statstutor

community project

encouraging academics to share statistics support resources

All stcp resources are released under a Creative Commons licence

stcp-gilchristsamuels-10

The following resources are associated:

Pearson Correlation worksheet

Simple Linear Regression worksheet

Simple Linear Regression – Additional Information

Research question type: When using one variable to predict or explain another variable in terms of a linear relationship

What kind of variables: Continuous (scale/interval/ratio)

Common Applications: Simple linear regression is the simplest model for predicting the value of one variable in terms of another

Definition

Simple linear regression estimates the coefficients b_0 and b_1 of a linear model which predicts the value of a single dependent variable (y) against a single independent variable (x) in the form:

$$y = b_0 + b_1x$$

b_0 is the intercept of the straight line (the value of y when it crosses the Y-axis) whilst b_1 is its slope.

Confidence Intervals

Obviously this model is subject to uncertainty, as the observed points do not normally all lie on a perfect straight line and are assumed to be a sample from a larger population. Thus the coefficients for the intercept and slope are only estimates of the true value. Confidence intervals can be calculated for these values to give a range of possible values.

The 95% confidence intervals for the example in the Simple Linear Regression worksheet have been plotted on the graph below:

- The bold line is the original model
- The thin solid lines are models using the upper and lower bounds for the constant
- The short dashed lines are models using the lower bound for the slope and the upper and lower bounds for the constant

BY NC SA

www.statstutor.ac.uk

© Mollie Gilchrist and Peter Samuels
Birmingham City University

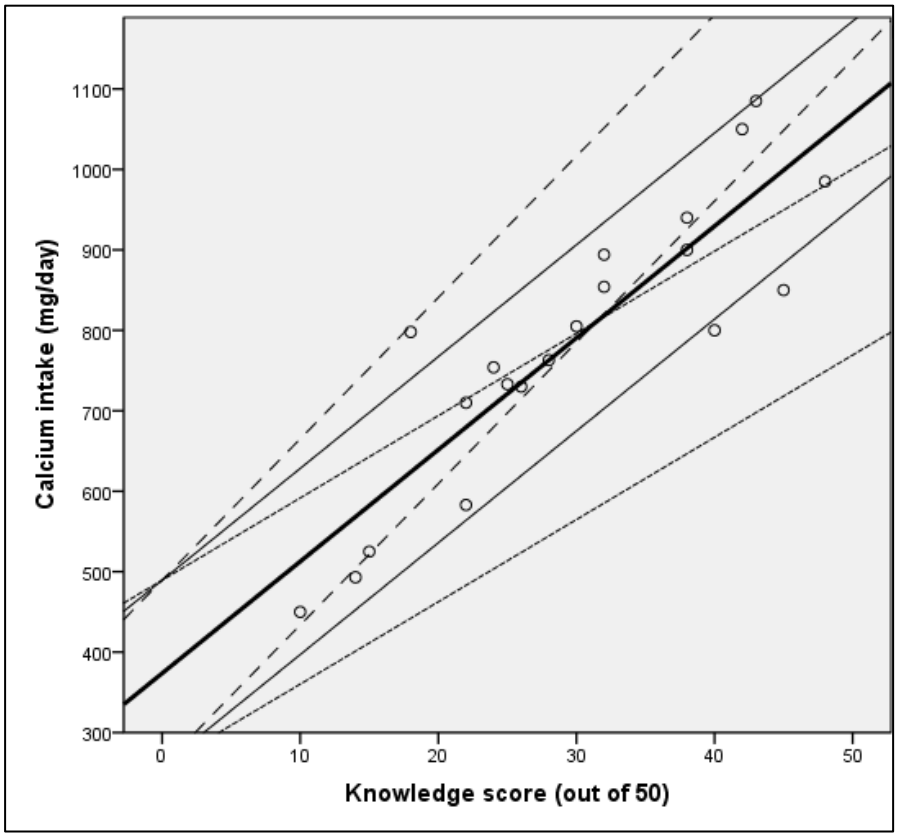
Reviewer: Ellen Marshall
University of Sheffield

Based on material provided by Loughborough University Mathematics Learning Support Centre and Coventry University Mathematics Support Centre

- The long dashed lines are models using the upper bound for the slope and the upper and lower bounds for the constant

These models emphasise the importance of only using the model to predict values within the region of the data.

Confidence interval model lines for the constant and slope



Validity of simple linear regression

Simple linear regression is based on the following assumptions:

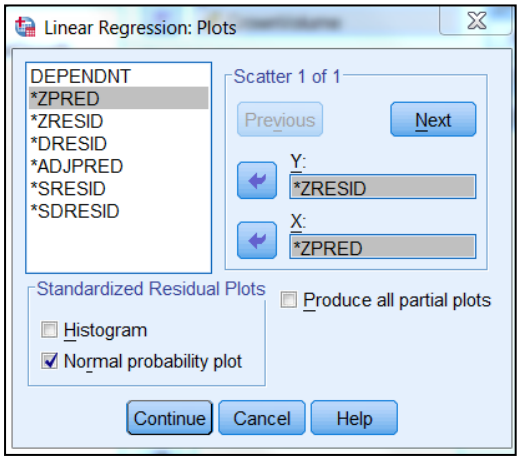
1. Both variables are continuous
2. The observations are random samples from normal distributions. However, according to Kleinbaum et al. (1998, p. 117), "normality is not necessary for the least-squares fitting of the regression model but it is required in general for inference making" (e.g. calculating the p-values and the confidence intervals of b_0 and b_1) "only extreme departures of the distribution of y from normality yield spurious results".
3. The data values are independent of each other, i.e. only one pair of readings per participant is used
4. There is a linear relationship between the two variables and a good theoretical rationale for assuming one variable depends on another in a linear way
5. The relationship between the two variables is **homoscedastic** (i.e. the variance of one variable is the same for all the values of the other variable). Assumptions 3, 4 and 5 can be evaluated simultaneously by looking for an approximate cigar shaped scatter plot (see Simple Linear Regression worksheet).
6. The residuals are normally distributed with constant variance. This can be tested in SPSS by looking at the standardised residuals against the standardised predicted values and a normal probability plot:
    - Select Analyze – Regression – Linear
    - Choose the dependent and independent variable as before
    - Select Plots…
    - Select *ZRESID for the Y variable (standardised residual)

BY NC SA

© Mollie Gilchrist and Peter Samuels
Birmingham City University

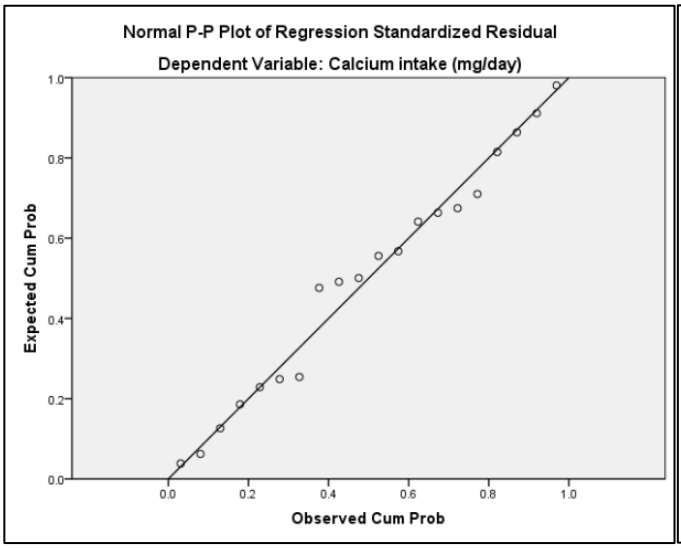
Based on material provided by Loughborough University Mathematics Learning Support Centre and Coventry University Mathematics Support Centre

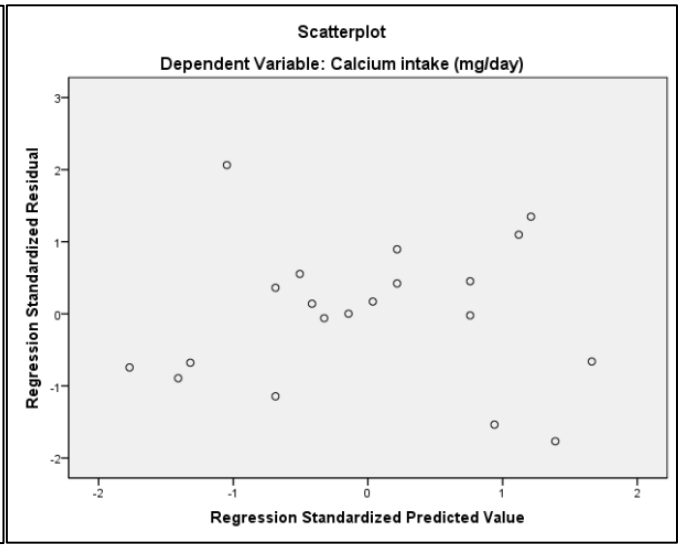
- Select *ZPRED for the X variable (standardised predicted value)
- Select Normal probability plot



The plotted points on the P-P plot should approximately fit the straight line. Any strong systematic curvature suggests some degree of non-normality. This one is fine.

The scatterplot of the standardised residuals against the predicted values should have a random pattern. Any discernible pattern (such as a 'U' shape) indicates a problem. This one is also fine.





Reference

Kleinbaum, D., Kupper L., Muller K. and Nizam, A. (1998) *Applied Regression Analysis and Multivariable Methods*. 3rd ed. Pacific Grove, CA: Duxbury.

BY NC SA

© Mollie Gilchrist and Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Based on material provided by Loughborough University Mathematics Learning Support Centre and Coventry University Mathematics Support Centre