stcp-marshall-regressionS

The following resources are associated:

Scatterplots and correlation, Checking normality in SPSS and the SPSS dataset Birthweight_reduced.sav'

Simple linear regression in SPSS

Dependent variable: Continuous (scale/interval/ratio)

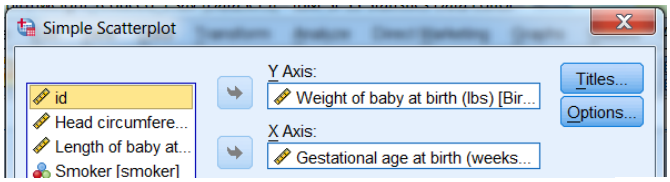
Independent variables: Continuous (scale/interval/ratio)

Common Applications: Regression is used to (a) *look for significant relationships* between two variables or (b) *predict* a value of one variable for a given value of the other.

Data: The data set 'Birthweight_reduced.sav' contains details of 42 babies and their parents at birth. The dependant variable is Birth weight (lbs) and the independent variable is the gestational age of the baby at birth (in weeks).

Before carrying out any analysis, investigate the relationship between the independent and dependent variables by producing a scatterplot and calculating the correlation coefficient.

For a scatterplot: *Graphs* → *Legacy Dialogs* → *Scatter/Dot*, then choose '*Simple Scatter*'.



Move the dependent 'Birth weight' to the *Y Axis* box and the independent 'Gestation' to the *X Axis* box.

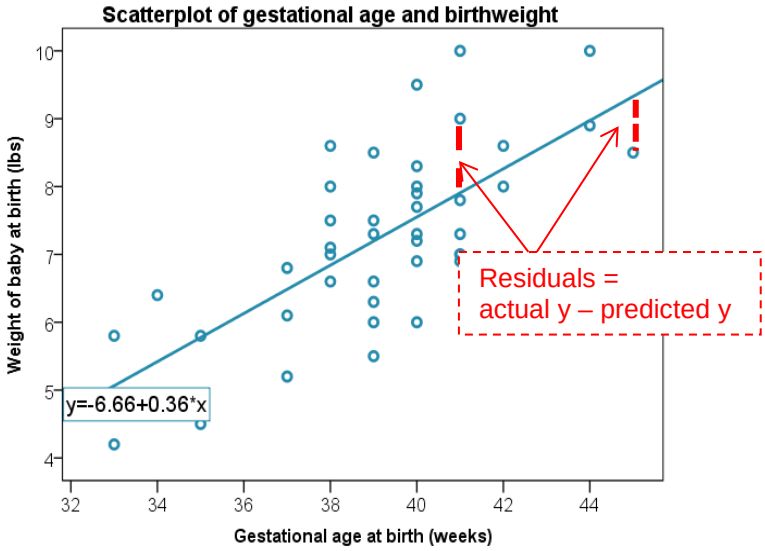
To calculate Pearson's correlation co-efficient use *Analyze* → *Correlate* → *Bivariate* and move both 'Birthweight' and 'Gestation' to the *variables* box.

Both the scatterplot and the Pearson's correlation co-efficient (r) of 0.706 suggest a strong positive linear relationship between gestational age and birthweight.

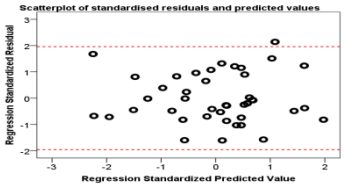
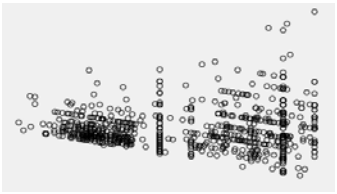
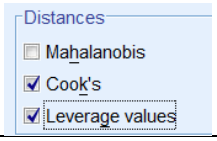
© Ellen Marshall
University of Sheffield

Reviewer: Jean Russell
University of Sheffield

Simple linear regression quantifies the relationship between two variables by producing an equation for a straight line of the form $y = a + \beta x$ which uses the independent variable (x) to predict the dependent variable (y). Regression involves estimating the values of the gradient (β) and intercept (a) of the line that best fits the data . This is defined as the line which minimises the sum of the squared residuals. A **residual** is the difference between an observed dependent value and one predicted from the regression equation.



Assumptions for regression

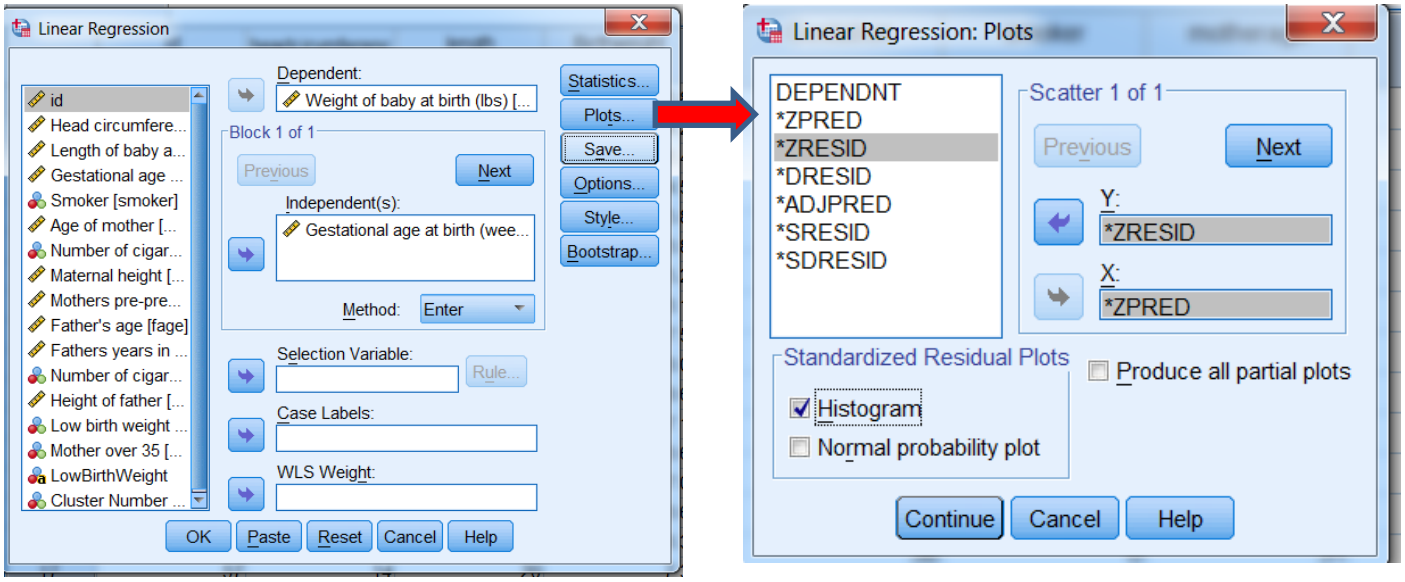
Assumptions	How to check	What to do if the assumption is not met
1) The relationship between the independent and dependent variable is linear	Scatterplot: scatter should form a line in the plot rather than a curve or other shape	Transform either the independent or dependent variable
2) Residuals should be approximately normally distributed	Request the histogram of residuals within the **Plots** menu	Transform the dependent variable
3) Homoscedasticity: Scatterplot of standardised residuals and standardised predicted values shows no pattern (scatter is roughly the same width as y increases) 	This shape is bad since the variation in the residuals (up and down) is not constant (variance is increasing) 	Transform the dependent variable
4) Independent observations (adjacent values are not related). This is only a possible problem if measurements are collected over time	Request the *Durbin Watson statistic* within the **Statistics** menu of regression. It should be between 1.5 – 2.5	If the Durbin-Watson Statistic is outside the range, use Time series (high level statistics)
5) No observations have a large overall influence (leverage). Look at individual Cook's and Leverage values. Interpretation of this is not included on this sheet.	If you wish to check leverage values, request columns with *Cook's and Leverage values* to be added to the dataset via the **Save** menu 	Run the regression with and without the observations and comment on the differences

Note: The **Further regression** resource contains more information on assumptions 4 and 5.

Steps in SPSS

Analyze → Regression → Linear

Move 'Weight of the baby at birth' to the *Dependent* box and 'Gestational age at birth' to the *Independent(s)* box. The plots for checking assumptions are found in the **Plots** menu. The histogram checks the normality of the residuals. There are a few options for the scatterplot of predicted values against residuals. Here the standardised residuals (ZRESID) and standardised predicted values (ZPRED) are used.



Output

The Coefficients table is the most important table. It contains the coefficients for the regression equation and tests of significance.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	**-6.660**	2.212		-3.011	**.004**
	Gestational age at birth (weeks)	**.355**	.056	.706	6.310	**.000**

a. Dependent Variable: Weight of baby at birth (lbs)

The 'B' column in the co-efficients table, gives us the values of the gradient and intercept terms for the regression line.

The model is: **Birth weight (y) = -6.66 + 0.355 *(Gestational age)**

The gradient $(\beta)$is tested for significance. If there is no relationship, the gradient of the line (β) would be 0 and therefore every baby would be predicted to be the same weight. The sig value against Gestational age is less than 0.05 and so there is significant evidence to suggest that the gradient is not 0 ($p < 0.001$).

The key information from the table below is the R^2 value of 0.499. This indicates that 49.9% of the variation in birth weight can be explained by the model containing only gestation. This is quite high so predictions from the regression equation are fairly reliable. It also means that 50.1% of the variation is still unexplained so adding other independent variables could improve the fit of the model.

Model Summary[b]

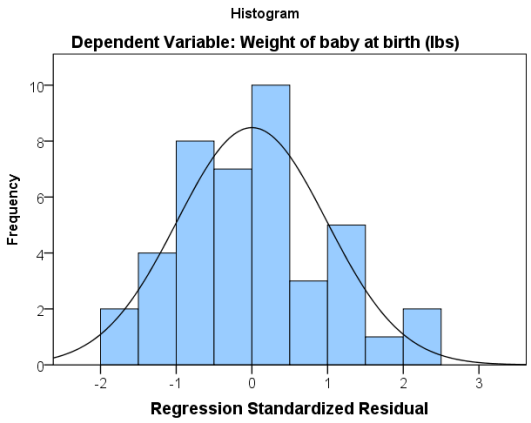
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.706[a]	.499	.486	.9530

a. Predictors: (Constant), Gestational age at birth (weeks)

b. Dependent Variable: Weight of baby at birth (lbs)

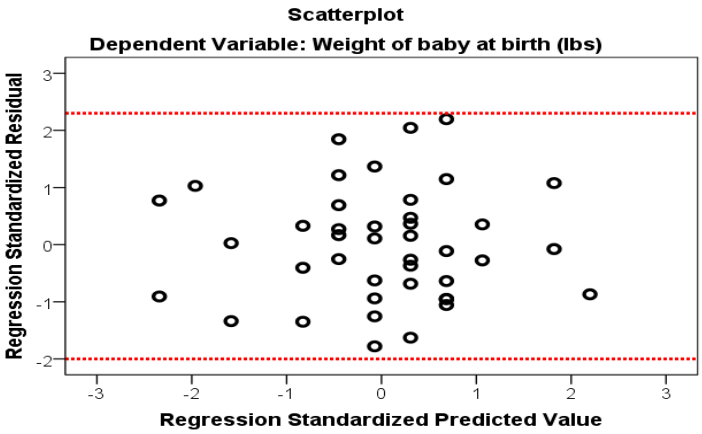
Checking the assumptions for this data

Normality of residuals



The residuals are approximately normally distributed

Homoscedasticity



There is no pattern in the scatter. The width of the scatter as predicted values increase is roughly the same so the assumption has been met.

Reporting regression

Simple linear regression was carried out to investigate the relationship between gestational age at birth (weeks) and birth weight (lbs). The scatterplot showed that there was a strong positive linear relationship between the two, which was confirmed with a Pearson's correlation coefficient of 0.706. Simple linear regression showed a significant relationship between gestation and birth weight ($p < 0.001$). The slope coefficient for gestation was 0.355 so the weight of baby increases by 0.355 lbs for each extra week of gestation. The R^2 value was 0.499 so 49.9% of the variation in birth weight can be explained by the model containing only gestation.
The scatterplot of standardised predicted values verses standardised residuals, showed that the data met the assumptions of homogeneity of variance and linearity and the residuals were approximately normally distributed.