Unit 4

Linear Correlation and Regression Models

Tony CHAN, Ph.D.

4.1 Linear Correlation

- To find out if there exists a linear relationship between a pair of variables *x* and *y*.
- To determine the degree of association (i.e. strength of correlation) of the linear relationship.

4.1 Linear Correlation (Cont'd)

With bivariate data, each observation consists of data on 2 variables, (x, y). Examples of linear relationships are as follows:

x	*y*
Size of house	Value of house
Mark in mock exam	Mark in real exam
Amount of fertilizer	Amount of growth
People's age	Their liquid assets
Average interest rate	Number of new housing starts
Height of father	Height of son when 18

These data are paired. We call them ordered pairs.

3

4.1.1 Scatter Diagram

- Bivariate data can be represented graphically using a scatter diagram. In treating problems involving bivariate data, it is a good idea to plot the data as points in the *xy*-plane. Each observation (x, y) is represented by a point.
- When the points are plotted, a visual pattern may evolve suggesting a particular type of relationship, such as a linear or curvilinear (non-linear) one.

4

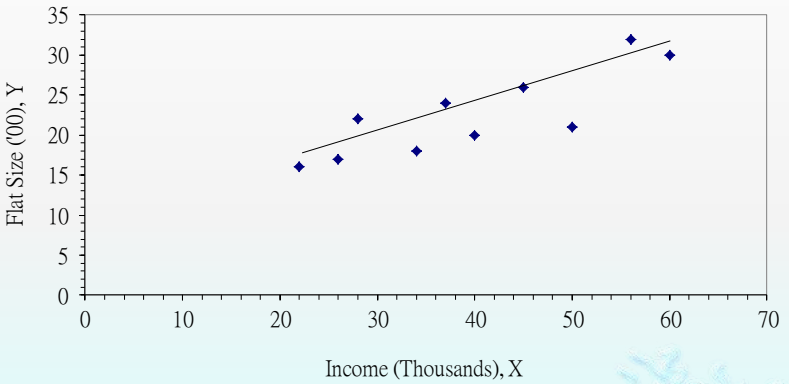
Example 4.1

A real estate developer is interested in determining the relationship between family income (X, \$000) of the local residents and their flat sizes of homes (Y, hundred meters).

Family Income, x	Flat Size, y
22	16
26	17
45	26
37	24
28	22
50	21
56	32
34	18
60	30
40	20

5

Scatter Diagram



6

4.1.1 Scatter Diagram

From the scatter diagram:

- residents with higher family incomes appear to live in flats of larger sizes.
- residents with lower family incomes tend to live in flats of smaller sizes.
- We see that the points are scattered in such a way that, more or less, a linear (straight-line) pattern is indicated. We expect a positive (direct) relationship between X and Y.

7

4.1.2 Pearson Correlation Coefficient, *r*

- It is often difficult to determine if a *significant* linear relationship exists between X and Y just *by inspecting* a scatter diagram of the data as *visual inspection may be misleading.*
- Correlation coefficient is a *quantitative measure* that characterizes the strength of linear relationship to a straight line.
- If (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) are n pairs of observations, r is defined by the following formula:

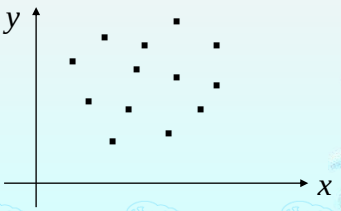
$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}, \text{ where } \begin{cases} S_{xy} = \sum xy - \frac{1}{n}\sum x \sum y; \\ S_{xx} = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2; \\ S_{yy} = \sum y^2 - \frac{1}{n}\left(\sum y\right)^2 \end{cases}$$

8

4.1.2 Pearson Correlation Coefficient (Cont'd)

- The range of r is $-1 \leq r \leq +1$.
- The larger $|r|$ (size of of r only) is, the stronger the linear relationship is.
- r close to zero indicates that there is no linear relationship between X and Y, and the scatter diagram typically appears to have a shotgun effect.

$r = 0$



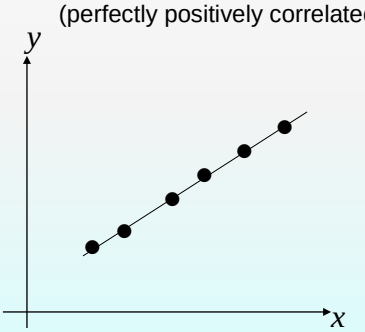
9

4.1.2 Pearson Correlation Coefficient (Cont'd)

- $r = 1$ or $r = -1$ implies that a perfect linear pattern exists between X and Y. That is, a single line will go through each point. We say that X and Y are perfectly correlated.

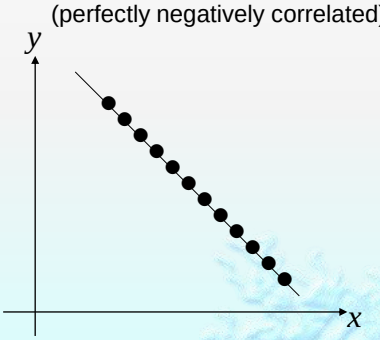
$r = +1$

(perfectly positively correlated)



$r = -1$

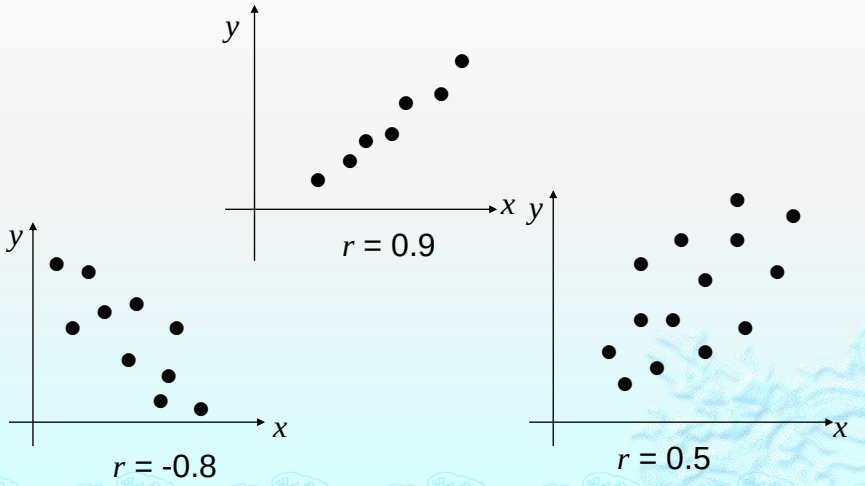
(perfectly negatively correlated)



10

4.1.2 Pearson Correlation Coefficient (Cont'd)

- Values of r = 0, 1 , or -1 are rare in practice. Several other values of the correlation coefficient are illustrated below:



11

4.1.2 Pearson Correlation Coefficient (Cont'd)

- Positive correlation ($r > 0$) indicates a linear relationship with positive slope
  - x and y are in the same direction.
  - High values of x tend to be associated with high values of y, or vice versa.

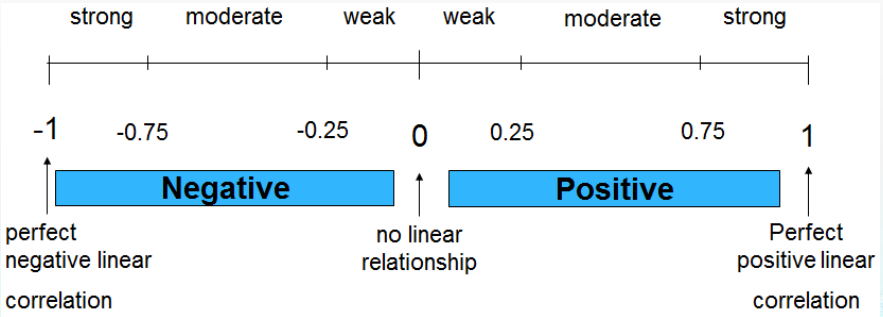
12

4.1.2 Pearson Correlation Coefficient (Cont'd)

- Negative correlation ($r < 0$) indicates a linear relationship with negative slope.
  - x and y are in the opposite direction.
  - High values of x are associated with low values of y, and vice versa. We say that a negative, or an inverse relationship exists.
- If $r = 0$, then there is no linear correlation between x and y but there can be non-linear relationship between them.

13

4.1.2 Pearson Correlation Coefficient (Cont'd)



14

Example 4.2 (Example 4.1 Revisited)

Compute the Pearson correlation coefficient, r of the following data:

Family Income, x	Flat Size, y
22	16
26	17
45	26
37	24
28	22
50	21
56	32
34	18
60	30
40	20

15

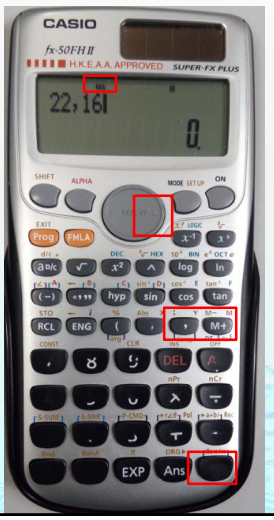
Example 4.2 (Cont'd)

Compute r using the built-in function of a CASIO calculator fx 50FH:

1. Enter regression mode "REG" by pressing "mode", "mode", "5", "1";

2. Data Input:
   "22" " , " "16" M+
   "26" " , " "17" M+
   .
   .
   .
   "40" " , " "20" M+

3. For r, press "shift" "2" "1" "→" "→" "3" "exe"



16

Example 4.2 (Cont'd)

$r = + 0.84$

- Conclusions
  - $r > 0 \Rightarrow$ a positive correlation
  - $r = 0.84 > 0.75 \Rightarrow$ very high correlation between income (x) and flat size (y).
  - x and y can be described by a linear relationship.

17

4.1.3 Spearman's Rank Correlation Coefficient, r_s

Instead of employing the precise values of the variables x and y or when such information is not available, we rank the data* in ascending order of magnitude using the numbers $1, 2, \cdots, n$.

*Examples of data with order but without numerical values (ordinal data):

- Ratings of singing contest to singers on a 10-point Likert scale (say)
- 5-point Likert scale scores of paintings given by judges
- Subject grades (A+, A, A-, B+,…)

18

4.1.3 Spearman's Rank Correlation Coefficient (Cont'd)

r_s could be computed in the following data types:

- Both variables are quantitative.
- Both variables are qualitative (ordinal: variable w/ order but w/o numerical value).
- One variable is quantitative, whereas the other is ordinal.

Remarks

- We can adopt the Spearman's method to quantitative data (data with both order and numerical values) to calculate the rank correlation coefficient. However, the result thus obtained is only approximate.
- Unlike the Pearson correlation coefficient (r), high value of r_s does not indicate linear relationship between x and y.

19

4.1.3 Spearman's Rank Correlation Coefficient (Cont'd)

- Rank of data

Consider the following datasets:

Dataset 1

Data value:	1,	2,	4,	5,	7,	10
Rank:	1,	2,	3,	4,	5,	6

Dataset 2

Data value:	1,	3,	3,	8,	9,	12
Rank:	1,	(2+3)/2=2.5,	2.5,	4,	5,	6

3 and 3 are equal and they are called tied observations or ties.

20

4.1.3 Spearman's Rank Correlation Coefficient (Cont'd)

Procedures:

1. Rank the values of x from 1 to n, where n is the numbers of pairs of values of x and y in the sample.
2. Rank the values of y from 1 to n.
3. Compute the value of d_i for each pair of obs, (x_i, y_i), where $d_i = rank(y_i) - rank(x_i), i = 1,2,\cdots,n.$
4. Calculate the sum of all d_i, i. e. $\sum d_i^2$.
5. Apply the formula: $r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$

21

Example 4.3

Two competitors rank the 8 photographs in a competition as follows:

Photograph	A	B	C	D	E	F	G	H
1st Competitor	2	5	3	6	1	4	7	8
2nd Competitor	4	3	2	6	1	8	5	7

Calculate Spearman's coefficient of rank correlation for the data.

22

Example 4.3 (Cont'd)

Photograph	Rank(x)	Rank(y)	|d| = |x-y|	d^2
A	2	4	2	4
B	5	3	2	4
C	3	2	1	1
D	6	6	0	0
E	1	1	0	0
F	4	8	4	16
G	7	5	2	4
H	8	7	1	1
			$\sum d^2$	30

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6(30)}{8(64-1)} = +0.64$$

It shows that a positive and moderate correlation exists between x and y.

23

4.2 Simple Linear Regression

From the magnitude of Pearson correlation coefficient, *r*, we can decide:

- if any 2 variables are linearly related;
- the closeness to a straight line of this relationship.
- Even for a strong linear correlation, *r* cannot tell us the way the 2 variables are linearly related as *r* by itself gives us neither the slope nor the *y*-intercept of a straight line.
- We must use regression theory to determine the linear relationship between x and y.

24

4.2 Simple Linear Regression (Cont'd)

- Once linear correlation confirms that there is a linear relationship between *x* and *y*, we can use linear regression theory to build up a linear regression model to describe the specific linear relationship.

- Regression can be used to predict the future values of *y* based on given values of *x*.

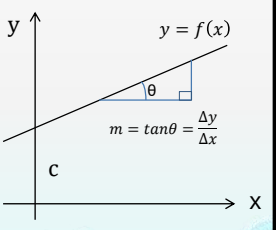
25

4.2.1 Types of Relationships

There are, in general, two relationships:

- Functional/mathematical relationship
  It describes the deterministic relationship between variables. That is, the relationship can be expressed as

$$y = f(x) = mx + c$$
or
$$y = f(x_1, x_2, \cdots, x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$



E.g. Suppose that your mobile phone service charge is calculated by the following functional relationship:

Charge, $y = 200 + 0.5x$,

where x = no. of calls beyond 3,000 minutes ($0.5/min), $200 is the basic monthly charge.

We can compute the charge for any value of x WITHOUT error

26

4.2.1 Types of Relationships (Cont'd)

- Statistical relationship

  Referring to Example 4.2, we can see that flat size (y) depends on household income (x_1). However, household income (x_1) is not the only factor affecting flat size. Suppose that the following factors may also affect the flat size:

  - family asset (x_2)
  - age of building (x_3)
  - direction (sea, hill, swimming pool, etc.) (x_4)
  - location (x_5)

27

4.2.1 Types of Relationships (Cont'd)

The appropriate model is given by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

where

y = the dependent/explained variable/response variable;
x = the independent/explanatory variable/ predictor/regressor;
β_i = the regression parameters of the model, $i = 1,2,\cdots,5$;
ε = the random error containing other factors not included into the model;

28

4.2.1 Types of Relationships (Cont'd)

If we study the relationship between household income (y) and flat size (x_1) only, then the other factors are intentionally or unintentionally ignored. The linear relationship between x_1 and y can be modeled by

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

where β_0 = y-intercept, β_1= slope, ε = random error.

All the ignored factors are contained in the random error, ε.

29

4.2.1 Types of Relationships (Cont'd)

Assume that y and x is statistically linearly related.

- If we only consider the effect of a single factor, x on y using regression technique, then such a regression model is called the simple linear regression model.
- If we use several x's to explain the variation of y, then such a regression model is called the multiple/general linear regression model.

30

4.2.3 Parameter Estimation

- Plausible Approach

Simple linear regression model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1,2,\cdots,n.$

Assumption: The errors ε_i's are very small and can be neglected.

The above regression model can then be approximated by

$$y_i = \beta_0 + \beta_1 x_i,\ i = 1, 2, \ldots, n. \qquad (*)$$

In order to estimate the values of β_0 and β_1, we can set up 2 equations so that they can be solved.

Suppose that we substitute (x_1, y_1) and (x_2, y_2) into (*). Then, we have

31

4.2.3 Parameter Estimation (Cont'd)

$$\begin{cases} y_1 \approx \beta_0 + \beta_1 x_1 \\ y_2 \approx \beta_0 + \beta_1 x_2 \end{cases}$$

From these 2 equations, we get

$$\begin{cases} \beta_0 = \dfrac{x_1 y_2 - x_2 y_1}{x_1 - x_2} \\ \beta_1 = \dfrac{y_1 - y_2}{x_1 - x_2} \end{cases}$$

32

4.2.3 Parameter Estimation (Cont'd)

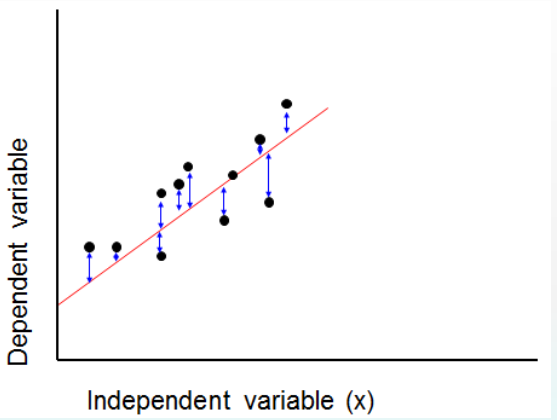
The plausible approach stated above is, in fact, problematic:

- The solutions depend upon the selection of (x_1, y_1) and (x_2, y_2). This means that different pairs of observations will yield different estimates. In other words, unique solutions cannot be obtained.
- The assumption of negligible and very small ε_i's is not justifiable.

We will resort to the estimation method devised by the great mathematician, Gauss.

33

4.2.3 Parameter Estimation (Cont'd)



- The least squares linear regression model selects the straight line such that the sum of squared errors (SSE) attains its minimum.

34

4.2.3 Parameter Estimation (Cont'd)

The principle of the method of least squares for the estimation of β_0 and β_1 is shown below:

Rewrite $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ as

$$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i), i = 1,2,\cdots,n.$$

SSE is given by

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

If we can find the pair of $\widehat{\beta_0}$ and $\widehat{\beta_1}$ such that *SSE* is a minimum, then $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are the best estimates, where $\widehat{\beta_0}$ and $\widehat{\beta_1}$ denote the estimates of β_0 and β_1, respectively.

35

4.2.3 Parameter Estimation (Cont'd)

For convenience, we will use the following symbols:

$$\widehat{\beta_0} = b_0, \quad \widehat{\beta_1} = b_1.$$

By calculus, we can find the least squares estimates as follows:

$$\begin{cases} b_0 = \bar{y} - b_1 \bar{x} \\ b_1 = \dfrac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} \end{cases}$$

where $\bar{x}$ = sample mean of x's;
$\bar{y}$ = sample mean of y's.
n = number of pairs of observations in the data

36

4.2.3 Parameter Estimation (Cont'd)

- The y-intercept b_0 and slope b_1 estimated by the method of least squares are referred to as the sample regression coefficients.

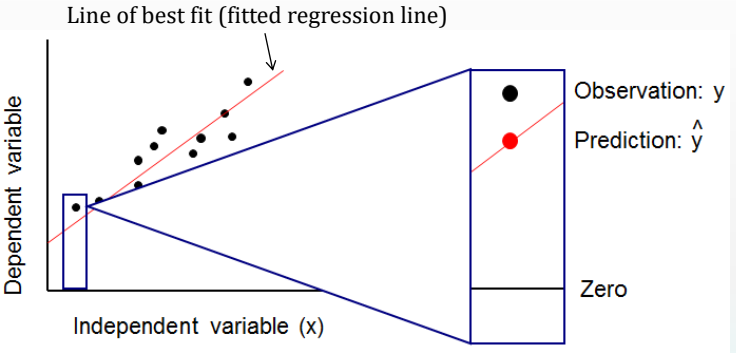
- The equation

$$\hat{y} = b_0 + b_1 x$$

obtained in this way is called the fitted regression equation of y on x.

37

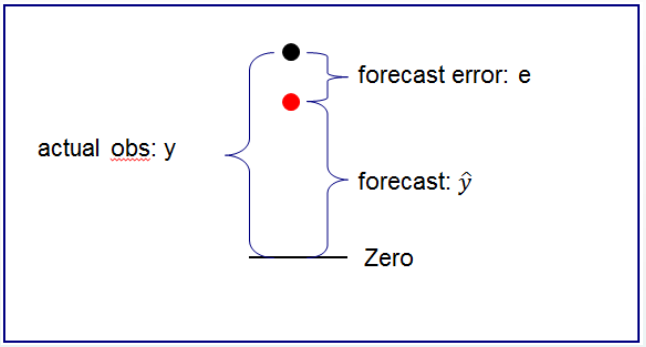
4.2.3 Parameter Estimation (Cont'd)



- The fitted regression equation will make a prediction for each observed data point.
- The actual observation is denoted by y, whereas the fitted value/forecast is denoted by $\hat{y}$.

38

4.2.3 Parameter Estimation (Cont'd)



$$\hat{y} = b_0 + b_1 x$$

For each observation, the variation can be described as:

y	=	$\hat{y}$	+	e
actual obs	=	fitted value/forecast	+	forecast error
		(explained part)		(unexplained part)

39

Example 4.3

Determine the fitted regression equation for the following data:

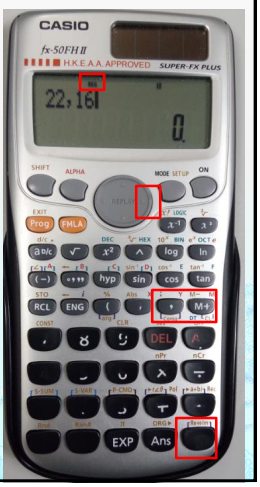
Family Income, x	Flat Size, y
22	16
26	17
45	26
37	24
28	22
50	21
56	32
34	18
60	30
40	20

40

Example 4.3 (Cont'd)

CASIO fx 50FH:

1. Enter regression mode, "REG", by pressing "mode" "mode" "5" "1";
2. Data Input:
   "22" "," "16" M+
   "26" "," "17" M+
   .
   .
   .
   "40" "," "20" M+



3. (i) For b_0, press "shift" "2" "1" "→" "→" "1" "exe".

   (ii) For b_1, press "shift" "2" "1" "→" "→" "2" "exe".

Example 4.3 (Cont'd)

$$b_0 = 8.51$$
$$b_1 = 0.35$$

The fitted regression equation is

$$\hat{y} = 8.51 + 0.35x$$

Example 4.3 (Cont'd)

Output Summary (Excel)

Regression Statistics	
Correlation coefficient	0.84325447
Coefficient of Determination	0.71107811
Adjusted R2	0.67496287
Standard Error	3.07841486
No. of Observation	10

ANOVA

	DF	*SS*	*MS*	*F*	*Significance*
Regression	1	186.5868958	186.5869	19.68914	0.00217591
Residual	8	75.81310419	9.476638		
Total	9	262.4			

	Coefficient	*Standard Error*	*t Statistic*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	8.51396348	3.320409102	2.5641309	0.033431	0.85708636	16.1708406	0.85708636	16.1708406
X1	0.35392052	0.079761317	4.4372452	0.002176	0.16999059	0.53785044	0.16999059	0.537850442

43

4.2.4 Fitting & Forecasting

- Fitting means the estimation of y values by giving known values of *x* lying within the data range of the observations.
- For *x* values lie outside the data range, the estimation is called forecasting, prediction or projection.
- Making point estimate of y based on the fitted regression equation is simply a matter of substituting a known or assumed value of *x* into the fitted regression equation, then calculating the estimated value of *y.*

44

Example 4.4 (Fitting)

From Example 4.3, we have obtained the following fitted regression equation:

$$\hat{y} = 8.51 + 0.35x$$

Find the fitted value of y when $x = 30$. (Note that $x = 30$ lies within the data range of x: [22, 60].)

Solution

When $x = 30$,

$$\widehat{y_{30}} = 8.51 + 0.35(30) = 19.01 hundred\ m^2 = 1901 m^2$$

OR

$$\widehat{y_{30}} = 19.13 hundred\ m^2 = 1913 m^2$$

(press "30" "shift" "2" "1" "→" "→" "→" "2")

45

Example 4.5 (Forecasting)

From Example 4.3, we have got the following fitted regression equation:

$$\hat{y} = 8.51 + 0.35x$$

Find the forecast of y when $x = 70$. (Note that $x = 70$ lies outside the data range [22, 60].)

46

Example 4.5 (Forecasting)

When $x = 70$,

$$\widehat{y_{70}} = 8.51 + 0.35(70) = 33.01 hundred\ m^2 = 3301 m^2$$

OR

$$\widehat{y_{70}} = 33.29 hundred\ m^2 = 3329 m^2$$

(press "70" "shift" "2" "1" "→" "→" "→" "2")

47

Remarks

- We can certainly build up a regression model relating x to y. However, it is untrue to relate any 2 dependent and independent variables that, actually, no relationship exists between them.

Example
There is no relationship between number of people drown last year in a city and the amount of crop yields in the same city. In fact, we can build a regression model for it, but this model does not make any sense.

- We CANNOT infer a causation between y and a set of x's from a statistical relationship obtained by regression or any other statistical techniques.

Example
There exists a statistical relationship between amount of crop yields and the amount of rainfall. However, based on common sense, we know that yield is the effect while rainfall is the cause. We cannot use the amount of yield to control the amount of rainfall.

48

4.2.5 Coefficient of Determination, R^2

- The proportion of total variation in y that can be explained by the fitted regression equation is known as the coefficient of determination and is denoted R^2.
- It is also a measure of the goodness of fit of a regression model.

Coefficient of Determination = (Pearson correlation coefficient)^2

i.e. $R^2 = (r)^2$

Explanation

R^2 is a measure of the total variation in y that can be explained by the fitted regression equation.

49

Example 4.6

From Example 4.3, the value of correlation coefficient, r is 0.84.

$R^2 = (0.84)^2 = 0.7056 = 70.56\%$

Conclusion

70.56% of the total variation in y can be explained by the fitted regression equation.

50

4.2.6 Partition of Total Sum of Squares

$Y_i - \bar{Y} \quad = \quad (\hat{Y}_i - \bar{Y}) \quad + \quad (Y_i - \hat{Y}_i), \quad i = 1, 2, \cdots, n.$

Total variation	explained variation	unexplained variation

Squaring the above equality on both sides and summing up from $i = 1$ to n, we get

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \quad = \quad \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \quad + \quad \sum(Y_i - \hat{Y}_i)^2$$

Sum of Squared Total (*SST*)	Sum of Squared Regression (*SSR*)	Sum of Squared Error (*SSE*)

The above equation can be rewritten as

$$SST = SSR + SSE$$

51

4.2.7 Formulae for *SST*, *SSR* and *SSE*

$$SST = S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$

$$SSR = \widehat{\beta_1^2} S_{xx} = \frac{S_{xy}^2}{S_{xx}}, \qquad \text{where} \begin{cases} S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2; \\ S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}). \end{cases}$$

$$SSE = SST - SSR$$

52

4.2.8 Analysis of Variance Table

Source of Variation	Degree of Freedom	Sum of Squares	Mean Squares	F-ratio
Regression	1	SSR	$MSR = \frac{SSR}{1}$	$F_0 = \frac{MSR}{MSE}$
Error	n-2	SSE	$MSE = \frac{SSE}{n-2} = \widehat{\sigma^2}$	-
Total	n-1	SST	-	-

53

4.2.9 Significance Test of Linearity

In order to test the significance of linearity of regression line, we may perform the usual F-test. The hypotheses are

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

The test statistic is

$$F = \frac{MSR}{MSE} \sim F(1, n-2)$$

The decision rule is as follows:

$$\begin{cases} \text{If } F \leq F_\alpha(1, n-2), \text{ then accept } H_0; \\ \text{If } F > F_\alpha(1, n-2), \text{ then reject } H_0. \end{cases}$$

54

Example 4.7 (Example 4.3 Revisited)

Family Income, x	Flat Size, y
22	16
26	17
45	26
37	24
28	22
50	21
56	32
34	18
60	30
40	20

55

Numerical Calculation of SST

CASIO fx-50FH/CASIO fx-50FHII

1) Enter regression mode (REG mode):
   mode mode **5** **1** → REG mode

2) Data input:
   22 , 16 M+
   26 , 17 M+
   ⋮
   60 , 30 M+
   40 , 20 M+

SST: shift 2 1 → 2 = x^2 × 10 =
SST = 262.4

CASIO fx-2650p/CASIO fx-3950p

1) Enter regression mode (REG mode):
   mode mode **2** **1** → REG mode

2) Data input:
   22 , 16 M+
   26 , 17 M+
   ⋮
   60 , 30 M+
   40 , 20 M+

SST: shift 2 → 2 = x^2 × 10 =
SST = 262.4

56

Numerical Calculation of SSR

$SSR = \widehat{\beta_1^2} S_{xx}$

CASIO fx-50FH/CASIO fx-50FHII

Find $\widehat{\beta_1}$: shift 2 1 → → 2 =
Result: $\widehat{\beta_1} = 0.353920515 \approx 0.35$

Find S_{xx}: shift 2 1 2 = x^2 × 10 =
Result: $S_{xx} = 1489.6$

$SSR = \widehat{\beta_1^2} S_{xx}$
$= (0.353920515)^2 \times 1489.6$
$= 186.5869$

$SSR = \widehat{\beta_1^2} S_{xx}$

CASIO fx-2650p/CASIO fx-3950p

Find $\widehat{\beta_1}$: shift 2 → → 2 =
Result: $\widehat{\beta_1} = 0.353920515 \approx 0.35$

Find S_{xx}: shift 2 2 = x^2 × 10 =
Result: $S_{xx} = 1489.6$

$SSR = \widehat{\beta_1^2} S_{xx}$
$= (0.353920515)^2 \times 1489.6$
$= 186.5869$

57

Numerical Calculation of SSR

SSE = SST – SSR
= 262.4 – 186.5869
= 75.8131

ANOVA Table

SV	DF	SS	MS = SS/DF	F-ratio =MSR/MSE
Regression	1	186.5869	186.5869/1= 186.5869	$F_0 = \frac{186.5869}{9.4766} = 19.69$
Error	8=n-2	75.8131	75.8131/8 = 9.4766	-
Total	9=n-1	262.4	-	-

58

Test for Linearity of Regression Line

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

Take α = 5%. $F_{5\%}(1, 8) = 5.32$.

$\because F_0 = 19.69 > 5.32,$
$\therefore$ reject H_0 at the 5% level.

We conclude that there is a linear relationship between family income and flat size.

59

4.2.10 Alternative Form of Coefficient of Determination

- When the regression model passes ***F test***, we only know that there is significant ***linear relationship*** between Y and X. But we do ***not*** know ***how strong is the linear relationship***.
- The coefficient of determination is a measure of the ***strength of linearity*** (i.e. ***goodness of fit***) a regression model.
- The value of R^2 shows the ***amount of variation*** that can be explained by ***y or the regression model*** under consideration.

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

60

4.2.10 Alternative Form of Coefficient of Determination (Cont'd)

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$ (Slide 8)

$$\begin{cases} SSR = \frac{S_{xy}^2}{S_{xx}} \\ SST = S_{yy} \end{cases}$$ (Slide 52)

$$R^2 = \frac{SSR}{SST} = SSR \times \frac{1}{SST} = \frac{S_{xy}^2}{S_{xx}} \times \frac{1}{S_{yy}} = \left(\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\right)^2 = (r)^2$$ (Slide 49)

(The above proof will not be examined.)

61

4.2.11 Distinctions between Regression & Correlation

Both regression and correlation analyses are used to investigate the statistical relationship between variables (dependent and independent variables). Nevertheless, they are not the same in many aspects as shown below:

Regression Analysis	Correlational Analysis
◆ It deals with the degree of influence on *Y* from a variable *X* or a set of variable *X*'s (independent variables).	◆ It deals with the degree of correlation between a variable (or a set of variables) and another variable (or another set of variables).
◆ The number of dependent variable is one while the number of independent variables is one or more than one.	

62

4.2.11 Distinctions between Regression & Correlation (Cont'd)

Regression Analysis

- The dependent variable Y is assumed to be a random variable having a probability distribution .
- The variable X's are nonrandom.
- The main use is to perform fitting or forecasting given known values of X.

Correlational Analysis

- Variables are not classified as dependent or independent variables.
- They are all random variables.
- It is used to measure the strength of association between Y's and X's.

63

Exercise 4.1

1. The yields of toy cars in 10 factories and their corresponding costs are listed in the following table:

Factory Number	Yield (x)	Cost (y)
1	39	144
2	47	220
3	45	138
4	47	145
5	65	162
6	46	142
7	67	170
8	42	124
9	67	158
10	56	154

(a) Construct the ANOVA Table.

(b) Test, at the 5% level of significance, if there is a linear relationship between yield and cost.

64

Solution

(a)

ANOVA Table

SV	DF	SS	MS	F-ratio
Reg	1	514.7749	514.7749/1= 514.7749	$F_0 = \frac{514.7749}{703.6656} = 0.73$
Error	8	5629.3251	5629.3251/8 = 703.6656	-
Total	9	6144.1	-	-

(b) H_0: $\beta_1 = 0 \leftrightarrow H_1$: $\beta_1 \neq 0$
Take α = 5%. $F_{5\%}(1, 8) = 5.32$.
$\because F_0 = 0.73 < 5.32$,
$\therefore$ do not reject H_0 at the 5% level.

We conclude that there is no linear relationship between yield and cost.

65

2. An instructor at OUHK asked a random sample of 8 students to record their study times in a beginning calculus course. He then made a table for total hours studied (x) over 2 weeks and test score (y) at the end of the 2 weeks. The results are shown below:

x	10	15	12	20	8	16	14	22
y	92	81	84	74	85	80	84	80

(a) Compute the sample Pearson correlation coefficient. Interpret your answer briefly.

(b) Construct the ANOVA Table.

(c) Find the estimate of σ^2, the population variance of error.

(d) At 5% level of significance, test if there is a linear relationship between x and y.

(e) How many percent of the total variation in y can be explained by x?

66

Solution

(a) r = -0.7750

The sign and magnitude of r show a strong (> 0.75) and negative linear correlation. A linear relationship exists between x and y.

(b) ANOVA Table

SV	DF	SS	MS	F-ratio
Regression	1	112.8884	112.8884	9.02
Error	6	75.1116	12.5186	-
Total	7	188.0000	-	-

67

Solution (Cont'd)

(c) $\widehat{\sigma^2} = MSE = 12.5186$

(d) At 5% level, $F_{5\%}(1, 6) = 5.99$.

$\because$ $F_0 = 9.02 > 5.99$,

$\therefore$ reject H_0 at the 5% level.

We conclude that there exists a linear relationship between x and y.

(e) $R^2 = \frac{SSR}{SST} = \frac{112.8884}{188} = 0.6005$ OR $R^2 = (r)^2 = (-0.7750)^2 = 0.6006$

Around 60% of the total variation in y can be explained by x.

68