ANALYSING THE DATA – TESTS OF ASSOCIATION

KEY QUESTIONS TO ASK

- How are the variables measured?
- Interval or ratio variables? – correlation analysis
- Interval or ratio variables involving a DV & an IV – regression analysis
- Ordinal variables? – Spearman's rank-order correlation

Tests of Association: Bivariate statistical techniques that measure whether two variables are associated/related

E.g. Are sales volume associated with advertising dollar expenditures? – if so, can one predict sales volume (DV) based on advertising expenditures (IV)?

Typically X is used as mathematical symbol for the IV & Y as the DV

CORRELATION	REGRESSION
Allows determination of the direction of the association, the strength of the association & the statistical significance of the association between 2 interval/ratio variables N.B. If between 2 ranking variables, use Spearman's correlation	Between 2 interval/ratio variables where 1 can be classified as the DV & the other the IV, thus prediction of values of the DV based upon values of the IV N.B. If categorical IV – create a dummy variable (0, 1)

Relationship: A consistent & systematic link between two variables

Association between variables includes:

- Presence of association – yes v. no
- Direction of association – positive vs. inverse (negative)
- Strength of association – strong v. moderate vs weak
- Type of association – linear v. curvilinear

TYPES OF POSSIBLE RELATIONSHIPS

PEARSON's CORRELATION COEFFICIENT

Correlation Analysis: Statistical measure of association, between two interval/ratio variables

Correlation Coefficient (r) – ranges from -1 to +1

⇒ Perfect positive linear relationship = +1

⇒ Perfect negative (inverse) linear relationship = -1

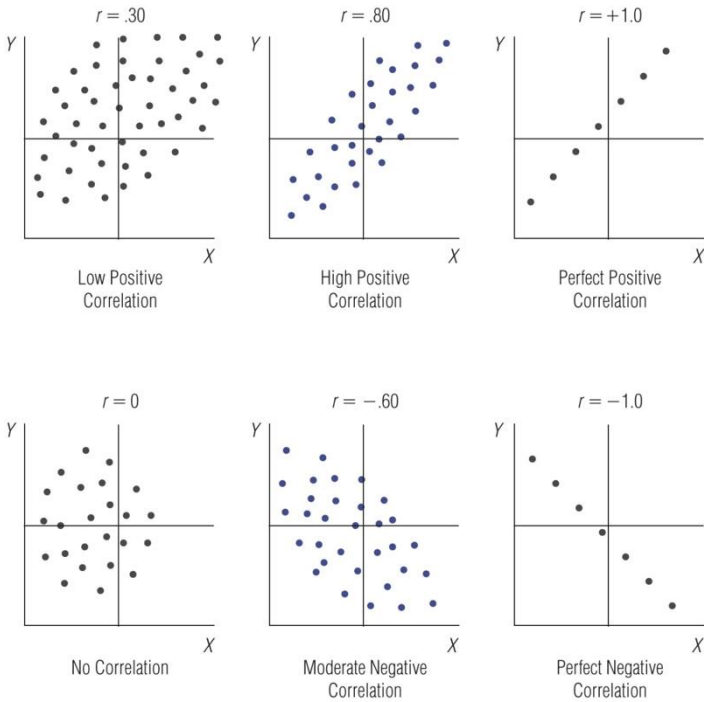
⇒ No correlation = 0

Correlation coefficient indicates the magnitude of the linear relationship & the direction

⇒ <0, variable X increases as Y decreases

⇒ >0, both variables increase together

They are useful because they can compare two correlations without regard for the amount of variance exhibited by each variable separately



Assessing the strength of the relationship:

Range of Coefficient	Description of Strength
± 0.81 to ± 1	Very strong

± 0.61 to 0.80	Strong
± 0.41 to ± 0.60	Moderate
± 0.21 to ± 0.40	Weak
± 0.01 to ± 0.20	Very weak
0	No relationship

Typically H_0 hypothesis being tested is that the population coefficient is 0, if $p<0.05$ this suggests there is evidence that the relationship observed in the sample is statistically significant (unlikely to have occurred by chance)

⇒ $H_0: \rho = 0, H_A: \rho \neq 0$

SPSS PROCEDURE
Analyse → Correlate → Bivariate
Move required variables into the 'variables' box
Check that "Pearson" is selected & [Ok]

Check for significance, if significant ($p<0.05$) explain the association/relationship in terms of strength & direction

COEFFIFICENT OF DETERMINATION (r^2)
Proportion of total variance of a variable that is accounted for by knowing the value of another variable
Measure obtained by squaring the correlation coefficient

⇒ I.E. If wish to know the proportion of variance in Y that is explained by X, can calculate the coefficient of determination by squaring the correlation coefficient

NON-PARAMETRIC CORRELATION

When data are ordinal (e.g. ranking data), a non-parametric correlation technique may be substituted for the Pearson's correlation technique
Common substitute is Spearman's rank-order correlation coefficient
Resulting output & interpretation similar to Pearson correlation analysis – examine strength & direction of relationship if it exists
N.B. Can also be used for analysing interval-scaled variables when intervals are not equal
SPSS Procedure – same as correlation but select "Spearman" option
Spearman correlation tests the monotonic association between two variables (i.e. not testing for a linear relationship but a general trend)
Assumptions of the Spearman test include that the data is ordinal, interval or ratio scaled and that there is a monotonic relationship

PREDICTIVE ANALYSIS USING REGRESSION

Value of Prediction – Prediction is a statement of what is believed will happen in the future based on past experience or prior observation
Two approaches to prediction

1) Extrapolation: Detects past patterns & use them for future predictions
2) Predictive Model: Uses relationships/associations found among variables to make predictions

REGRESSION ANALYSIS
Information obtained from a regression provides the same information as a correlation

coefficient but it also allows prediction of values of one variable based upon values of another variable

Bivariate Linear Regression: Measure of linear association that investigates straight-line relationships between an interval/ratio-scaled DV & an interval-ratio-scaled IV

⇒ Can also include a categorical dummy (0 vs. 1) IV

$$Y = \alpha + \beta X$$

Where:

- Y is the continuous dependent variable
- X is the independent variable
- α is the Y intercept (regression line intercepts Y-axis)
- β is the slope of the intercept (rise over run)

Regression examines the relationship between 2 variables by producing a line of best fit through the data

⇒ Hence, purpose of regression is to estimate the slope of the line

Regression makes the assumption that the DV & IV are causally linked

N.B. Multiple regression is an extension of Bivariate Regression where there are 2 or more IV

SPSS OUTPUT

Model summary table:

- R^2 value shows how well the straight line model fits the scatter of points
- The higher the r^2 value, the better the model fits the data & therefore the better the prediction of the DV

$$r^2 = \frac{regression\ sum\ of\ squares\ (RSS)}{Total\ Sum\ of\ Squares\ (TSS)}$$

Standard Error of Estimate is another measure of how well the model fits the data

⇒ Useful in constructing CI for predicted value (point estimate) of the DV

The ANOVA table:

F-test can be applied to a regression to determine whether more variability is explained by the regression or unexplained by the regression (relative magnitudes of the mean square)

- To determine whether the straight line model applied is appropriate for the variables concerned
- Regression ANOVA – F statistic

$$F\ Statistic = \frac{Regression\ Mean\ Square\ (MSR)}{Residual\ Mean\ Square\ (MSE)}$$

- When the significance/p-value of the F statistic is:
  - >0.05 → the IV do not explain the variation in the DV to that level of confidence
  - <0.05 → the IV(s) collectively do a good job explaining the variation in the DV

Regression coefficients table:

- The 'unstandardised coefficients', B, are the coefficients of the estimated regression model, these are the values of α (constant) & β (slope) for the regression equation
- The 'standardised coefficients', Beta:
  - IV generally measured in different units (e.g. dollars, years etc.)
  - Standardised coefficients are an attempt to make the regression coefficients comparable

- In multiple regression, used to determine which IV contribute most in explaining the variables in the DV

SPPS Procedure:
- [Analyse] → [Regression] → [Linear]
- Specify the [Dependent] & [Independent] variable(s) & place them in their respective box
- Default method [Enter]
- [Ok]

MULITPLE REGRESSION ANALYSIS

Multiple Regression Analysis: When the effects of two or more metric scaled independent variables on a single, metric-scaled DV are investigated

Equation:

$$Y = \alpha + \beta_i X_i + \cdots + \beta_n X_n$$

Where:
- X = independent variable m
- α = y-intercept, constant
- β = slope for independent variable m
- Y = dependent variable

Relationship is linear, however with more than one independent variable there are multiple independent dimensions

To investigate these relationships, examine the regression coefficients which describe the average amount of change in Y given a unit change in the value of X (whichever X you are considering)

⇒ Each regression coefficient describes the relationship of that variable to the dependent variable

⇒ If the IVs are measured in different units use the standardised coefficients of beta values to identify which IV has a higher influence/impact over the DV

Adjusted R^2 is similar to R^2 & is only used when comparing different regression equations with different degrees of freedom, means that if the study was replicated many times with samples drawn from the same population, would on average account for the R^{2}% of the variance in the DV with the IVs

⇒ Hence, adjusted R^2 will always be slightly lower than R^2

The F-statistic (ANOVA table) allows determination of the overall usefulness of the model – <u>if the significance level (p-value) of the F statistic is <0.05 there is some explanatory power in the regression model</u>

To derive the regression equation:
- Use the unstandardized coefficients (B) in the 'coefficients' output table
- Beta is a standardised coefficient where all variables are re-scaled so that they have a mean of zero & a standard deviation of one (enables comparison)
- For both coefficients & beta, p-value is the statistical significance of that coefficient (if p<0.05, the IV has an effect on the DV)

The coefficients β_1, β_2…. are coefficients of partial regression

⇒ The coefficient is the unit change in the DV in each unit change in the IV (holding other IVs constant)

N.B. The coefficient with the largest number does not mean that the IV has the greatest influence on the DV – this coefficient is a function of its respective measurement scale
Instead use the standardised coefficients (beta) & look for the highest beta for greatest relative influence

⇒ These are actually partial correlations – (i.e. correlation between each IV & the DV with the effects of all other IVs removed)

N.B. Regression equation – use unstandardized coefficient (B), relative influence – use standardised (beta)