)



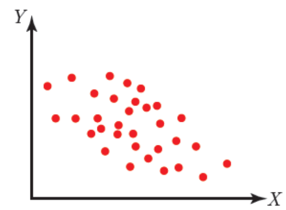
Very Weak correlation

($\mathbf{r}$ = 0.25)



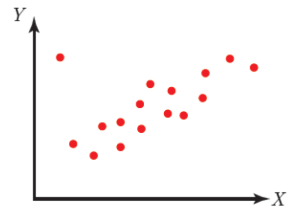
Maximum negative correlation

($\mathbf{r}$ = -1.0)



Weak negative correlation

($\mathbf{r}$ = -0.45)



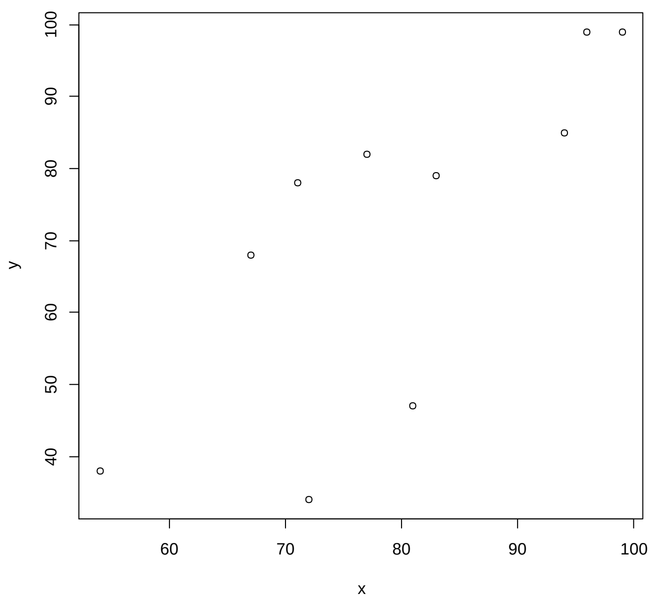
Strong correlation & outlier

($\mathbf{r}$ = 0.7)

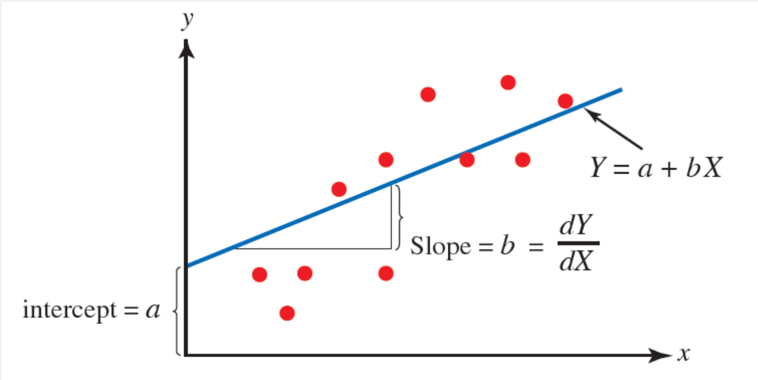
Example 1:

The results of a class of 10 students on <u>midterm marks (X)</u> and on the <u>final marks (Y)</u> are as follows:

i	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Midterm (X)	*77*	*54*	*71*	*72*	*81*	*94*	*96*	*99*	*83*	*67*
Final (Y)	*82*	*38*	*78*	*34*	*47*	*85*	*99*	*99*	*79*	*68*



By using regression, we can describe the relationship between these variables



$$Y = a + bX \implies \boxed{Y = \beta_0 + \beta_1 X}$$

i	*Midterm* (X)	*Final* (Y)	X^2	Y^2	XY
1	*77*	*82*	*5929*	*6724*	*6314*
2	*54*	*38*	*2916*	*1444*	*2052*
3	*71*	*78*	*5041*	*6084*	*5538*
4	*72*	*34*	*5184*	*1156*	*2448*
5	*81*	*47*	*6561*	*2209*	*3807*
6	*94*	*85*	*8836*	*7225*	*7990*
7	*96*	*99*	*9216*	*9801*	*9504*
8	*99*	*99*	*9801*	*9801*	*9801*
9	*83*	*79*	*6889*	*6241*	*6557*
10	*67*	*68*	*4489*	*4624*	*4556*
total	*794*	*709*	*64862*	*55309*	*58567*
mean	*79.4*	*70.9*			

$$E(\beta_1) = b_1 = \frac{\sum x_i y_i - \bar{x}\sum y_i}{\sum x_i^2 - n\bar{x}^2} = \frac{58567 - 79.4(709)}{64862 - 10(79.4)^2} = 1.24967$$

$$E(\beta_0) = b_0 = \bar{y} - b_1\bar{x} = 70.9 - 1.24967(79.4) = -28.3238$$

The regression model is $\boxed{\hat{Y} = -28.32 + 1.25\,X}$

If a student got in his midterm 50 what the expected mark in his final exam?

$$\hat{Y} = -28.32 + 1.25\,X = -28.32 + 1.25\,(50) = 34.18$$

By using R:

```
> x=c(77,54,71,72,81,94,96,99,83,67)
> y=c(82,38,78,34,47,85,99,99,79,68)

> model=lm(y~x)
> model

Call:
lm(formula = y ~ x)
Coefficients:
  (Intercept)          x
     -28.32          1.25
```

$\boxed{\hat{Y} = -28.32 + 1.25\,X}$

- ***Interpret the coefficients:***

b_1 = *The changes in value of* y *when* x *increase by one unit.*
b_o = *The value of* y *when* $x = 0$ *or the intersection with* y *axis.*

b_1 = *The changes in final mark when midterm mark increase by one mark.*
b_o = *The final mark when midterm mark =0 or the intersection with* y *axis.*

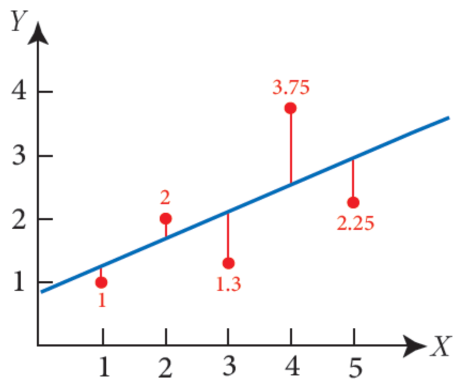
Example 2:

the data given below

X	*1*	*2*	*3*	*4*	*5*
Y	*1*	*2*	*1.3*	*3.75*	*2.25*

Find the regression line

$$\boxed{\hat{Y} = 0.785 + 0.425\, X}$$



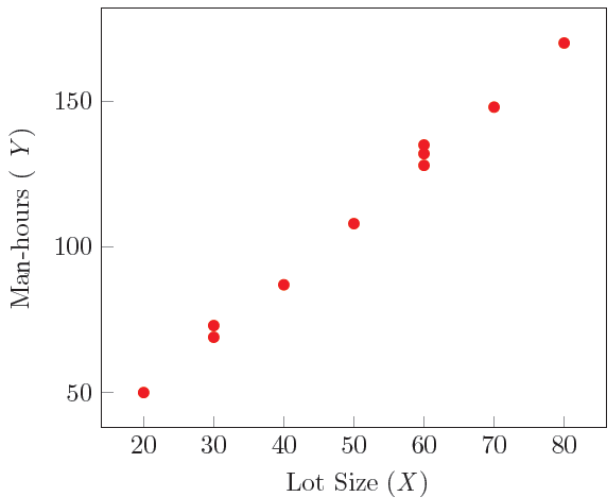
Example 3:

A certain spare part is manufactured by Westwood Company once a month in lots which vary in size as demand fluctuates. Let X represents the and Y the number of Man-hours labour for recent production runs. The data is given in the table below.

i	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Lot size (X)	*30*	*20*	*60*	*80*	*40*	*50*	*60*	*30*	*70*	*60*
Man-hours (Y)	*73*	*50*	*128*	*170*	*87*	*108*	*135*	*69*	*148*	*132*

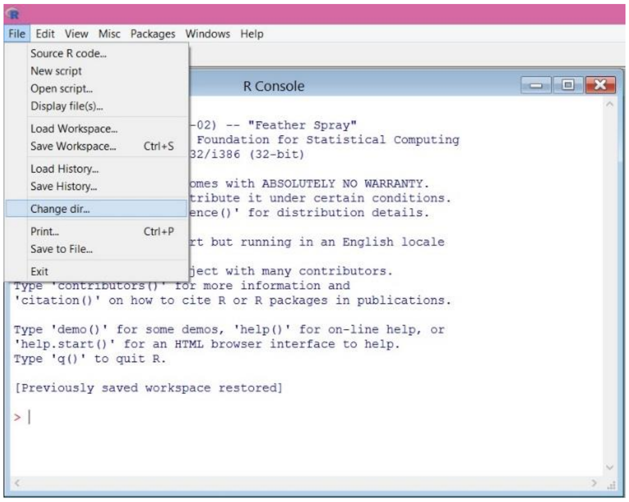
Find the regression line

$$\boxed{\hat{Y} = 10 + 2\,X}$$



- *Reading files in R (Excel, txt, …) :*

File → **Change dir …**



Excel

Text

> data=read.csv("filename.csv") > data x y 1 71 82 2 64 91 3 43 100 4 67 68 5 56 87 6 73 73 7 68 78 8 56 80 9 76 65 10 65 84 11 45 116 12 58 76 13 45 97 14 53 100 15 49 105 16 78 77	> data=read.table("filename.txt",header=TRUE) > data x y 1 71 82 2 64 91 3 43 100 4 67 68 5 56 87 6 73 73 7 68 78 8 56 80 9 76 65 10 65 84 11 45 116 12 58 76 13 45 97 14 53 100 15 49 105 16 78 77

```
>x=data$x
> x
 [1] 71 64 43 67 56 73 68 56 76 65 45 58 45 53 49 78
> y=data$y
> y
 [1]  82  91 100  68  87  73  78  80  65  84 116  76  97 100 105  77
> model=lm(y~x)
> model

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)           x
   148.051       -1.024
```

لحساب القيمة $t_{(1-\alpha/2, n-2)}$ باستخدام R

$$= t_{(1-\ 0.05/2\ ,\ 16-2)}$$
$$= t_{(0.975\ ,\ 14)}$$
$$= 2.145$$

```
> qt(0.975,14)
[1]   2.144787
```

لحساب القيمة $F_{(1-\alpha\ ,1\ ,n-2)}$ باستخدام R

$$= F_{(1-\ 0.05\ ,1\ ,\ 16-2)}$$
$$= F_{(0.95\ ,1\ ,14)}$$
$$= 4.60011$$

```
> qf(0.95,1,14)
[1] 4.60011
```

Question 1:

Consider a company that markets and repairs small computers. 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 calls was taken. The data consist of the length of service calls in minutes (the response variable) and the number of components repaired (the predictor variable). The data are presented in the table below:

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

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.162      3.355    1.24    0.239
x             15.509      0.505   30.71 8.92e-13 ***
---

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

Analysis of Variance Table

Response: y
          Df     Sum Sq     Mean Sq     F value     Pr(>F)
x          1     27419.5    27419.5     943.2       8.916e-13 ***
Residuals 12     348.8      29.1
```

(a) Estimate the regression line and interpret the coefficients.

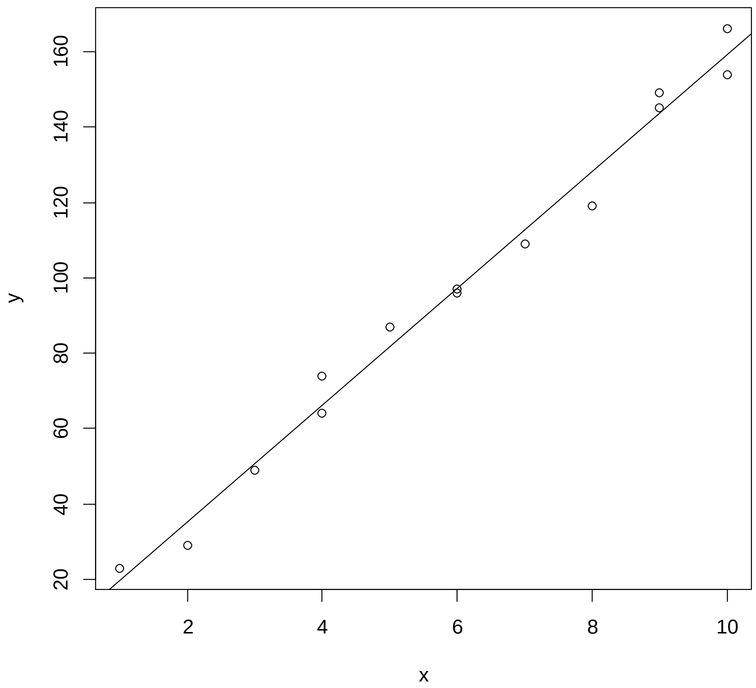
$$\hat{y} = b_o + b_1 x$$

$$\hat{y} = 4.16 + (15.51)x$$
$$(\widehat{service\ time}) = 4.16 + (15.51)(\#of\ units)$$

$b_1 =$ *The changes in service time when number of units increase by one.*

$b_o =$ *The intersection with y axis.*

```
> y=c(23,29,49,64,74,87,96,97,109,119,149,145,154,166)
>  x=c(1,2,3,4,4,5,6,6,7,8,9,9,10,10)
> model=lm(y~x)
> summary(model)
> with(plot(x,y),abline(model))
```



(b) Construct 90% confidence intervals for the model coefficients and explain the results.

```
> confint(model,level=0.90)
                  5 %      95 %
(Intercept)    -1.81810   10.14141
     x         14.60875   16.40879
```

$\beta_0 \in (-1.82 , 10.14\ )$ $\qquad$ $\beta_1 \in (14.61 , 16.41\ )$

We are 90% sure that, when the number of units increase by one the service time increase somewhere between $(\ 14.41\ , 16.61\ )$

(c) Test the linearity by using two different approaches.

1. *Using T-test :*

$$H_0: \beta_1 = 0 \quad vs \quad H_a: \beta_1 \neq 0$$

Since p-value = 0.000 <0.05 then the decision:
We reject H_0 (there is a linear association).

2. *Using F-test :*

$$H_0: \beta_1 = 0 \quad vs \quad H_a: \beta_1 \neq 0$$

Since p-value = 0.000 <0.05 then the decision:
We reject H_0 (there is a linear association).

(d) Calculate the residual at Units=4 and Minutes=64

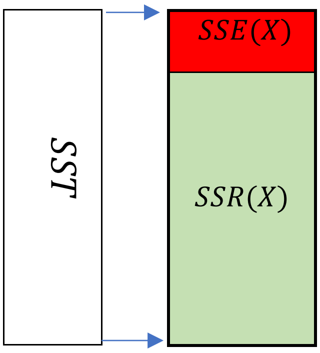
```
> summary(model)$res
        1          2          3          4          5          6          7
3.3295739 -6.1791980 -1.6879699 -2.1967419  7.8032581  5.2944862 -1.2142857
        8          9         10         11         12         13         14
-0.2142857 -3.7230576 -9.2318296  5.2593985  1.2593985 -5.2493734  6.7506266
```

(e) Estimate the standard deviation of the residuals.

$$S = \sqrt{MSE} = \sqrt{29.1} = 5.39$$

ANOVA:

	df	SS	MS	F
Regression (R)	1	SSR	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Error (E)	$n-2$	SSE	$MSE = \frac{SSE}{n-2}$	
Total (T)	$n-1$	SST		



Question 2:

A linear regression was run on a set of data. You are given only the following partial information:

Predictor	Coef	SE Coef	T
Constant	293.89	5.62	$\frac{293.89}{5.62} = 52.29$
X	$0.13 \times -13.13 = -1.7069$	0.13	-13.13

Analysis of Variance

Source	DF	SS	MS	F
Regression	1	7621.667	$172.3969 \times 44.21 = 7621.667$	$(-13.13)^2 = 172.3969$
Residual Error	5	$5 \times 44.21 = 221.05$	44.21	
Total	6	$7621.667 + 221.21 = 7842.717$		

(a) Compute the 90% Confidence intervals for β_0 and β_1

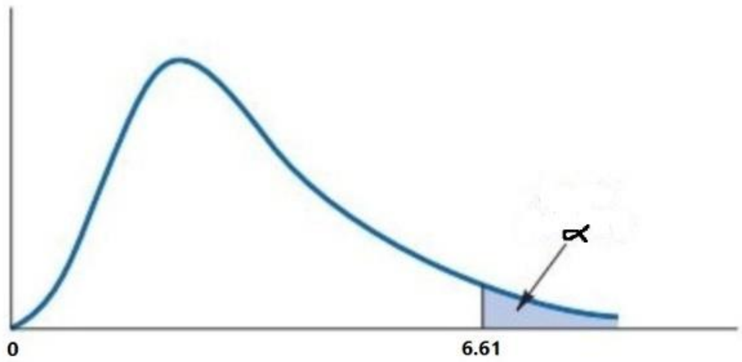
$$\left(b_0 \pm t_{(1-0.1/2, 7-2)}\ S(b_0)\right) \qquad \left(b_1 \pm t_{(1-0.1/2, 7-2)}\ S(b_1)\right)$$

$$(293.89 \pm 2.015 \times\ 5.62) \qquad (-1.7069 \pm 2.015 \times\ 0.13)$$

$$(282.5657\,, 305.2143) \qquad (-1.96885\,, -1.44495)$$

(b) Give the F-statistic and test $H_0: \beta_1 = 0 \ \ vs \ \ H_1: \beta_1 \neq 0$

$$F - statistics = 172.3969$$



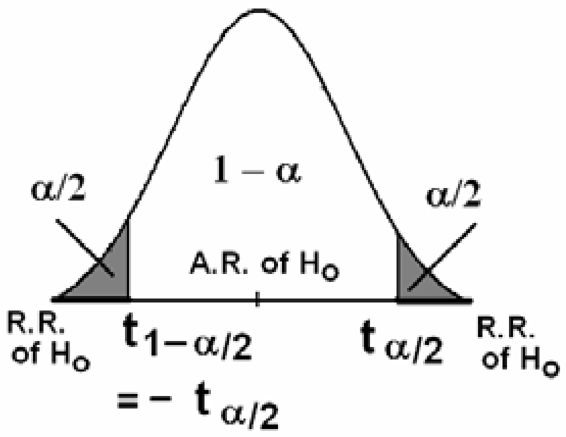
$$F_{(1-\alpha, 1, n-2)} = F_{(0.95, 1, 5)} = 6.60789 < 172.3969$$

$$We\ reject\ H_0$$

(c) *Test* $H_0: \beta_0 = 0 \;\; vs \;\; H_1: \beta_0 \neq 0$

$$T = \frac{b_0}{S(b_0)} = 52.2936$$

$$t_{1-\frac{\alpha}{2},n-2} = t_{0.975\,,5} = 2.57058$$



The decision: we reject H_0

(d) *Compute the coefficient of determination and hence the correlation coefficient.*

$$R^2 = \frac{SSR}{SST} = \frac{7621.667}{7842.717} = 0.9718 \;\; \Longrightarrow \;\; r = -\sqrt{0.9718} = -0.9858$$

b_1 اشارة

Problem 3:

The computer repair data gives the length of time of service calls in minutes (y) and the number of components repaired in a computer (x). Some summary measures for this data are:

$$\boldsymbol{n = 14 \qquad \sum x_i = 84 \qquad \sum y_i = 1361}$$
$$\boldsymbol{S_{XX} = 114 \quad S_{YY} = 27768.36 \quad S_{XY} = 1768}$$

a. ***Find the point estimate of the intercept and slop to model the length of service call as a linear function of the number of units serviced.***

$$\hat{y} = b_o + b_1 x$$

$$\boxed{\bar{y} = \frac{\sum y}{n} = \frac{1361}{14} = 97.21 \qquad \bar{x} = \frac{\sum x}{n} = \frac{84}{14} = 6}$$

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{1768}{114} = 15.51$$

$$\begin{aligned} b_0 &= \bar{y} - b_1 \bar{x} \\ &= 97.21 - 15.51 \times 6 \\ &= 4.16 \end{aligned}$$

$$\hat{y} = 4.16 + (15.51)x$$
$$(\widehat{time\ of service}) = 4.16 + (15.51)(\#of\ components)$$

b. *Show that the error sum of square can be written as*

$$SSE = S_{YY} - \hat{\beta}_1 S_{XY}$$

$$\begin{aligned} SSE &= S_{yy} - \frac{S_{xy}^2}{S_{xx}} \\ &= S_{yy} - \frac{S_{xy} S_{xy}}{S_{xx}} \\ &= S_{yy} - b_1 S_{xy} \end{aligned}$$

c. Give 95% confidence interval for the slop.

$$\left(b_1 \pm t_{(1-{\alpha}/{2},n-2)}\ S(b_1)\right)$$

- $SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 27768.36 - \frac{(1768)^2}{114} = 348.85$
- $MSE = \frac{SSE}{n-2} = \frac{348.85}{12} = 29.07$
- $Var(b_1) = \frac{MSE}{S_{xx}} = \frac{29.07}{114} = 0.255 \implies S(b_1) = 0.505$
- $t_{(1-{\alpha}/{2},n-2)} = t_{(1-{0.05}/{2},14-2)} = t_{(0.975,12)} = 2.178813$

$$\left(b_1 \pm t_{(1-{\alpha}/{2},n-2)}\ S(b_1)\right)$$

$$(15.51 \pm 2.178813 \times\ 0.505)$$

$$(\ 14.41\ , 16.61\ )$$

When the number of components increase by one the service time increase somewhere between $(\ 14.41\ , 16.61\ )$

Problem 4:

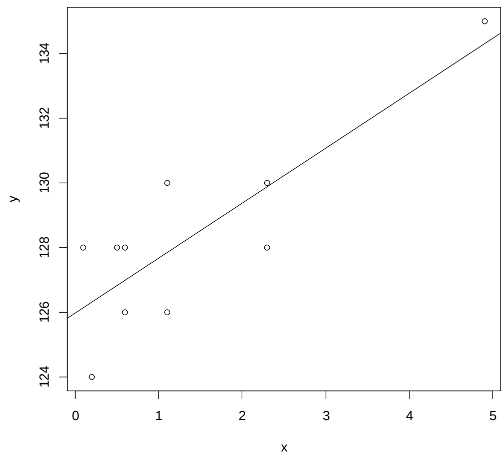
It is of interest to study the effect of population size in various cities in certain country on ozone concentration. The following data consists of the population in million and the amount of ozone present per hour in (parts per billion). The data is gives as follows.

i	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Ozone Y	*126*	*135*	*124*	*128*	*130*	*128*	*126*	*128*	*128*	*130*
Population X	*0.6*	*4.9*	*0.2*	*0.5*	*1.1*	*0.1*	*1.1*	*2.3*	*0.6*	*2.3*

a. *Fit the linear regression model relating ozone concentration to the population size and explain the estimated model.*

```
> y=c(126,135,124,128,130,128,126,128,128,130)
> x=c(0.6,4.9,0.2,0.5,1.1,0.1,1.1,2.3,0.6,2.3)
> model=lm(y~x)
> summary(model)
Coefficients:
             Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)  125.9677   0.7741       162.732    2.27e-15 ***
x            1.7024     0.3969       4.289      0.00266 **
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.742 on 8 degrees of freedom
Multiple R-squared:  0.6969,    Adjusted R-squared:  0.659
F-statistic: 18.39 on 1 and 8 DF,  p-value: 0.002656

> with(plot(x,y),abline(model))
```

$$\hat{y} = b_o + b_1 x$$
$$\hat{y} = 125.97 + (1.7)x$$
$$(\widehat{Ozone}) = 125.97 + (1.7)(Pop)$$



b_1 =the changes in Ozone concentration when the population increase by one million.

b_o =the Ozone concentration when the population =0, and its the intersection with y axis.

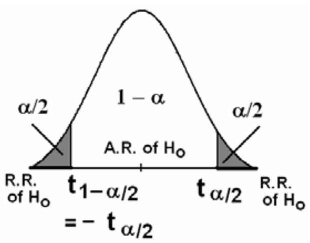
b. Test the hypothesis $H_0: \beta_1 = 0.$

Since p-value=0.00266 < 0.05

The decision: reject H_0 (there is positive linear association between population size and Ozone concentration).

c. Test the hypothesis $H_0: \beta_0 = 0.6$.

$$T = \frac{b_0 - \beta_0^{(0)}}{S(b_0)} = \frac{125.9677 - 0.6}{0.7741} = 161.95$$



$161.95 \notin (-2.306\,, 2.306)$

The decision: reject H_0

$$t_{1-\frac{\alpha}{2},n-2} = t_{0.975\,,8} = 2.306$$

d. *Construct 90% confidence interval for the coefficients.*

```
> confint(model,Level=0.90)
                  2.5 %       97.5 %
(Intercept)  124.1826711  127.752742
x              0.7870611    2.617747
```

$\boxed{\beta_0 \in (124.18\,, 127.75)}$

$\boxed{\beta_1 \in (0.787\quad, 2.618\;)}$

e. *Find the coefficient of determination and the correlation and interpret the result.*

$$R^2 = 0.6969 \qquad\qquad r = \sqrt{0.6969} = 0.8348$$

The model explain 69.69% of variation in Ozone concentration(Y) by using population size(X)

f. *Construct 95% confidence interval for the mean Y when X=11*

```
> newx=data.frame(x=11)
> predict(model,newx,level=0.95,int="confidence")
       fit      lwr      upr
1 144.6941 135.7883    153.6
```

$\boxed{Y \in (135.79\,, 153.6\;)}$

Problem 5:

Suppose a sample of size 12 is used to estimate a simple linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$ and obtain a 95% level confidence interval for the slope coefficient of (-0.045,-0.021). Based on the given information, complete the following statements (keep three decimal digits during the calculations):

$$\beta_1 \in (-0.045, -0.021) \quad , \quad \alpha = 0.05 \quad , \quad n = 12$$

(a) *The point estimate for the slope is:*

$$b_1 = \frac{(-0.045)+(-0.021)}{2} = -0.033$$

(b) *The standard error for the slope is:*

$$b_1 - t_{\left(1-\frac{\alpha}{2}, n-2\right)} \quad S(b_1) = -0.045$$

$$b_1 - t_{\left(1-\frac{0.05}{2}, 12-2\right)} \; S(b_1) = -0.045$$

$$-0.033 - (2.22814) \; S(b_1) = -0.045$$

$$-(2.22814) \; S(b_1) = -0.045 + 0.033$$

$$-(2.22814) \; S(b_1) = -0.012$$

$$\boxed{S(b_1) = 0.00539}$$

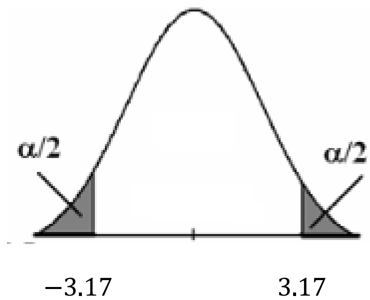
(c) *The value of the test statistic for testing the slope is equal to 0 is:*

$$T = \frac{b_1}{S(b_1)} = \frac{-0.033}{0.00539} = -6.122$$

(d) *The decision of the test in (c) at 1% level of significance is:*

$$H_0: \beta_1 = 0 \quad vs \quad H_1: \beta_1 \neq 0$$

$$\alpha = 0.01 \quad \Longrightarrow \quad t_{(1-\alpha/_2, n-2)} = t_{(0.995,10)} = 3.16927$$



$$T = -6.122 \notin (-3.17, 3.17) \quad \textit{Then we reject } H_0$$

(e) *The probability that the true (population) slope is between -0.045 and -0.021 is:*

0.95

Question:

An experiment is conducted, relating weekly sales for a food delivery (Y) service to the amount of advertising (X) during the week. The data is given below:

Week	*1*	*2*	*3*	*4*	*5*	*6*
X	*2*	*2*	*4*	*4*	*6*	*6*
Y	*20*	*30*	*40*	*50*	*70*	*60*

Calculate the confident coefficient when the CI of the slop is $(6.240199, 13.7598)$

```
> x=c(2,2,4,4,6,6)
> y=c(20,30,40,50,70,60)
> model=lm(y~x)
> summary(model)

Call:
lm(formula = y ~ x)

 Coefficients:
              Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   5.000      6.614        0.756     0.49177
x            10.000      1.531        6.532     0.00284 **
---

Residual standard error: 6.124 on 4 degrees of freedom
Multiple R-squared:  0.9143,    Adjusted R-squared:  0.8929
F-statistic: 42.67 on 1 and 4 DF,  p-value: 0.002838
```

$$b_1 \pm t_{(1-\alpha/_2, n-2)} \; S(b_1)$$

$$10 + t_{(1-\alpha/_2, 4)} \; (1.531) = 13.7598$$

$$t_{(1-\alpha/_2, 4)} = \frac{13.7598 - 10}{1.531} = 2.45578$$

```
> pt(2.45578,4)
[1] 0.9649958
```

$$\Longrightarrow 96.5\%$$

Question:

For the following data

i	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
X	*1*	*0*	*2*	*0*	*3*	*1*	*0*	*1*	*2*	*0*
Y	*16*	*9*	*17*	*12*	*22*	*13*	*8*	*15*	*19*	*11*

a. Obtain the regression function and interpret the coefficients.

```
> x=c(1,0,2,0,3,1,0,1,2,0)
> y=c(16,9,17,12,22,13,8,15,19,11)
> model=lm(y~x)
> model
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)           x
       10.2         4.0
```

$$\hat{y} = 10.2 + 4x$$

b_1 = *The changes in value of* y *when* x *increase by one unit.*

b_o = *The value of* y *when* $x = 0$ *or the intersection with* y *axis.*

b. Verify that fitted regression line goes through the point $(\bar{X}, \bar{Y})$

```
> x=c(1,0,2,0,3,1,0,1,2,0)
> y=c(16,9,17,12,22,13,8,15,19,11)
> mean(x)
[1] 1
> mean(y)
[1] 14.2
```

$$\hat{y} = 10.2 + 4x$$
$$\hat{y} = 10.2 + 4(1)$$
$$\hat{y} = 14.2$$

c. Conduct ANOVA table for testing the linearity of the model

```
> x=c(1,0,2,0,3,1,0,1,2,0)
> y=c(16,9,17,12,22,13,8,15,19,11)
> model=lm(y~x)
> anova(model)

Analysis of Variance Table
Response: y
          Df  Sum Sq  Mean Sq  F value  Pr(>F)
x          1  160.0   160.0    72.727   2.749e-05 ***
Residuals  8   17.6     2.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0: \beta_1 = 0$ *(The model is not linear)*

$H_1: \beta_1 \neq 0$ *(The model is linear)*

p – value = 2.749e-05 *** < 0.05, *then we reject* H_0

d. Obtain a 95% confident interval for β_1

```
> confint(model)
             2.5 %      97.5 %
(Intercept)  8.670370   11.729630
x            2.918388   5.081612
```

$$\beta_1 \in (2.91\ , 5.08)$$

e. Estimate the point estimator and 95% CI for the mean of Y for X=4, and interpret your result

```
> newx=data.frame(x=4)
> predict(model,newx,level=0.95,int="confidence")
  fit      lwr       upr
 26.2     22.77964   29.62036
```

The point estimate of Y when X=4 is 26.2

The 95% CI of Y when X=4 is (22.78 , 29.62)

Question:

A second-hand cars dealer has 10 cars for sale. He decides to investigate the relation between the cars age X (in years) and the milage Y (in thousands miles) by using the simple linear regression model. The dealer reported the following:

The mean and standard deviation of the cars milage are given, respectively, by 40.6 and 11.87153. The correlation coefficient between X and Y is 0.9687105. The estimated simple regression model is $\hat{Y} = 8.892 + 7.7337X$

$n = 10$	$\bar{Y} = 40.6$	$S_Y = 11.87153$	$r_{XY} = 0.9687105$	$b_0 = 8.892$	$b_1 = 7.7337$

a. Obtain S_{YY} *,* S_{XX} *and* S_{XY}

$$S_Y^2 = \frac{\sum(\bar{Y}_i - \bar{Y})^2}{n-1} = \frac{S_{YY}}{10-1} = (11.87153)^2 \implies S_{YY} = 9(11.87153)^2 \implies \boxed{S_{YY} = 1268.399}$$

- $b_1 = \frac{S_{XY}}{S_{XX}} \implies S_{XY} = b_1 S_{XX}$
- $r_{XY} = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}} \implies S_{XY} = r_{XY}\sqrt{S_{XX}}\sqrt{S_{YY}}$

$$b_1 S_{XX} = r_{XY}\sqrt{S_{XX}}\sqrt{S_{YY}}$$

$$7.7337 S_{XX} = 0.9687105\sqrt{S_{XX}}\sqrt{1268.399}$$

$$7.7337\sqrt{S_{XX}} = 0.9687105\sqrt{1268.399}$$

$$S_{XX} = \left(\frac{0.9687105\sqrt{1268.399}}{7.7337}\right)^2 \implies \boxed{S_{XX} = 19.9}$$

$$b_1 = \frac{S_{XY}}{S_{XX}} \implies S_{XY} = b_1 S_{XX} \implies S_{XY} = 7.7337(19.9) \implies \boxed{S_{XY} = 153.9}$$

b. *Construct 90% CI for the slop*

- $SSE = S_{YY} - \frac{S_{XY}^2}{S_{XX}} = 1268.399 - \frac{153.9^2}{19.9} = 78.187$
- $MSE = \frac{SSE}{n-2} = \frac{78.187}{8} = 9.773$
- $S(b_1) = \sqrt{\frac{MSE}{S_{XX}}} = \sqrt{\frac{9.773}{19.9}} = 0.7$

$$\left(b_1 \pm t_{(1-\alpha/_2, n-2)}\ S(b_1)\right)$$

$$\left(7.7337 \pm t_{(0.95,8)}\ 0.7\right)$$

$$(7.7337 \pm (1.859548)\ 0.7)$$

$$(6.432\ , 9.035)$$

c. *Compute the 95% CI for car milage with age 7 years.*

- $\hat{Y}_h = 8.892 + 7.7337X_h = 8.892 + 7.7337(7) \Longrightarrow \hat{Y}_h = 63.0279$
- $b_o = \bar{Y} - b_1\bar{X} \Longrightarrow 8.892 = 40.6 - 7.7337\bar{X} \quad \Longrightarrow \bar{X} = 4.1$
- $S(\hat{Y}_h) = \sqrt{MSE\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}}\right)} = \sqrt{9.773\left(\frac{1}{10} + \frac{(7-4.1)^2}{19.9}\right)} = 2.26$

$$\left(\hat{Y}_h \pm t_{(1-\alpha/_2, n-2)}\ S(\hat{Y}_h)\right)$$

$$\left(63.0279 \pm t_{(0.975,8)}\ 2.26\right)$$

$$(63.0279 \pm (2.306)\ 2.26)$$

$$(57.816\ ,68.239)$$

d. *Use ANOVA for testing the significance of the linearity*

	df	*SS*	*MS*	*F*
Regression (R)	1	$SSR = b_1^2 S_{XX}$ $= 1190.221302$	$MSR = \frac{1190.221302}{1}$	$F = 121.787$
Error(E)	$n - 2 = 8$	$SSE = 78.187$	$MSE = 9.773$	
Total (T)	$n - 1 = 9$	$SST = S_{YY} = 1268.399$		

$$F_{0.95,1,8} = 5.31 < 121.787 \implies Reject\ H_0$$

e. *What proportion of the total variation in milage is explained by age?*

$$R^2 = \frac{SSR}{SST} = \frac{1190.221302}{1268.399} = 0.9384$$

Muscle mass p.56

i	x	y	xy	x^2	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})y_i$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
1	71	82	5822	5041	10.56	-4.19	866.13	111.57	17.54	-44.23
2	64	91	5824	4096	3.56	4.81	324.19	12.69	23.16	17.14
3	43	100	4300	1849	-17.44	13.81	-1743.75	304.07	190.79	-240.86
4	67	68	4556	4489	6.56	-18.19	446.25	43.07	330.79	-119.36
5	56	87	4872	3136	-4.44	0.81	-386.06	19.69	0.66	-3.61
6	73	73	5329	5329	12.56	-13.19	917.06	157.82	173.91	-165.67
7	68	78	5304	4624	7.56	-8.19	589.88	57.19	67.04	-61.92
8	56	80	4480	3136	-4.44	-6.19	-355.00	19.69	38.29	27.46
9	76	65	4940	5776	15.56	-21.19	1011.56	242.19	448.91	-329.73
10	65	84	5460	4225	4.56	-2.19	383.25	20.82	4.79	-9.98
11	45	116	5220	2025	-15.44	29.81	-1790.75	238.32	888.79	-460.23
12	58	76	4408	3364	-2.44	-10.19	-185.25	5.94	103.79	24.83
13	45	97	4365	2025	-15.44	10.81	-1497.44	238.32	116.91	-166.92
14	53	100	5300	2809	-7.44	13.81	-743.75	55.32	190.79	-102.73
15	49	105	5145	2401	-11.44	18.81	-1200.94	130.82	353.91	-215.17
16	78	77	6006	6084	17.56	-9.19	1352.31	308.44	84.41	-161.36
total	967	1379	81331	60409	0	0	-2012.3125	1965.94	3034.44	-2012.31

$$\bar{x}=\frac{\sum x_i}{n}=\frac{967}{16}=60.44\ ,\quad \bar{y}=\frac{\sum y_i}{n}=\frac{1379}{16}=86.19$$

- ***Obtain the estimated regression function:***

$$b_1=\frac{n\sum x_iy_i-\sum x_i\sum y_i}{n\sum x_i^2-(\sum x_i)^2}=\frac{16(81331)-967(1379)}{16(60409)-(967)^2}=-1.02$$

$$b_1=\frac{\sum x_iy_i-\bar{x}\sum y_i}{\sum(x_i-\bar{x})^2}=\frac{81331-60.44(1379)}{1965.94}=-1.02$$

$$b_1=\frac{\sum(x_i-\bar{x})y_i}{\sum(x_i-\bar{x})^2}=\frac{-2012.31}{1965.94}=-1.02$$

$$b_1=\frac{S_{xy}}{S_{xx}}=\frac{-2012.31}{1965.94}=-1.02$$

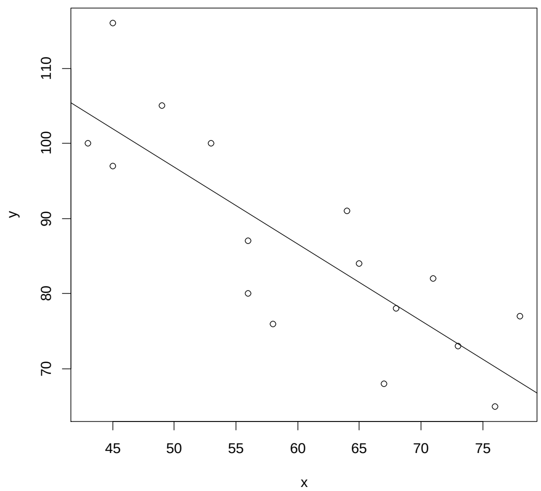
$$b_o=\bar{y}-b_1\bar{x}=86.19-(-1.02)(60.44)=148.05$$

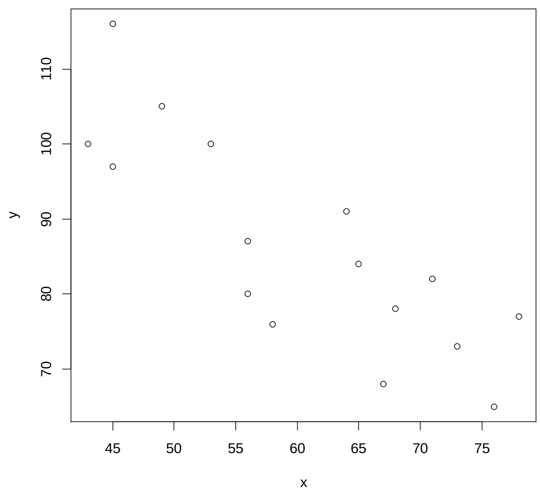
$$\hat{y}=b_o+b_1x$$

$$\hat{y}=148.05+(-1.02)x$$

- Using R:

```
> y=c(82,91,100,68,87,73,78,80,65,84,116,76,97,100,105,77)
> x=c(71,64,43,67,56,73,68,56,76,65,45,58,45,53,49,78)

> plot(x,y)

> b1=(n*sum(x*y)-(sum(x)*sum(y)))/(n*sum(x^2)-(sum(x))^2)
> b1
[1] -1.023589
> b0=mean(y)-b1*(mean(x))
> b0
[1] 148.0507
------------------------------------ or -------------------------------------
> model=lm(y~x)
> model

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)           x
   148.051       -1.024

> with(plot(x,y),abline(model))
```





- ***Interpret the coefficients:***

$b_1 =$ *The changes in value of muscle mass when the age increase by one year.*
$b_o =$ *The muscle mass when the age$= 0$, which has no meaning here.*
So, it is just the intersection with y axis.

- ***A point estimate of the difference in the mean muscle mass for a woman differing in age by one year.***

$$b_1 = -1.02$$

- ***A point estimate of the mean muscle mass for women aged X=60 years.***

$$\hat{y} = b_o + b_1 x = 148.05 + (-1.02)(60) = 86.64$$

- ***The value of the residual for the 8th observation.***

$$y_8 = 80$$
$$\hat{y}_8 = b_o + b_1 x_8$$
$$= 148.05 + (-1.02)(56) = 90.73$$

$$e_8 = y_8 - \hat{y}_8$$
$$= 80 - \ 90.73 = -10.73$$

```
> e=model$res
> e
         1          2          3          4          5          6 
 6.6241615  8.4590367 -4.0363376 -11.4701955 -3.7296773 -0.3286600 
         7          8          9         10         11         12 
-0.4466063 -10.7296773 -5.2578922  2.4826260 14.0108409 -12.6824988 
        13         14         15         16 
-4.9891591  6.1995549  7.1051979  8.7892863 
```

- ***A point estimate of σ^2:***

$$MSE = \frac{SSE}{n-2} = \frac{\Sigma y_i^2 - b_o \Sigma y_i - b_1 \Sigma x_i y_i}{n-2} = \frac{121887 - 148.05(1379) - (-1.02)(81331)}{16-2} = 69.62$$

```
> mse=sum(e^2)/14
> mse
  [1] 69.61829
```

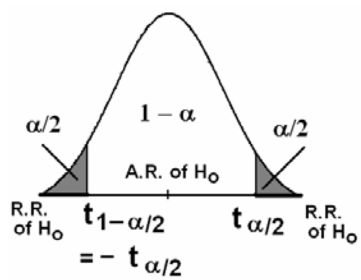
- ***Conduct a test to decide whether or not there is a negative linear association between amount of muscle mass and age.***

$H_0: \beta_1 = 0 \quad vs \quad H_a: \beta_1 \neq 0$

$$T = \frac{b_1 - 0}{S(b_1)} = \frac{-1.02}{\sqrt{\frac{MSE}{S_{xx}}}} = \frac{-1.02}{\sqrt{\frac{69.618}{1965.94}}} = -5.44$$

$$t_{1-\frac{\alpha}{2}, n-2} = t_{0.975\,,14}$$
$$= 2.145$$

$$-5.44 \notin (-2.145\,, 2.145)$$



The decision: reject H_0 *(there is a negative linear association).*

- ***The P-value is :***

$$P - value = 2 \times P(T < -5.44) \approx 0$$

```
> summary(model)

Call:
lm(formula = y ~ x)
Residuals:
    Min       1Q    Median      3Q      Max
-12.6825  -5.0563  -0.3876  6.7444  14.0108

Coefficients:
```

b_o $\quad S(b_0)$ $\quad \frac{b_0}{S(b_0)}$

```
            Estimate   Std. Error    t     value Pr(>|t|)
(Intercept) 148.0507    11.5629    12.804   4.05e-09 ***
x            -1.0236     0.1882    -5.439   8.72e-05 ***
```

b_1 $\quad S(b_1)$ $\quad \frac{b_1}{S(b_1)}$ $\quad R^2$

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 8.344 on 14 degrees of freedom
Multiple R-squared:  0.6788,  Adjusted R-squared:  0.6559
F-statistic: 29.59 on 1 and 14 DF,  p-value: 8.721e-05
```

- ***Estimate with 95 percent confidence interval the difference in expected muscle mass for women whose ages differ by one year.***

$$\left(b_1 \pm t_{(1-\alpha/_2, n-2)} \ S(b_1)\right)$$

$$(-1.02 \pm 2.145 \times \ 0.1882)$$

$$(-1.427, -0.619 \ )$$

```
> confint(model,level=0.95)

                  2.5 %        97.5 %
(Intercept)   123.250671   172.8506799
x              -1.427198    -0.6199802
```

- ***Find the 95% <u>confidence interval</u> of the mean muscle mass for women aged X=60 years:***

$$S(\hat{y}_h) = \sqrt{MSE \times \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right)}$$

$$= \sqrt{69.618 \times \left(\frac{1}{16} + \frac{(60 - 60.44)^2}{1965.94}\right)} = 2.09$$

$$\hat{y}_h \pm t_{(1-\alpha/2, n-2)}\ S(\hat{y}_h)$$

$$86.64 \pm (2.145)(2.09)$$

$$(82.16\,, 91.11)$$

```
> newx=data.frame(x=60)
> newx
   x
1  60
> predict(model,newx,level=0.95,int="confidence")
         fit      lwr      upr
1   86.63532  82.15794  91.1127
```

- ***Find the 95% <u>prediction interval</u> of the mean muscle mass for women aged X=60 years:***

$$S(\hat{y}_{h(new)}) = \sqrt{MSE + S^2(\hat{y}_h)}$$

$$= \sqrt{69.618 + 4.3681} = 8.6$$

$$\hat{y}_h \pm t_{(1-\alpha/2, n-2)}\ S(\hat{y}_{h(new)})$$

$$86.64 \pm (2.145)(8.6)$$

$$(68.19\,, 105.08)$$

```
> newx=data.frame(x=60)
> newx
   x
1  60
> predict(model,newx,level=0.95,int="predict")
    fit        lwr        upr
1   86.63532   68.18813   105.0825
```

- ***Find the determination coefficient, what does it mean, and find the correlation coefficient:***

$SSR = b_1^2 S_{xx} = (-1.02)^2(1965.94) = 2059.8$

$SST = \Sigma(y_i - \bar{y})^2 = 3034.44$

$$R^2 = \frac{SSR}{SST} = \frac{2059.8}{3034.44} = 0.679$$

The estimated regression function explains 67.9 % of the changes in muscle mass by using the age.

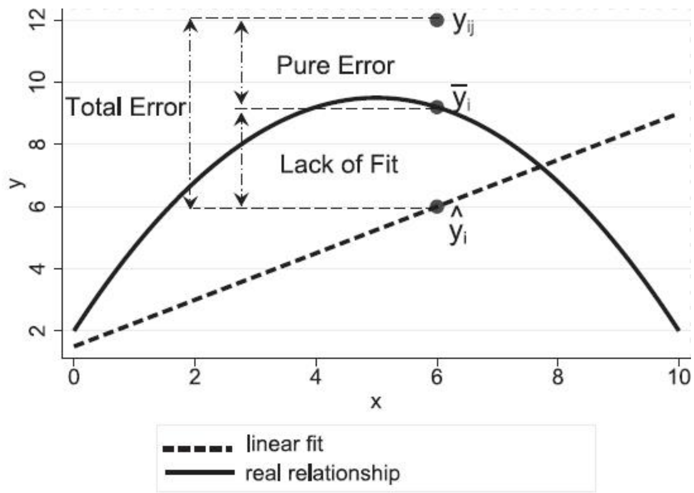
```
> summary(model)$r.squared
[1] 0.6788017
```

- ***The correlation coefficient:***

$$r = -\sqrt{R^2} = -\sqrt{0.679} = -0.824$$

```
> cor(x,y)
[1] -0.8238943
```

- ***Lack of Fit Test:***



H_0: *The regression function is linear (no lack of fit)*

H_1: *The regression function is not linear (lack of fit)*

Source		DF	SS	MS	F
Regression		1	$SSR = \Sigma(\hat{Y}_{ij} - \bar{Y})^2$	$MSR = \frac{SSR}{1}$	$\frac{MSR}{MSE}$
Error		$n-2$	$SSE = \Sigma(Y_{ij} - \hat{Y}_{ij})^2$	$MSE = \frac{SSE}{n-2}$	
	Lack of fit	$c-2$	$SSLF = \Sigma(\bar{Y}_j - \hat{Y}_{ij})^2$	$MSLF = \frac{SSLF}{c-2}$	$\frac{MSLF}{MSPE}$
	Pure Error	$n-c$	$SSPE = \Sigma(Y_{ij} - \bar{Y}_j)^2$	$MSPE = \frac{SSPE}{n-c}$	
Total		$n-1$	$SST = \Sigma(Y_{ij} - \bar{Y})^2$		

$c = \#$ *of different X's*

Example:

id	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
x	1.3	1.3	2	2	2.7	3.3	3.3	3.7	3.7	4	4	4	4.7	4.7	5	5.3	5.3	5.3	5.7	6	6	6.3	6.7
y	2.3	1.8	2.8	1.5	2.2	3.8	1.8	3.7	1.7	2.8	2.8	2.2	3.2	1.9	1.8	3.5	2.8	2.1	3.4	3.2	3	3	5.9

	1		2		3	4		5			6		7		8		9		10	11		12	13
id	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
x	1.3	1.3	2	2	2.7	3.3	3.3	3.7	3.7	4	4	4	4.7	4.7	5	5.3	5.3	5.3	5.7	6	6	6.3	6.7
y	2.3	1.8	2.8	1.5	2.2	3.8	1.8	3.7	1.7	2.8	2.8	2.2	3.2	1.9	1.8	3.5	2.8	2.1	3.4	3.2	3	3	5.9

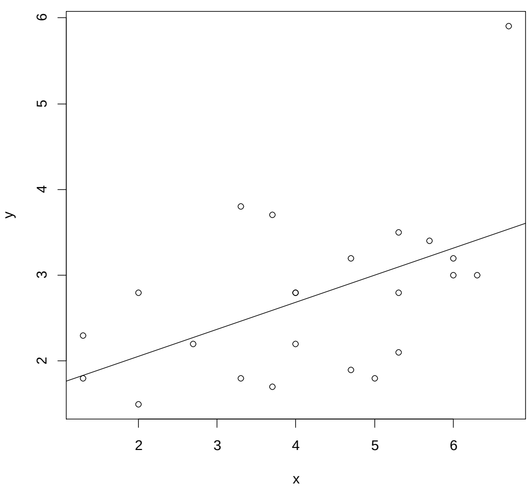
$c = 13$

```
> data=read.csv("LOF1.csv")
> x=data$X
> y=data$Y
> model=lm(y~x)
> with(plot(x,y),abline(model))

> install.libraries(EnvStats)
> install.packages("EnvStats")
> library(EnvStats)

> anovaPE(model)
```

$SSE(X) = 15.2782$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	5.4992	5.4992	7.7948	0.01906
Lack of Fit	11	8.2232	0.7476	1.0596	0.46751
Pure Error	10	7.0550	0.7055		

Since p-value = 0.46751 > 0.05 we accept H_0 (The regression function is linear (no lack of fit)).



Question 1:

Consider the following data:

X	10	85	20	25	30	35
Y	73	85	90	86	75	61
	78	87	92	87	76	63

a. Estimate the simple linear regression model.

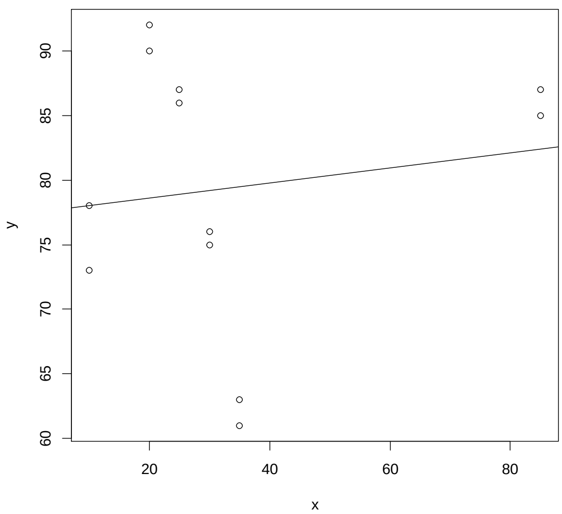
```
> x=c(10,10,85,85,20,20,25,25,30,30,35,35)
> y=c(73,78,85,87,90,92,86,87,75,76,61,63)
> model=lm(y~x)
> model
 > with(plot(x,y),abline(model))
Call:
lm(formula = y ~ x)
Coefficients:
(Intercept)          x
  77.42737      0.05822
> anovaPE(model)
             Df   Sum Sq    Mean Sq   F value    Pr(>F)
x             1   23.53     23.532    7.2406     0.03601 *
Lack of Fit   4   1099.88   274.971   84.6065    2.087e-05 ***
Pure Error    6   19.50     3.250
```

$$\boxed{Y = 77.43 + 0.058X_1}$$

b. Preform the lack of fit for the model.

H_0: No lack of fit. Vs H_1: There is lack of fit

Since p-value= 2.087e-05 < 0.05, We Reject H_0 (There is lack of fit).



<u>Question 2:</u> *(without package "EnvStats")*

Consider the following data:

X	10	85	20	25	30	35
Y	73 78	85 87	90 92	86 87	75 76	61 63

**** Preform the lack of fit for the model.***

```
> x=c(10,10,85,85,20,20,25,25,30,30,35,35)
> y=c(73,78,85,87,90,92,86,87,75,76,61,63)
> model=lm(y~x)
> anova(model)
Analysis of Variance Table
            Df  Sum Sq   Mean Sq   F value   Pr(>F)
x            1   23.53    23.532    0.2102    0.6564
Residuals   10   1119.38  111.938
```

X_j	10	85	20	25	30	35
Y_j	73 78	85 87	90 92	86 87	75 76	61 63
$\bar{Y}_j$	$\frac{73+78}{2}=75.5$	$\frac{85+87}{2}=86$	$\frac{90+92}{2}=91$	$\frac{86+87}{2}=86.5$	$\frac{75+76}{2}=75.5$	$\frac{61+63}{2}=62$

i	X_j	Y_j	$\bar{Y}_j$	$(Y_{ij}-\bar{Y}_j)^2$
1	10	73	75.5	$(73-75.5)^2=6.25$
2	10	78	75.5	$(78-75.5)^2=6.25$
3	85	85	86	$(85-86)^2=1$
4	85	87	86	$(87-86)^2=1$
5	20	90	91	$(90-91)^2=1$
6	20	92	91	$(92-91)^2=1$
7	25	86	86.5	0.25
8	25	87	86.5	0.25
9	30	75	75.5	0.25
10	30	76	75.5	0.25
11	35	61	62	1
12	35	63	62	1
				$SSPE=\sum(Y_{ij}-\bar{Y}_j)^2=19.5$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	23.53	23.532	0.2102	0.6564
Residuals	10	1119.38	111.938		
Lack of fit	$c-2=4$	$\boxed{1099.88}$	274.97	84.61	$\boxed{1-\text{pf}(84.61,4,6)=2.086708e-05}$
Pure Error	$n-c=6$	19.5	3.25		

$\boxed{1119.38-19.5=1099.88}$

H_0: *No lack of fit.* *Vs* H_1: *There is lack of fit*

Since p-value= 2.087e-05 < 0.05, We Reject H₀ (There is lack of fit).

Question 3:

An experiment is conducted, relating weekly sales for a food delivery (Y) service to the amount of advertising (X) during the week. The data is given below:

Week	1	2	3	4	5	6
X	2	2	4	4	6	6
Y	20	30	40	50	70	60

Test the lack of fit

```
> x=c(2,2,4,4,6,6)
> y=c(20,30,40,50,70,60)

> model=lm(y~x)

> anovaPE(model)
             Df   Sum Sq   Mean Sq   F value    Pr(>F)
x             1   1600     1600      32         0.01094 *
Lack of Fit   1    0        0         0          1.00000
Pure Error    3   150      50
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H_0: *The regression function is linear (no lack of fit)*

H_1: *The regression function is not linear (lack of fit)*

p-value = 1 >0.05

We Accept H_0 (no lack of fit)

Question 4:

Use the data in file "hw4Q1".

1. *Find the regression model* $Y = \beta_o + \beta_1 X_1$.

```
> df=read.csv("LOF2.csv")
> x=df$x
> y=df$y
> model=lm(y~x)
> model

Call:
lm(formula = y ~ x)
Coefficients:
(Intercept)          x
     -43.06      15.68
>summary(model)$r.squared
[1] 0.9548736
```

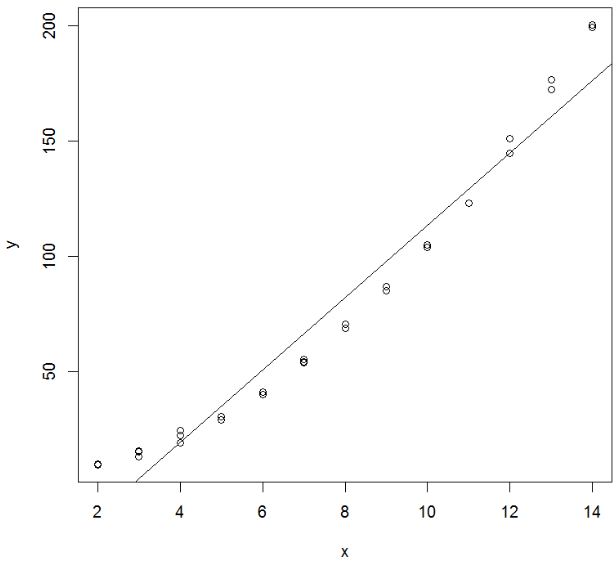
$\boxed{Y = -43.06 + 15.68X_1}$

$\boxed{R^2 = 0.9548736}$

2. *Plot the regression model with the data by using the R code*

```
with(plot(x,y),abline(model))
```

do you need a lack of fit test? Why ?



From the plot, the regression line does not represent the relationship between X and Y properly. And yes, we need a lack of fit test.

3. *Test the model for a lack of fit.*

```
>install.libraries("EnvStats")
>install.packages("EnvStats")
>library(EnvStats)
>anovaPE(model)
             Df   Sum Sq   Mean Sq   F value    Pr(>F)
x             1   97578    97578     31708.26   < 2.2e-16 ***
Lack of Fit  11   4559     414       134.68     2.784e-14 ***
Pure Error   17   52       3
---
```

H_0: The regression function is linear (no lack of fit)

H_1: The regression function is not linear (lack of fit)

p-value = 2.784e-14 <0.05

We reject H_0 (there is a lack of fit)

4. *What is the value of c?*

$$c - 2 = 11 \implies c = 13$$

5. *Use the transformation X^2. Find the model and their R^2.*

```
> x2=x^2
> model1=lm(y~(x2))
> model1

Call:
lm(formula = y ~ (x2))

Coefficients:
(Intercept)          x2
      5.829       0.990
> summary(model1)$r.squared
[1] 0.9993502
```

$\boxed{Y = 5.829 + 0.99\,X_1}$

$\boxed{R^2 = 0.9993502}$

6. *Use the transformation X^4. Find the model and their R^2.*

```
> x3=x^4
> model2=lm(y~(x3))
> model2

Call:
lm(formula = y ~ (x3))

Coefficients:
(Intercept)          x3
  34.385526    0.004919
> summary(model2)$r.squared
[1] 0.9242395
```

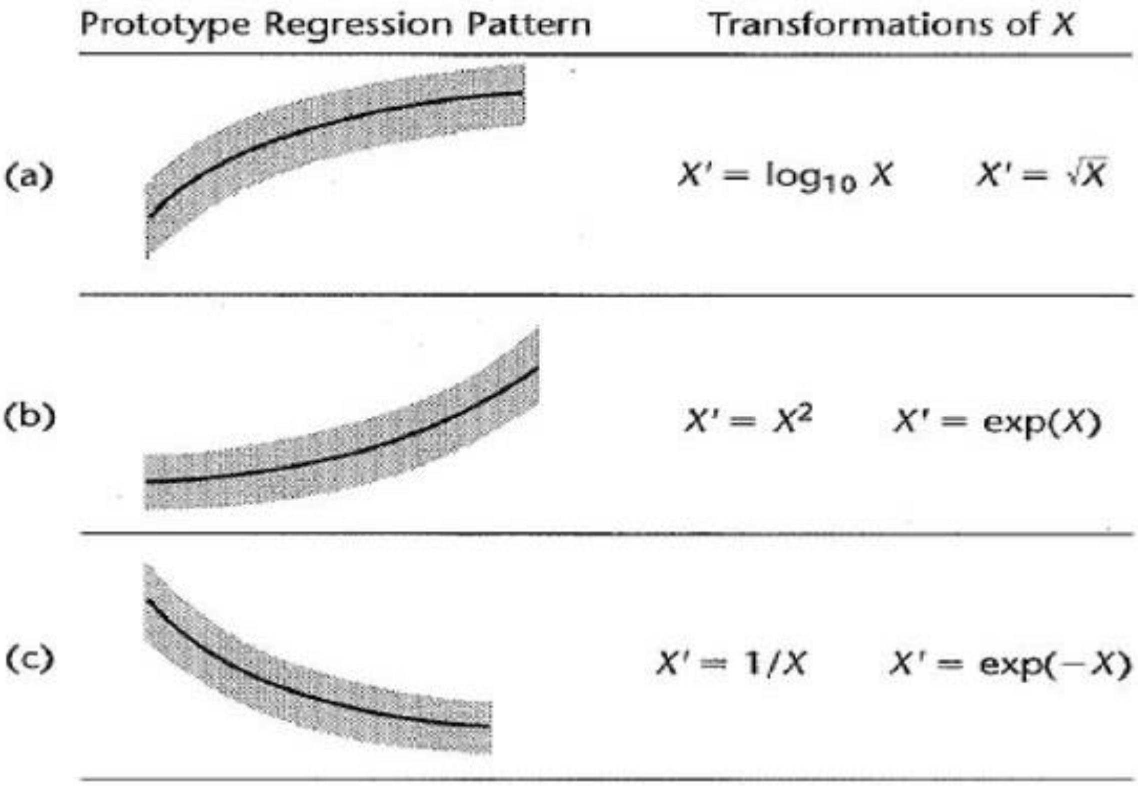
$\boxed{Y = 34.385526 + 0.004919\,X_1}$

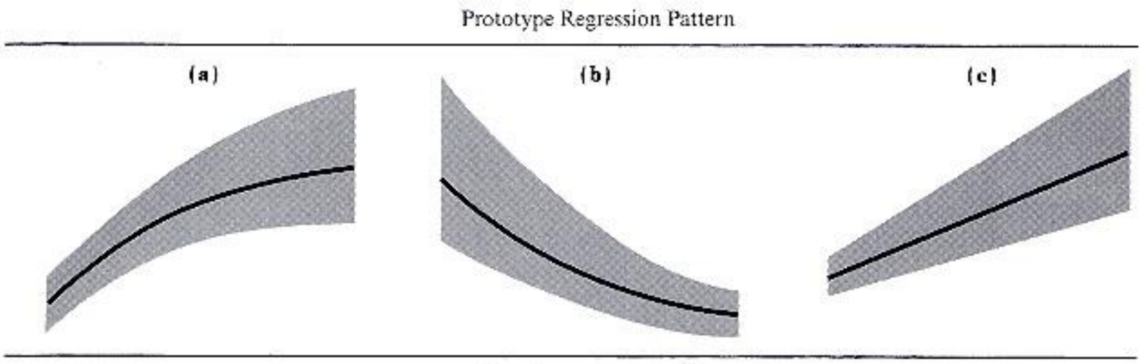
$\boxed{R^2 = 0.9243295}$

7. *Compare between the three models, what is the best model?*

The best model is the one with the highest R^2 (model 1)

Transformation





Transformations on Y

$Y' = \sqrt{Y}$

$Y' = \log_{10} Y$

$Y' = 1/Y$

Note: A simultaneous transformation on X may also be helpful or necessary.

Problem 1

A marketing researcher studied annual sales of a product that had been introduced 10 years ago. The data are as follows, where X is the year (coded) and Y is sales in thousands

X	0	1	2	3	4	5	6	7	8	9
Y	98	135	162	178	221	232	283	300	374	395

(a) Use the transformation $Y' = \sqrt{Y}$ and obtain the estimated linear regression function for the transformed data.

Before transformation

```
> x=c(0,1,2,3,4,5,6,7,8,9)
> y= c(98,135,162,178,221,232,283,300,374,395)
> model1=lm(y~x)
> summary(model1)

Call:
lm(formula = y ~ x)
Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   91.564     8.814        10.39     6.38e-06 ***
x             32.497     1.651        19.68     4.62e-08 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 15 on 8 degrees of freedom
Multiple R-squared:  0.9798,    Adjusted R-squared:  0.9772
F-statistic: 387.4 on 1 and 8 DF,  p-value: 4.62e-08
```

$$\boxed{\hat{y} = 91.564 + 32.497x}$$

After transformation

```
> sy=sqrt(y)
> model2=lm(sy~x)
> summary(model2)

Call:
lm(formula = sy ~ x)
Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)  10.26093   0.21290      48.20     3.80e-11 ***
x            1.07629    0.03988      26.99     3.83e-09 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.3622 on 8 degrees of freedom
Multiple R-squared:  0.9891,    Adjusted R-squared:  0.9878
F-statistic: 728.4 on 1 and 8 DF,  p-value: 3.826e-09
```

$$\boxed{\widehat{\sqrt{y}} = 10.26093 + 1.07629x}$$

(b) Does the regression line appear to be a good fit to the transformed data?

By checking the T-test and R^2. Yes, the regression line appears to be a good fit to the transformed data

(c) Find the value of Y when X=8.5 under both models.

```
> newx=data.frame(x=8.5)

> predict(model1,newx,level=0.95,int="confidence")
       fit         lwr         upr
1  367.7879    349.0385    386.5372
```

$$\hat{y} = 91.564 + 32.497(8.5)$$
$$\hat{y} = 367.7879$$

```
> predict(model2,newx,level=0.95,int="confidence")
       fit         lwr         upr
1  19.40941    18.95655    19.86228
```

$$\widehat{\sqrt{y}} = 10.26093 + 1.07629(8.5)$$
$$\widehat{\sqrt{y}} = 19.40941$$
$$\hat{y} = 376.7252$$

For model 1:

$$\hat{y} \in (349.0385, 386.5372)$$

For model 2:

$$\hat{y} \in (18.95655^2, 19.86228^2)$$

$$\hat{y} \in (359.35, 394.51)$$

Problem 2

Use the following table gives the level of inventories Y and sales X

X	*315*	*314*	*366*	*387*	*398*	*415*	*443*
Y	*516*	*544*	*560*	*620*	*634*	*633*	*680*

We consider the following transformation of Y given by:

$$Y_1 = \log Y \ , \ Y_2 = \frac{1}{Y}, \ Y_3 = \sqrt{Y}$$

a. *Regress Y and X and find the value of Multiple R-square.*

```
>x=c(315,314,366,387,398,415,443)
>y=c(516,544,560,620,634,633,680)
> model=lm(y~x)
> model
lm(formula = y ~ x)
Coefficients:
   (Intercept)          x
     159.317        1.164
> summary (model)$r.squared
[1] 0.9299734
```

b. *Regress Y_1 and X and find the value of Multiple R-square.*

```
> y1=log10(y)
> model1=lm(y1~x)
> model1
lm(formula = y ~ x)
Coefficients:
   (Intercept)          x
   2.4533011    0.0008536
> summary (model1)$r.squared
[1] 0.9286431
```

c. Regress Y_2 and X and find the value of Multiple R-square.

```
> y2=1/y
> model2=lm(y2~x)
> model2
lm(formula = y ~ x)
Coefficients:
(Intercept)          x
 2.942e-03    -3.333e-06
> summary (model2)$r.squared
[1] 0.9248624
```

d. Regress Y_3 and X and find the value of Multiple R-square.

```
> y3=sqrt(y)
> model3=lm(y3~x)
> model3
lm(formula = y3 ~ x)
Coefficients:
    (Intercept)          x
     15.42231      0.02391
> summary (model3)$r.squared
[1] 0.9296231
```

e. What is the best transformation of Y with the justification?

Transformation	R^2
$Y_1 = \log Y$	*0.9286431*
$Y_2 = \frac{1}{Y}$	*0.9248624*
$Y_3 = \sqrt{Y}$	*0.9296231*

$\sqrt{Y}$ is the best transformation.

f. What is the expected Y for each model when X = 462.

```
> newx=data.frame(x=462)
> predict(model ,newx,level=0.95,int="confidence")
       fit        lwr        upr
1  697.2862   661.8838   732.6885
```

$\hat{y} = 697.2862$

```
> predict(model1,newx,level=0.95,int="confidence")
       fit        lwr        upr
1  2.847651   2.821436   2.873867
```

$\widehat{y_1} = \widehat{\log y} = 2.847651$

$\hat{y} = 10^{2.847651}$

$\hat{y} = 704.13$

```
> predict(model2,newx,level=0.95,int="confidence")
        fit        lwr        upr
1  0.001402   0.001297   0.001507
```

$\widehat{y_2} = \frac{1}{\hat{y}} = 0.001402$

$\hat{y} = \frac{1}{0.001402}$

$\hat{y} = 713.12$

```
> predict(model3,newx,level=0.95,int="confidence")
        fit        lwr        upr
1  26.46671   25.73795   27.19547
```

$\widehat{y_3} = \widehat{\sqrt{y}} = 26.46671$

$\hat{y} = 26.46671^2$

$\hat{y} = 700.49$

Problem 3:

We study the production Y (in cubic meter per second) of a manufacturing process, as a function of temperature X (in Celsius). For temperature increasing by 100 degrees from 100 to 600 °C , the production increased from 49 to 68. From the data we have summarize the following:

$$\sum_{i=1}^{6} Y_i = 369, \qquad \sum_{i=1}^{6} Y_i^2 = 2291$$

$$\sum_{i=1}^{6} X_i Y_i = 134600, \qquad \sum_{i=1}^{6} Y_i \sqrt{X_i} = 6825.639$$

$X = 100\,, 200\,, 300\,, 400\,, 500\,, 600$		$n = 6$
$\sum_{i=1}^{6} X_i = 2100$	$\sum_{i=1}^{6} X_i^2 = 910000$	$\bar{X} = 350$
$\sum_{i=1}^{6} Y_i = 369$	$\sum_{i=1}^{6} Y_i^2 = 22911$	$\bar{Y} = 61.5$
$\sum_{i=1}^{6} X_i Y_i = 134600$		

(a) We propose the regression model (model-1) as

$$Y = \beta_0 + \beta_1 X + \varepsilon\,, \qquad E(\varepsilon) = 0 \quad and \quad Var(\varepsilon) = \sigma^2$$

i) Estimate the model and interpret the slop.

$$b_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{6(134600) - (2100 \times 369)}{6(910000) - (2100)^2} = 0.031143$$

$$b_0 = \bar{y} - b_1\bar{x} = (61.5) - (0.031143)(350) = 50.6$$

$$\hat{Y} = 50.6 + 0.03\,X$$

b_1 : *The change in (production) when (the temperature) increase by one degree.*

ii) *Calculate the coefficient of determination and explain the result.*

$$SSE = \Sigma y_i^2 - b_o \Sigma y_i - b_1 \Sigma x_i y_i$$

$$= 22911 - (50.6)(369) - (0.031143)(134600) = 47.77143$$

$$SST = \Sigma(y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2 = 22911 - (6)(61.5)^2 = 217.5$$

$$SSR = SST - SSE = 217.5 - 47.77143 = 169.7286$$

$$R^2 = \frac{SSR}{SST} = \frac{169.7286}{217.5} = 0.78$$

iii) *Calculate 95% C.I for β_0.*

$$S_{XX} = \sum x_i^2 - n\bar{x}^2 = 910000 - (6)(350)^2 = 175000$$

$$MSE = \frac{SSE}{n-2} = \frac{47.77143}{4} = 11.94286$$

$$S(b_o) = \sqrt{MSE \times \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} = \sqrt{11.94286 \times \left(\frac{1}{6} + \frac{(350)^2}{175000}\right)} = 3.217$$

$$t_{(1-\alpha/_2, n-2)} = t_{(1-0.5/_2, 6-2)} = t_{(0.975,4)} = 2.776445$$

$$b_o \pm t_{(1-\alpha/_2, n-2)}\ S(b_o)$$

$$(50.6) \pm (2.776445)(3.217)$$

$$(41.67, 59.53)$$

iv) *Use T- test for testing the linearity of the estimated model*

$H_0: \beta_1 = 0 \quad vs \quad H_1: \beta_1 \neq 0$

$$S(b_1) = \sqrt{\frac{MSE}{S_{xx}}} = \sqrt{\frac{11.94286}{175000}} = 0.00827$$

$$T = \frac{b_1}{S(b_1)} = \frac{0.031143}{0.00827} = 3.77$$

$$3.77 \notin (-2.77\ , 2.77)$$

We reject H_0 (there is a linear relationship)

(b) *We propose the regression model (model-2) as*

$$Y = \alpha_0 + \alpha_1\sqrt{X} + \varepsilon\ , \qquad E(\varepsilon) = 0 \ \ and \ \ Var(\varepsilon) = \sigma^2$$

$\sqrt{X} = \sqrt{100}\ , \sqrt{200}, \sqrt{300}, \sqrt{400}, \sqrt{500}, \sqrt{600}$		$n = 6$
$\sum_{i=1}^{6} \sqrt{X_i} = 108.3182$	$\sum_{i=1}^{6} \sqrt{X_i}^2 = 2100$	$\overline{\sqrt{X}} = 18.05304$
$\sum_{i=1}^{6} Y_i = 369$	$\sum_{i=1}^{6} Y_i^2 = 22911$	$\bar{Y} = 61.5$
	$\sum_{i=1}^{6} Y_i \sqrt{X_i} = 6825.639$	

v) *Calculate the point estimate of α_0 and α_1.*

$$\alpha_1 = \frac{n\sum y_i\sqrt{x_i} - \sum\sqrt{x_i}\sum y_i}{n\sum\sqrt{x_i}^2 - \left(\sum\sqrt{x_i}\right)^2} = \frac{6(6825.639) - (108.3182\times 369)}{6(2100) - (108.3182)^2} = 1.135211$$

$$\alpha_0 = \bar{y} - \alpha_1\overline{\sqrt{x}} = (61.5) - (1.135211)(18.05304) = 41.006$$

$$\hat{Y} = 41.006 + 1.135211\sqrt{X}$$

vi) *Construct the ANOVA table and use it for testing the model.*

$$SSE = \Sigma y_i^2 - \alpha_0 \Sigma y_i - \alpha_1 \sum y_i \sqrt{x_i}$$

$$= 22911 - (41.006)(369) - (1.135211)(6825.639) = 31.24626$$

$$SST = \Sigma(y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2$$

$$= 22911 - (6)(61.5)^2 = 217.5$$

$$SSR = SST - SSE = 217.5 - 31.24626 = 186.2537$$

source	*SS*	*df*	*MS*	*F*
Regression	186.2537	1	186.2537	*23.84*
Error	31.24626	4	7.811565	
Total	217.5	5		

$H_0: \alpha_1 = 0 \quad vs \quad H_1: \alpha_1 \neq 0$

$$F_{1-\alpha,1,n-2} = F_{0.95,1,4} = 7.71 < 23.84$$

We reject H_0 (there is a linear relationship)

vii) *Calculate the percentage of production of the total variation explained by* $\sqrt{X}$.

$$R^2 = \frac{SSR}{SST} = \frac{186.2537}{217.5} = 0.86$$

viii) *Compute the expected production when the temperature is 550* ^0C.

$$\hat{Y} = 41.006 + 1.135211\sqrt{X}$$

$$\hat{Y} = 41.006 + 1.135211\sqrt{550}$$

$$\hat{Y} = 67.63\ m^3/s$$

(c) Which model is better for such data (model-1 or model-2)? Why?

Model 1	*Model 2*
$R^2 = 0.78$	$R^2 = 0.86$

Model 2 is better than Model 1

Problem 4:

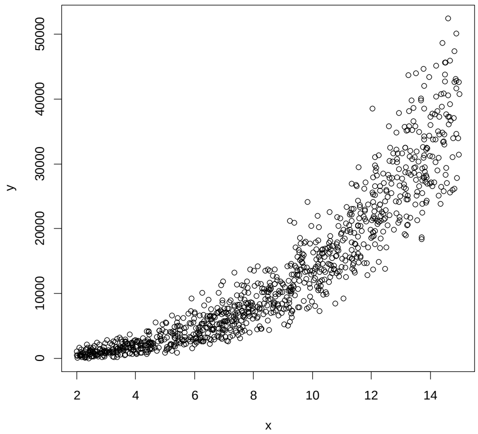
Use the data in file "d5".

8. ***Find the regression model (model1):*** $Y = \beta_o + \beta_1 X_1$.

```
> plot(x,y)
> d5=read.csv("d5.csv")
> x=d5$x
> y=d5$y
> model1=lm(y~x)
> summary(model1)
Coefficients:
             Estimate    Std. Error   t value    Pr(>|t|)
(Intercept)  -11203.6    382.3        -29.30     <2e-16 ***
x             2815.6     41.1          68.51     <2e-16 ***

Residual standard error: 4785 on 998 degrees of freedom
Multiple R-squared:  0.8246,    Adjusted R-squared:  0.8245
F-statistic:  4693 on 1 and 998 DF,  p-value: < 2.2e-16
```

$$\boxed{Y = -11203.6 + 2815.6\, X_1}$$

9. ***Plot the scatter plot for (model1) in R using the code*** "**plot(x,y)**".



When X increase the variation become bigger and bigger.

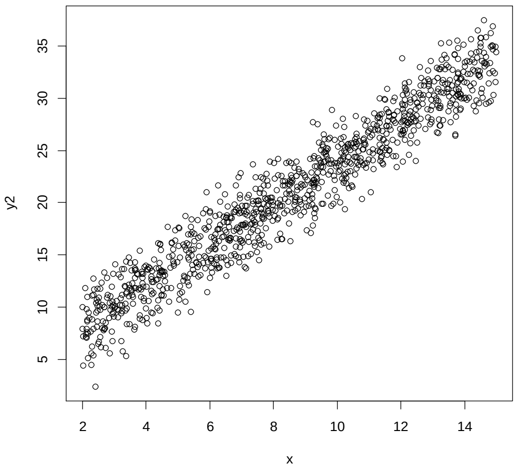
10. Use the transformation $Y^{\frac{1}{3}}$ and find the regression model

(model2): $Y^{\frac{1}{3}} = \alpha_o + \alpha_1 X_1$.

```
> y2=y^(1/3)
> model2=lm(y2~x)
> summary(model2)
Coefficients:
              Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   4.1170     0.1628       25.29     <2e-16 ***
x             1.9764     0.0175       112.96    <2e-16 ***

Residual standard error: 2.037 on 998 degrees of freedom
Multiple R-squared:  0.9275,    Adjusted R-squared:  0.9274
F-statistic: 1.276e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

$$\boxed{Y^{\frac{1}{3}} = 4.12 + 1.98\ X_1}$$

11. Plot the scatter plot for (model2).



12. Compare between scatter plot before and after transformation in terms of MSE.

$$MSE_{model1} = 4785^2 = 22896225$$

$$MSE_{model2} = 2.037^2 = 4.15$$

There is a huge decreasing in MSE after transformation and we can see this different form the two plots.

13.Find a point estimate for Y when X = 10 by using (model2).

```
> newx=data.frame(x=10)
> newx
   x
1  10
> predict(model2,newx,level=0.95,int="predict")
        fit          lwr          upr
1      23.88059  19.88102   27.88017
```

$$Y^{\frac{1}{3}} = 23.88059$$

$$Y = 23.88059^3$$

$$Y = 13618.684$$

Simple linear regression by matrices

$$Y_{n\times 1} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \;,\quad X_{n\times 2} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \;,\quad \epsilon_{n\times 1} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix} \;,\quad \beta_{2\times 1} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

$$Y_{n\times 1} = X_{n\times 2}\, \beta_{2\times 1} + \epsilon_{n\times 1}$$

- *Coefficients:*

$$\boxed{\hat{\beta}_{2\times 1} = (X'X)^{-1} X'Y}$$

$$\hat{\beta}_{2\times 1} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

$$X'X = \begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{bmatrix}$$

$$\Rightarrow (X'X)^{-1} = \frac{1}{\Delta} \begin{bmatrix} \Sigma x_i^2 & -\Sigma x_i \\ -\Sigma x_i & n \end{bmatrix} \quad ; \Delta = n\Sigma x_i^2 - (\Sigma x_i)^2$$

$$X'Y = \begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{bmatrix} \times \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \Sigma y_i \\ \Sigma x_i y_i \end{bmatrix}$$

- *Variance for coefficients:*

$$\boxed{Var(\hat{\beta}) = MSE(X'X)^{-1}}$$

$$Var(\hat{\beta}) = \begin{bmatrix} Var(b_0) & Cov(b_0, b_1) \\ Cov(b_0, b_1) & Var(b_1) \end{bmatrix}$$

Multiple linear regression (2 variables)

$$Y_{n\times 1} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \;,\quad X_{n\times 3} = \begin{bmatrix} 1 & x_{11} & x_{21} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{bmatrix} \;,\quad \epsilon_{n\times 1} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix} \;,\quad \beta_{3\times 1} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

$$Y_{n\times 1} = X_{n\times 3}\,\beta_{3\times 1} + \epsilon_{n\times 1}$$

- *Coefficients:*

$$\boxed{\hat{\beta}_{3\times 1} = (X'X)^{-1}X'Y}$$

$$\hat{\beta}_{3\times 1} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

$$X'X = \begin{bmatrix} n & \Sigma x_1 & \Sigma x_2 \\ \Sigma x_1 & \Sigma x_1^2 & \Sigma x_1 x_2 \\ \Sigma x_2 & \Sigma x_1 x_2 & \Sigma x_2^2 \end{bmatrix} \qquad X'Y = \begin{bmatrix} 1 & \dots & 1 \\ x_{11} & \dots & x_{1n} \\ x_{21} & \dots & x_{2n} \end{bmatrix} \times \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \Sigma y \\ \Sigma x_1 y \\ \Sigma x_2 y \end{bmatrix}$$

- *Variance for coefficients:*

$$\boxed{Var(\hat{\beta}) = MSE(X'X)^{-1}}$$

$$Var(\hat{\beta}) = \begin{bmatrix} Var(b_0) & Cov(b_0, b_1) & Cov(b_0, b_2) \\ Cov(b_1, b_0) & Var(b_1) & Cov(b_1, b_2) \\ Cov(b_2, b_0) & Cov(b_2, b_1) & Var(b_2) \end{bmatrix}$$

Question 1:

An experiment is conducted, relating weekly sales for a food delivery (Y) service to the amount of advertising (X) during the week. The data is given below:

Week	1	2	3	4	5	6
X	2	2	4	4	6	6
Y	20	30	40	50	70	60

Assume the regression model: $Y = X\beta + \varepsilon$

a. *calculate the following:* $X'X,\ X'Y,\ (X'X)^{-1},\ \beta,\ \ Y',\ e',$ *SSE, SST, SSR,* R^2 *and* r_{xy}

```
> x=c(2,2,4,4,6,6)
> y=c(20,30,40,50,70,60)

> sum(x)
[1] 24
> sum(x^2)
[1] 112
> sum(y)
[1] 270
> sum(x*y)
[1] 1240
```

$$X'X = \begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{bmatrix} = \begin{bmatrix} 6 & 24 \\ 24 & 112 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} \Sigma y_i \\ \Sigma x_i y_i \end{bmatrix} = \begin{bmatrix} 270 \\ 1240 \end{bmatrix}$$

```
> xtx=matrix(c(6,24,24,112),2,2)
> solve(xtx)
        [,1]     [,2]
[1,]  1.1667  -0.2500
[2,] -0.2500   0.0625
```

$$(X'X)^{-1} = \frac{1}{\Delta}\begin{bmatrix} \Sigma x_i^2 & -\Sigma x_i \\ -\Sigma x_i & n \end{bmatrix} \quad ; \Delta = n\Sigma x_i^2 - (\Sigma x_i)^2$$

$$(X'X)^{-1} = \begin{bmatrix} 1.1667 & -0.2500 \\ -0.2500 & 0.0625 \end{bmatrix}$$

```
> xtx=matrix(c(6,24,24,112),2,2)
> xty=matrix(c(270,1240),2,1)
> b=solve(xtx)%*%(xty)
> b
      [,1]
[1,]    5
[2,]    0
```

$$\hat{\beta}_{2\times 1} = (X'X)^{-1}X'Y = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 5 \\ 10 \end{bmatrix}$$

$$Y' = [20 \quad 30 \quad 40 \quad 50 \quad 70 \quad 60]$$

```
>model=lm(y~x)
>model$res
 1  2  3  4  5  6
-5  5 -5  5  5 -5
```

$$e' = [-5 \quad 5 \quad -5 \quad 5 \quad 5 \quad -5]$$

```
>anova(model)

Analysis of Variance Table
Response: y
           Df  Sum Sq  Mean Sq  F value  Pr(>F)
     x      1  1600    1600.0   42.667   0.002838**
Residuals   4   150      37.5
```

$SSR = 1600$
$SSE = 150$
$SST = 1750$

```
>summary(model)

Call:
lm(formula = y ~ x)
Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)   5.000     6.614       0.756    0.49177
   x         10.000     1.531       6.532    0.00284**

Residual standard error: 6.124 on 4 degrees of freedom
Multiple R-squared:  0.9143,   Adjusted R-squared:  0.8929
F-statistic: 42.67 on 1 and 4 DF,  p-value: 0.002838

>cor(x,y)
[1] 0.9561829
```

$R^2 = 0.9143$
$r_{xy} = 0.9561829$

Question 2:

i	1	2	3	4	5	6
X_{1i}	7	4	16	3	21	8
X_{2i}	33	41	7	49	5	31
Y_i	42	33	75	28	91	55

Assume that regression model

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$ *with independent normal error*

By using matrix forms obtain

a. $X'X$ and $X'Y$

$$X'X = \begin{bmatrix} n & \Sigma x_1 & \Sigma x_2 \\ \Sigma x_1 & \Sigma x_1^2 & \Sigma x_1 x_2 \\ \Sigma x_2 & \Sigma x_1 x_2 & \Sigma x_2^2 \end{bmatrix} \qquad X'Y = \begin{bmatrix} \Sigma y \\ \Sigma x_1 y \\ \Sigma x_2 y \end{bmatrix}$$

```
> x1=c(7,4,16,3,21,8)
> x2=c(33,41,7,49,5,31)
> y =c(42,33,75,28,91,55)
> n=6
> m1=c(n,sum(x1),sum(x2),
+      sum(x1),sum(x1^2) ,sum(x1*x2),
+      sum(x2),sum(x1*x2),sum(x2^2))

> xtx=matrix(m1,3,3)
> xtx
        [,1]   [,2]   [,3]
[1,]    6      59     166
[2,]    59     835    1007     X'X
[3,]    166   1007    6206

> m2=c(sum(y),sum(x1*y),sum(x2*y))
> xty=matrix(m2,3,1)
> xty
      [,1]
[1,]   324
[2,]   4061        X'Y
[3,]   6796
```

b. $\widehat{\beta}$

```
> b=solve(xtx)%*%(xty)
> b
           [,1]
[1,] 33.9321033
[2,]  2.7847614
[3,] -0.2644189
```

c. SSE and SSR

```
> sse=(sum(y^2))-t(b)%*%(xty)
> sse
         [,1]
[1,] 62.07354
> sst=(sum(y^2))-n*(mean(y))^2
> sst
[1] 3072
> ssr=sst-sse
> ssr
         [,1]
[1,] 3009.926
```

d. $Var(\widehat{\beta})$

```
> mse=sse/(n-3)
> mse
         [,1]
[1,] 20.69118

> vb=mse[1,1]*solve(xtx)
> vb
           [,1]         [,2]          [,3]
[1,] 715.47114  -34.1589166  -13.5949371
[2,] -34.15892   1.6616664     0.6440674
[3,] -13.59494   0.6440674     0.2624678
```

$Var(\hat{\beta})$

e. Find $\widehat{Y}_h$ and $Var(\widehat{Y}_h)$ when $X_{b1} = 10$ and $X_{b2} = 30$

$$\boxed{Var(\hat{Y}_h) = MSE(\hat{X}_h'(X'X)^{-1}\hat{X}_h)}$$

```
> newx=data.frame(x1=10, x2=30)
> predict(model,newx,level=0.95)
       1
53.84715
```

$\longrightarrow \hat{Y}_h$

```
> xh=matrix(c(1,10,30),3,1)
> mse*(t(xh)%*%solve(xtx)%*%xh)
        [,1]
[1,] 5.42462
```

$\longrightarrow Var(\hat{Y}_h)$

Question 3:

y	*85*	*152*	*41*	*93*	*101*	*38*	*203*	*78*	*117*	*44*	*121*	*112*	*50*	*82*	*48*	*127*	*140*	*155*	*39*	*90*
x1	*7*	*18*	*5*	*14*	*11*	*5*	*23*	*9*	*16*	*5*	*17*	*12*	*6*	*12*	*8*	*15*	*17*	*21*	*6*	*11*
x2	*5.1*	*17*	*3.2*	*7*	*11*	*4*	*22*	*7*	*11*	*4.8*	*11*	*9.5*	*3.8*	*6.5*	*4.6*	*14*	*13*	*15*	*3.6*	*9.6*

```
>data=read.csv("dataX12.csv")
>y=data$y
> y
 [1]  85 152  41  93 101  38 203  78 117  44 121 112  50  82  48 127 140 155
39
[20]  90
>x1=data$x1

> x1
 [1]  7 18  5 14 11  5 23  9 16  5 17 12  6 12  8 15 17 21  6 11
>x2=data$x2

> x2
 [1]  5.11 16.72  3.20  7.03 10.98  4.04 22.07  7.03 10.62  4.76 11.02  9.51
[13]  3.79  6.45  4.60 13.86 13.03 15.21  3.64  9.57

>model=lm(y~x1+x2)
> model
Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)         x1          x2
     7.427       3.483       5.150
```

$$\boxed{\hat{Y} = 7.427 + 3.483\, X_1 + 5.150 X_2}$$

Question 4:

In a certain country, the driver's insurance is based on the deriver experience. A random sample of eight drivers insured with a company and having similar auto insurance policies was selected. Summary of the data is given below:

$$X'X = \begin{bmatrix} 8 & 90 \\ 90 & 1396 \end{bmatrix}, \quad X'Y = \begin{bmatrix} 474 \\ 4739 \end{bmatrix} \quad and \quad Y'Y = 29642$$

a. Estimate the simple linear regression equation and interpret the model.

```
> xtx=matrix(c(8,90,90,1396),2,2)
> xty=matrix(c(474,4739),2,1)
> yty=matrix(c(29642),1,1)

> b=solve(xtx)%*%(xty)
> b
        [,1]
[1,] 76.660365
[2,] -1.547588
```

defining the matrices

$$\overbrace{\hat{\beta}_{2\times 1} = (X'X)^{-1}X'Y} = \begin{pmatrix} 76.66 \\ -1.55 \end{pmatrix}$$

$$\hat{Y} = 76.66 - 1.55\,X$$

b_1 = the insurance decrease by 1.55 when the driver experience increase by one year.

b_o = the insurance is 76.66 when the driver have no experience.

b. Calculate SSE, SSTO and SSR

```
> sse=(yty)-t(b)%*%(xty)
> sse
         [,1]
[1,] 639.0065

> sst=(yty)-8*(474/8)^2
> sst
       [,1]
[1,] 1557.5

> ssr=sst-sse
> ssr
         [,1]
[1,] 918.4935
```

$\boxed{SSE = Y'Y - b'X'Y} = 639.0065$

$\boxed{SST = Y'Y - n\,(\bar{Y})^2} = 1557.5$

$\boxed{SSR = SST - SSE} = 918.4923$

c. Calculate the variances of the estimators of the in part (a).

```
> mse=sse/6
> mse
         [,1]
[1,] 106.5011

> vb=mse[1,1]*solve(xtx)
> vb
          [,1]       [,2]
[1,] 48.460077 -3.1242170
[2,] -3.124217  0.2777082
```

$$\overbrace{Var(\hat{\beta}) = MSE(X'X)^{-1}} = \begin{bmatrix} 48.46 & -3.1242 \\ -3.1242 & 0.2777 \end{bmatrix}$$

$$Var(\hat{\beta}_0) = 48.46, \qquad Var(\hat{\beta}_1) = 0.2777$$

d. Estimate 95% confidence interval for the slop of the model.

```
> LB=b[2,1]-qt(0.975,6)*sqrt(vb[2,2])
> LB
[1] -2.837062

> UB=b[2,1]+qt(0.975,6)*sqrt(vb[2,2])
> UB
[1] -0.2581138
```

$$\hat{\beta}_1 - t_{(1-\alpha/_2, n-2)}\ s(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{(1-\alpha/_2, n-2)}\ s(\hat{\beta}_1)$$
$$-2.837 \leq \beta_1 \leq -0.2581$$

e. Test the slope coefficient using t test.

```
> T=b [2,1]/sqrt(vb[2,2])
> T
    [1] -2.93671
```

Hypothesis: $H_0: \beta_1 = 0 \quad vs \quad H_1: \beta_1 \neq 0$

T-statistic $= T_0 = \frac{\widehat{\beta}_1}{s(\widehat{\beta}_1)} = \frac{-1.55}{\sqrt{0.2777}} = -2.93671$

The critical value $= t_{\left(1-\frac{\alpha}{2},\ 6\right)} = \pm 2.446912$

$Since\ \ -2.93671 \notin (-2.446912, 2.446912).$ *We reject* H_0

f. What is the Monthly Auto Insurance Premium of the derivers with 21 years' experience?

```
> xh=matrix(c(1,21),2,1)
> xh
     [,1]
[1,]    1
[2,]   21

> yh=t(xh)%*%b
> yh
        [,1]
[1,] 44.16102
```

$\boxed{\hat{Y} = X_h'\beta} = 44.16102$

g. Constant 90% prediction interval of the mean of the Monthly Auto Insurance Premium with 24 years 'experience?

```
> xh=matrix(c(1,24),2,1)
> yh=t(xh)%*%b
> yh
        [,1]
[1,] 39.51825

> vyhnew=t(xh)%*%vb%*%(xh)+mse
> vyhnew
        [,1]
[1,] 164.9587

> LB=yh-qt(0.95,6)*sqrt(vyhnew)
> LB
        [,1]
[1,] 14.56078

> UB=yh+qt(0.95,6)*sqrt(vyhnew)
> UB
   [,1]
[1,] 64.47573
```

$\boxed{\hat{Y} = X_h'\beta} = 39.51825$

$\boxed{V(Y_{h(new)}) = X_h'\, S^2\{b\}\, X_h + MSE}$

$$\hat{Y} - t_{(1-\alpha/2, n-2)}\ S(Y_{h(new)}) \le \ \mathrm{E}(\hat{Y}_{(new)}) \le \hat{Y} + t_{(1-\alpha/2, n-2)}\ S(Y_{h(new)})$$
$$14.56078 \le \ \mathrm{E}(\hat{Y}_{(new)}) \le 64.47573$$

Question 5:

Suppose we have a sample of student,

X: represent their score in programming exam.

Y: represent their score in ICDL exam.

$$\sum X = 261 \qquad \sum Y = 784$$
$$\sum X^2 = 7363 \qquad \sum XY = 22116$$

$$Y' = [\,89\,,\ 41\,,\ 68\,,\ 63\,,\ 102\ ,\ k\ ,\ 103\,,\ 77\ ,\ 45\ ,108\ ]$$

By matrices:

a. *Find the value of (k).*

$$89 + 41 + 68 + 63 + 102 + k + 103 + 77 + 45 + 108 = 784$$

$$k + 696 = 784$$

$$k = 784 - 696$$

$$k = 88$$

b. *Find the simple regression model* $Y = \beta_0 + \beta_1 X$.

$$X'X = \begin{bmatrix} n & \sum X \\ \sum X & \sum X^2 \end{bmatrix} = \begin{bmatrix} 10 & 261 \\ 261 & 7363 \end{bmatrix} \qquad X'Y = \begin{bmatrix} \sum Y \\ \sum XY \end{bmatrix} = \begin{bmatrix} 784 \\ 22116 \end{bmatrix}$$

```
> xtx=matrix(c(10,261,261,7363),2,2)
> xtx
      [,1]   [,2]
[1,]   10    261
[2,]  261   7363

> xty=matrix(c(784,22116),2,1)
> xty
      [,1]
[1,]  784
[2,]  22116

> y=c(89,41,68,63,102,88,103,77,45,108)
> sum(y*y)
[1] 66570

> yty=matrix(c(66570),1,1)
> b=solve(xtx)%*%(xty)
> b

[1,] 0.05736068

[2,] 3.00163369
```

$$Y = \quad b_0 \ + \ b_1\, X + \varepsilon$$
$$Y = 0.06 + 3\ \ X\ + \varepsilon$$

c. Find MSE.

```
> n=10
> sse=(yty)-t(b)%*%(xty)
> mse=sse/(n-2)
> mse
         [,1]
[1,]  17.61232
```

$\boxed{MSE = 17.61232}$

d. Find the variance matrix for the coefficients.

```
> vb=mse[1,1]*solve(xtx)
> vb
           [,1]          [,2]
[1,]  23.539569   -0.83441905
[2,]  -0.834419    0.03197008
```

$$\boxed{Var(b) = \begin{bmatrix} 23.54 & -0.83 \\ -0.83 & 0.03 \end{bmatrix}}$$

e. Find 95 % confident interval for the slop.

```
> LB=b[2,1]-qt(0.975,8)*sqrt(vb[2,2])
> UB=b[2,1]+qt(0.975,8)*sqrt(vb[2,2])
> LB
[1] 2.589316
> UB
[1] 3.413951
```

$\boxed{2.59 < \beta_1 < 3.41}$

Question 6:

To investigate the linear model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ *, we assume the following data:*

$$X'X = \begin{bmatrix} \boxed{} & 171 & 25500 \\ 171 & 2271 & 262600 \\ 25500 & 262600 & 46150000 \end{bmatrix}, \quad X'Y = \begin{bmatrix} \boxed{} \\ 18410 \\ 3371000 \end{bmatrix}$$

$$Y' = [60\ \ 70\ \ 80\ \ 90\ \ 100\ \ 120\ \ 110\ \ 110\ \ 130\ \ 130\ \ 140\ \ 180\ \ 160\ \ 170\ \ 190]$$

(a) Complete the matrix.

$$X'X = \begin{bmatrix} 15 & 171 & 25500 \\ 171 & 2271 & 262600 \\ 25500 & 262600 & 46150000 \end{bmatrix} \quad X'Y = \begin{bmatrix} \sum Y_i = 1840 \\ 18410 \\ 3371000 \end{bmatrix}$$

(b) Estimate the coefficients of the model and interpret the results.

```
> xtx=matrix(c(15,171,25500,171,2271,262600,25500,262600,46150000),3,3)
> xty=matrix(c(1840,18410,3371000),3,1)
> y=c(60,70,80,90,100,120,110,110,130,130,140,180,160,170,190)
> yty=t(y)%*%(y)
> sum(y)
[1] 1840
> b=solve(xtx)%*%(xty)
> b
          [,1]
[1,] 66.6457413      β̂0
[2,] -3.2154776      β̂1
[3,]  0.0545161      β̂2
```

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

$$\hat{Y} = 66.645 - 3.215\, X_1 + 0.0545\, X_2$$

$\hat{\beta}_0$ = *the value of Y is 66.645. When* $X_1 = X_2 = 0$ *or the intersection with Y axis.*

$\hat{\beta}_1$ = *the value of Y decrease by 3.215. When* X_1 *increase by one unit. With no changes in* X_2 *.*

$\hat{\beta}_2$ = *the value of Y increase by 0.055. When* X_2 *increase by one unit. With no changes in* X_1 *.*

(c) Find the variances of the coefficients.

```
> sse=(yty)-t(b)%*%(xty)
>sse
[,1]
[1,] 795.0056

> n=15
> mse=sse/(n-3)
> mse
[,1]
[1,] 66.25046

>vb=mse[1,1]*solve(xtx)
>vb
           [,1]           [,2]           [,3]
[1,]  1428.0624925  -47.61882779  -0.5181124460
[2,]  -47.6188278     1.67314240   0.0167911791
[3,]   -0.5181124     0.01679118   0.0001921724
```

$\boxed{\mathrm{SSE = Y'Y - b'X'Y}} = 795.0056$

$$Var(\hat{\beta}_0) = 1428.06 \qquad Var(\hat{\beta}_1) = 1.67 \qquad Var(\hat{\beta}_2) = 0.00019$$

(d) Discuss the efficiency of the estimated model by using ANOVA.

```
> sst=(yty)-n*(sum(y)/n)^2
> sst
     [1,] 22293.33
> ssr=sst-sse
> ssr
 [1,] 21498.33
>msr=ssr/(2)
>f=msr/mse
>f
 [1,] 162.2504
>qf(0.95,2,(n-3))
[1] 3.885294
```

$\boxed{SST = Y'Y - n\,(\bar{Y})^2} = 22293.33$

$\boxed{SSR = SST - SSE} = 21498.33$

$$H_0: \beta_1 = \beta_2 = 0 \qquad vs \qquad H_1: at\ least\ one\ \beta_i \neq 0 \qquad i = 1,2$$

$$F = \frac{MSR}{MSE} = 162.2504 > F_{1-\alpha,2,n-2} = 3.885294$$

We reject H_0 (the is a relation between X_1, X_2 and Y) *(e) Calculate 95% confidence interval of β_1.*

```
> LB=b[2,1]-qt(0.975,(n-3))*sqrt(vb[2,2])
> LB
[1] -6.033772

> UB=b[2,1]+qt(0.975,(n-3))*sqrt(vb[2,2])
> UB
[1] -0.3971831
```

$$\hat{\beta}_1 - t_{(1-\alpha/2, n-p)}\ S(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{(1-\alpha/2, n-p)}\ S(\hat{\beta}_1)$$

$$-6.034 \leq \beta_1 \leq -0.397$$

Question 7:

A forensic study related Hand (X_1) *and foot* (X_2) *length* (Y) *for a sample of n=75 adults (each variable in 100s of mms). Consider the following three models* M_1, M_2 *and* M_3

$$M_1\!: Y = \beta_0 + \beta_1 X_1 \;,\quad X'X = \begin{bmatrix} 75 & 142.185 \\ 142.185 & 270.1992 \end{bmatrix} \;,\quad X'Y = \begin{bmatrix} 1199.70 \\ 2276.80 \end{bmatrix} ,Y'Y = 19208.28$$

$$M_2\!: Y = \beta_0 + \beta_2 X_2 \;,\; X'X = \begin{bmatrix} 75 & 176.062 \\ 176.062 & 414.390402 \end{bmatrix} \;,\; X'Y = \begin{bmatrix} 1199.70 \\ 2819.37 \end{bmatrix}$$

$$M_2\!: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \;,\; X'X = \begin{bmatrix} 75 & 142.185 & 176.062 \\ 142.185 & 270.1992 & 334.2884 \\ 176.062 & 334.2884 & 414.390402 \end{bmatrix} \quad X'Y = \begin{bmatrix} 1199.70 \\ 2276.80 \\ 2819.37 \end{bmatrix}$$

a. *Calculate SST, SSR, SSE and* $Y'\left(\frac{1}{n}J\right)Y$ *for each model, where J is a square matrix of ones of dimension* $(n \times n)$.

```
------M1-------
xtx=matrix(c(75,142.185,142.185,270.1992),2,2)
xty=matrix(c(1199.7,2276.8),2,1)
yty=matrix(c(19208.28),1,1)
b=solve(xtx)%*%(xty)
sse=(yty)-t(b)%*%(xty)
sst=(yty)-75*(1199.75/75)^2
ssr=sst-sse
sse
ssr
sst
YnY=19208.28-sst
```

$\boxed{SST = Y'Y - Y'\left(\frac{1}{n}J\right)Y}$

```
------M2-------
xtx=matrix(c(75,176.062,176.062,414.390402),2,2)
xty=matrix(c(1199.7,2819.37),2,1)
yty=matrix(c(19208.28),1,1)
b=solve(xtx)%*%(xty)
sse=(yty)-t(b)%*%(xty)
sst=(yty)-75*(1199.75/75)^2
ssr=sst-sse
sse
ssr
sst
YnY=19208.28-sst
------M3-------
xtx=matrix(c(75,142.185,176.062,142.185,270.1995,334.2884,176.062,334.2884,414.390402),3,3)
xty=matrix(c(1199.7,2276.8,2819.37),3,1)
yty=matrix(c(19208.28),1,1)
b=solve(xtx)%*%(xty)
sse=(yty)-t(b)%*%(xty)
sst=(yty)-75*(1199.75/75)^2
ssr=sst-sse
sse
ssr
sst
YnY=19208.28-sst
```

	SSE	SSR	SST	$Y'\left(\frac{1}{n}J\right)Y$
M1	8.88	7.40	16.28	19192
M2	9.13	7.15	16.28	19192
M3	6.85	9.43	16.28	19192

b. *Calculate the* $SSR(X_1|X_2)$ *and* $SSR(X_2|X_1)$

$$SSR\left(\overbrace{X_1,\dots,X_n}|\underbrace{X_{n+1},\dots,X_m}\right) = SSR\left(\overbrace{X_1,\dots,X_m}\right) - SSR\left(\underbrace{X_{n+1},\dots,X_m}\right)$$

$$SSR(X_1,\dots,X_n|X_{n+1},\dots,X_m) = SSE(X_{n+1},\dots,X_m) - SSE(X_1,\dots,X_m)$$

$$SSR(X_1|X_2) = SSR(X_1,X_2) - SSR(X_2) = 9.43 - 7.15 = 2.28$$

$$SSR(X_2|X_1) = SSR(X_1,X_2) - SSR(X_1) = 9.43 - 7.40 = 2.03$$

c. *Use Partial F to test* $H_0\colon \beta_2 = 0 \quad vs \quad H_1\colon \beta_2 \neq 0$

$$F^* = \frac{MSR(X_2|X_1)}{MSE(X_2,X_1)} = \frac{2.03/1}{6.85/72} = 21.34 > F_{0.95,1,72} = 3.97$$

$$We\ reject\ H_0$$

d. *When p=4 Prove that*

$$\boldsymbol{SSR(X_1,X_2,X_3) = SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1,X_2)}$$

$$= \cancel{SSR(X_1)}$$
$$+SSR\cancel{(X_1,X_2)} - \cancel{SSR(X_1)}$$
$$+SSR(X_1,X_2,X_3) - \cancel{SSR(X_1,X_2)} = SSR(X_1,X_2,X_3)$$

Question 8:

Use the data in file "hwX".

a. *Find the regression model* $Y = \beta_o + \beta_1 X_1 + \beta_2 X_2$.

```
> df=read.csv("hwX.csv")
> x1=df$x1
> x2=df$x2
> y=df$y
> model=lm(y~x1+x2)
> summary(model)

Call:

lm(formula = y ~ x1 + x2)

Coefficients:

              Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)  -3.9265     4.2678       -0.920     0.388
x1            3.1605     0.2376       13.303     3.17e-06 ***
x2            0.4023     0.4821        0.834     0.432
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.53 on 7 degrees of freedom
Multiple R-squared:  0.9751,    Adjusted R-squared:  0.968
F-statistic: 137.2 on 2 and 7 DF,  p-value: 2.426e-06
```

$$\boxed{Y = -3.93 + 3.16\, X_1 + 0.402 X_2}$$

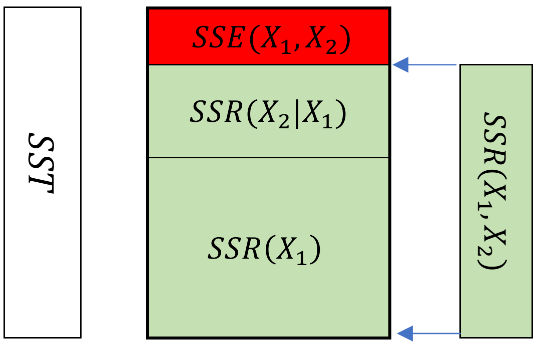
b. *Test,* $H_0: \beta_2 = 0 \quad vs \quad H_1: \beta_2 \neq 0$

p-value = 0.432 >0.05, we accept H_0

c. Find the ANOVA table and F-test.

```
> anova (model)
Analysis of Variance Table
Response: y
              Df   Sum Sq   Mean Sq    F value     Pr(>F)
x1             1   640.88   640.88     273.7595    7.192e-07 ***   SSR(X1)
x2             1   1.63     1.63       0.6962      0.4316          SSR(X2|X1)
Residuals      7   16.39    2.34
```

	df	*SS*	*MS*	*F*
Regression (R)	*2*	*641.51*	*320.755*	*137.07*
Error (E)	*7*	*16.39*	*2.34*	
Total (T)	*9*	*657.9*		



$H_0: \beta_1 = \beta_2 = 0 \quad vs \quad H_1: at\ least\ one\ \neq 0$

p-value =2.426e-06 <0.05, we reject H_0

d. What is the different between T-test in (a) and F-test in (b)?

T-test: to test only β_2 *(if* X_2 *can be removed from the model)*
F-test: to test the whole model.

Question 9:

Y	26	13	8	21	8	4	11	3	28	19
X_1	6	3	3	2	5	2	6	3	10	4
X_2	21	15	20	14	25	21	22	24	22	13
X_3	35	23	22	31	25	21	21	21	30	23

f. Find the regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$.

```
> y=c(26,13,8,21,8,4,11,3,28,19)
> x1=c(6,3,3,2,5,2,6,3,10,4)
> x2=c(21,15,20,14,25,21,22,24,22,13)
> x3=c(35,23,22,31,25,21,21,21,30,23)
>
> model=lm(y~x1+x2+x3)
> summary(model)

Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
     Min       1Q   Median      3Q     Max
-1.32735 -0.00617  0.23515  0.42670  0.64115

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.98386    2.26151   0.877    0.414
x1           2.08013    0.14384  14.461 6.85e-06 ***
x2          -1.09913    0.07661 -14.348 7.18e-06 ***
x3           0.97684    0.06576  14.855 5.86e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.854 on 6 degrees of freedom
Multiple R-squared:  0.9939,    Adjusted R-squared:  0.9908
F-statistic: 325.6 on 3 and 6 DF,  p-value: 4.965e-07
```

$$\boxed{Y = 1.98 + 2.08\,X_1 - 1.1\,X_2 + 0.98\,X_3}$$

g. Find R^2.

$$R^2 = 0.9939$$

h. Test $H_0: \beta_2 = 0$ vs $H_1: \beta_2 \neq 0$.

$$T = -14.348$$

$$P - value = 7.18e - 06 < 0.05$$

$$Then\ we\ reject\ H_0$$

---

i. Find 90 % C.I for the coefficients.

```
> confint(model, level=0.90)
                   5 %          95 %
(Intercept)  -2.4106558    6.3783751
x1            1.8006274    2.3596407
x2           -1.2479872   -0.9502645
x3            0.8490554    1.1046216
```

$$-2.41 < \beta_0 < 6.38$$

$$1.80 < \beta_1 < 2.36$$

$$-1.25 < \beta_2 < -0.95$$

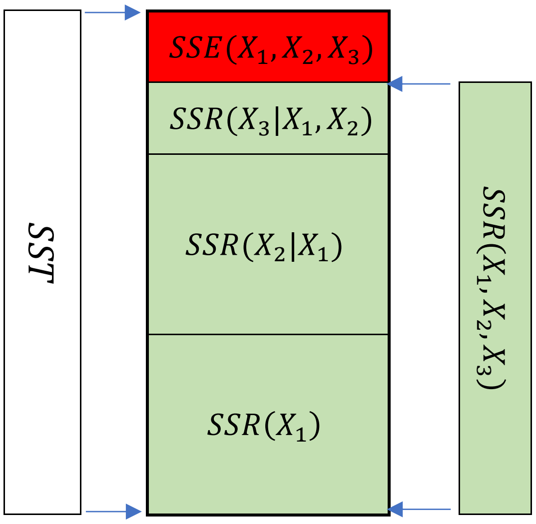
$$0.85 < \beta_3 < 1.10$$

j. Find the ANOVA table.

> model=lm(y~x1+x2+x3)
> anova(model)
Analysis of Variance Table
Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)		
x1	1	262.944	262.944	360.50	1.380e-06	***	$SSR(X_1)$
x2	1	288.633	288.633	395.72	1.047e-06	***	$SSR(X_2\|X_1)$
x3	1	160.946	160.946	220.66	5.855e-06	***	$SSR(X_3\|X_1,X_2)$
Residuals	6	4.376	0.729				$SSE(X_1,X_2,X_3)$

	df	*SS*	*MS*	*F*
Regression	*3*	*712.523*	*237.5*	*325.6*
Error	*6*	*4.374*	*0.729*	
Total	*9*	*716.894*		



> model7=lm(y~x2+x3+x1)
> anova(model7)
Analysis of Variance Table
Response: y

	Df	Sum Sq	MeanSq	F value	Pr(>F)			
x2	1	92.51	92.51	126.83	2.930e-05	***	$SSR(X_2)$	
x3	1	467.47	467.47	640.92	2.502e-07	***	$SSR(X_3\|X_2)$	712.523
x1	1	152.54	152.54	209.13	6.851e-06	***	$SSR(X_1\|X_2,X_3)$	
Residuals	6	4.376	0.729				$SSE(X_1,X_2,X_3)$	

k. ***Calculate $SSR(X_2, X_3|X_1)$ and $SSR(X_1, X_3|X_2)$:***

$$SSR\left(\overbrace{X_1, \dots, X_n | \underbrace{X_{n+1}, \dots, X_m}}\right) = SSR\left(\overbrace{X_1, \dots, X_m}\right) - SSR\left(\underbrace{X_{n+1}, \dots, X_m}\right)$$

$$SSR(X_1, \dots, X_n | X_{n+1}, \dots, X_m) = SSE(X_{n+1}, \dots, X_m) - SSE(X_1, \dots, X_m)$$

- $SSR(X_2, X_3|X_1) = SSR(X_1, X_2, X_3) - SSR(X_1)$

  $= \quad 712.523 \quad - 262.944 = 449.579$

- $SSR(X_1, X_3|X_2) = SSR(X_1, X_2, X_3) - SSR(X_2)$

  $= \quad 712.523 \quad - \quad 92.51 \quad = 620.013$

```
> model=lm(y~x2+x1+x3)
> anova(model)
Analysis of Variance Table
Response: y
           Df   Sum Sq   Mean Sq   F value    Pr(>F)
x2          1   92.51    92.51     126.83     2.930e-05 *** SSR(X2)
x1          1   459.07   459.07    629.39     2.641e-07 *** SSR(X1|X2)
x3          1   160.95   160.95    220.66     5.855e-06 *** SSR(X3|X1,X2)
Residuals   6   4.376    0.73
```

Question 10:

y	85	152	41	93	101	38	203	78	117	44	121	112	50	82	48	127	140	155	39	90
x1	7	18	5	14	11	5	23	9	16	5	17	12	6	12	8	15	17	21	6	11
x2	5.1	17	3.2	7	11	4	22	7	11	4.8	11	9.5	3.8	6.5	4.6	14	13	15	3.6	9.6

1. Estimate the liner model for the given data and interpret its coefficients.

```
>data=read.csv("dataX12.csv")
>y=data$y
>x1=data$x1
>x2=data$x2

> y
 [1]  85 152  41  93 101  38 203  78 117  44 121 112  50  82  48 127 140 155  39
[20]  90

> x1
 [1]  7 18  5 14 11  5 23  9 16  5 17 12  6 12  8 15 17 21  6 11

> x2
 [1]  5.11 16.72  3.20  7.03 10.98  4.04 22.07  7.03 10.62  4.76 11.02  9.51
[13]  3.79  6.45  4.60 13.86 13.03 15.21  3.64  9.57

>model=lm(y~x1+x2)
> model
Call:
lm(formula = y ~ x1 + x2)

Coefficients:

(Intercept)        x1         x2

     7.427       3.483       5.150
```

$$\boxed{\begin{aligned}\hat{Y} &= b_0 + b_1X_1 + b_2X_2 \\ \hat{Y} &= 7.427 + 3.483\,X_1 + 5.150X_2\end{aligned}}$$

b_0 : *the value of* y *when* $x_1 = 0$ *and* $x_2 = 0$.

b_1 : *the changes in* y *when* x_1 *Increase by one unit with no changing in* x_2.

b_2 : *the changes in* y *when* x_2 *Increase by one unit with no changing in* x_1.

2. Discuss the efficiency of the model by two different approaches.

```
> summary(model)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q   Median      3Q     Max
-10.980  -4.426  -1.066   1.894  26.876

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.4274     4.8896   1.519 0.147136
x1            3.4827     0.9655   3.607 0.002174 **
x2            5.1501     1.0461   4.923 0.000129 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.83 on 17 degrees of freedom
Multiple R-squared:  0.9667,	Adjusted R-squared:  0.9627
F-statistic: 246.5 on 2 and 17 DF,  p-value: 2.785e-13
```

1. *Using (T-test):*

$H_0: \beta_1 = 0 \quad vs \quad H_a: \beta_1 \neq 0 \quad P - value \ = 0.002174 < 0.05 \ \ Then\ we\ reject\ H_0$

$H_0: \beta_2 = 0 \quad vs \quad H_a: \beta_2 \neq 0 \quad P - value \ = 0.000129 < 0.05 \ \ Then\ we\ reject\ H_0$

2. *Using ANOVA table (F-test):*

$$H_0: \beta_1 = \beta_2 = 0 \qquad vs \qquad H_a: at\ least\ \beta_i \neq 0 \ ; i = 1,2$$

$$P - value = 2.785\text{e} - 13 < 0.05$$

$$Then\ we\ reject\ H_0$$

3. Write the ANOVA table that factorize the sum square regression X_1 and X_2 given X_1.

```
> anova(model)
Analysis of Variance Table
Response: y
           Df   Sum Sq   Mean Sq   F value    Pr(>F)
x1          1   36542    36542     468.714    8.163e-14 ***
x2          1   1890     1890      24.237     0.0001287 ***
Residuals  17   1325     78
```

$SSR(X_1) = 36542$

$SSR(X_2|X_1) = 1890$

4. Use partial F to test whether you can remove X_2 from model.

$H_0: \beta_2 = 0 \quad vs \quad H_a: \beta_2 \neq 0$

$$MSR(X_2|X_1) = \frac{SSR(X_2|X_1)}{1} = \frac{1890}{1} = 1890$$

$$F^* = \frac{MSR(X_2|X_1)}{MSE(X_2, X_1)} = \frac{1890}{78} = 24.23 \Longrightarrow F_{0.95,1,17} = 4.45 < F^* = 24.23$$

$P - value = 0.0001287 < 0.05$ *Then we reject H_0 (we can't remove X_2 from model)*

5. Is $SSR(X_1|X_2) = SSR(X_2)$? Explain?

```
> model2=lm(y~x2+x1)
> anova(model2)

Analysis of Variance Table
Response: y
          Df   Sum Sq   Mean Sq   F value    Pr(>F)
x2         1   37417    37417     479.938    6.719e-14 ***
x1         1   1015     1015      13.013     0.002174 **
Residuals 17   1325     78
```

$SST(X_1, X_2) = 37417 + 1015 + 1325 = 39757$

$$SSR(X_2) = 37417 \Rightarrow \frac{SSR(X_2)}{SST(X_2)} = \frac{37417}{39757} = 0.941$$

$$SSR(X_1|X_2) = 1015 \Rightarrow \frac{SSR(X_1|X_2)}{SST(X_2, X_1)} = \frac{1015}{39757} = 0.026$$

بإضافتنا المتغير X_1 في نموذج معطى فية X_2 استطعنا تفسير 2.6% اضافة لما كان قد تم تفسيرة من التغيرفي Y وقد كان 94.1%

6. *Calculate R^2 , $r^2_{Y\,2.1}$, $r_{Y\,1.2}$, $r^2_{Y\,2}$.*

$\boldsymbol{R^2} = 96.67\%$

وهذا يعني ان النموذج فسر 96.67% من التغير في Y باستخدام X_1 و X_2 ويعتبر مؤشر جيد.

$$r^2_{Y2.1} = \frac{SSR(X_2|X_1)}{SSE(X_1)} = \frac{1890}{SST - SSR(X_1)} = \frac{1890}{39757 - 36542} = \frac{1890}{3215} = 0.59$$

وهو معامل التحديد الجزئي بين Y و X_2 علما أن X_1 في النموذج وهو يقيس :

التخفيض النسبي في تغير Y المتبقي بعد أن كان X_1 في النموذج والذي يتم اكتسابة بضم X_2 أيضا إلى النموذج.

$$r_{Y1.2} = \sqrt{\frac{SSR(X_1|X_2)}{SSE(X_2)}} = \sqrt{\frac{1015}{SST - SSR(X_2)}} = \sqrt{\frac{1015}{39757 - 37417}} = \sqrt{\frac{1015}{2340}} = 0.66$$

وهو معامل الارتباط الجزئي بين Y و X_1 علما أن X_2 في النموذج وهو يقيس :

التخفيض النسبي في تغير Y المتبقي بعد أن كان X_2 في النموذج والذي يتم اكتسابة بضم X_1 أيضا إلى النموذج

$$r^2_{Y2} = 0.9411$$

```
> summary(lm(y~x2))$r.squared
 [1]  0.9411451
```

وهو معامل التحديد بين Y و X_2

7. ***Estimate the corresponding standard model and discuss its coefficient.***

```
> model5=lm(scale(y)~scale(x1)+scale(x2))
> model5

Call:
lm(formula = scale(y) ~ scale(x1) + scale(x2))

Coefficients:
(Intercept)    scale(x1)    scale(x2)
 9.679e-17    4.235e-01    5.779e-01
```

$$Y' = 0.423\,X'_1 + 0.578\,X'_2$$

إذن الزيادة بمقدار انحراف معياري واحد في X_2 مع ثبات X_1 يؤدي إلى زيادة <u>أكبر</u> في Y من الزيادة اللتي $0.423 < 0.578$ يؤدي

إليها زيادة انحراف معياري واحد في X_1 مع ثبات X_2

Question 11:

Use the data in file "d6".

a. ***Find the model*** $\boldsymbol{Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4}$.

```
> d6=read.csv("d6.csv")
> y=d6$y
> x1=d6$x1
> x2=d6$x2
> x3=d6$x3
> x4=d6$x4
> model1=lm(y~x1+x2+x3+x4)
> summary(model1)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.0443  -2.6966  -0.0322   2.2315  13.1724 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  11.5751     6.8292   1.695 0.092232 .  
x1            0.9175     0.3710   2.473 0.014549 *  
x2            2.0948     0.3750   5.586 1.11e-07 ***
x3            2.2067     0.5933   3.719 0.000285 ***
x4            4.0288     0.2209  18.234  < 2e-16 ***

Residual standard error: 3.999 on 145 degrees of freedom
Multiple R-squared:  0.7245,	Adjusted R-squared:  0.7168 
F-statistic: 95.31 on 4 and 145 DF,  p-value: < 2.2e-16
```

$$\boxed{Y = 11.58 + 0.92\, X_1 + 2.09\, X_2 + 2.21\, X_3 + 4.03\, X_4}$$

b. ***Test if we can remove*** $\boldsymbol{X_3}$ ***from the model.***

$$H_0: \beta_3 = 0 \quad vs \quad H_1: \beta_3 \neq 0$$

$$P\text{-}value = 0.000285 < 0.05$$

We reject H_0 *(we can't remove* X_3 *from the model)*

c. *Find the ANOVA table and F-test.*

```
> anova(model1)
Analysis of Variance Table

Response: y
           Df   Sum Sq  Mean Sq  F value    Pr(>F)
x1          1     20.7     20.7   1.2921  0.257535
x2          1    586.8    586.8  36.6852  1.136e-08 ***
x3          1    172.0    172.0  10.7548  0.001302 **
x4          1   5317.9   5317.9 332.4893  < 2.2e-16 ***
Residuals 145   2319.2     16.0
```

$SSR(X_1)$

$SSR(X_2|X_1)$

$SSR(X_3|X_1,X_2)$

$SSR(X_4|X_1,X_2,X_3)$

	df	*SS*	*MS*	*F*
Regression (R)	*4*	*6097.4*	*1524.35*	*95.3*
Error (E)	*145*	*2319.2*	*36.765*	
Total (T)	*149*	*8416.6*		

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

$H_1: at\ least\ one\ \ \beta_i \neq 0$

P-value = 2.2e-16 < 0.05

We reject H_0 (not all betas equal to zero)

d. *Find $R^2_{Y34.12}$.*

$$R^2_{Y34.12} = \frac{SSR(X_3,X_4|X_1,X_2)}{SSE(X_1,X_2)}$$

$$= \frac{SSR(X_1,X_2,X_3,X_4) - SSR(X_1,X_2)}{SST - SSR(X_1,X_2)}$$

$$= \frac{6097.4-(20.7+586.8)}{8416.6-(20.7+586.8)} = 0.70$$

Calculating the AIC:

$$AIC = -2(logL) + 2(p+1)$$

$$AIC = \overbrace{n\left(log(2\pi)\right) + 1 + log\left(\frac{SSE}{n}\right)}^{-2(logL)} + 2(p+1)$$

- *Example:*

for the data [cars] *in R,*

Find the model $\widehat{Distance} = b_0 + b_1(Speed)$ *and calculate AIC*

```
> cars
> head(cars)
  speed dist
1     4    2
2     4   10
3     7    4
4     7   22
5     8   16
6     9   10

> x=cars$speed
> y=cars$dist

> model=lm(y~x)
> model

Call:
lm(formula = y ~ x)
Coefficients:
(Intercept)          x
   -17.579       3.932
```

the mode is $\boxed{\widehat{Distance} = -17.579 + 3.932(Speed)}$

```
> AIC(model)
[1] 419.1569
```

$\boxed{AIC = 419.1569}$

---------------*Or*-----------------

$$AIC = -2\,\underbrace{(logL)} + 2(p+1)$$

```
> logL=logLik(model)
> aic=-2*logL+2*(3)
> aic
'log Lik.' 419.1569 (df=3)
```

Problem:

Consider a company that markets and repairs small computers. 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 calls was taken. The data consist of the length of service calls in minutes (the response variable) and the number of components repaired (the predictor variable). The data are presented in the table below:

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

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.162      3.355    1.24    0.239
x             15.509      0.505   30.71 8.92e-13 ***
---
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

Analysis of Variance Table
Response: y
          Df   Sum Sq     Mean Sq    F value     Pr(>F)
x          1   27419.5    27419.5    943.2       8.916e-13 ***
Residuals 12   348.8      29.1
```

Find the model and its AIC:

$$\hat{y} = b_o + b_1 x$$

$$\hat{y} = 4.16 + (15.51)x$$

$$(\widehat{service\ time}) = 4.16 + (15.51)(\#of\ units)$$

Finding AIC for the model:

$$\boxed{AIC = \overbrace{n\left(log(2\pi)\right) + 1 + log\left(\frac{SSE}{n}\right)}^{-2(logL)} + 2(p+1)}$$

```
> n=14
> p=2
> sse=348.8

> aic=n*(log(2*pi)+1+log(sse/n))+2*(p+1)
> aic
[1] 90.74646
```

$\boxed{AIC = 90.74646}$

- ***Model selection:***

Question:

Use the data in file "d.sw".

e. Find the model $\boldsymbol{Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4}$.

```
> df=read.csv("d.sw.csv")
> head(df)
         y    x1   x2   x3   x4
1   19.740497   11   5   41   16
2   -1.502785   14   5   26   13
3   -6.515215    8   5   10   15
4   -2.679700   13   8   23   18
5   42.994423    2   3   46   11
6  -14.988794   13   6   12   13

> model=lm(y~.,data=df)
> summary(model)

Call:
lm(formula = y ~ ., data = df)

Residuals:
    Min        1Q    Median      3Q      Max
-1.89489  -0.66960  0.04773  0.67010  1.90896

Coefficients:
             Estimate   Std. Error  t value   Pr(>|t|)
(Intercept) -0.257179   0.730359   -0.352    0.726
x1          -2.032584   0.028174   -72.144   <2e-16 ***
x2           0.036195   0.047240    0.766    0.447
x3           1.011206   0.009607   105.258   <2e-16 ***
x4           0.002141   0.037902    0.056    0.955
---

Residual standard error: 0.8708 on 55 degrees of freedom
Multiple R-squared:  0.9965,    Adjusted R-squared:  0.9962
F-statistic:  3863 on 4 and 55 DF,  p-value: < 2.2e-16
```

$$\hat{Y} = -0.26 - 2.03X_1 + 0.04X_2 + 1.01X_3 + 0.002X_4$$

b. Select the best combination of predictor variables for building an optimal predictive model.

```
> step(model, direction="backward")

Start:  AIC=-11.82
y ~ x1 + x2 + x3 + x4

        Df   Sum of Sq    RSS      AIC
- x4     1      0.0       41.7   -13.813
- x2     1      0.4       42.2   -13.179
<none>                    41.7   -11.816
- x1     1   3947.1     3988.8   259.815
- x3     1   8402.1     8443.8   304.810

Step:  AIC=-13.81
y ~ x1 + x2 + x3

        Df   Sum of Sq    RSS      AIC
- x2     1      0.4       42.2   -15.179
<none>                    41.7   -13.813
- x1     1   3955.3     3997.0   257.937
- x3     1   8543.5     8585.2   303.807

Step:  AIC=-15.18
y ~ x1 + x3

        Df   Sum of Sq    RSS      AIC
<none>                    42.2   -15.179
- x1     1   3968.7     4010.9   256.146
- x3     1   8708.9     8751.1   302.955

Call:
lm(formula = y ~ x1 + x3, data = df)

Coefficients:
(Intercept)         x1          x3
  -0.008352   -2.033702    1.010219
```

طريقة (Backward) تضيف جميع المتغيرات المستقلة وتبدأ بحذف اقل (AIC)

Removing X4

Removing X2

$$\boxed{\hat{Y} = -0.0084 - 2.03X_1 + 1.01X_3}$$

```
> modelstart=lm(y~1,data=df)
> summary(modelstart)
```

طريقة (Forward) تحذف جميع المتغيرات المستقلة وتبدأ بإضافة المتغير المستقلة واحدا تلو الآخر حسب (AIC)

```
Call:
lm(formula = y ~ 1, data = df)

Residuals:
    Min       1Q      Median     3Q       Max
-26.908   -12.287   -0.324    7.613    32.117

Coefficients:
             Estimate   Std. Error  t value    Pr(>|t|)
(Intercept)  11.920     1.823       6.539      1.62e-08 ***
---
Residual standard error: 14.12 on 59 degrees of freedom

> step(modelstart, direction="forward", scope=formula(model))
Start:  AIC=318.69
y ~ 1

          Df   Sum of Sq    RSS      AIC
+ x3       1    7749.9      4010.9   256.15
+ x1       1    3009.8      8751.1   302.96
<none>                     11760.8   318.69
+ x4       1     209.3     11551.5   319.61
+ x2       1     90.1      11670.7   320.23
```

Adding X3

```
Step:  AIC=256.15
y ~ x3

          Df   Sum of Sq   RSS      AIC
+ x1       1    3968.7     42.2     -15.179
<none>                     4010.9   256.146
+ x2       1     13.9      3997.0   257.937
+ x4       1     6.9       4004.0   258.042
```

Adding X1

```
Step:  AIC=-15.18
y ~ x3 + x1

          Df   Sum of Sq   RSS      AIC
<none>                     42.155   -15.179
+ x2       1    0.44283    41.712   -13.813
+ x4       1    0.00007    42.155   -13.179

Call:
lm(formula = y ~ x3 + x1, data = df)

Coefficients:
(Intercept)         x3          x1
  -0.008352     1.010219    -2.033702
```

$$\boxed{\hat{Y} = -0.0084 + 1.01X_3 - 2.03X_1}$$

طريقة (Both) يمكن ان يكون هناك حذف أوإضافة حسب كل خطوة

```
> step(model, direction="both")
Start:  AIC=-11.82
y ~ x1 + x2 + x3 + x4

        Df   Sum of Sq    RSS      AIC
- x4     1       0.0      41.7   -13.813
- x2     1       0.4      42.2   -13.179
<none>                    41.7   -11.816
- x1     1    3947.1    3988.8   259.815
- x3     1    8402.1    8443.8   304.810

Step:  AIC=-13.81
y ~ x1 + x2 + x3

        Df   Sum of Sq    RSS      AIC
- x2     1       0.4      42.2   -15.179
<none>                    41.7   -13.813
+ x4     1       0.0      41.7   -11.816
- x1     1    3955.3    3997.0   257.937
- x3     1    8543.5    8585.2   303.807

Step:  AIC=-15.18
y ~ x1 + x3

        Df   Sum of Sq    RSS      AIC
<none>                    42.2   -15.179
+ x2     1       0.4      41.7   -13.813
+ x4     1       0.0      42.2   -13.179
- x1     1    3968.7    4010.9   256.146
- x3     1    8708.9    8751.1   302.955

Call:
lm(formula = y ~ x1 + x3, data = df)

Coefficients:
(Intercept)           x1           x3
  -0.008352    -2.033702     1.010219
```

$$\boxed{\hat{Y} = -0.0084 - 2.03X_1 + 1.01X_3}$$

طريقة (Both) يمكن ان يكون هناك حذف أوإضافة حسب كل خطوة

```
> step(modelstart, direction="both", scope=formula(model))
Start:  AIC=318.69
y ~ 1

         Df  Sum of Sq    RSS     AIC
+ x3      1    7749.9   4010.9  256.15
+ x1      1    3009.8   8751.1  302.96
<none>                 11760.8  318.69
+ x4      1     209.3  11551.5  319.61
+ x2      1      90.1  11670.7  320.23

Step:  AIC=256.15
y ~ x3

         Df  Sum of Sq    RSS     AIC
+ x1      1    3968.7     42.2  -15.18
<none>                  4010.9  256.15
+ x2      1      13.9   3997.0  257.94
+ x4      1       6.9   4004.0  258.04
- x3      1    7749.9  11760.8  318.69

Step:  AIC=-15.18
y ~ x3 + x1

         Df  Sum of Sq    RSS     AIC
<none>                    42.2  -15.179
+ x2      1       0.4     41.7  -13.813
+ x4      1       0.0     42.2  -13.179
- x1      1    3968.7   4010.9  256.146
- x3      1    8708.9   8751.1  302.955

Call:
lm(formula = y ~ x3 + x1, data = df)

Coefficients:
(Intercept)          x3          x1
  -0.008352    1.010219   -2.033702
```

$$\boxed{\hat{Y} = -0.0084 + 1.01X_3 - 2.03X_1}$$

more data in R:

```
trees
rock
randu
sleep
Orange
airquality
```