

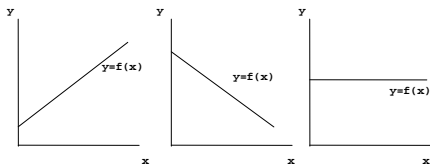
X. Regression Analysis

A. Definitions

1. **Dependent Variable** - also called the response variable, this is the phenomena or characteristic we wish to estimate or predict. The dependent variable is usually denoted y .
2. **Independent Variable** - also called a predictor variable, this is the phenomena used to estimate or predict the value of the dependent variable. Independent variables are usually denoted x .

3. **Functional Relationship** - The unique value of the dependent variable (usually denoted y) can be precisely determined given the value of the independent variable (usually denoted x) - this is often referred to as a *deterministic* relationship between x and y .

Classes of Functional Relationships Between x & y



Positive/Direct Relationship

Negative/Inverse Relationship

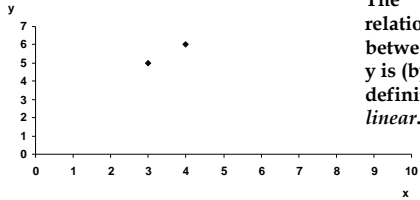
No Relationship

4. **Slope** - Change in the value of the dependent variable that corresponds to a one-unit change in the independent variable. Usually denoted B_1 .
5. **Y-Intercept** - Value of the dependent variable when the independent variable is zero. Denoted B_0 .
6. **Linear Relationship** - The slope is constant at all values of the independent variable .

Suppose you had the following two observations on variables x and y:

x	y
3	5
4	6

The resulting graphical display would look like this:

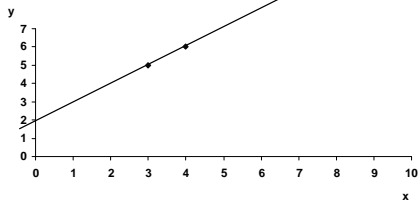


The relationship between x and y is (by definition) *linear*.

We could find the line on which both points lie by solving the equation $y = B_0 + B_1x$ simultaneously for our two observations:

$$\begin{aligned}
 5 &= B_0 + 3B_1 \\
 -[6 &= B_0 + 4B_1] \\
 \hline
 -1 &= -B_1 \rightarrow B_1 = \textcircled{1} \\
 \text{so } 5 &= B_0 + 3B_1 = B_0 + 3(\textcircled{1}) \rightarrow B_0 = 5 - 3 = \textcircled{2}
 \end{aligned}$$

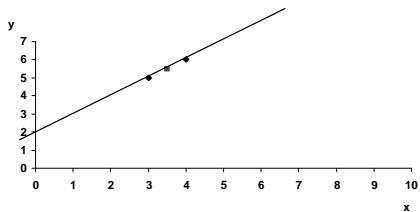
The resulting line is $y = 2 + x$



Such a line based on two points will *always* be linear and functional (why?). What happens if we introduce a third point (3.5, 5.5) into our data set?

Since the new observation (3.5, 5.5) satisfies our original formula (falls on the line through points (3, 5) and (4, 6))
 $y = 2 + x = 2 + 3.5 = 5.5$

we still have a functional linear relationship.

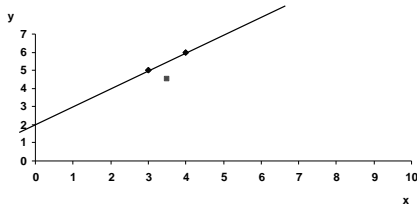


On the other hand, what happens if we introduce a third point (3.5, 4.5) into our data set?

Since the new observation (3.5, 5.5) does not satisfy our original formula (does not fall on the line through points (3, 5) and (4, 6))

$$y = 2 + x = 2 + 3.5 = 5.5 \neq 4.5$$

we no longer have a *functional linear* relationship.

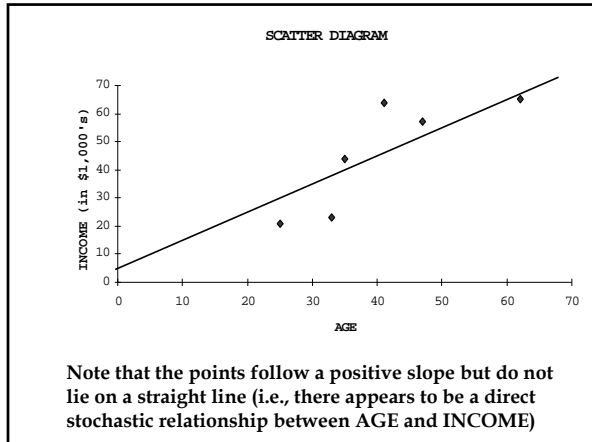


7. **Statistical Relationship** - The relationship(s) between values of the dependent variable and corresponding values of the independent variable(s) is (are) not deterministic. Thus the value of y is *estimated* given the value of x . The estimated value of the dependent variable is denoted \hat{y} (y-hat), and the population slope and y-intercept are usually denoted β_1 and β_0 .

8. **Scatter Diagram** - Scatter Diagram - graphical simultaneous presentation of the values of two variables on a Cartesian coordinate system

- Example: suppose we have collected the following sample of 6 observations on age and income:

	INCOME
AGE	(in \$1,000's)
25	21
47	57
35	44
62	65
41	64
33	23



9. Regression Analysis - Statistical methods for *estimating* the relationship between the dependent variable and the independent variable. The estimates of β_1 and β_0 are usually denoted b_1 and b_0 .

10. Linear Regression - Indicates that the relationship(s) between the dependent variable and the independent variable(s).

11. Simple Regression - Indicates that the relationship is between the dependent variable and a single independent variable. Thus we would have

$$\hat{y} = b_0 + b_1x_i$$

as our *estimate* of the value of the dependent variable y when $x = x_i$.

12. Regression Error - also called the *residual*, this is the difference between our estimate of the value of the dependent variable y when $x = x_i$ (i.e., \hat{y}_i) and the actual value of the dependent variable y when $x = x_i$ (i.e., y_i). The residual for the i^{th} observation is usually denoted e_i , so we have that

$$e_i = y_i - \hat{y}_i$$

which by substitution is

$$e_i = y_i - (b_0 + b_1x_i)$$

Now consider 1st quarter 2007 sales data for a small sample (four units) from a restaurant chain:

Unit	1 st Quarter 2007 Sales in 1,000's (Y)
1	\$123
2	\$225
3	\$193
4	\$111

How would we ordinarily summarize these data?

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{143 + 225 + 173 + 111}{4} = 163.0$$

What error do we make when we use the sample mean (163.0) to estimate the sales of each unit?

Unit	1 st Quarter 2007 Sales in \$1,000's	Error (y _i - \bar{y})
1	\$123	123-163=-40
2	\$225	225-163= 62
3	\$193	193-163= 30
4	\$111	111-163=-52

How could we do better (i.e., make smaller errors)?
What if we knew the associated values of some variable that is directly related to the restaurants' quarterly sales?

What if we also knew how much each unit spent on advertising and promotion during the 1st quarter of 2007? Could that help us?

1 st Quarter 2007 (\$1,000's)		
Unit	Sales (Y)	Ad Expenses
1	\$123	\$22
2	\$225	\$35
3	\$193	\$28
4	\$111	\$15

How would we ordinarily summarize the relationship expressed by data? What would we estimate 1st quarter 2007 sales per unit to be based on these sample data?

The sample covariance:

$$s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$= \frac{(22-25)(123-163) + (35-25)(225-163) + (28-25)(193-163) + (15-25)(111-163)}{4-1}$$

$$= \frac{1350}{3} = 450.000$$

and the sample correlation coefficient (for which we need the sample standard deviations s_x and s_y):

$$s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(22-25)^2 + (35-25)^2 + (28-25)^2 + (15-25)^2}{4-1}} = \sqrt{\frac{218}{3}} = 8.525$$

$$s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{(123-163)^2 + (225-163)^2 + (193-163)^2 + (111-163)^2}{4-1}} = \sqrt{\frac{9048}{3}} = 54.918$$

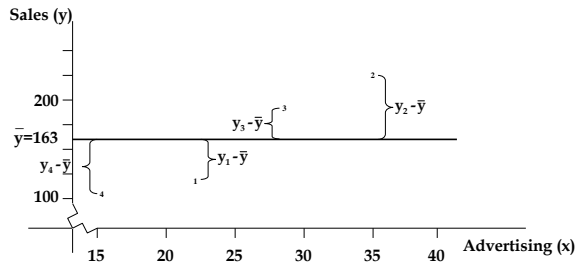
The sample Pearson's correlation coefficient is:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{450.000}{(8.525)(54.918)} = 0.96123$$

which suggests a strong positive relationship between advertising expenses and sales (as expected):

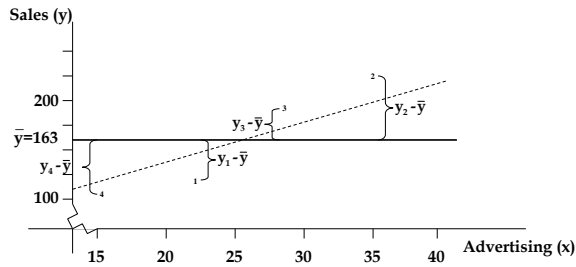
Could we learn more by looking at a graph?

If we use $\bar{y} = 163$ as our estimate of sales for each of the four units (and any other units), our estimate would look like this



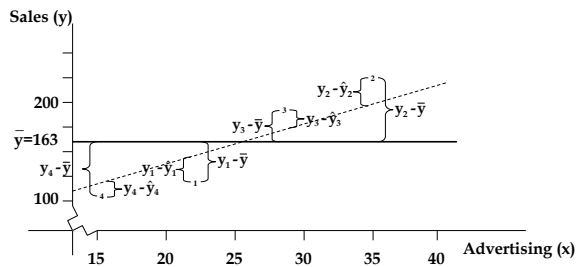
The errors we make if we use \bar{y} to estimate y (1st Quarter Sales) for each unit are:

However, the graph suggests a positive relationship between Advertising (x) and Sales (y) – can we use this information to improve our ability to estimate Sales?



If we estimate the relationship between Advertising and sales to look like the green line, what errors will we make?

Are the estimates of Sales at the Advertising levels of the four units in our sample green more accurate than the sample mean \bar{y} ?



Knowing the relationship between Advertising (x) and Sales (y) enables us to much more accurately estimate Sales at different levels of Advertising!

13. Ordinary Least Squares (OLS) - Criteria for fitting an estimated regression line to sample data in which the sum of the squared differences between actual and estimated values of the dependent variable are minimized, i.e.,

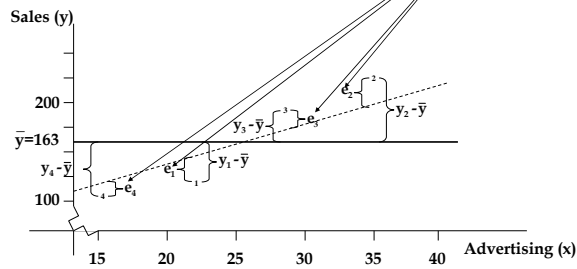
$$\min \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

which is equivalent to

$$\min \sum_{i=1}^n e_i^2$$

that is, we wish to find the regression line that minimizes the squared regression errors we would commit using the line to estimate the values of y_i for the values of x_i in our sample data.

i.e., for our example we minimize the sum of



Knowing the relationship between *Advertising* (x) and *Sales* (y) enables us to much more accurately estimate *Sales* at different levels of *Advertising*!

B. Ordinary Least Squares Estimate of the Regression Line

- Recall that our objective is to

$$\min \sum_{i=1}^n e_i^2, \text{ i.e., } \min \sum_{i=1}^n (y_i - \hat{y})^2$$

which by another substitution can be rewritten as

$$\min \sum_{i=1}^n e_i^2, \text{ i.e., } \min \sum_{i=1}^n [y_i - (b_0 + b_1 x_i)]^2$$

A little differential calculus can be applied to this expression to derive the *normal equations*:

$$\sum_{i=1}^n y_i = n b_0 + b_1 \sum_{i=1}^n x_i$$

and

$$\sum_{i=1}^n x_i y_i = b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2$$

We can solve these equations simultaneously to obtain the equations necessary to produce the estimates b_1 and b_0 .

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}}$$

and $b_0 = \bar{y} - b_1 \bar{x}$

Example: Find the estimated regression line for the sample of six observations we have collected on age and income:

	INCOME	
AGE	(in \$1,000's)	
25	21	
47	57	
35	44	
62	65	
41	64	
33	23	

Which is the independent variable and which is the dependent variable for this problem?

our summary calculations are:

i	Age (x_i)	Income (y_i)	$x_i y_i$	x_i^2
1	25	21	525	625
2	47	57	2679	2209
3	35	44	1540	1225
4	62	65	4030	3844
5	41	64	2624	1681
6	33	23	759	1089
Σ	243	274	12157	10673

so:

$$\sum_{i=1}^n x_i y_i = 12157 \quad \sum_{i=1}^n x_i = 243$$

$$\sum_{i=1}^n y_i = 274 \quad \sum_{i=1}^n x_i^2 = 10673$$

which results in

$$b_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}}$$

$$= \frac{12157 - \frac{(243)(274)}{6}}{10673 - \frac{243^2}{6}} = \frac{1060}{831.5} = 1.275$$

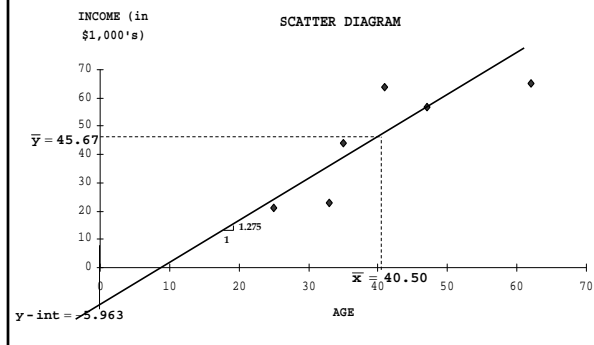
and

$$b_0 = \bar{y} - b_1 \bar{x} = \frac{274}{6} - 1.275 \left(\frac{243}{6}\right) = -5.963$$

so our estimated regression equation is

$$\hat{y}_i = b_0 + b_1 x_i = -5.963 + 1.275 x_i$$

Note that the estimated regression line does i) cut through the sample points on the scatter diagram and ii) goes through the point (\bar{x}, \bar{y}) .



C. Using the Ordinary Least Squares Estimate of the Regression Line to Estimate Values of the Dependent Variable y_i

- Recall that our generic estimated regression equation is

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

so to estimate the value of the dependent variable y_i for a corresponding value of the independent variable x_i , substitute x_i into the estimated regression equation and solve.

Example: Recall that our estimated regression equation for the previous problem is

$$\hat{y}_i = b_0 + b_1x_i = -5.963 + 1.275x_i$$

so to estimate the value of the dependent variable y_i for a corresponding value of the independent variable x_i , substitute x_i into the estimated regression equation and solve.

Thus, for our first observation ($x_1 = 25$) our regression estimated value of the dependent variable income is

$$\hat{y}_1 = -5.963 + 1.275(25) = 25.91$$

i.e., we estimate that a twenty-five year-old will earn an annual salary of \$25,910 in this organization.

We could estimate the value of the dependent variable y_i for each corresponding value of the independent variable x_i in our sample:

i	Age (x_i)	Income (y_i)	\bar{y}	$y_i - \bar{y}$	\hat{y}_i	$e_i = y_i - \hat{y}_i$
1	25	21	45.67	-24.67	25.29	-4.29
2	47	57	45.67	11.33	52.79	4.21
3	35	44	45.67	-1.67	37.79	6.21
4	62	65	45.67	19.33	71.54	-6.54
5	41	64	45.67	18.33	45.29	18.71
6	33	23	45.67	-22.67	35.29	-12.29

Note that the sum of the regression-estimates is exactly equal to the sum of the observed values of the dependent variable y_i , and the sum of the errors is 0 (which is why we square the errors e_i).

Also note that we can estimate the value of the dependent variable y for any value of the independent variable x that is within the range of values for x in our sample

For example, what is the estimated value of income (the dependent variable y) for a fifty year-old employee in this organization?

The value of the independent variable x (50) is within the range of values for x in our sample (25 - 62), so

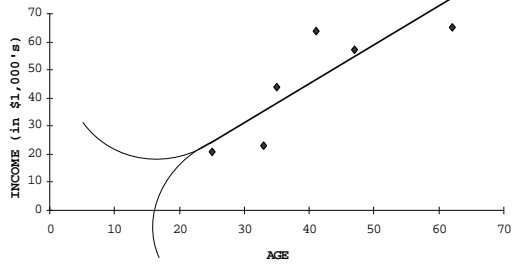
$$\hat{y}_i = -5.963 + 1.275(50) = 57.78$$

i.e., we estimate that a fifty year-old will earn an annual salary of \$57,780 in this organization.

Why can't we estimate estimate the value of the dependent variable y for a value of the independent variable x that is outside the range of values for x in our sample?

We have no idea what the relationship between the dependent variable y and the independent variable x is outside the range of values for x in our sample

SCATTER DIAGRAM



An attempt to do so is called *extrapolation* - this is one of the dangers of using regression analysis!
