statstutor

community project

encouraging academics to share statistics support resources

stcp-karadimitriou-regressionMultR

The following resources are associated: Dataset 'Birthweight_reduced.csv' and Multiple regression script file.

Scatterplots, Correlation, Simple linear regression and Checking normality in R resources.

Multiple linear regression in R

Dependent variable: Continuous (scale)

Independent variables: Continuous (scale) or binary (e.g. yes/no)

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

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 variables on this sheet are gestational age of the baby at birth (in weeks) and variables relating to the mother (mothers' height and weight as well as whether or not she smokes).

Birthweight	Gestation	smoker	moth		mheight	mppwt
5.80	33	0		0	58	99
4.20	33	1		7	63	109
6.40	34	0	26	0	65	140

Mother smokes = 1

Weight of mother before pregnancy

The **Simple linear regression in R** resource should be read before using this sheet.
Open the birthweight reduced dataset from a csv file and call it birthweightR then attach the data so just the variable name is needed in commands.

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

Tell R that 'smoker' is a factor and attach labels to the categories e.g. 1 is smoker.

```
smoker<-factor(smoker,c(0,1),labels=c('Non-smoker','Smoker'))
```

Assumptions for regression

All the assumptions for simple regression (with one independent variable) also apply for multiple regression with one addition. If two of the independent variables are highly related, this leads to a problem called multicollinearity. This causes problems with the analysis and interpretation. To investigate possible multicollinearity, first look at the correlation coefficients for each pair of continuous (scale) variables. Correlations of 0.8 or above suggest a strong relationship and only one of the two variables is needed in the regression analysis.

© Sofia Maria Karadimitriou and Ellen Marshall
University of Sheffield

Reviewer: Jim Bull
University of Swansea

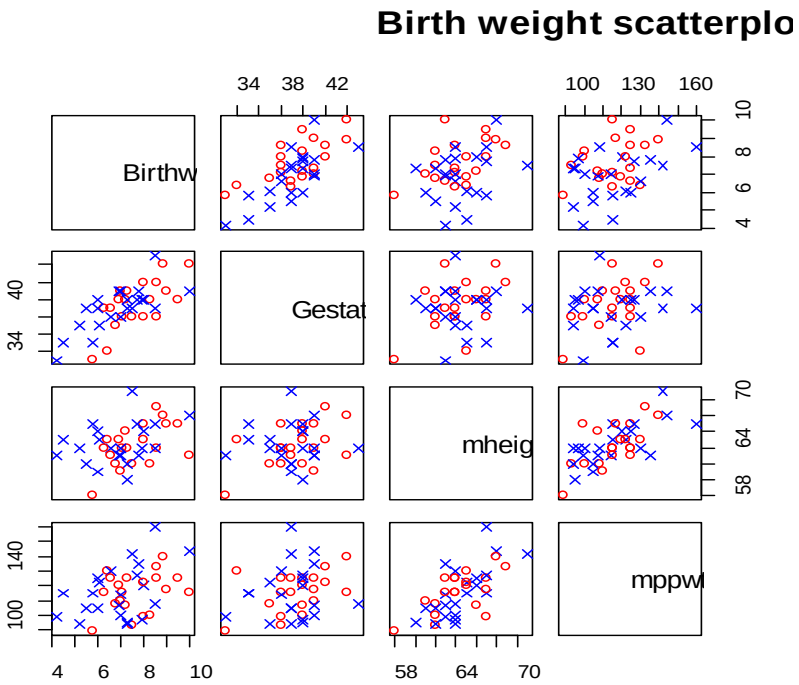
First produce a table of Pearson's correlation coefficients rounded to two decimal places:
`round(cor(cbind(Birthweight,Gestation,mheight,mppwt)),2)`

	Birthweight	Gestation	mppwt	mheight
Birthweight	1.00	0.71	0.39	0.37
Gestation	0.71	1.00	0.25	0.23
mppwt	0.39	0.25	1.00	0.67
mheight	0.37	0.23	0.67	1.00

Gestational age has the strongest relationship with birthweight (r = 0.71) and weight and height are moderately related to birthweight. Maternal weight and height are strongly related to each other (r = 0.69) but this is not above 0.8. R also provides a measure of multicollinearity called the Variance Inflation Factor (VIF) which assesses the relationships between each independent variable and all the other variables.

Scatterplots should be produced for each independent with the dependent so see if the relationship is linear (scatter forms a rough line). Binary variables can be distinguished by different markers on scatterplots which helps to investigate patterns within groups.

To produce multiple scatterplots identifying smokers with red circles:
`pairs(~Birthweight+Gestation+mheight+mppwt,main='Birth weight scatterplots',col=c('red','blue')[smoker],pch=c(1,4)[smoker])`



There are no non-linear patterns between any pair of variables. The relationship between gestational age and birthweight is the strongest but there is also a moderate/ strong relationship between two of the independent variables (weight and height of the mother). The babies of smokers (shown using red circles) tend to be lighter at each gestational age.

Steps in R

Fit the regression model using the `lm(dependent~Independent1+ independent 2)` command and give it a name (reg2). Then request the regression output using `summary()`.

```
reg2<-lm(Birthweight~Gestation+smoker+mppwt)
summary(reg2)
```

Output

The Coefficients table contains the coefficients for the regression equation (model), tests of significance for each variable and R squared value.

```
Call:
lm(formula = Birthweight ~ Gestation + smoker + mppwt)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3824 -0.6246 -0.1033  0.7158  1.9276

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.164740   2.107344  -3.400   0.0016 **
Gestation    0.313421   0.052887   5.926 7.19e-07 ***
smokerSmoker -0.665279   0.267762  -2.485   0.0175 *
mppwt        0.019819   0.008764   2.261   0.0295 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8622 on 38 degrees of freedom
Multiple R-squared:  0.6104,    Adjusted R-squared:  0.5796
F-statistic: 19.84 on 3 and 38 DF,  p-value: 6.635e-08
```

P-value for gestation after controlling for smoking and weight of mother
p < 0.001
*** = highly significant

The last column contains the p-values for each of the independent variables. The hypothesis being tested for each is that the coefficient (B) is 0 after controlling for the other variables. For example, the effects of gestational age and smoking are removed before assessing the relationship between the weight of the mother and the weight of the baby. A p-value < 0.05, provides evidence that the coefficient is different to 0. Gestational age (p < 0.001), smoker (p = 0.017) and mothers' pre-pregnancy weight (p = 0.03) are all significant predictors of birthweight. If the independent value is significant, explain the relationship between the independent and dependent variables using the *Estimate* column.

The *Estimate* column in the coefficients table, gives us the coefficients for each independent variable in the regression model. The model is:
Birthweight (y) = -7.165 + 0.313 *(Gestation) – 0.665*(Smoker) + 0.02*(mppwt)
For gestation, there is a 0.313 lb increase in birthweight for each extra week of gestation. For each extra pound (lb) a mother weighs, the baby's weight increases by 0.02 lbs. A binary variable such as Smoker coded as 0 and 1, the coefficient only applies for the group coded as 1. Here smokers have babies who weigh 0.665 lbs less than non-smokers.

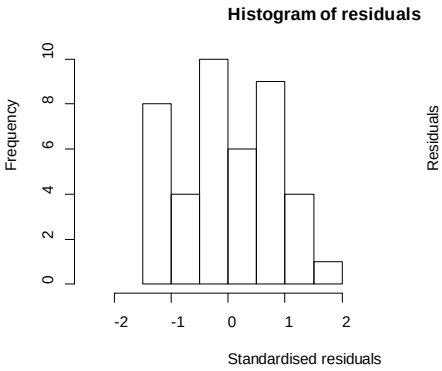
The R^2 value increases with the number of independent variables so it is better to use the adjusted R squared value especially when comparing models. The adjusted R^2 indicates that 57% of the variation in birth weight can be explained by the model containing gestation, smoker and pre-pregnancy weight which is quite high so predictions from the regression equation are fairly reliable.

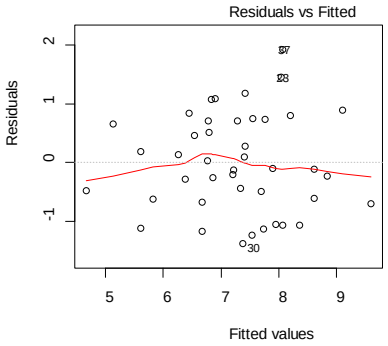
Checking the assumptions
Plots will help check the assumptions of normality and homoscedasticity.
First produce a histogram of standardised residuals to check the assumption of normality.

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

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, run the regression with and without the extreme values to see if the coefficients of the model change much. 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 of homoscedasticity has been met.

Collinearity statistics measure the relationship between multiple independent variables by giving a score for each independent. The "tolerance" is an indication of the percent of variance in an independent that cannot be accounted for by the other independent variables, hence very small values indicate that an independent variable is redundant. The VIF, which stands for *variance inflation factor*, is (1 / tolerance). The VIF scores should be close to 1 but under 5 is fine and 10+ suggests high collinearity so the variable may not be needed. All the values in this analysis have scores close to 1.

To calculate the Variance Inflation Factors the library car must be loaded. `library(car)`
If this command does not work, you will need to select Packages --> Install package(s) then the UK (London) CRAN mirror and choose car from the list. For Rstudio, use Tools → Install packages.
`library(car)`
Calculate the VIF for each variable.
`vif(reg2)`

```
> library(car)
> vif(reg2)
Gestation    smoker     mppwt
 1.077986  1.010494  1.068478
```

The VIF scores should be close to 1 but under 5 is fine and 10+ suggests high collinearity so the variable may not be needed. All the values in this analysis have scores close to 1.

Note: Some institutions are using early versions of `car` where the `vif` command does not work. If this is the case, try installing the `usdm` package instead and `library(usdm)`
You will need to format independent variables as a data frame.
`independents<-data.frame(cbind(Gestation,smoker,mppwt))`
Then request the VIF scores for the independent variables using the command
`vif(independents)`

Reporting regression

Multiple linear regression was carried out to investigate the relationship between gestational age at birth (weeks), mothers' pre-pregnancy weight and whether she smokes and birth weight (lbs). There was a significant relationship between gestation and birth weight ($t = 5.926$, $p < 0.001$), smoking and birth weight ($t = 2.485$, $p = 0.017$) and pre-pregnancy weight and birth weight ($t = 2.261$, $p = 0.03$). For gestation, there was a 0.313 lb increase in birthweight for each extra week of gestation. For each extra pound (lb) a mother weighs, the baby's weight increases by 0.02 lbs and smokers have babies who weigh 0.665 lbs less than non-smokers. The adjusted R^2 value was 0.58 so 58% of the variation in birth weight can be explained by the model containing gestation, pre-pregnancy weight and whether the mother smokes or not. The data met the assumptions of homogeneity of variance and linearity and the residuals were approximately normally distributed.