REG No. SST S04 | SBN S04

Write on both sides of the paper

Question (1)

Do not write in either margin

Simple Linear Regression Model

The model is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1\ 2\ 3\ 4 \dots n$$

Simple since there is only one x_i

It is linear in β_0 and β_1

$E(y_i) = \beta_0 + \beta x_i$ is linear

$y = \beta_0 + \beta_1 x_1 + e^{\beta_2 x_2} + \varepsilon_i$ is not linear

$y = \beta_0 + e^{\beta_1 x_1} + \varepsilon_i$ is not linear also.

- A simple linear regression model assumes that the following specification is true in the population

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

- Where other unobserved factors determining y are captured by the error term ε.
- To estimate the parameters β_0 & β_1 in the model Assumptions 1-5, Linearity, identification, exogeneity, spherical error terms and data generation are required for Ordinary least squares. (OLS)
- Assumptions 1-6, ie including normality are required to use Maximum likelihood to estimate β_0 & β_1 (MLE)

Assumptions summarized

i $E(\varepsilon_i) = 0$ for all $i = 1 2 \dots n \Rightarrow E(y_i) = \beta_0 + \beta_1 x_i$

ii $Var(\varepsilon_i) = \sigma^2$ for all $i = 1 2 \dots n \Rightarrow Var(y_i) = \sigma^2$ Constant

iii Covariance$(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j \Rightarrow Cov(y_i\, y_j) = 0$ uncorrelatedness

iv Normality $\varepsilon_i \sim N(0, \sigma^2)$, ε_i and y_i are independent and uncorrelated.

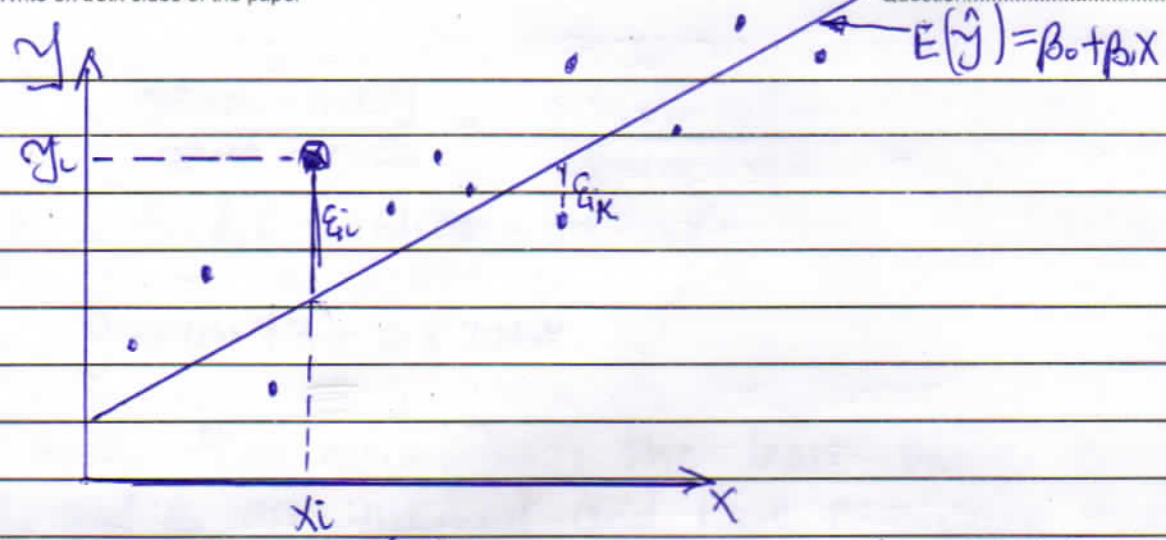
Ordinary least squares estimator

- Least squares Method does not require any distributional assumptions. (It does not require normality)
- MLE estimation requires the normality assumption (6)
- ordinary least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ Minimizes the sum of squared residuals (SSR)

$$Q = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\varepsilon}'\hat{\varepsilon} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

REG No.

Write on both sides of the paper — Question 2 — Do not write in either margin



$\hat{y}_i$ estimates $E(y_i) = \beta_0 + \beta_1 x_i$ not $\beta_0 + \beta_1 x + \varepsilon_i$.

The minimum is acquired by.

$$\frac{\partial Q}{\partial \hat{\beta}_0} = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\partial Q / \partial \hat{\beta}_1 = -2 \sum x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

Simplify

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i = \sum y_i$$

$$\hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum x_i y_i$$

Solving for

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \quad \text{OR} \quad \hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Note

$$S_{xx} = \sum_i (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

$$S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

Example: Is a student performance in final exam (y) determined by continuous assessment (x)

y	95	80	0	0	79	77	72	66	98	90	0	95	35	50	72	55	75	66	$\sum y =$
x	96	77	0	0	78	64	89	47	90	93	18	86	0	30	59	77	74	67	$\sum x =$
x^2																			$\sum x^2 =$
xy																			$\sum xy =$
y^2																			$\sum y^2 =$

REG No.

Write on both sides of the paper — Question 3 — Do not write in either margin

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{81{,}195 - 18(58.056)(61.389)}{80199 - 18(58.056)^2} = 0.8726$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 61.389 - 0.8726(58.056) = 10.73$$

$$\hat{y} = 10.73 + 0.8726x.$$

With the three assumptions the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased and have minimum variance among all linear unbiased estimators "MVUE"

$$E(\hat{\beta}_1) = \beta_1$$

$$E(\hat{\beta}_0) = \beta_0$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum(x_i-\bar{x})^2} = \frac{\sigma^2}{S_{xx}}$$

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

If $\sigma^2 = E(y_i - E(y_i))^2$ using $\hat{y}$ as estimate for $E(y_i)$ weestimate $\hat{\sigma}^2$ by

$$S^2 = \frac{\sum(y_i - \hat{y})^2}{n-2}.$$

Note

$$Var(y_i) = E(y_i - E(y_i))^2 = E(y_i - \beta_0 - \beta_1 x_i)^2 = E(\hat{\varepsilon}_i^2) = \sigma^2$$

$$S^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)}{n-2} = \frac{SSE}{n-2}$$

$\hat{\varepsilon} = (y_i - \hat{y}_i)$ is refered to residuals

$$E(S^2) = E\frac{SSE}{n-2} = \frac{(n-2)\sigma^2}{n-2} = \sigma^2 \quad \text{unbiased}$$

S^2 is an unbiased estimators of σ^2

$$SSE = \sum(y_i - \hat{y}_i)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}} \qquad S^2 = \frac{SSE}{n-2}$$

REG No.

Write on both sides of the paper

Question ④

Do not write in either margin

Hypothesis Testing and Confidence Intervals for $\hat{\beta}_1$

$H_0: \beta_1 = 0$ Vs $H_1\ \beta_1 \neq 0$:

$H_0: \beta_1 = 0$ There is no linear relationship between y and x

Assuming $\varepsilon \sim N(0, \sigma^2)$ or $y_i \sim N(\beta_0 + \beta_1 x_1, \sigma^2)$

(i) $\hat{\beta}_1$ is $N\left[\beta_1, \sigma^2/S_{xx}\right]$

(ii) $\frac{(n-2)S^2}{\sigma^2}$ is $\chi^2(n-2)$

(iii) $\hat{\beta}_1$ and S^2 are independent.

Test statistics for $H_0: \beta_1 = 0$ Vs $\beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1}{S/\sqrt{S_{xx}}} \sim t(n-2) \quad \text{If } \beta = 0$$

Note

$$t = \frac{\text{Estimate}}{\text{Standard error of estimate}} = \frac{\hat{\beta}_1}{S_e(\hat{\beta}_1)}$$

$$\text{Reject } H_0 \text{ If } |t| \geq t_{\alpha/2}(n-2)$$

Detailed Hypothesis

Case 1: $H_0: \beta_1 \leq 0$ Vs $H_1: \beta_1 > 0$

Case 2: $H_0: \beta_1 \geq 0$ Vs $H_1: \beta_1 < 0$

Case 3: $H_0: \beta_1 = 0$ Vs $H_1: \beta_1 \neq 0$

Test statistic (T.S) $t = \frac{\hat{\beta}_1 - 0}{S_e/\sqrt{S_{xx}}}$

Rejection Region For d.f. $n-2$ and type 1 error α

Case 1. Reject H_0 If $t > t_\alpha$

2. Reject H_0 if $t < -t_\alpha$

3. Reject H_0 if $|t| > t_{\alpha/2}$

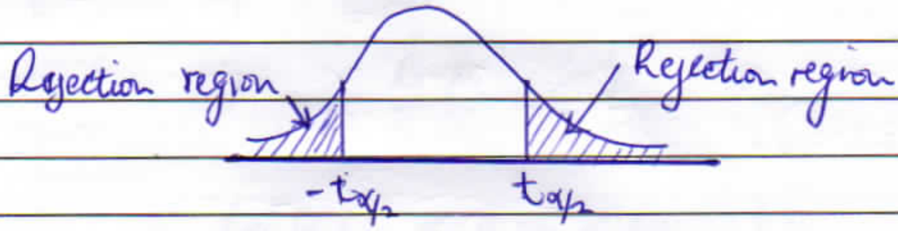
REG No.

Write on both sides of the paper

Question. ⑤

Do not write in either margin

For case 3: $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$



Example on Marks above: Test $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1}{S/\sqrt{S_{xx}}} = \frac{0.8726}{13.8547/139.753} = 8.8025$$

$t = 8.8025 > t_{0.025}(16) = 2.120$ Reject H_0.

Confidence interval for Slope $\hat{\beta}_1$

~~95%~~ A $100(1-\alpha)\%$ Confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{\alpha/2}(n-1)\frac{S}{\sqrt{S_{xx}}} \quad \text{Where } S = \frac{S_{YY} - \frac{S_{xy}^2}{S_{xx}}}{n-2}$$

$$\beta_1 \pm t_{\alpha/2}(n-1)\frac{S}{\sqrt{S_{xx}}}$$

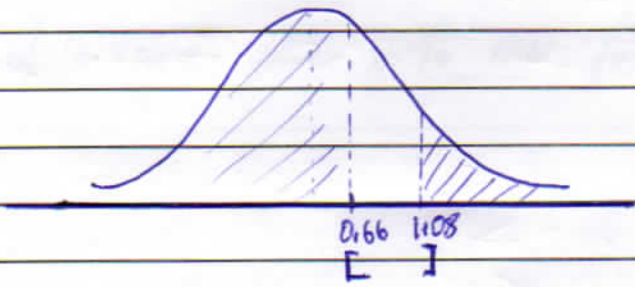
A 95% Confidence interval for β_1 for the Marks example

$$\hat{\beta}_1 \pm t_{0.025}^{(16)} \frac{S}{\sqrt{(x_i - \bar{x})^2}}$$

$0.8726 \pm 2.120(0.09914)$

0.8726 ± 0.2102

$(0.6624, 1.0828)$



Confidence interval for Slope β_1

$$\left| \frac{\hat{\beta}_1 - \beta_1}{S_e \sqrt{1/S_{xx}}} \right| < t_{\alpha/2} \quad \text{is a } (1-\alpha)100\% \text{ Confidence I.}$$

$$-t_{\alpha/2} < \frac{\hat{\beta}_1 - \beta_1}{S_e\sqrt{1/S_{xx}}} < t_{\alpha/2}$$

$$-t_{\alpha/2} S_e \sqrt{\frac{1}{S_{xx}}} < \hat{\beta}_1 - \beta_1 < t_{\alpha/2} S_e \sqrt{1/S_{xx}}$$

$$\hat{\beta}_1 - t_{\alpha/2} S_e \sqrt{\frac{1}{S_{xx}}} < \beta < \hat{\beta}_1 + t_{\alpha/2} S_e \sqrt{1/S_{xx}} \quad \text{is the}$$

$(1-\alpha)100\%$ Confidence interval for β_1 where t has $n-2$ df.

F test for H_0 $\beta_1 = 0$ (Recall ANOVA)

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

Test Statistics $F = \dfrac{SS(\text{Regression})/n-1}{SS\,\text{Error}/n-2} = \dfrac{MS\,Reg}{MSE}$

Rejection Region: With $df_1 = 1$ $df_2 = n-2$, reject H_0 if $F > F_\alpha$

$$SS\,Reg = \sum(\hat{y} - \bar{y})^2$$
$$SSE = \sum(y_i - \bar{y})^2$$

Confidence Interval for intercept β_0

$$\sigma_{\hat{\beta}_0} = S_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

The $(1-\alpha)100\%$ Confidence interval for β_0 is given by

$$\beta_0 \pm t_{\alpha/2} S_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \qquad \text{for } t \text{ with } n-2 \text{ df.}$$

Q Determine the 95% CI for $\hat{\beta}_1$ for example 1.1.

REG No.

Write on both sides of the paper Question 7

Do not write in either margin

Inference about the Correlation coefficient ρ_{yx}

The t statistic for β_1 can be expressed in terms of r as follows

$$t = \frac{\hat{\beta}_1}{\delta / \sqrt{\sum (x_i - \bar{x})^2}}$$

$$t = r_{yx} \frac{\sqrt{n-2}}{\sqrt{1-r_{yx}^2}}$$

The sample correlation r_{yx} is the basis for estimation and significance testing of the population correlation ρ_{yx}

Hypothesis

Case 1: $H_0\ \rho_{yx} \leq 0$ vs. $H_1\ \rho_{xy} > 0$

Case 2: $H_0\ \rho_{yx} \geq 0$ vs. $H_a\ \rho_{yx} < 0$

Case 3: $H_0\ \rho_{yx} = 0$ vs. $H_1\ \rho_{yx} \neq 0$

Time Series T.S. $t = r_{yx} \frac{\sqrt{n-2}}{\sqrt{1-r^2_{yx}}}$

R.R With n-2 df and type I error probability α,

1. $t > t_\alpha$
2. $t < -t_\alpha$
3. $|t| > t_{\alpha/2}$

Check assumptions and draw Conclusions

Exercise

Soil PH :	3.3	3.4	3.4	3.5	3.6	3.6	3.7	3.7	3.8	3.8
Growth Red :	17.78	21.59	23.84	15.13	23.45	20.87	17.78	20.09	17.78	12.46

QI Examine the scatter plot and decide whether a straight line is reasonable model. Is regression significant?

QII Identify the Least squares model estimates for $Y = \beta_0 + \beta_1 X + \epsilon$

QIII Predict the growth retardation for a soil PH of 4.0

REG No.

Write on both sides of the paper Question 8

Do not write in either margin

Simple Linear Regression.

The Model: The simple linear regression mode for n observation can be written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1,2,3,\ldots,n$$

Linear in β_0 and β_1

$y = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$ is Linear in β_0 & β_1

$y = \beta_0 + e^{\beta_1 x_i} + \varepsilon_i$ is not Linear ~~in~~

Model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1\,2\ldots n$$

Assumption

(i) $E(\varepsilon_i) = 0$ for all $i = 1\,2\ldots n$ or $E(y_i) = \beta_0 + \beta_1 x_i$

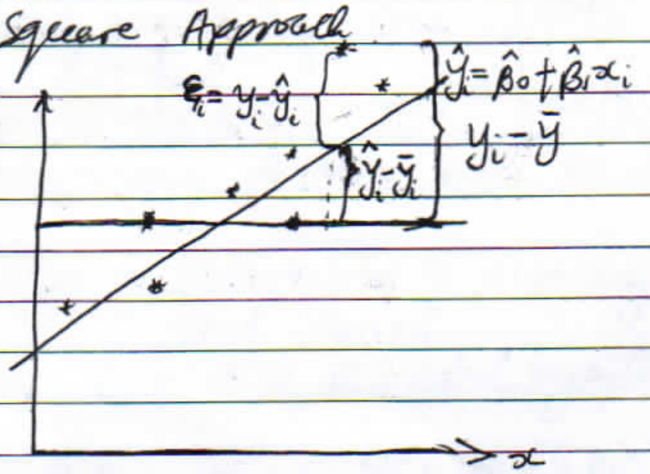
(ii) $Var(\varepsilon_i) = \sigma^2$ for all $i = 1\,2\ldots n$ or $Var(y_i) = \sigma^2$; Homoscedestic

(iii) $Cov(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$ or $Cov(x_i, y_i) = 0$

$$Var(y_i) = E[y_i - Ey_i]^2 = E(y_i - \beta_0 - \beta_1 x_i)^2 = E(\varepsilon_i^2) = \sigma^2$$

Estimation of β_0, β_1 and σ^2

Least square Approach



Total variation $= y_i - \bar{y}$

error (Residual) $= y_i - \hat{y}$

explained by regression $= \hat{y}_i - \bar{y}_i$

$y_i - \hat{y}_i = (y_i - \bar{y}) - (\hat{y}_i - \bar{y})$ from the sum

$$(y_i - \bar{y}) = (y_i - \hat{y}) + (\hat{y}_i - \bar{y}_i)$$

Sum of squares

$$\sum (y_i-\bar{y})^2 = \sum(\hat{y}_i-\bar{y})^2 + \sum(y_i-\hat{y})^2 + \underbrace{2\sum(\hat{y}_i-\bar{y})(y_i-\hat{y}_i)}_{0}$$

(Sum of squares about mean SST corrected) = (Sum of squares due to regression $SSR(b_1/b_0)$) + (Sum of square about regression Residual SS)

$$SST = SSR + SSE.$$

Therefore to test if regression is significant ie
$H_0: \beta_1 = 0$ Vs H_1 $\beta_1 \neq 0$
We have the anova.

ANOVA Table.

Source of variation	Degrees of freedom	Sum of Squares SS	Mean Square	F
Due to regression	1	$\sum(\hat{y}_i-\bar{y})^2$	MS_{reg}	$F_c = \frac{MS_{Reg}}{S^2}$
About regression Residual	$n-2$	$\sum(y_i-\hat{y}_i)^2$	$S^2 = \frac{SSE}{n-2}$	
Total corrected for $\bar{y}$	$n-1$	$\sum(y_i-\bar{y})^2$		

An accompanying p value will be available (given)

Pvalue = probability of observing a value Larger than that observed.

In this case it is a 2 sided hypothesis

$$p\text{-value} = 2 * P(F \geq F_c(1, n-2))$$

- If ~~#~~ F_c is Large reject H_0 $\beta_1 = 0$: $\Rightarrow$ implying that most variation is caused by (explained by) regression.
- Reject H_0 if pvalue is too small

REG No.

Write on both sides of the paper

Question (10)

Do not write in either margin

Coefficient of Determination

Coefficient of Determination is defined as

$$r^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

SSR - Regression sum of squares $= \sum(\hat{y}_i - \bar{y})^2$
SST - Total sum of squares $= \sum(y_i - \bar{y})^2$

$$SST = SSR + SSE$$

$$\sum(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\sum(\hat{y} - \bar{y}) = \sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2$$

$$SSR = SST - SSE$$

$$r^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}.$$

proportion of variation in y explained by the model or accounted by regression.

$$r = \frac{S_{xy}}{\sqrt{S_x^2 S_y^2}} = \frac{\sum(x_i - \bar{x})\sum(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

$$r^2 = \frac{SSR}{SST} = \frac{14873.0}{17944} = 0.8288$$

$$t = \frac{\hat{\beta}_1}{s/\sqrt{\sum_i (x_i - \bar{x})^2}} = \frac{\sqrt{n-2}\, r}{\sqrt{1-r^2}}$$

The fit of the regression is "good" if the sum $\sum \varepsilon_i^2$ ($= SSE$) is small ie the unexplained part of the variance of y is small. If SSE is small then R^2 is close to 1

REG No.

Write on both sides of the paper

Question F11

in either margin

Coefficient of Determination (Multiple Correlation Coefficient)

$$r^2 = \frac{SSR}{SST} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} \quad \text{Where } SST = SSR + SSE$$

$$SST \sum (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y})^2$$

r^2 gives the proportion of variation in y that is explained by the model or accounted for by regression on x.

Sample Correlation Coefficient r between x and y

$$r = \frac{S_{xy}}{\sqrt{S_{xx}^2 S_{yy}^2}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum (x_i - \bar{x})^2][\sum_{i=1}^{n} (y_i - \bar{y})^2]}}$$

For Example 1.1 $\quad r^2 = \frac{SSR}{SST} = \frac{45\,873.0}{17\,944.3} = 0.8288$

$$r = \sqrt{0.8288} = 0.91$$

Exercise

y	x	x^2	y^2	xy
25	10			
55	18			
50	25			
75	40			
110	50			
138	63			
90	42			
60	30			
10	5			
100	55			

(i) Determine the least square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$

(ii) Determine r^2 and explain how well your model fits the data

(iii) Use ANOVA to test whether regression is significant.