



community project

encouraging academics to share statistics support resources

All stcp resources are released under a Creative Commons licence

stcp-karadimitriou-regressionR

The following resources are associated: Scatterplots, Correlation and Checking normality in R, the Excel dataset 'Birthweight reduced.csv' and the Simple linear regression in R script

Simple linear regression in R

Dependent variable: Continuous (scale)

Independent variables: Continuous (scale)

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.csv' 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). These variables are called 'Birthweight' and 'Gestation'.

Open the birthweight reduced dataset which is saved as a csv file and call it `birthweightR`.

You will need to change the command depending on where you have saved the file.

```
birthweightR<-read.csv("D:\\\\Birthweightreduced.csv",header=T)
```

Tell R to use the birthweight dataset until further notice using `attach(birthweightR)`. This means that 'Gestation' can be used instead of `birthweightR$Gestation`.

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

A scatterplot shows the relationship between two continuous variables.

```
plot(Gestation,Birthweight,main='Scatterplot of gestational age and birthweight',xlab='Gestation (weeks)',ylab='Birthweight(lbs)')
```

Calculating Pearson's correlation coefficient gives a measure of the strength.

```
cor(Gestation,Birthweight)
```

Both the scatterplot and the Pearson's correlation coefficient (r) of 0.706 suggest a strong positive linear relationship between gestational age and birthweight. This means that as gestation increases so does birthweight.

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).

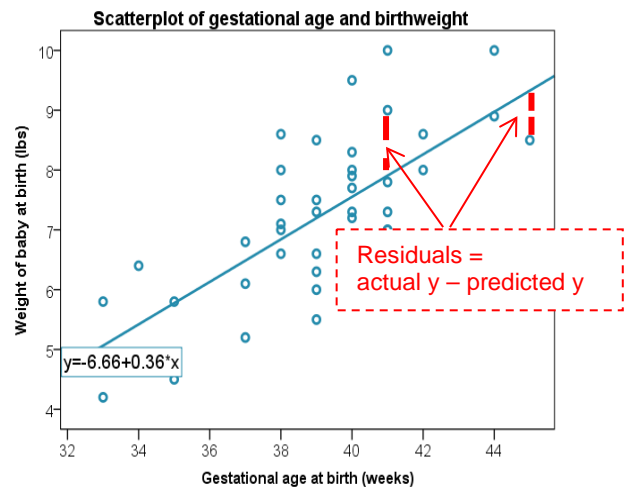
Regression involves estimating the values of the gradient (β) and intercept (a) of the line that best fits the data. This is defined as the line which minimises the sum of the squared residuals.

A **residual** is the difference between an observed dependent value and one predicted from the regression equation.

If you wish to add a regression line to your scatterplot use:

```
abline(lm(Birthweight~Gestation), col='red', lwd=2)
```

For more on this see the **Scatterplots in R** resource.



Steps in R

Fit the regression model using the `lm(dependent~Independent)` command and give it a name (`reg1`). Then request the regression output using `summary(reg1)`.

```
reg1<-lm(Birthweight~Gestation)
```

Output

Call:

```
lm(formula = Birthweight ~ Gestation)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.69661	-0.78430	-0.02426	0.60604	2.09279

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-6.66019	2.21162	-3.011	0.00449 **
Gestation	0.35530	0.05631	6.310	1.73e-07 ***

P-value for gestation
p < 0.001
*** = highly significant

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.953 on 40 degrees of freedom

Multiple R-squared: 0.4988, Adjusted R-squared: 0.4863

F-statistic: 39.82 on 1 and 40 DF, p-value: 1.733e-07

The Coefficients table is the most important table. It contains the coefficients for the regression equation (*Estimate*) and p-values for tests of significance.

The gradient (β) 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 p-value against Gestational age (p < 0.001) is less than 0.05 and so there is significant evidence to suggest that the gradient is not 0 and therefore, gestation is a significant predictor of birthweight.

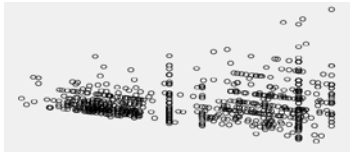
The *Estimate* column in the coefficients table, gives the values of the gradient and intercept terms for the regression line.

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

The gestation coefficient can be interpreted as, with a unit increase of the gestational age, the expected birthweight will increase by 0.355. This means that for each extra week of gestation, a baby weighs an extra 0.355lbs.

Another important value in the output is the multiple R squared 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.

Assumptions for regression

Assumptions	How to check	What to do if the assumption is not met
1) The relationship between the independent and dependent variables 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 from the model	Transform the dependent variable
3) Homoscedasticity: Scatterplot of standardised residuals against 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) 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 the plot of leverage values for the fitted model	Run the regression with and without the observations and comment on the differences
5) Independent observations (adjacent values are not related). This is only a possible problem if measurements are collected over time	Request the <i>Durbin Watson statistic</i> It should be between 1.5 – 2.5	If the Durbin-Watson Statistic is outside the range, use Time series (high level statistics)

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

Checking the assumptions for this data

The most important assumptions to check are the assumptions of normality and homoscedasticity. To check them, first tell R to display 2 plots next to each other.

```
par(mfrow=c(1,2))
```

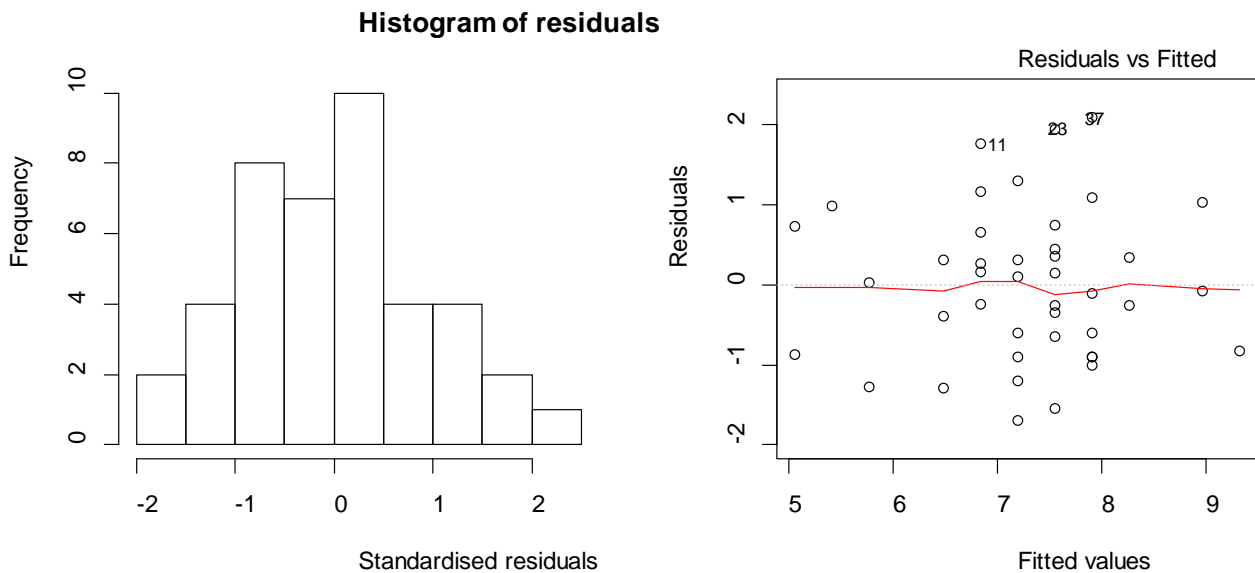
Produce a histogram of standardised residuals to check the assumption of normality.

```
hist(resid(reg1),main='Histogram of residuals',xlab='Standardised Residuals',ylab='Frequency')
```

R produces four diagnostic plots using `plot(reg1)` but we only want the first one.

To produce only the fitted values and residuals plot to check the assumption of homoscedasticity.

```
plot(reg1, which = 1)
```



The residuals are approximately normally distributed so the assumption of normality has been met. We expect 5% of standardised residuals to be outside ± 1.96 but if there are more than this or if there extreme residuals outside ± 3 , they could be influencing the model. There is no pattern in the scatter of the fitted values and residuals. The width of the scatter as predicted values increase is roughly the same so the assumption has been met. The Durbin-Watson statistic and checks for influential data points are discussed on the '**Further regression**' sheet.

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 ($t = 6.31$, $p < 0.001$). 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 showed that 49.9% of the variation in birth weight can be explained by the model containing only gestation. The scatterplot of standardised predicted values versus standardised residuals, showed that the data met the assumptions of homogeneity of variance and linearity and the residuals were approximately normally distributed.