Simple Linear Regression Analysis

DEAKIN
UNIVERSITY AUSTRALIA
Worldly

LEARNING OBJECTIVES

Upon completing this session, you should be able to do the following:

- Calculate and interpret the correlation between two variables.
- Recognize regression analysis applications for purposes of description and prediction.
- Calculate the simple linear regression equation for a set of data and know the basic assumptions behind regression analysis.
- Determine whether a regression model is significant.
- Calculate and interpret confidence intervals for the regression analysis.
- Recognize some potential problems if regression analysis

  Is used incorrectly.

PURPOSE OF REGRESSION AND CORRELATION

Explanation (Description)

- Regression helps to explain or understand the variation in a (dependent) variable.
- We do this by finding other (independent) variables that are related to the dependent variable.

We wish to know:

- The direction of that relationship
- The strength of that relationship

Prediction

- We can make use of the explanatory (independent) variables to help predict the likely outcome of the dependent variable.
- For example, knowing the number of customers a fast food restaurant has… may enable management to forecast sales.

PURPOSE OF REGRESSION AND CORRELATION

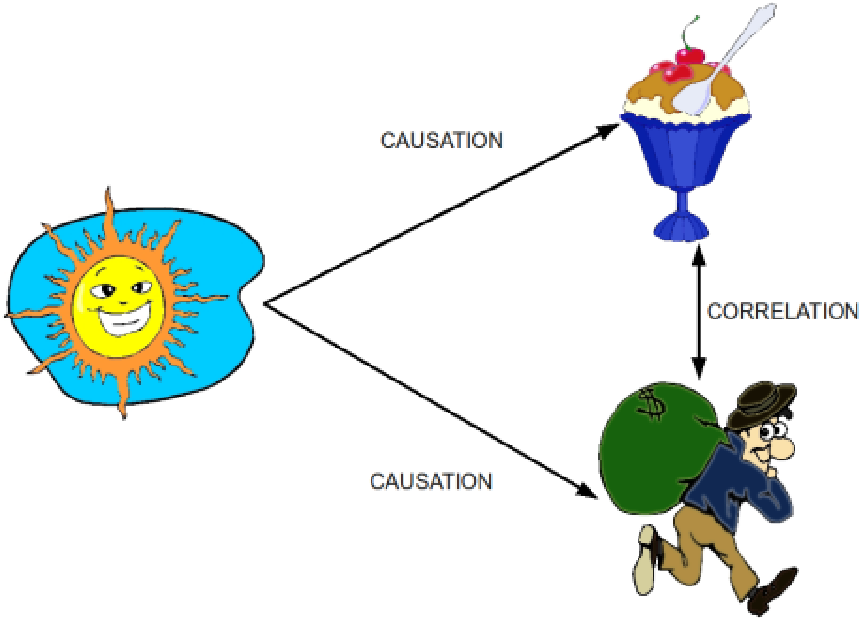
Control

- In a situation where we have some control over the value of the independent variable, this in-turn, enables some form of control over the dependent variable.
- For example, by varying advertising expenditure up or down, to a certain extent, we may be able to control the movement in sales.

CAUSATION V. CORRELATION

Be wary of Causality v. Correlation

- Correlation does NOT imply cause and effect. Just because two variables are correlation it does not mean one causes the other.



CONCEPTS IN REGRESSION AND CORRELATION

1. Scatter diagram:
   Graphical representation of the possible relationship between two variables.

2. Correlation:
   Measures the strength and direction of a linear relationship between two variables.

3. Regression:
   Gives the mathematical model of the relationship.

SIMPLE VS MULTIPLE LINEAR REGRESSION

- Simple Linear regression:
  The model involves only one independent variable.

- Multiple regression:
  Involves the use of more than one independent variable to help explain the variation in the dependent variable (covered next week).

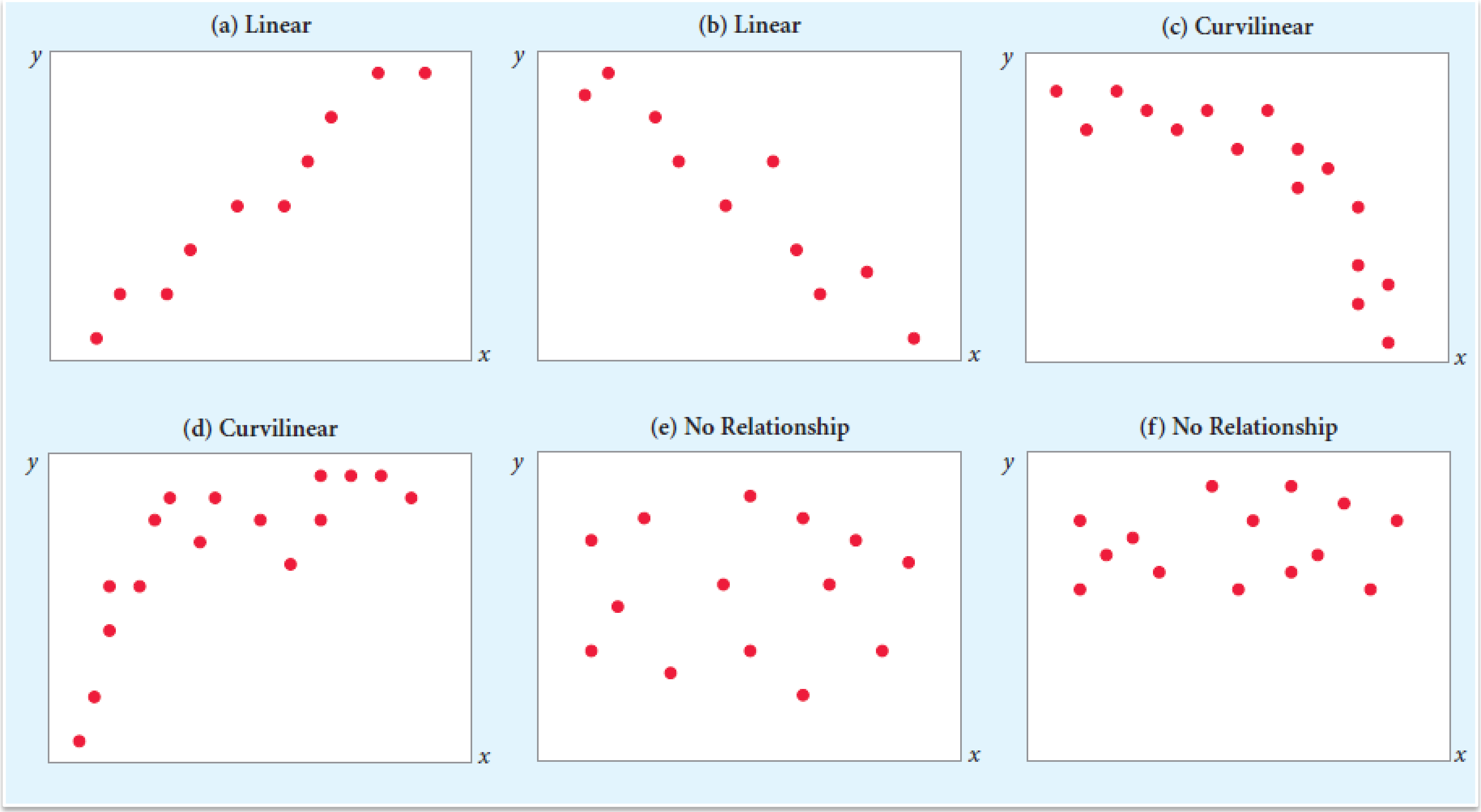
SCATTER DIAGRAMS

Scatter Plot

A two-dimensional plot showing the values for the joint occurrence of two quantitative variables. The scatter plot may be used to graphically represent the relationship between two variables. It is also known as a scatter diagram.

- The vertical (y) axis always contains the dependent variable.
- Look For
  - No relationships
  - Linear relationships
  - Non-linear relationships

TWO-VARIABLE (BI-VARIATE) RELATIONSHIPS



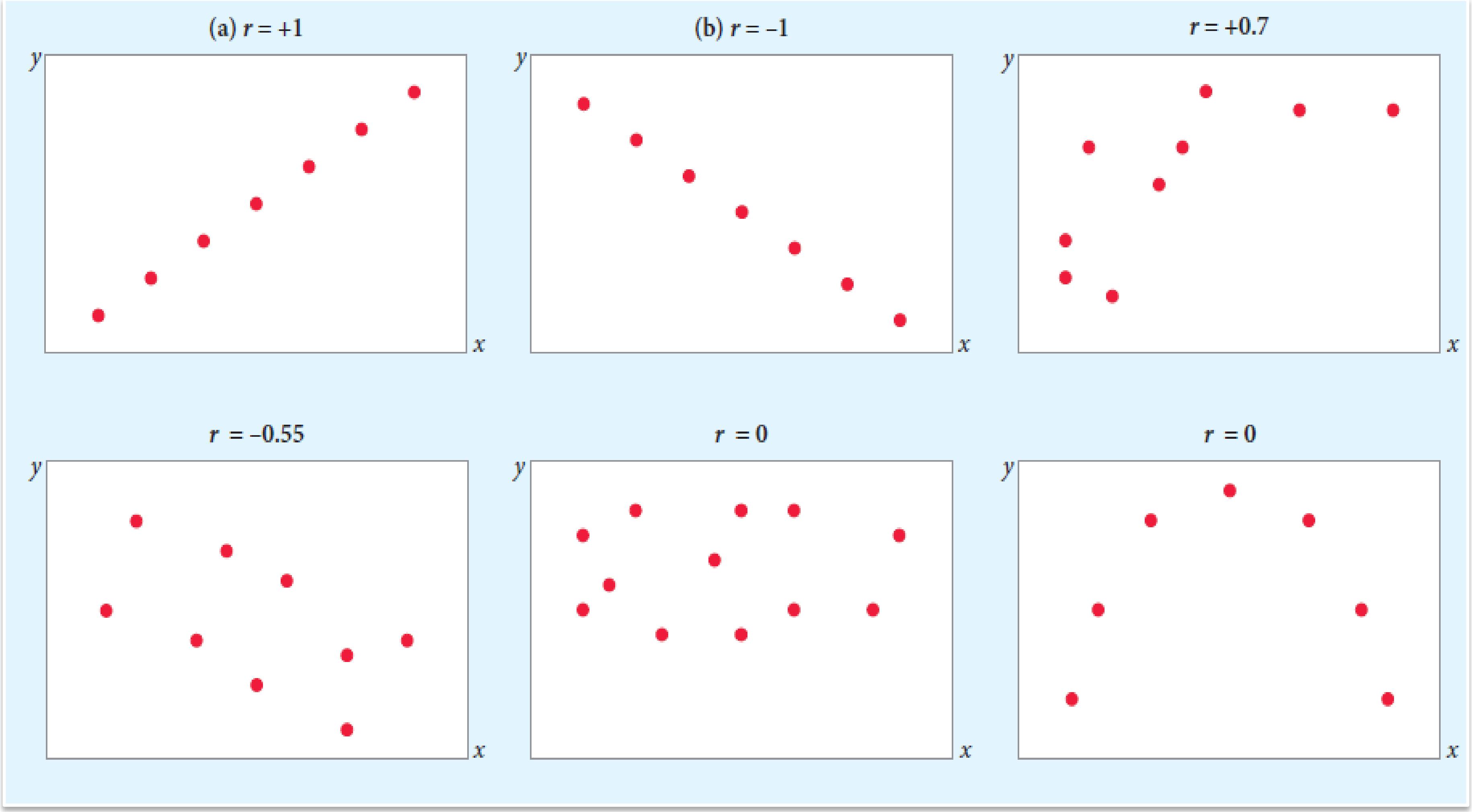
CORRELATION COEFFICIENT

Correlation Coefficient *r*

A quantitative measure of the strength of the linear relationship between two variables. The correlation ranges from -1.0 to + 1.0. A correlation of ±1.0 indicates a perfect linear relationship, whereas a correlation of 0 indicates no linear relationship.

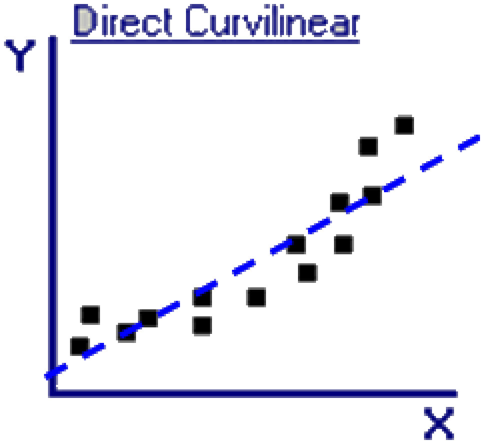
The sign of *r* provides the direction of the relationship.

CORRELATION BETWEEN TWO VARIABLES

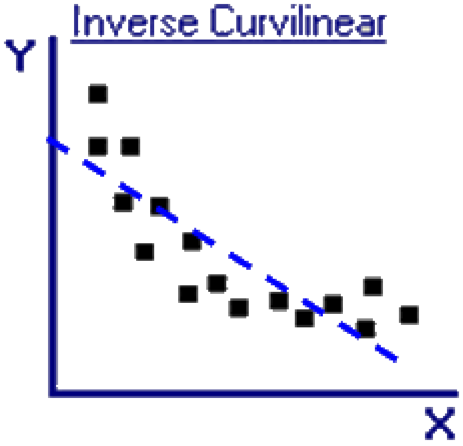


CAUTION: NON-LINEAR RELATIONSHIPS

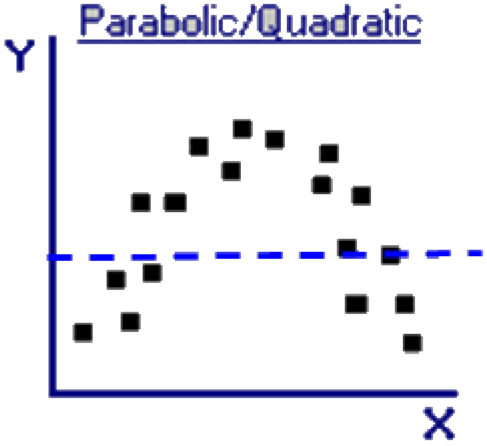
- Before interpreting *r* a scatter plot must always be drawn.
- In the following, *r* would be poor indicators of the actual strengths of each relationship.



Direct Curvilinear



Inverse Curvilinear



Parabolic/Quadratic

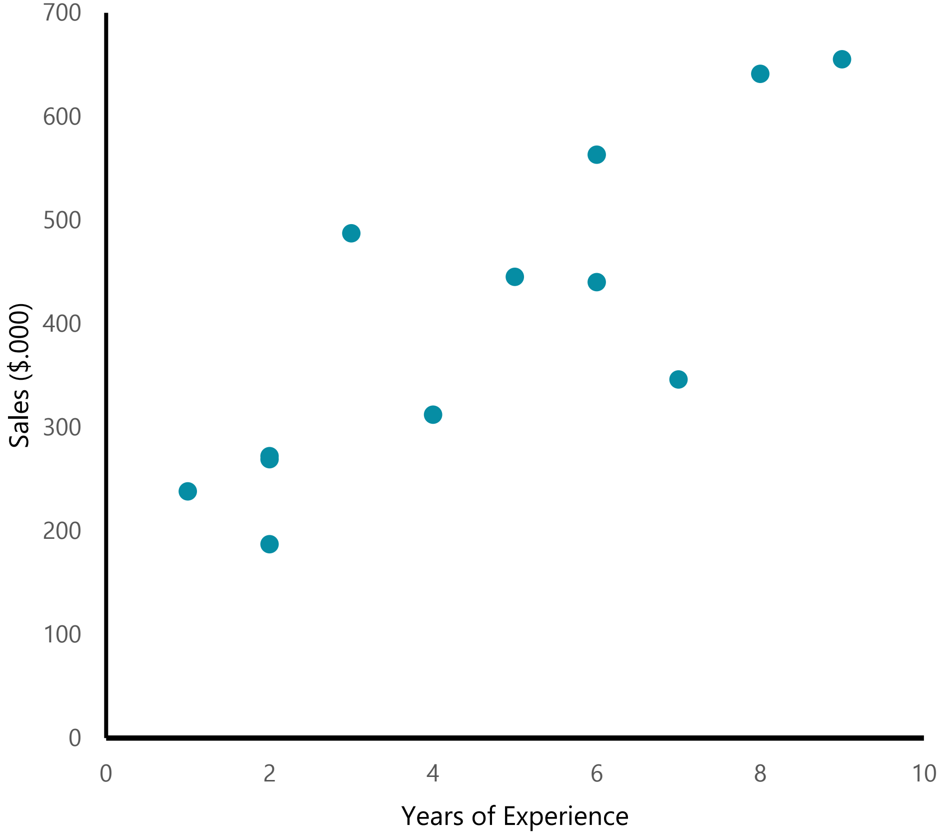
EXAMPLE:
SALES V. YEARS OF EXPERIENCE

- BLITZ is studying the relationship between sales (on which commissions are paid) and number of years a sales person is with the company. A random sample of 12 sales representatives is collected.

Sales	487	445	272	641	187	440	346	238	312	269	655	563
Years with BLITZ	3	5	2	8	2	6	7	1	4	2	9	6

SALES V. YEARS OF EXPERIENCE

ARE SALES AND YEARS OF EXP RELATED?

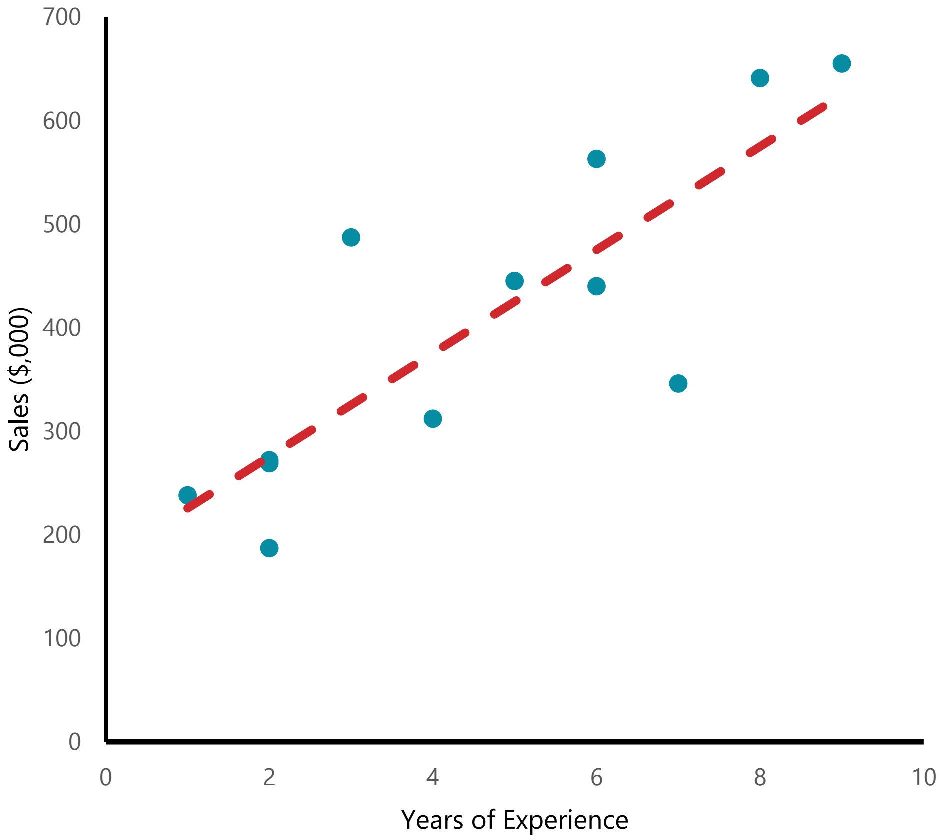


$r = +0.832$

There appears to be a strong positive linear relationship between years of experience and sales.

CRICOS Provider Code: 00113B

REGRESSION: LINE OF BEST FIT



- Is there one line or curve which is closest to the set of data?
- The "method of least squares" provides us with the line of best fit through the points on a scatter diagram

CRICOS Provider Code: 00113B

THE REGRESSION MODEL

The estimated simple linear regression equation, is given by:

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

Where:

$\hat{y}$ is the dependent variable

x is the independent variable

b_0 is the Y-intercept (i.e. where the line cuts the vertical axis)

b_1 is the slope of the line

MEANING OF REGRESSION COEFFICIENTS

The regression coefficients “b0" and "b1" can be interpreted in three ways.

- Geometrically (i.e. graphically)
- Algebraically (i.e. in equation form)
- Practically (i.e. practical interpretation)

We explain using the previous example:

$$y = 175.83 + 49.91x$$

INTERPRETING INTERCEPT COEFFICIENT

Geometrically:

On the graph, b_0 is where the line cuts the vertical axis.

Our example: The line cuts the Y axis at 175.83.

Algebraically:

b_0 is the value of Y when X = 0.

Our example: Y = 175.83 when X = 0 years of experience.

Practically:

b_0 will not always have a useful interpretation as X = 0 may be well outside the range of X values used for the regression equation. Sometimes it is useful.

Our example: The average sales for an agent with no years of experience is 175.83 ($,000)

INTERPRETING REGRESSION COEFFICIENT

Geometrically:

On the graph, b_1 is the slope of the regression line.

Our example: The slop is 49.91.

Algebraically:

b_1 is the change in the value of Y when X changes by one unit.

Our example: If X increases by 1, Y increases by 49.91.

Practically:

b_1 indicates the impact on Y from a change in X.

Our example: For each year of experience gained by an agent, the amount of sales increases by an average of 49.91 ($,000)

HOW WELL DOES THE LINE FIT THE DATA?

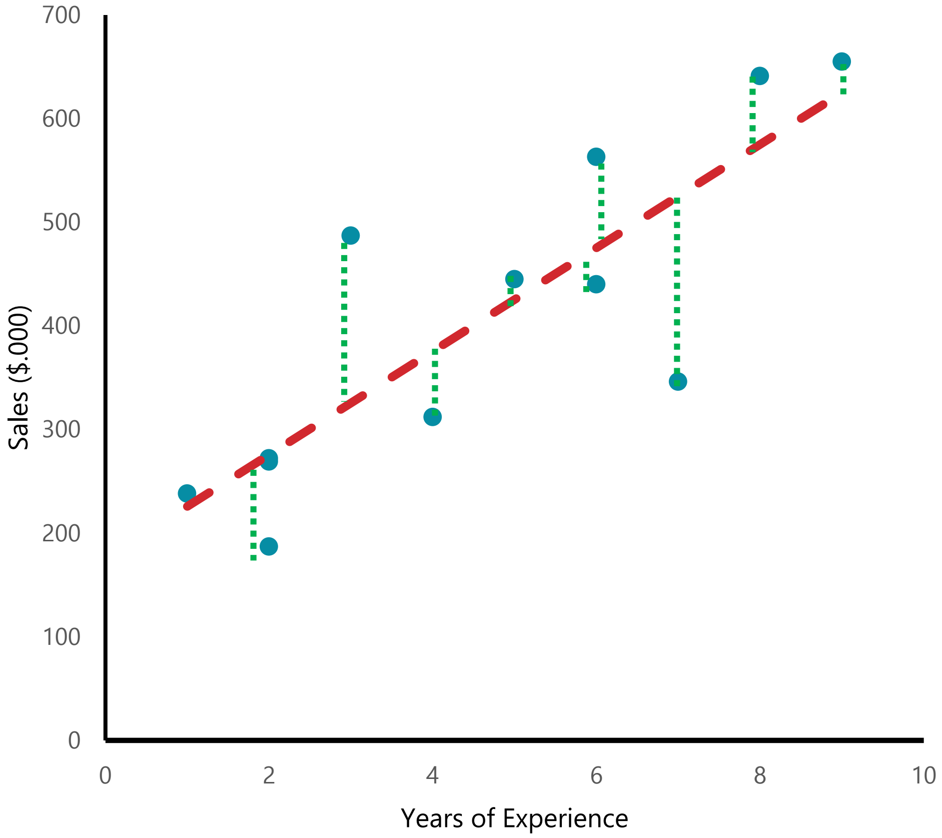
In all practical situations the regression line does not perfectly fit to the data.

There will be small variations (errors) between the line $\hat{y}$ and the actual points y_i

$$y - \hat{y}$$

These variations are called residuals (error terms).

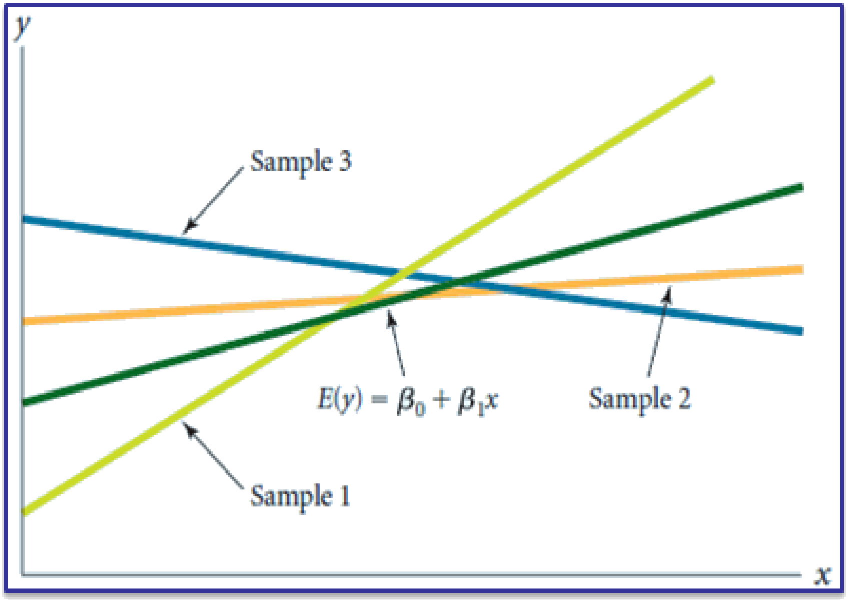
RESIDUALS

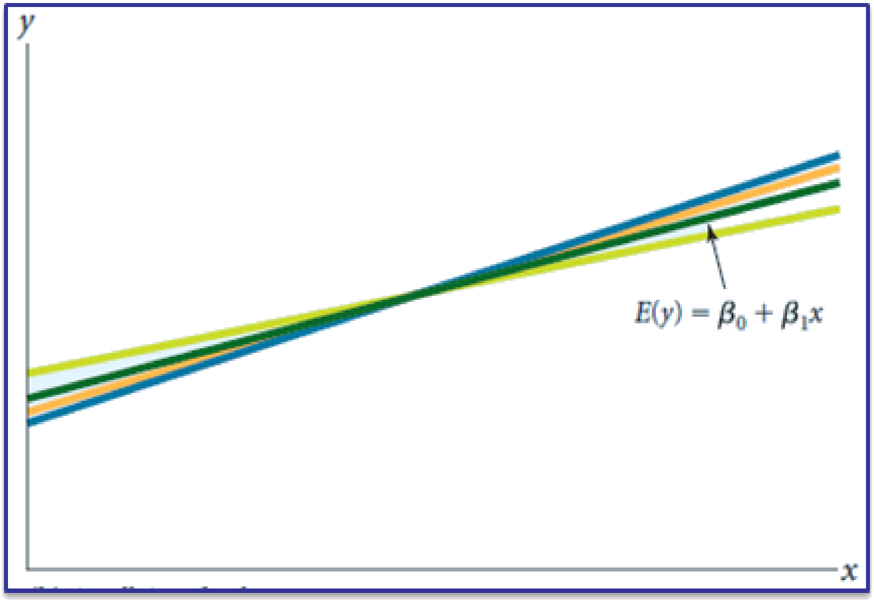

700
600
500
400
300
200
100
0
Sales ($.000)
0
2
4
6
8
10
Years of Experience

GOODNESS OF FIT

- We need to obtain measures of these residuals and hence how well the line fits the data.
- To measure the variation around the line. We use Standard Error of the Estimate, s_{yx}.
- For how well the line fits the data we use the Coefficient of Determination, R^2.

STANDARD ERROR





Large Standard Error

Small Standard Error

STANDARD ERROR OF THE ESTIMATE

Example:

Sales vs. Years of Experience

$s_{yx} = 92.10$

- Interpretation:

  “We estimate that the average variation of each individual Sales figures around the regression line is 92.10 ($,000)”.

- As a rough approximation using the empirical rule, we could also say that the maximum deviation from the line will be:

  ± (3 × 92.10) or ± 276.3 ($,000)

COEFFICIENT OF DETERMINATION

- The portion of the total variation in the dependent variable that is explained by its relationship with the independent variable.
- Normally expressed as a percentage.
- It provides an absolute measure of the strength of the relationship.

NOTE: Coefficient of Determination for the Single Independent Variable Case

$$R^2 = r^2$$

COEFFICIENT OF DETERMINATION

Example:

Sales vs. Years of Experience

$R^2 = 0.693$ or 69.3 %

- Interpretation:

  “Approximately 69% of the variation in sales is explained by or attributed to variation in the years of experience.

  The remaining 31% of variation would be the result of other factors not included in the model, e.g. negotiation skills, education, gender etc.”

LINEAR REGRESSION ASSUMPTIONS

- Linearity

  The underlying relationship between X and Y is linear.

  The error ε is a normally distributed.

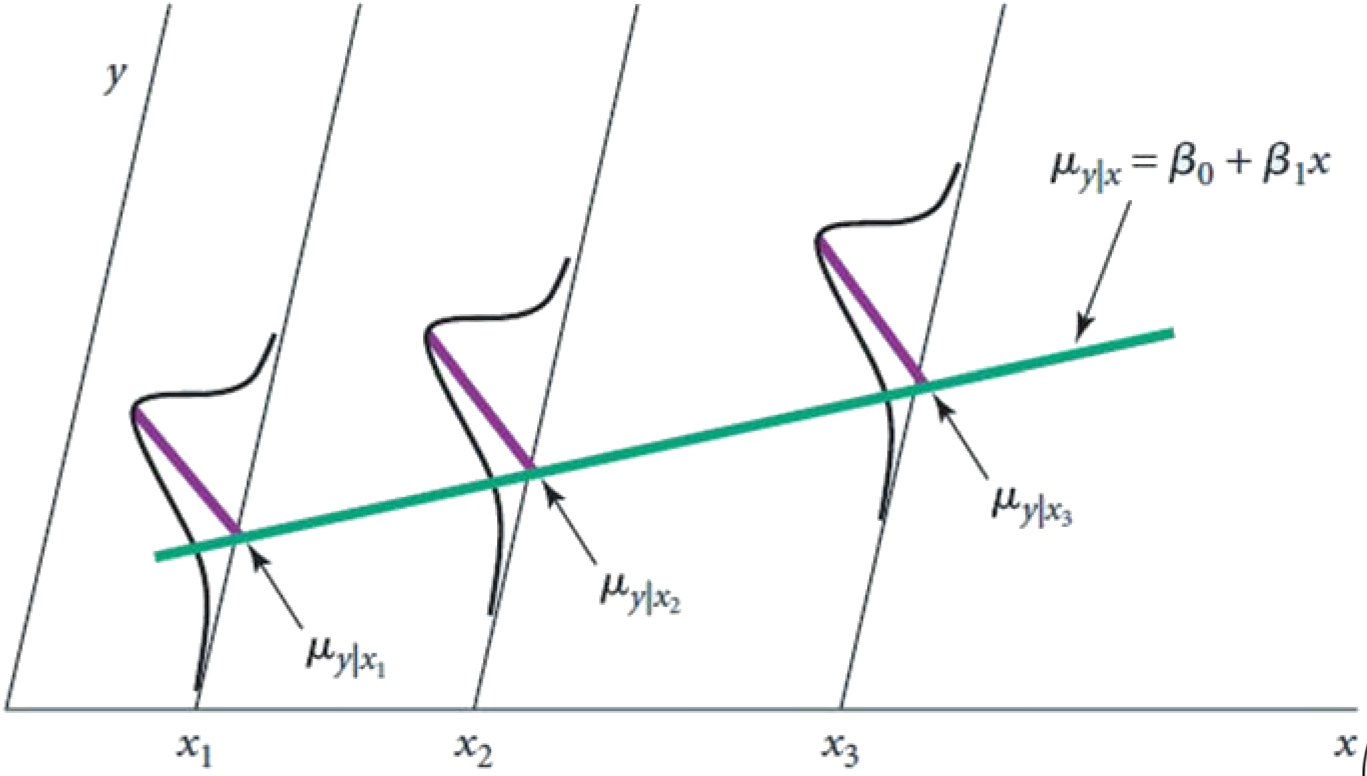
- Homoscedasticity (Constant Variance)

  The variance of ε, is the same for all values of the independent variable.

- Independence of error terms

  The values of ε are independent.

LINEAR REGRESSION ASSUMPTIONS



CONSEQUENCES OF VIOLATING ASSUMPTIONS

- Non-normality

  Error not normally distributed.

- Heteroscedasticity

  Variance not constant.

  Usually happens in cross-sectional data

- Autocorrelation

  Errors are not independent.

  Usually happens in time-series data

- Consequences of Any Violation of the Assumptions

  Predictions and estimations obtained from the sample regression line will not be accurate.

  Hypothesis testing results will not be reliable.

- It is Important to Verify the Assumptions

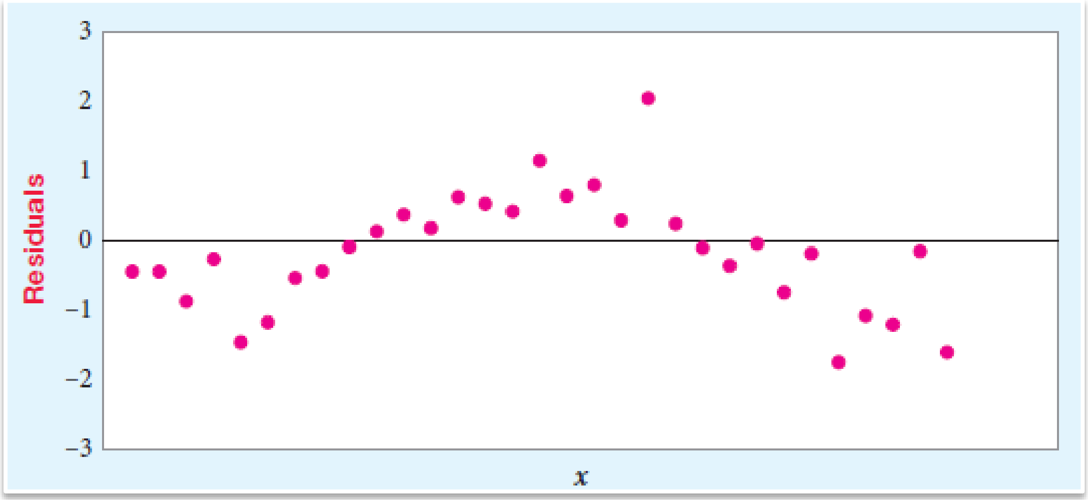
RESIDUAL ANALYSIS

- Purposes
  - Examine linearity
  - Evaluate violations of assumptions

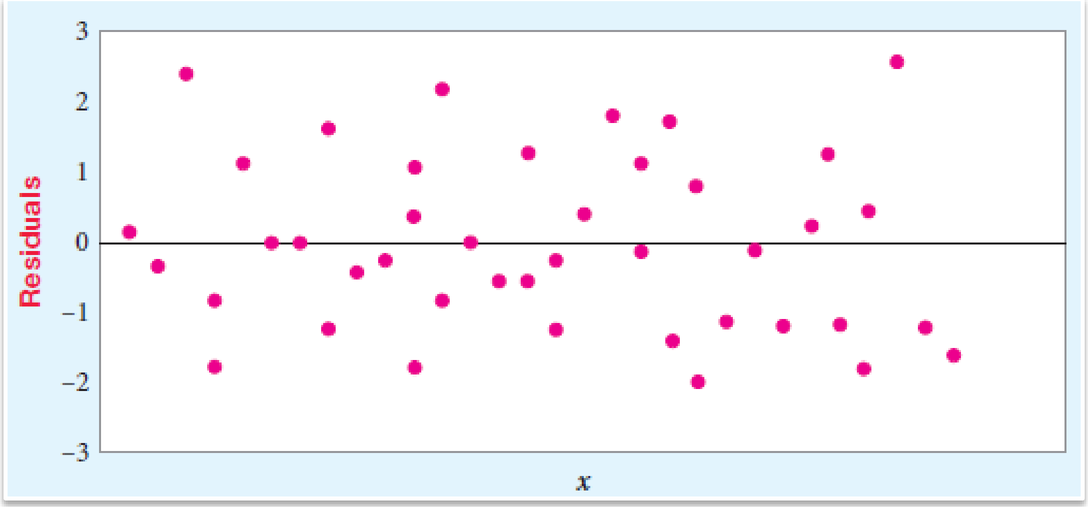
- Graphical Analysis of Residuals
  - Plot residuals vs. X and time

CHECKING FOR LINEARITY

- Nonlinear Pattern:

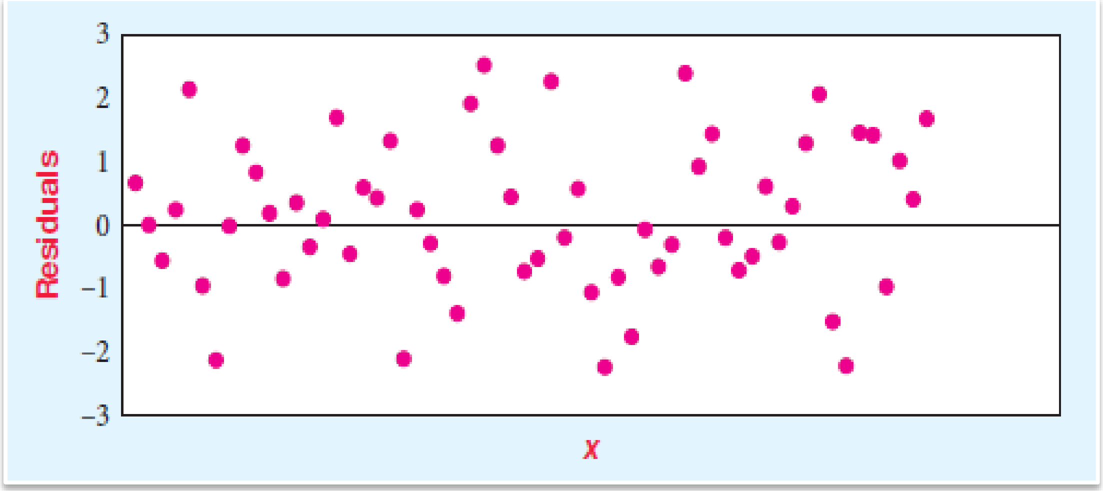


- Linear Pattern:

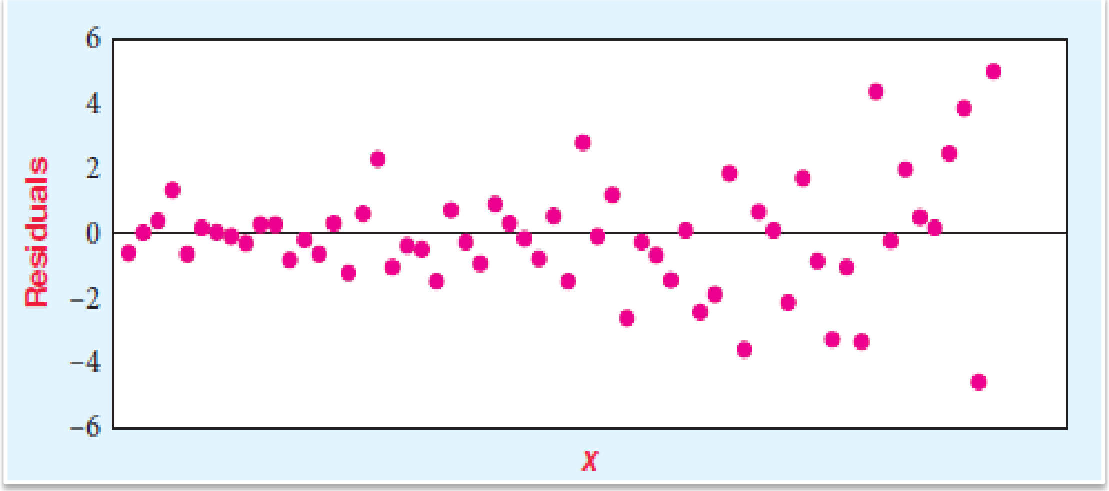


CONSTANT AND NON-CONSTANT VARIANCES

- Constant Variances:

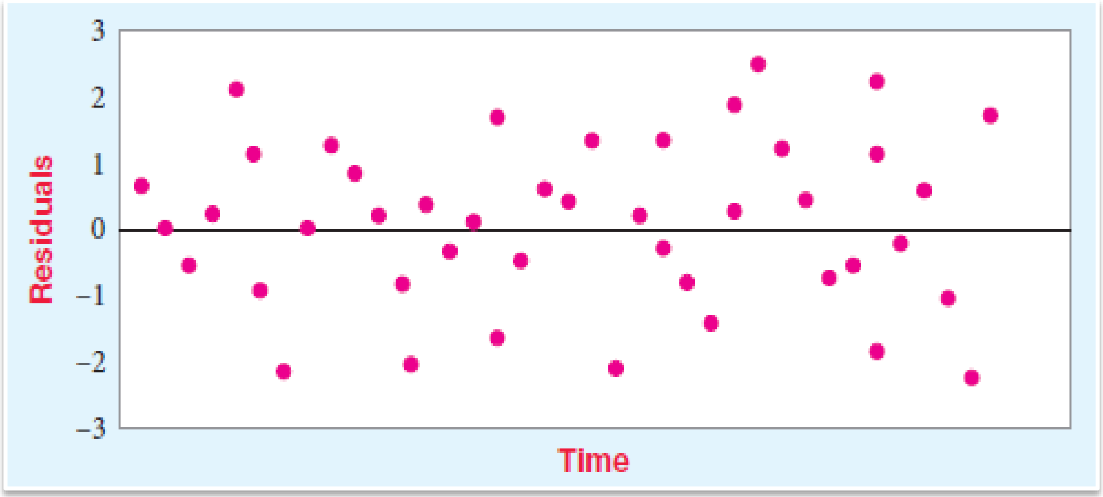


- Non-constant Variances:

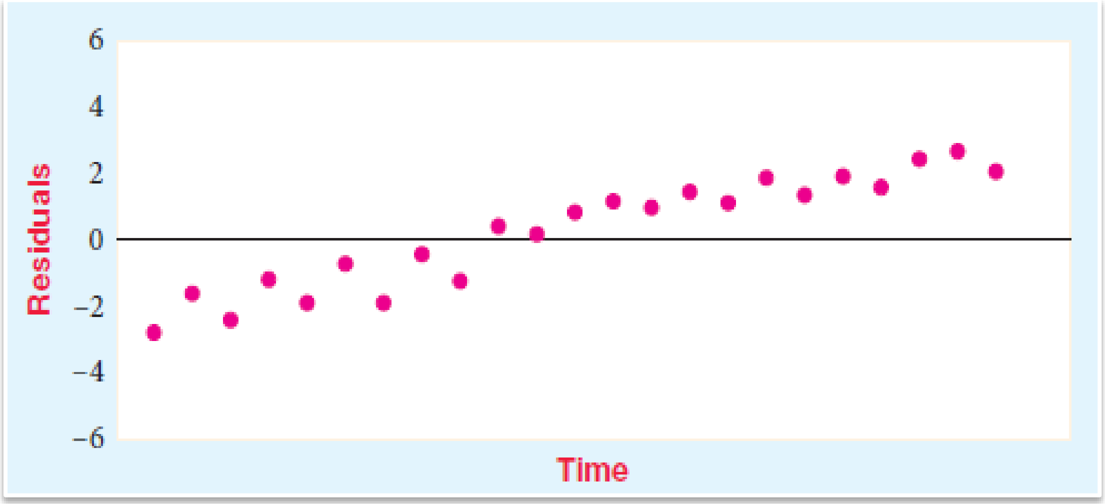


ARE THE RESIDUALS INDEPENDENT?

- Independent Residuals:



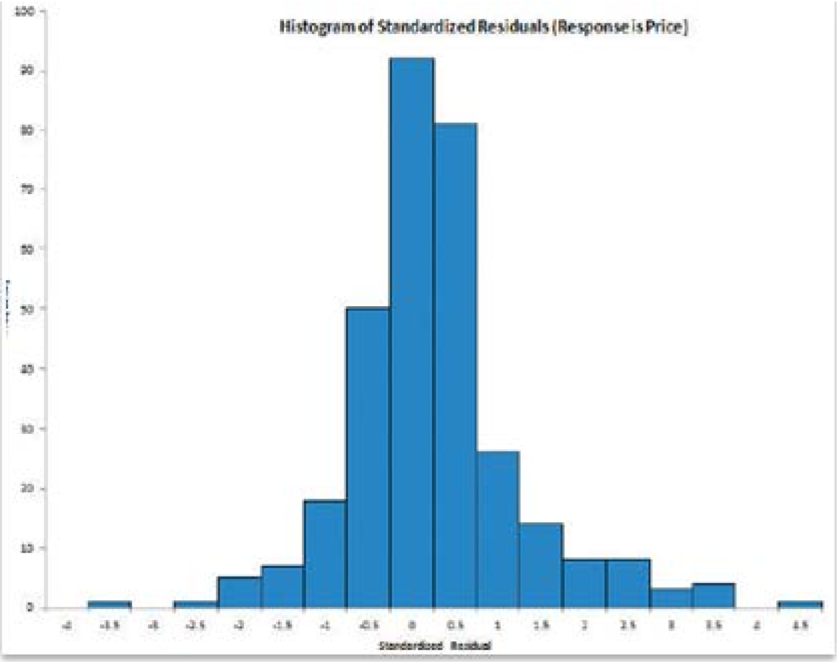
- Residuals NOT independent:

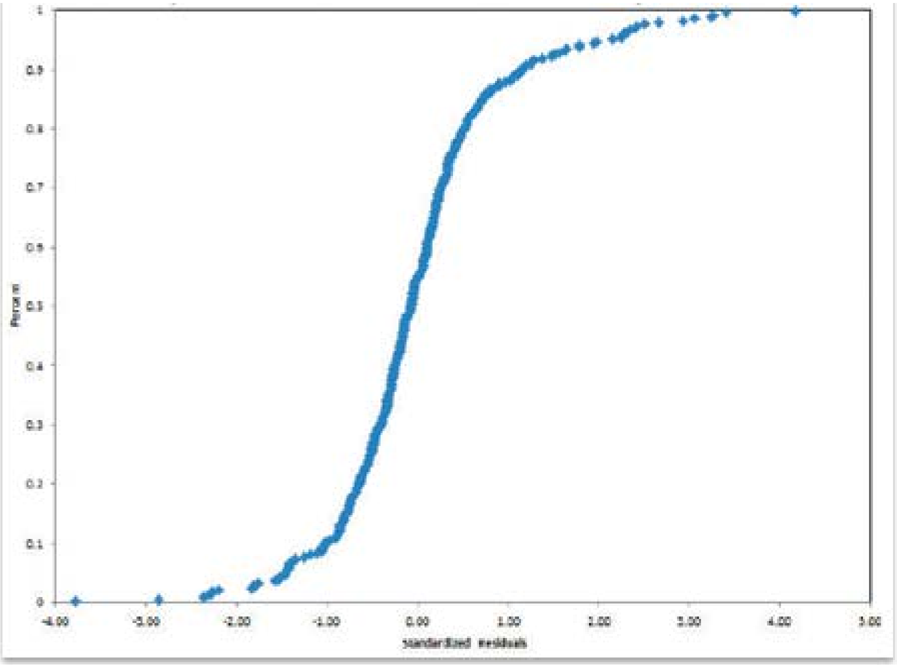


NORMALITY ASSUMPTION

Histogram of Standardized Residuals

Normal Probability Plot of Residuals





CORRECTIVE ACTIONS

- Approaches that may work if the model is not appropriate:
  - Transforming some of the independent variable
    - Raising *x* to a power
    - Taking the square root of *x*
    - Taking the log of *x*
    - If the normality assumption is not met, transforming the dependent variable (*y*) may help.
  - Remove some variables from the model

    (only when performing multiple regression)

STANDARDISED RESIDUALS

- Standardized Residual for Observation *i*

$$\frac{y - \hat{y}}{se_{(y - \hat{y})}}$$

- Standardized residuals can also be used to detect bi-variate outliers as well as to examine the assumption of regression.
- If the error term is normally distributed, 95% of the standardized residuals will be between -2 and +2.

USING THE REGRESSION EQUATION

- The regression equation coefficients (b_0 and b_1) define the nature of the relationship between the variables.
- The regression equation is also used for estimation or prediction.
- Example:

  For an agent with 5 years of work experience, the estimated sales figure would be:

$$y = 175.83 + 49.91 \times (5)$$

This is a point estimate →

$$= 423.35 \ (\$,000)$$

PREDICTIONS VS. EXTRAPOLATION

- Prediction is when we use the regression model with a value of X contained in the range of X values from the sample.
- Extrapolation is when we use the model with a value of X outside the range of original X values.
- Extrapolation should be used with caution as there is no guarantee that the same model holds outside the original range of data.

INFERENCES IN CORRELATION AND REGRESSION

- From sample data we obtain a sample regression line and associated results.
- This provides an estimate of the population regression line and other parameters.
- We use hypothesis tests and confidence intervals to make inferences about these population parameters.

INFERENCES IN CORRELATION AND REGRESSION

- The sample correlation coefficient is *r*. The corresponding population value is ρ “rho”.

- The estimated simple linear regression equation based on sample data is given by:

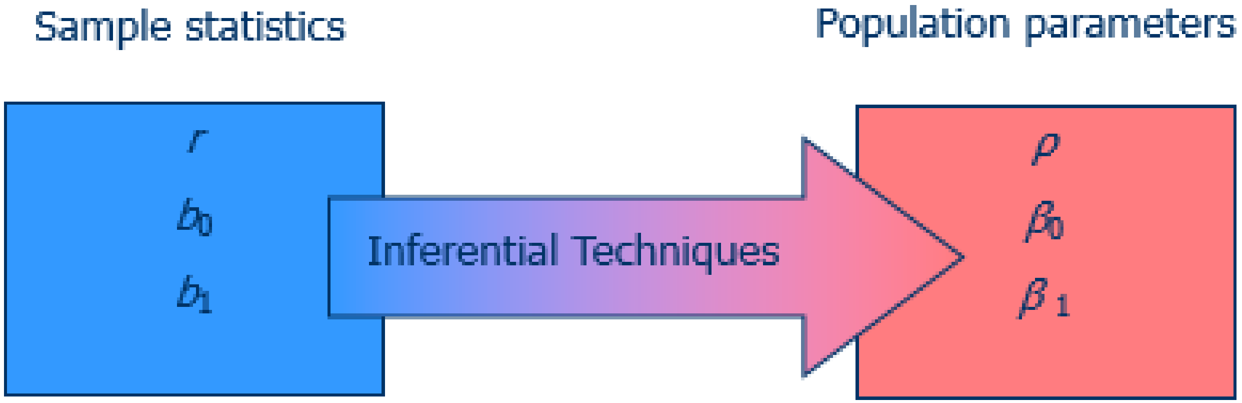
$$\hat{y} = b_0 + b_1 x$$

- The corresponding equation for the population is given by:

$$Y = \beta_1 + \beta_2 X$$

INFERENCES IN CORRELATION AND REGRESSION

- As we did in topics two and three for μ and π, we can use confidence intervals and hypothesis tests to estimate/test the corresponding population parameters.



INFERENCES ABOUT THE SLOPE T-TEST

- We want to determine whether the population parameter slope β_1 could be different from zero.
- That is, a linear relationship exists between Y (Sales) and X (Years of Experience) in the population.

H_0: $\beta_1 = 0$ (NO linear relationship)

H_1: $\beta_1 \neq 0$ (linear relationship does exist)

INFERENCES ABOUT THE SLOPE T-TEST

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	175.82	54.98	3.19	0.009	53.30	298.35
Years of Experience	49.910	10.50	4.75	0.000	26.50	73.31

This is a two-tail test, the p-value is = 0.000

Decision:

P-value < α (.05) so **reject H_0**

Conclusion:

There is sufficient evidence that the years of experience predicts the amount of sales ($,000) i.e. linear relationship exists, as $\beta_1 \neq 0$.

USES FOR REGRESSION ANALYSIS

- Earlier we calculated point estimates
- We can improve on the point estimate by calculating intervals
  - Confidence intervals for an Average value of Y, given X
  - Prediction interval for a Particular value of Y, given X

Question: Which of the interval estimates would you expect to be wider?

CONFIDENCE INTERVAL FOR *E*(Y) | X$_p$

$$\hat{y} \pm ts_e\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Where:

$\hat{y}$ = Point estimate of the DV

t = Critical value with n-2 *df*

n = sample size

x_p = specific value of the IV

$\bar{x}$ = Mean of the IV observations in the sample

s_e = Estimate of the standard error of the estimate

A rough approximation to the confidence interval estimate can be obtained by:

$$\hat{Y} \pm t \times \frac{S_e}{\sqrt{n}}$$

CONFIDENCE INTERVAL FOR E (Y) | X_P

EXAMPLE

- Calculate the 95% confidence interval for average, or expected value - E(Y) - of the amount of sales for all agents with 5 years of work experience :

$s_e = 10.502$ (\$,000)

$n = 12$

$t_{(df = 10)} = 2.228$

Point estimate → $\hat{y} = 175.83 + 49.91 \times (5)$

$$\hat{Y} \pm t \times \frac{S_e}{\sqrt{n}} \quad 425.38 \pm 2.228 \times \frac{10.502}{\sqrt{12}}$$ OR 418.618 to 432.14 (\$,000)

PREDICTION INTERVAL FOR A PARTICULAR Y | X_P

$$\hat{y} \pm t s_e \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Where:

$\hat{y}$ = Point estimate of the DV

t = Critical value with n-2 *df*

n = sample size

x_p = specific value of the IV

$\bar{x}$ = Mean of the IV observations in the sample

s_e = Estimate of the standard error of the estimate

A rough approximation to the confidence interval estimate can be obtained by:

$$\hat{Y} \pm t \times S_e$$

PREDICTION INTERVAL FOR A PARTICULAR Y | X_P

EXAMPLE

- Calculate the 95% confidence interval for of the amount of sales made by a particular agent with 5 years of work experience :

$s_e = 10.502$ (\$,000)

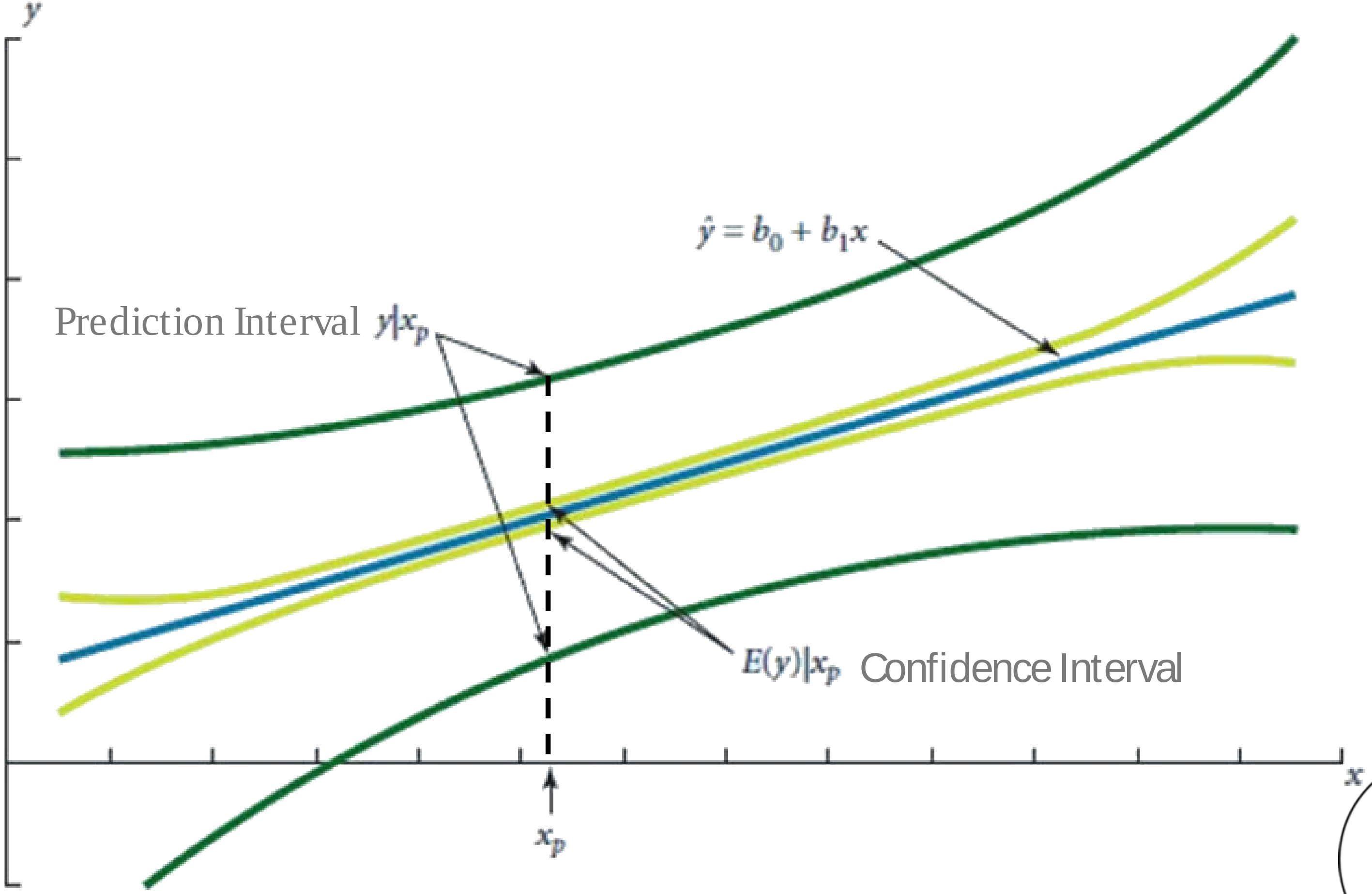
$n = 12$

$t_{(df=10)} = 2.228$

Point estimate → $\hat{y} = 175.83 + 49.91 \times (5)$

$\hat{Y} \pm t \times S_e$ $425.38 \pm 2.228 \times 10.502$ OR 401.98 to 448.77 (\$,000)

CONFIDENCE AND PREDICTION INTERVALS



PITFALLS OF REGRESSION ANALYSIS

- Causation vs. Correlation
- Extrapolation
- Lacking an awareness of the assumptions underlining least-squares regression
- Not knowing how to evaluate the assumptions
- Not knowing what the alternatives to least-squares regression are if a particular assumption is violated
- Using a regression model without knowledge of the subject matter

STRATEGY FOR AVOIDING THE PITFALLS

- Start with a scatter plot to observe possible relationship between X on Y
- Perform residual analysis to check the assumptions
- If there is violation of any assumption, use alternative methods or take corrective actions
- If there is no evidence of assumption violation, then test for the significance of the regression coefficients and construct confidence intervals and prediction intervals
- NOTE:

  Confidence and prediction errors may simply be too wide for the model to be used in many situations

QUESTIONS?

DEAKIN
UNIVERSITY AUSTRALIA
Worldly