The University of Sydney
STAT2012 Statistical Tests

Semester 2	Computer Practice Week 10	2015

Useful R commands

- Regression model:

  If $Y_i = \alpha + \beta x_i + \epsilon_i,\ i = 1, \ldots, n$ where $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ can be obtained by

  ```
  lsfit(x,y)$coef
  ```

  where `x` and `y` are `n` × `1` vectors of observations.

- Model checking:

  The fitted line plot, residual plot, and normal qq plot to check the linearity of y on x, the equality of variance and the normality of residual assumptions respectively can be obtained by

  ```
  par(mfrow=c(2,2))
  plot(x, y)
  abline(lsfit(x,y))
  title("Fitted line plot")
  plot(x,lsfit(x,y)$res)
  abline(h=0)
  title("Residual plot")
  qqnorm(lsfit(x,y)$res)
  qqline(lsfit(x,y)$res)
  ```

Important points

- You will perform the regression analysis to build simple linear regression models to independent and dependent variables.
- You will test the sensitivity of the regression model by inserting an outlier and refit the model.
- You will check the linearity of the model (between dependent and independent variables) using the scatter plots of all pairs of variables.

Practice Problems

1. The data `fuel` contain information on makes of cars taken from the April 1990 issue of Consumer Reports. Open the data set `fuel`.

```
fuel=read.csv("http://www.maths.usyd.edu.au/u/UG/IM/STAT2012/r/fuel.csv")
attach(fuel)
fuel
library(MASS)
```

Note that we add `library(MASS)` to do the pairwise plots of all variables in the data.

(a) Create two vectors x and y which correspond to `Weight` and `Fuel` respectively.

```
x=Weight
y=Fuel
```

(b) Find the least squares regression line for `Fuel` on `Weight`. State the intercept and slope estimates and the regression model.

2. Check the model assumptions using the following graphs:

   (a) The *fitted line plot* of `Fuel` against `Weight` and comment on whether the *linearity* assumption between y and x is suitable.

   (b) The *residual plot* against the `Weight` and comment on whether the *equality of variance* assumption is satisfied by checking if there is a systematic pattern on the residuals.

   (c) The *normal qq-plot* of the residuals and comment on whether the *normality* assumption on residuals is satisfied.

3. Check the slope and y-intercept estimates in (a).

4. Suppose the last observation is entered incorrectly as `y1[60]=52.63158`.

```
y1=y
y1[60]=52.63158
lsfit(x,y1)$coeff
```

   (a) Refit the regression model and plot the two regression lines into a *fitted line plot* to compare the two regression models. Include also the residuals plot of the new regression model.

   (b) Comment on the effect of one observation on

      1. the increase or drop for the new slope,
      2. the improvement or worsening of the model fit and
      3. the more random or systematic pattern of residuals.

   Hence comment on whether regression model is sensitive to outliers.

5. Provide the scatter plots between all pairs of variables in the `fuel` data using `plot(fuel)`. Note that for each scatter plot in the lower triangle say, the two variables can be identified by looking vertically up and horizontally across to the names on the diagonal boxes.

   Comment on whether linear regression models should be built between the response variable `Fuel` and the predictors `Weight`, `Disp.` and `Mileage`.