Linear Regression Models

SPSS for Windows® Intermediate & Advanced Applied Statistics

Zayed University Office of Research SPSS for Windows® Workshop Series

Presented by

Dr. Maher Khelifa
Associate Professor
Department of Humanities and Social Sciences
College of Arts and Sciences

Bi-variate Linear Regression

(Simple Linear Regression)

Understanding Bivariate Linear Regression

- Many statistical indices summarize information about particular phenomena under study.

- For example, the Pearson (r) summarizes the magnitude of a linear relationship between pairs of variables.

- However, one major scientific research objective is to "explain", "predict", or "control" phenomena.

Understanding Bivariate Linear Regression

- To explain, predict, and control phenomena, we must not view variables in isolation.

- How variables do or do not relate to other variables provide us with valuable clues which allow us to:
  - Explain
  - Predict, and
  - Control

- The examination of these relationships leads to the formation of **networks of variables** that provide the basis for the **development of theories** about a phenomenon.

Understanding Bivariate Linear Regression

- Linear regression analyses are statistical procedures which allow us to move from description to explanation, prediction, and possibly control.

- Bivariate linear regression analysis is the simplest linear regression procedure.

- The procedure is called **simple linear regression** because the model:
  - explores the predictive or explanatory relationship for only 2 variables, and
  - Examines only linear relationships.

Understanding Bivariate Linear Regression

- Simple linear regression focuses on explaining/ predicting one of the variables on the basis of information on the other variable.

- The regression model thus examines changes in one variable as a function of changes or differences in values of the other variable.

Understanding Bivariate Linear Regression

- The regression model labels variables according to their role:
  - Dependent Variable (Criterion Variable): The variable whose variation we want to explain or predict.
  - Independent Variable (Predictor Variable): Variable used to predict systematic changes in the dependent/criterion variable.

Understanding Bivariate Linear Regression

- To summarize:
  - The regression analysis aims to determine how, and to what extent, the criterion variable varies as a function of changes in the predictor variable.
  - The criterion variable in a study is easily identifiable. It is the variable of primary interest, the one we want to explain or predict.

Understanding Bivariate Linear Regression

- Several points should be remembered in conceptualizing simple linear regression:
  - Data must be collected on two variables under investigation.
  - The dependent and independent variables should be quantitative (categorical variables need to recoded to binary variables).
  - The criterion variable is designated as **Y** and the predictor variable as **X**.
  - The data analyzed are the same as in correlational analysis.
  - The test still examines covariability and variability but with different assumptions and intentions.

Understanding Bivariate Linear Regression

- The relationship between X & Y explored by the linear regression is described by the **general linear model**.
- The model applies to both experimental and non-experimental settings.
- The model has both explanatory and predictive capabilities.
- The word linear indicates that the model produces a straight line.

Understanding Bivariate Linear Regression

- The mathematical equation for the general linear model using population parameters is:
  - $\mathbf{Y = \beta_0 + \beta_1 X + \varepsilon}$
- Where :
  - Y and X represent the scores for ***individual$_i$*** on the criterion and predictor variable respectively.
  - The parameters $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ are constants describing the functional relationship in the population.
  - The value of $\boldsymbol{\beta_1}$ identifies the change along the Y scale expected for every unit changed in fixed values of X (represents the **slope** or degree of steepness).
  - The values of $\boldsymbol{\beta_0}$ identifies an adjustment constant due to scale differences in measuring X and Y (the intercept or the place on the Y axis through which the straight line passes. It is the value of Y when X = 0).
  - ∑ (Epsilon) represents an error component for each individual. The portion of Y score that cannot be accounted for by its systematic relationship with values of X.

Understanding Bivariate Linear Regression

- The formula $\mathbf{Y = \beta_0 + \beta_1 X + \varepsilon}$ can be thought of as:
  - $Y_i = Y' + \varepsilon_i$ (where $\alpha + \beta 1 X i$ define the predictable part of any Y score for fixed values of X. Y' is considered the predicted score).

- The mathematical equation for the sample general linear model is represented as:
  - $\mathbf{Y_i = b_0 + b_1 X_i + e_i}$.

- In this equation the values of *a* and *b* can be thought of as values that maximize the explanatory power or predictive accuracy of X in relation to Y.

- In maximizing explanatory power or predictive accuracy these values **minimize prediction error**.

- If Y represents an individual's score on the criterion variable and Y' is the predicted score, then Y-Y' = error score (*e*) or the discrepancy between the actual and predicted scores.

- **In a good prediction Y' will tend to equal Y**.

Understanding Bivariate Linear Regression

- The general mathematical equation defines a straight line that may be fitted to the data points in a scatter diagram.

- The extent to which the data points do not lie on the straight line indicates individual errors.

- The straight line defined by the equation is called the **Best Fitting Straight Line** for the data.

- The formula minimizes error scores **across all individuals** to enhance prediction. The test uses **the principle of least squares** which selects among many possible lines the one that best fits the data (and minimizes the sum of squared vertical distances from the observed data points to the line).

- The method of least squares produces the smallest variability among error scores.

Obtaining a Bivariate Linear Regression

- For a bivariate linear regression data are collected on a predictor variable (X) and a criterion variable (Y) for each individual.
- Indices are computed to assess how accurately the Y scores are predicted by the linear equation.
- The significance test evaluates whether X is useful in predicting Y.
- The test evaluates the **null hypothesis** that:
  - the population slope = 0, or
  - the population correlation coefficient = 0

Test Assumptions

- There are two sets of assumptions to be considered for the:
  - Fixed-effects model (appropriate for experimental studies).
  - Random-effects model (more appropriate for non-experimental studies).

Fixed-Effects Model

- The following are assumptions for a fixed-effects model:

  - 1) **Normality Assumption: R**
    - The dependent variable is normally distributed in the population for each level of the independent variable.
      - With a moderate or large sample size the test yields accurate p values even when this assumption is violated.

  - 2) **Homogeneity of Variance Assumption**: **NR**
    - The population variances of the dependent variable are the same for all levels of the independent variable.
      - To the extent that this assumption is violated the resulting p values for the F test is not to be trusted.

  - 3) **Assumption of Independence: NR**
    - The cases represent a random sample from the population and the scores are independent of each other from one individual to the next.
      - The significance test will yield inaccurate p values if the independence assumption is violated.

Random-Effects Model

- The following are assumptions for the Random-effects model:

  - 1). **Normality Assumption: R**
    - The predictor and criterion variables are normally distributed in the population.
      - The significance test yields valid p values when the sample is moderate to large in size even if this assumption is violated.
      - If X and Y are normally distributed the only type of relationship that exists between these variables is linear.

  - 2). **Assumption of Independence: NR**
    - The cases represent a random sample from the population and the scores on each variable are independent of other scores on the same variable.
      - The significance test will yield inaccurate p values if the independence assumption is violated.

SPSS Output

- The SPSS reports statistic of strength of relationship that are useful for regression analyses with bivariate and multiple predictors.

- Several correlational indices are presented in the output:

  - The multiple correlation coefficient (multiple R), for simple linear regression the R is equal to the Pearson product moment correlation coefficient (*r)*,

  - Its squared value (R^2), and

  - The adjusted R^2

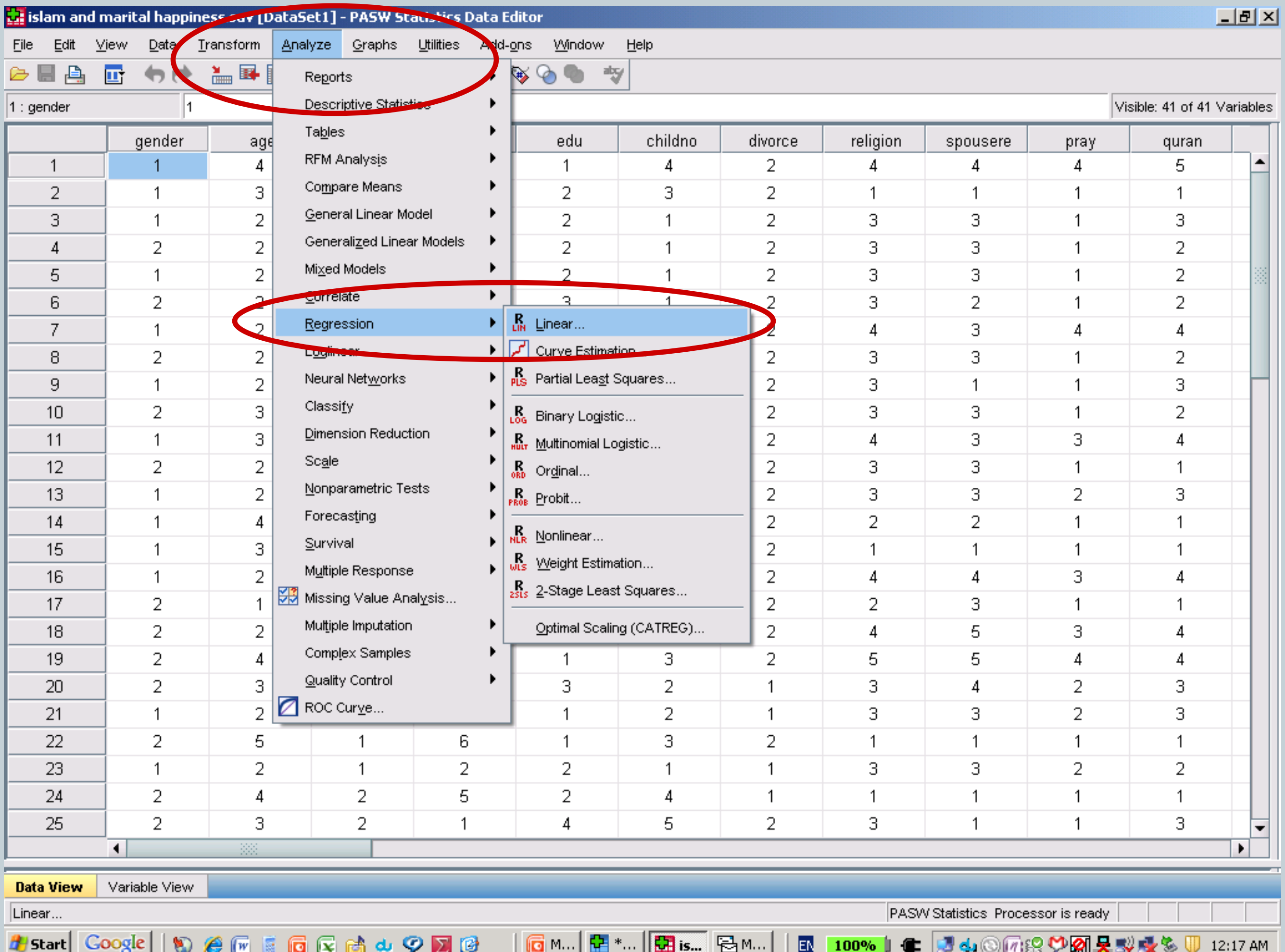
SPSS Output

- There is considerable redundancy among these statistics for the simple linear regression case where:
  - R = r
  - $R^2 = r^2$
  - Adjusted R^2 is approximately equal to R^2
  - Accordingly the indices we need to report include: r and r^2 (this index is obtained by squaring the obtained *r value*).
  - *The r^2 indicates how well we can predict Y from X.*
  - r^2 indicates the proportion of Y variance that is accounted for by its linear relationship with X.
  - Can be conceptualized as the proportion reduction in error that we achieve by including X in the regression equation in comparison with not including it.

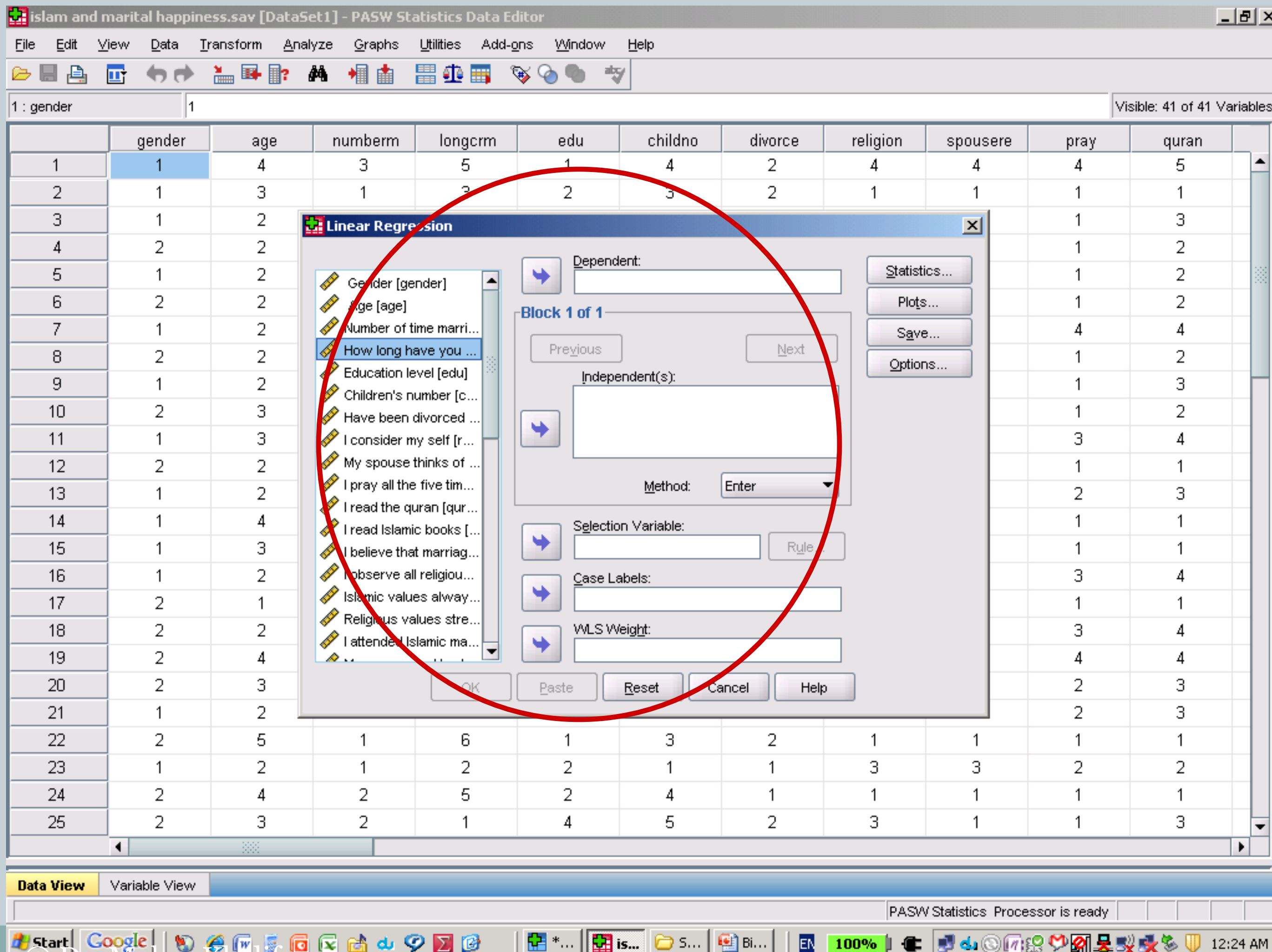
SPSS Output

- The Pearson *r* ranges in value from -1 to + 1.
- By convention relationships of .10, .30, and .50 regardless of the sign, are interpreted as small, medium, and large coefficients, respectively.
- The interpretation of the strength of relationship depends on the research context.

Conducting A Bivariate Linear Regression



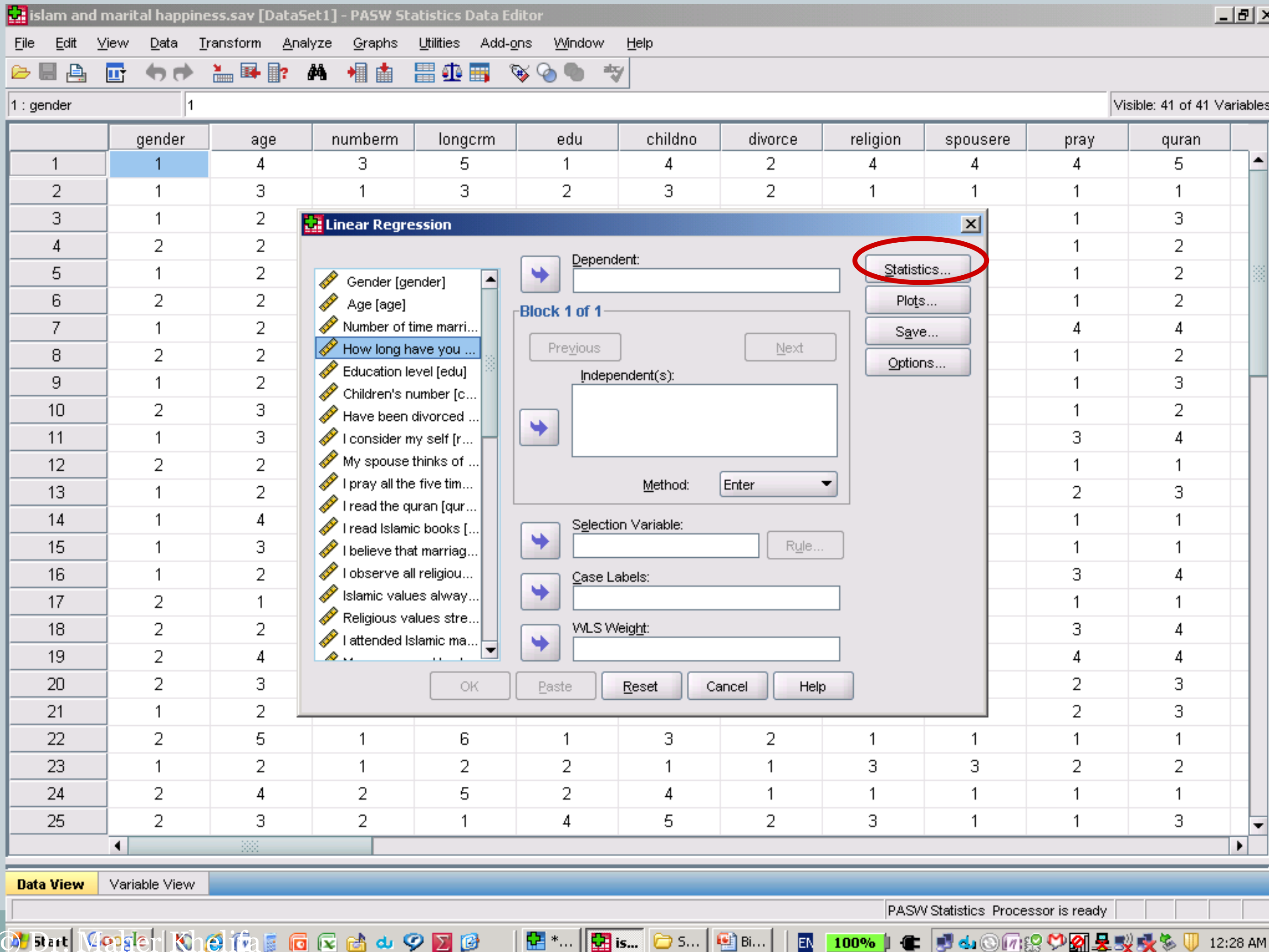
Click Analyze, Regression, Linear.



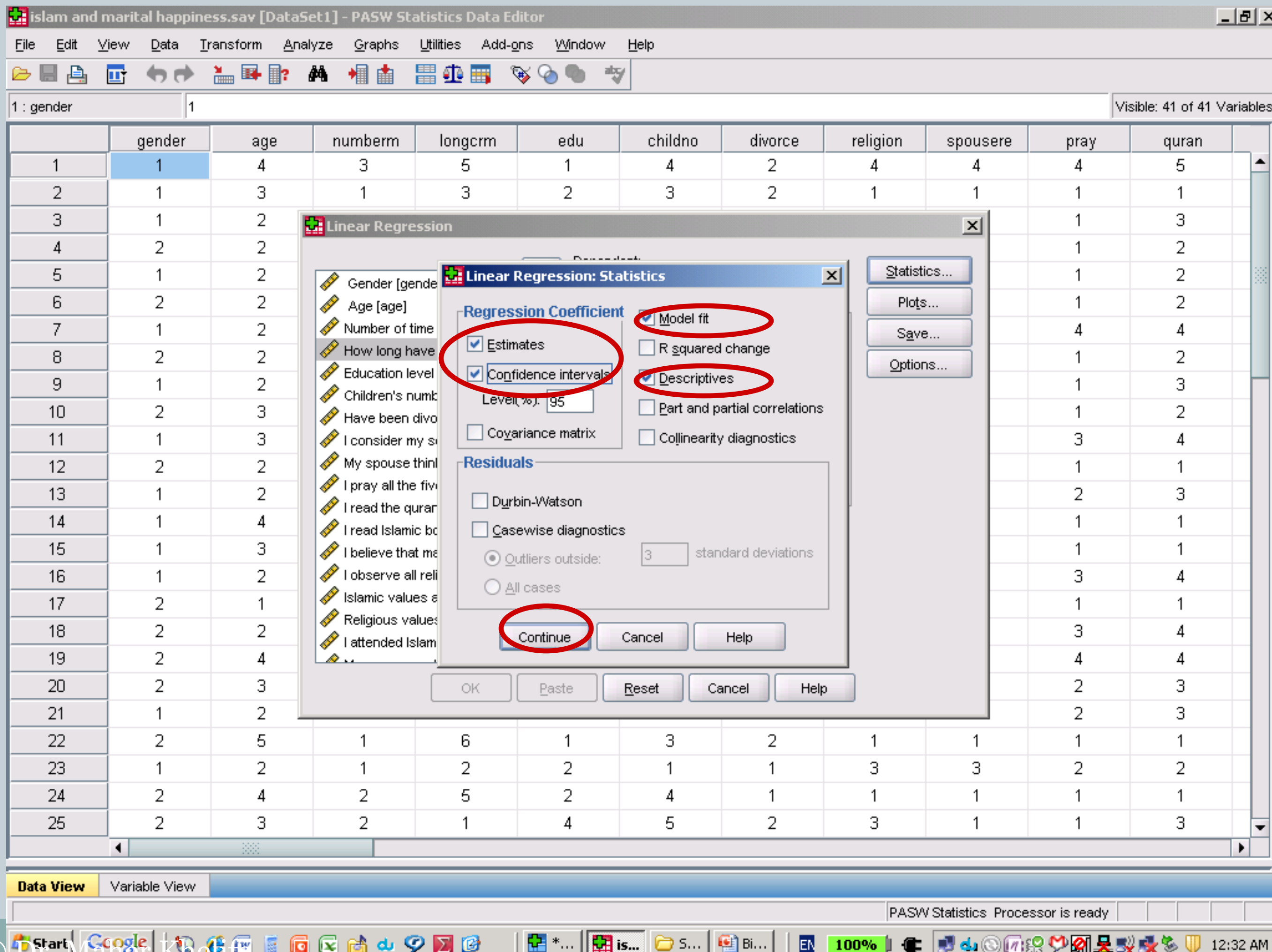
A linear regression dialog box appears.

- Chose the variable to move to the **Dependent box** and to the **Independent box.**

- The variable you want predicted is your dependent variable



Click statistics



Select:

- Estimates
- Confidence Interval
- Model Fit
- Descriptives

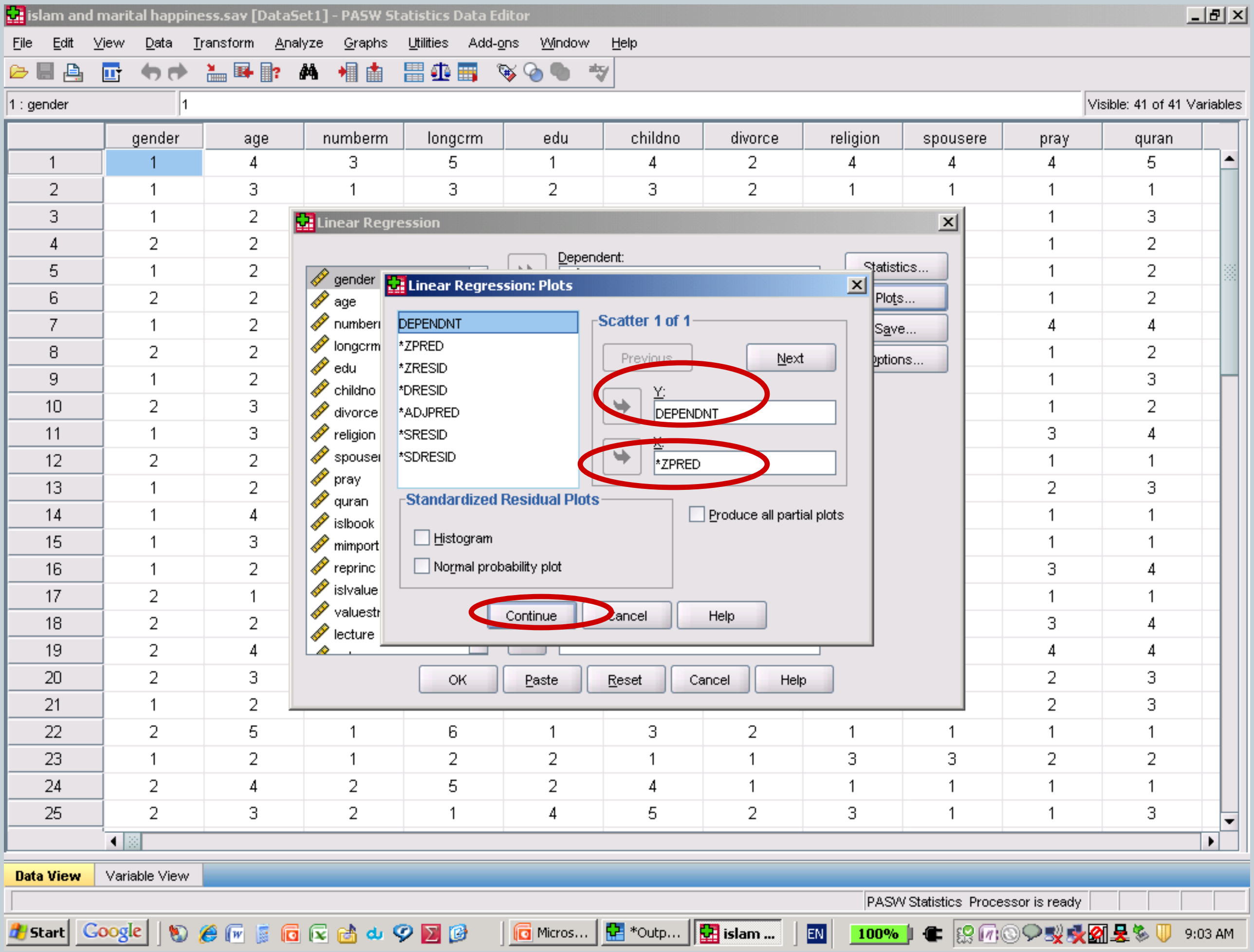
Then press Continue

Then click OK.

Linear Regression Plots

- Plots are very important in linear regression:
  - They can validate the assumptions of normality, equality of variance and linearity.
  - They also help in detecting unusual observations, outliers, and other types of relationships.
  - Typically scatterlpots are used especially plotting of Y versus X and plots of the Residuals (ZRESID) against Predicted Values [ZPRED]) of the model.

How to Obtain Scatterplots



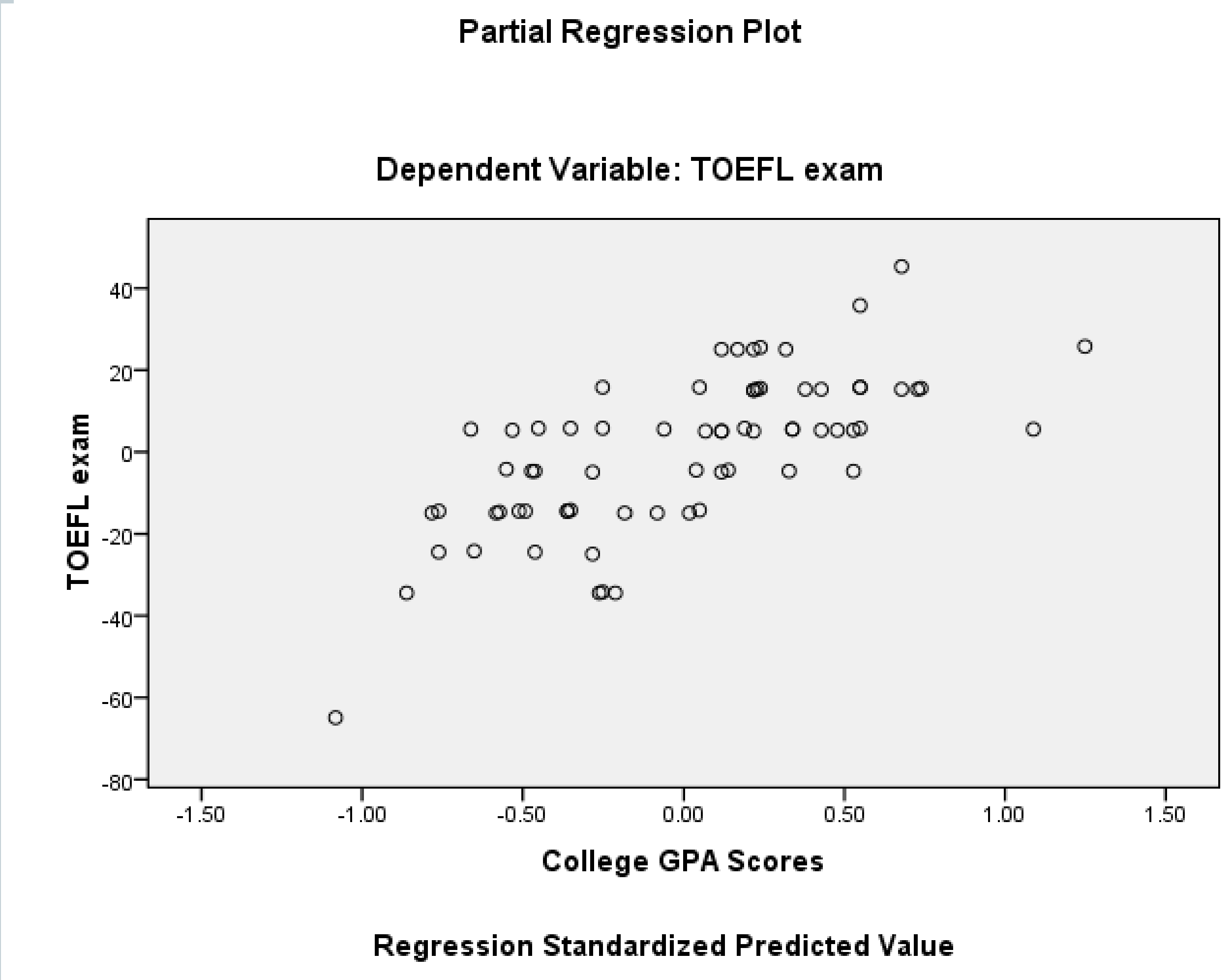
In the regression box click Plots then choose :

ZRESID and move it to the Y box

ZPRED and move it to the X box

The scatterplot will depict the relationship between standardized predicted and residual values of the dependent variable

Example of a Scatterplot



Multiple Linear Regression

Understanding Multiple Linear Regression

- Multiple Linear Regression extends bivariate linear regression by incorporating multiple independent variables (predictors).

  - **$Y = \beta_0 + \beta_1 X + \varepsilon$ (The simple linear model with 1 predictor)**

- When adding a second predictor, the model is expressed as:
  - **$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$**

- When adding more than 2 predictors, the model is expressed as:
  - **$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$**

- **ε:**In model building, a **residual** is what is left after the model is fit. It is the difference between the observed values and the values predicted by the model.

Assumptions

- The assumptions for the Multiple Linear Regression are the same as for the Simple Linear Regression model (see slides 15-17):
  - Normality Assumption (R)
  - Homogeneity of variance assumption (NR), and
  - Assumption of independence (NR).

- Conducting regression analysis without considering possible violations of the necessary assumptions can lead to results that are difficult to interpret and apply.

Number of Cases

- When conducting a regression analysis, the cases-to-Independent Variables ratio is ideally 20 cases for every Independent Variable in the model.

- The lowest acceptable ratio is 5 cases for every Independent Variable included in the model.

Screening

- In order to identify problems, the data needs to be screened first.
- Look for missing data.
- Look for outliers.
- Examine if the relationships are other than linear.

Missing Data

- If some variables have a lot of missing values, you may not want to include those variables in the analysis if possible.
- If only a few cases have missing values, then delete those cases.
- If there are missing values for several cases on different variables, then retain those cases to avoid data loss.
- If there are not too much missing data, and you are satisfied that the missing data are random (there is no pattern in terms of what is missing), then there should be no worry.

Outliers

- Outliers are atypical points suspiciously different from others which have a substantial effect on the model's goodness of fit.

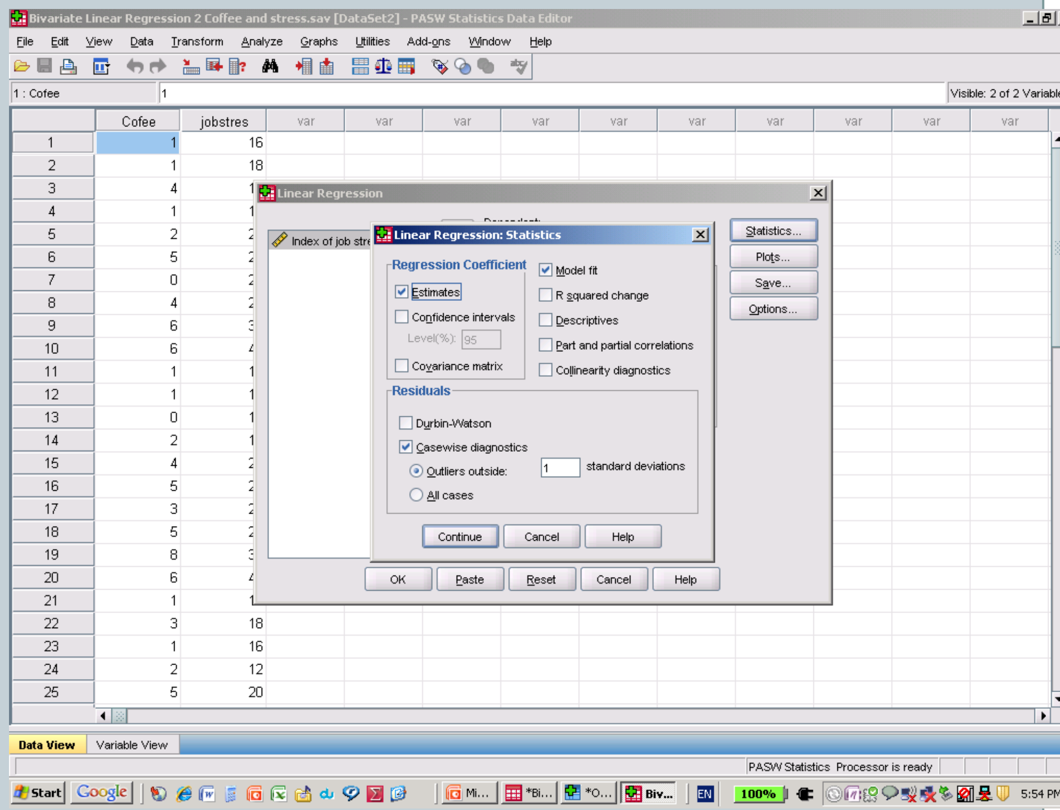
- An outlier is frequently defined as a value that is at least 3 standard deviations above or below the mean.

- Examine outliers carefully to see if they result from errors in gathering, coding, or entering data. Correct if it is the case.

- Consider interaction with other variables in case of no apparent reasons for the outlier.

Outliers

- Use SPSS regression diagnostic to identify outliers among independent variables and in the dependent variable.
  - Go to Analyze, Regression, Linear
  - Enter dependent and independent variables in their respective boxes
  - Press Statistics, then select Casewise diagnostics
  - Write Outliers outside 1 Standard deviation, Press continue and press OK



Outliers

Casewise Diagnostics[a]

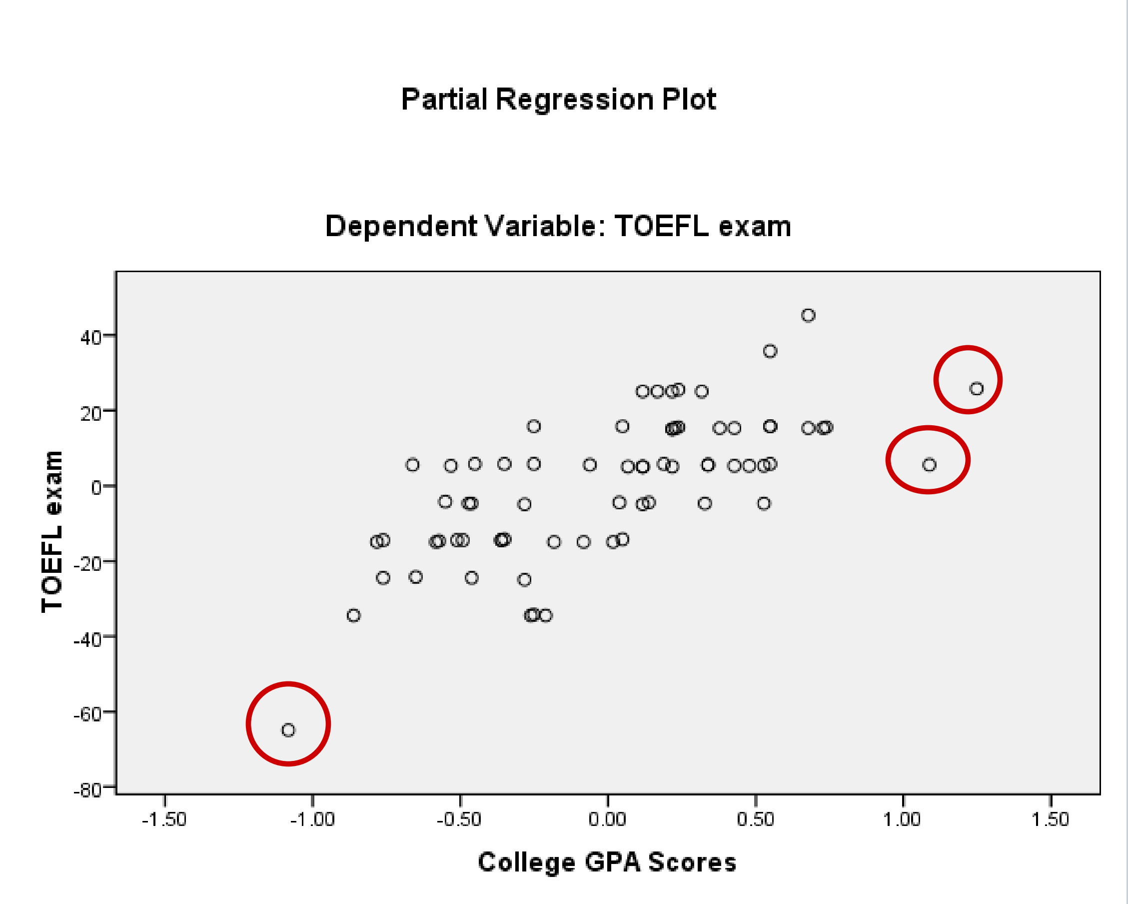
Case Number	Std. Residual	Number of cups of coffee for day 1	Predicted Value	Residual
3	1.263	4	1.91	2.092
6	1.020	5	3.31	1.690
7	-2.280	0	3.78	-3.777
13	-1.152	0	1.91	-1.908
16	1.020	5	3.31	1.690
19	1.702	8	5.18	2.821
25	1.302	5	2.84	2.157
28	-1.535	1	3.54	-2.544
29	1.702	8	5.18	2.821

a. Dependent Variable: Number of cups of coffee for day 1

Outliers

- Outliers can also be spotted readily on residual plots since they are cases with very large positive or negative residuals (error).

- In general, standardized residual values greater than an absolute value of 3 are considered outliers.



Outliers

- Distances to determine outliers can also be estimated using ***Mahalanobis Distance*** and Cook's Distance.

- **Large Mahalanobis** distance identifies cases that have an X value far from the mean (outliers).

- ***Cook's distance*** identifies **influential cases** as it considers changes in all residuals when the particular case is omitted. Non influential outliers can be deleted.

- Observation with **large influence on estimates** of the parameters considerably affect the regression line if they are omitted.

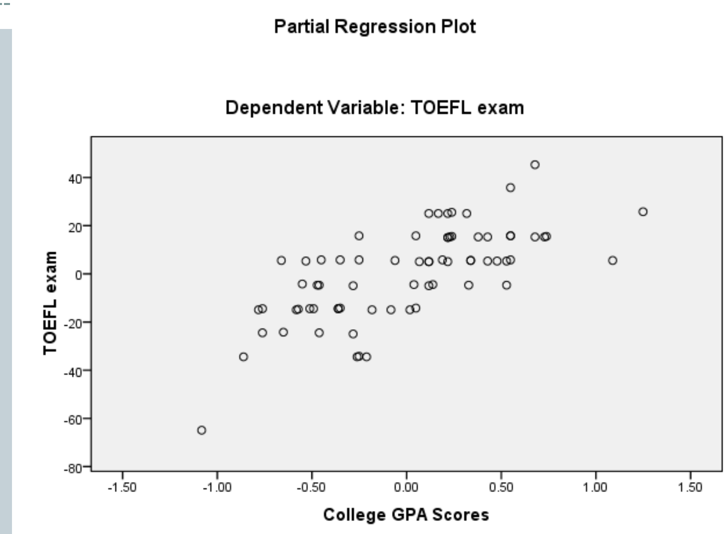
- Outliers can be deleted if necessary and possible

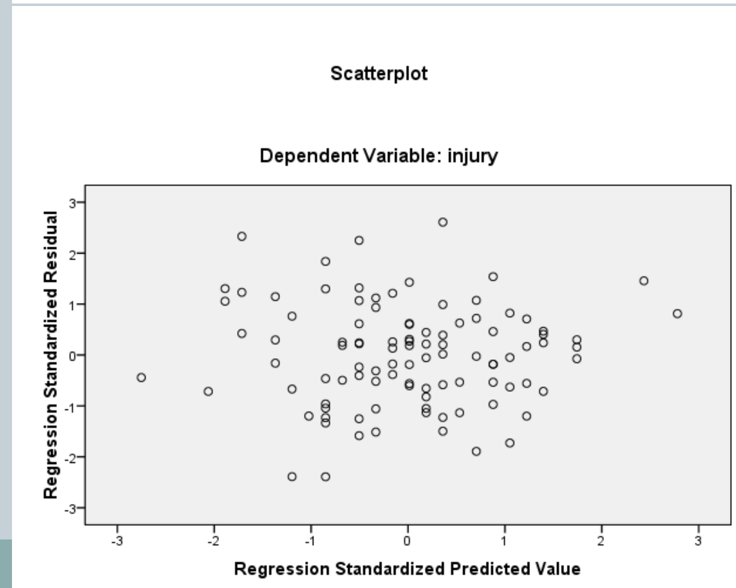
- Outlier can also considered as "missing," but retain the case for other variables

- Alternatively, retain the outlier, but reduce how extreme it is by recoding the value so that it is the highest/or lowest non-outlier value.

Linearity

- A scatterplot is a good means for judging how well a straight line fits the data (see upper scatterplot).
- Another method is to plot the residuals against predicted scores.
- If the assumptions of linearity and homogeneity of variance are met, there should be no relationship between the predicted and residual values.
- If the assumptions are met the residuals would be randomly distributed in a band clustered around the horizontal line through 0 (see lower scatterplot).
- Systematic patterns between the residuals and predicted values suggest possible violations of the assumption of linearity.

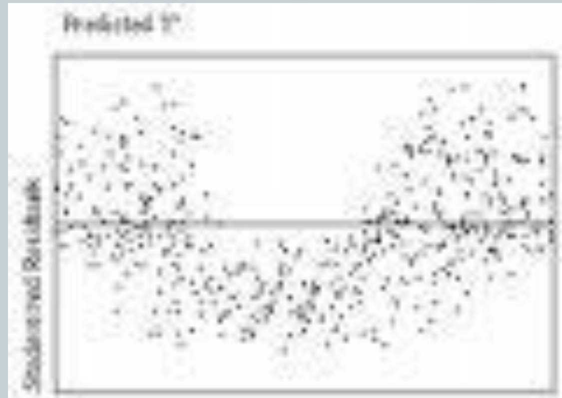


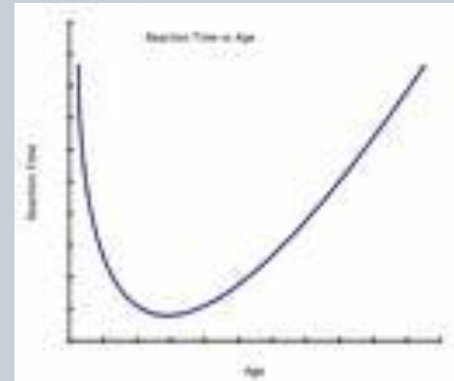


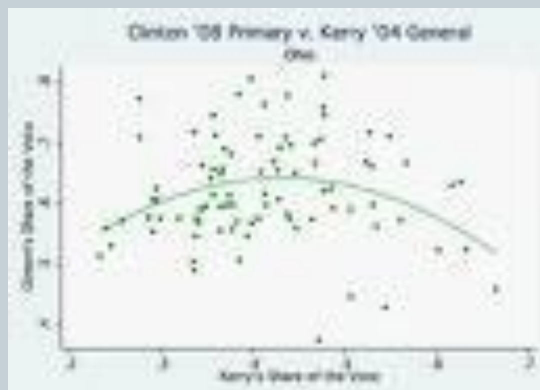
Non Linear Relationships

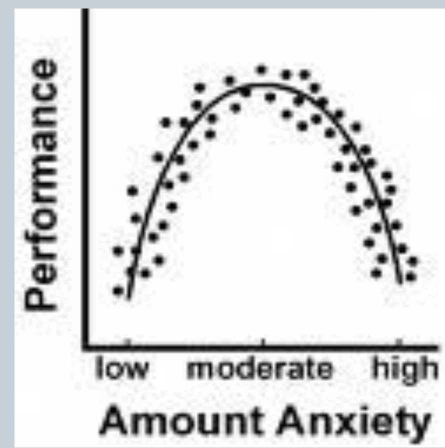
- Scan the plots for relationships other than linear.

- The attached plots indicate non-linear relationships (curvilinear).









Non-Linear relationships

- If the plot indicates that a straight line is a not good summary measure of the relationship, you should consider other methods of analysis including non-linear regression.

- Transforming the data to achieve linearity (co-axing a non-linear relation to linearity) is also used although not preferred.

- To achieve linearity, you can transform either the dependent or the independent variables, or both.
  - Transformation include
    - altering the scale of the variable.
    - Using the square root of Y to diminish curvature.

The Correlation Matrix

- The first step in calculating an equation with several independent variables is to calculate a correlation matrix for all variables.
- The matrix displays correlations between the dependent variable and each independent variable, and correlations between the independent variables.
- Watch for any large inter-correlations between the independent variables as they can substantially affect outcomes of the multiple regression analysis.

Collinearity

- This issue is referred to as collinearity.
- When highly intercorrelated independent variables are included in the equation, results appear anomalous.
- The overall regression may be significant, while none of the individual coefficients are significant.
- The signs of the regression coefficient may be counterintuitive.
- The unique contribution of each independent variable becomes difficult to assess.
- Collinearity inflates the variances of the estimates
- Makes individual coefficients quite unreliable without adding to the overall fit of the model.

Collinearity

- A commonly used measure of collinearity is **Tolerance**. If tolerance of a variable is small, it is almost a linear combination of the other independent variables.

- Another measure is **Variance Inflation Factor (VIF)**. As VIF increases, so does the variance of the regression coefficient.

- **Eigen Values and Condition Indexes** are also used. The smaller the Eigen value and the larger the Condition indexes indicate dependencies among the variables.

Building A Model

- Researchers often try to build a model from available data.
- A variety of regression models can be constructed from the same set of variables.
- For example, 7 different equations can be built with 3 independent variables:
  - 3 with only one independent variable.
  - 3 with 2 independent variables, and
  - 1 with all three.
- As the number of independent variables increases so does the number of potential models.

Building A Model

- Although there are procedures for computing all possible regression models several variable selection methods are frequently used:
  - Forward selection
  - Backward elimination, and
  - Stepwise regression

- Other methods include:
  - Enter regression
  - Remove

Forward Selection

- The first variable considered for entry into the equation is the one **with the largest positive or negative correlation** with the dependent variable.

- This variable is entered into the equation only if it satisfies one of the 2 criteria for entry:
  - FIN or *F*-to-enter: the default value is **3.84**, or
  - PIN or Probability of *F*-to-enter: the default is **.05** (another value may be specified). By default SPSS used the PIN criterion.
    - The SPSS output generally displays *t* values and their probabilities. The *F* values can be obtained by squaring the *t* value (since $t^2 = F$)

- If the first variable is entered, the independent variable not in the equation with the **largest partial correlation in absolute value** is considered next.

- The procedure stops when there are no variables that meet the entry criterion

Backward Elimination

- A variable selection procedure that starts with all variables in the equation and sequentially removes them.

- Instead of entry criteria, 2 removal criteria are used by SPSS:
  - F-to –remove (FOUT): The minimum F value that a variable must have to remain in the equation. The default FOUT is **2.71**.
  - Probability of F- to-remove (POUT): the maximum POUT that a variable can have. The default POUT value is **.10**.

- The variable with **the smallest partial correlation with the dependent variable** is considered first for removal. It will be removed if it meets the criterion for removal.

- After the first variable is removed, the variable remaining in the equation **with the smallest partial correlation** is considered next.

- The procedure stops when there are no variables in the equation that satisfy the removal criteria.

- Forward-Selection and Backward-elimination procedures may generate different results even with similar entry and removal criteria.

Stepwise Regression

- Stepwise regression is really a combination of backward and forward procedures.
- The first variable is selected in the same manner as in forward selection.
- If the variable fails to meet the entry requirement (FIN or PIN), the procedure terminates with no independent variable in the equation.
- However, if it passes the criterion, the second variable is selected based on **the highest partial correlation**. It also enters the equation if it passes entry criteria.
- Variables in the equation are then removed if the their probability of F becomes large (POUT exceeds .10).
- The method terminates when no more variables are eligible for inclusion or removal.

Other Variable Selection Methods

- **Enter regression**: A procedure for variable selection in which all variables in a block are entered in a single step.

- **Remove:** A procedure for variable selection in which all variables in a block are removed in a single step.

The Optimal Number of Independent Variables

- Variable selection and/or model building is seldom a simple process.
- More is not better and may be worse.
- Adding more independent variables increases the R^2 but does not necessarily decrease the standard error of the estimate.
- Each time a variable is added to the equation it may affect the degree of freedom of the residual sum of squares and the F value of the overall regression.

The Optimal Number of Independent Variables

- Including a large number of independent variables in a regression model is never a good strategy, unless there are strong reasons to suggest it.
- The observed increase in R^2 does not reflect a better fit of the model in the population. R^2 never decreases as independent variables are added (adjusted R^2 is needed).
- Including irrelevant variables increases the standard errors of all estimates and does improve prediction.
- Including many variables or **Overfitting** is considered poor.
- Such models usually perform poorly when applied to a new sample drawn from the same population.
- A model with many variables is often difficult to Interpret.

R^2 Change

- One way to assess the relative importance of independent variables is to consider the increases in R^2 when a variable is entered in an equation that already contains the other independent variables.

- A large change in R^2 indicates that a variable provides unique information about the dependent variable that is not available from other independent variables in the equation.

Goodness of Fit

- An important part of multiple regression analysis is to establish how well the model actually fits the data.
- This includes the detection of possible violations of the required test assumptions.
- A commonly used measure of goodness of fit of a linear model is R^2 or the coefficient of determination.
- If all observations fall on the regression line $R^2 = 1$.
- In case of absence of linear relationship $R^2 = 0$.
- The sample R^2 tends to overestimate how well the model fits the population. The adjusted R^2 corrects R^2 to more closely reflect the goodness of fit of the model in the population.

Practice Regression Analyses

- Forward Selection
- Backward Elimination
- Stepwise regression
- Conducting Multiple Regression with **one set of Predictors**.
- Conducting Multiple Regression with **two unordered sets of Predictors**.
- Conducting Multiple Regression with **two ordered sets of predictors**.

Bibliographical References

- Almar, E.C. (2000). Statistical Tricks and traps. Los Angeles, CA: Pyrczak Publishing.
- Bluman, A.G. (2008). Elemtary Statistics (6th Ed.). New York, NY: McGraw Hill.
- Chatterjee, S., Hadi, A., & Price, B. (2000) *Regression analysis by example*. New York: Wiley.
- Cohen, J., & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences (2nd Ed.). Hillsdale, NJ.: Lawrence Erlbaum.
- Darlington, R.B. (1990). Regression and linear models. New York: McGraw-Hill.
- Einspruch, E.L. (2005). An introductory Guide to SPSS for Windows (2nd Ed.). Thousand Oak, CA: Sage Publications.
- Fox, J. (1997) *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage Publications.
- Glassnapp, D. R. (1984). Change scores and regression suppressor conditions. Educational and Psychological Measurement (44), 851-867.
- Glassnapp. D. R., & Poggio, J. (1985). Essentials of Statistical Analysis for the Behavioral Sciences. Columbus, OH: Charles E. Merril Publishing.
- Grimm, L.G., & Yarnold, P.R. (2000). Reading and understanding Multivariate statistics. Washington DC: American Psychological Association.
- Hamilton, L.C. (1992) *Regression with graphics*. Belmont, CA: Wadsworth.
- Hochberg, Y., & Tamhane, A.C. (1987). Multiple Comparisons Procedures. New York: John Wiley.
- Jaeger, R. M. Statistics: A spectator sport (2nd Ed.). Newbury Park, London: Sage Publications.

Bibliographical References

- Keppel, G. (1991). Design and Analysis: A researcher's handbook (3rd Ed.). Englwood Cliffs, NJ: Prentice Hall.
- Maracuilo, L.A., & Serlin, R.C. (1988). Statistical methods for the social and behavioral sciences. New York: Freeman and Company.
- Maxwell, S.E., & Delaney, H.D. (2000). Designing experiments and analyzing data: Amodel comparison perspective. Mahwah, NJ. : Lawrence Erlbaum.
- Norusis, J. M. (1993). SPSS for Windows Base System User's Guide. Release 6.0. Chicago, IL: SPSS Inc.
- Norusis, J. M. (1993). SPSS for Windows Advanced Statistics. Release 6.0. Chicago, IL: SPSS Inc.
- Norusis, J. M. (1994). SPSS Professional Statistics 6.1 . Chicago, IL: SPSS Inc.
- Norusis, J. M. (2006). *SPSS Statistics 15.0 Guide to Data Analysis*. Upper Saddle River, NJ.: Prentice Hall.
- Norusis, J. M. (2008). *SPSS Statistics 17.0 Guide to Data Analysis*. Upper Saddle River, NJ.: Prentice Hall.
- Norusis, J. M. (2008). *SPSS Statistics 17.0 Statistical Procedures Companion*. Upper Saddle River, NJ.: Prentice Hall.
- Norusis, J. M. (2008). *SPSS Statistics 17.0 Advanced Statistical Procedures Companion*. Upper Saddle River, NJ.: Prentice Hall.
- Pedhazur, E.J. (1997). *Multiple regression in behavioral research, third edition*. New York: Harcourt Brace College Publishers.

Bibliographical References

- SPSS Base 7.0 Application Guide (1996). Chicago, IL: SPSS Inc.
- SPSS Base 7.5 For Windows User's Guide (1996). Chicago, IL: SPSS Inc.
- SPSS Base 8.0 Application Guide (1998). Chicago, IL: SPSS Inc.
- SPSS Base 8.0 Syntax Reference Guide (1998). Chicago, IL: SPSS Inc.
- SPSS Base 9.0 User's Guide (1999). Chicago, IL: SPSS Inc.
- SPSS Base 10.0 Application Guide (1999). Chicago, IL: SPSS Inc.
- SPSS Base 10.0 Application Guide (1999). Chicago, IL: SPSS Inc.
- SPSS Interactive graphics (1999). Chicago, IL: SPSS Inc.
- SPSS Regression Models 11.0 (2001). Chicago, IL: SPSS Inc.
- SPSS Advanced Models 11.5 (2002) Chicago, IL: SPSS Inc.
- SPSS Base 11.5 User's Guide (2002). Chicago, IL: SPSS Inc.
- SPSS Base 12.0 User's Guide (2003). Chicago, IL: SPSS Inc.
- SPSS 13.0 Base User's Guide (2004). Chicago, IL: SPSS Inc.
- SPSS Base 14.0 User's Guide (2005). Chicago, IL: SPSS Inc..
- SPSS Base 15.0 User's Guide (2007). Chicago, IL: SPSS Inc.
- SPSS Base 16.0 User's Guide (2007). Chicago, IL: SPSS Inc.
- SPSS Statistics Base 17.0 User's Guide (2007). Chicago, IL: SPSS Inc.