Summary of Issues in ANOVA, Regression and General Linear Modelling

1. **Model Assumptions :** The key model assumption is that the residual (measurement) errors are independent and identically distributed Normal random quantities. If this assumption is not met, then none of the hypothesis tests based on the Student and Fisher-F distributions are valid.

   The validity of this model assumption can be checked by the inspection of the *residuals*, $\widehat{e}_i$, or *standardized residuals*, $\widehat{z}_i$ where

   $$\widehat{e}_i = y_i - \widehat{y}_i \qquad\qquad \widehat{z}_i = \frac{\widehat{e}_i}{s} = \frac{y_i - \widehat{y}_i}{s}$$

   for $i = 1, \dots, n$. Plots of the residuals can be used to check for

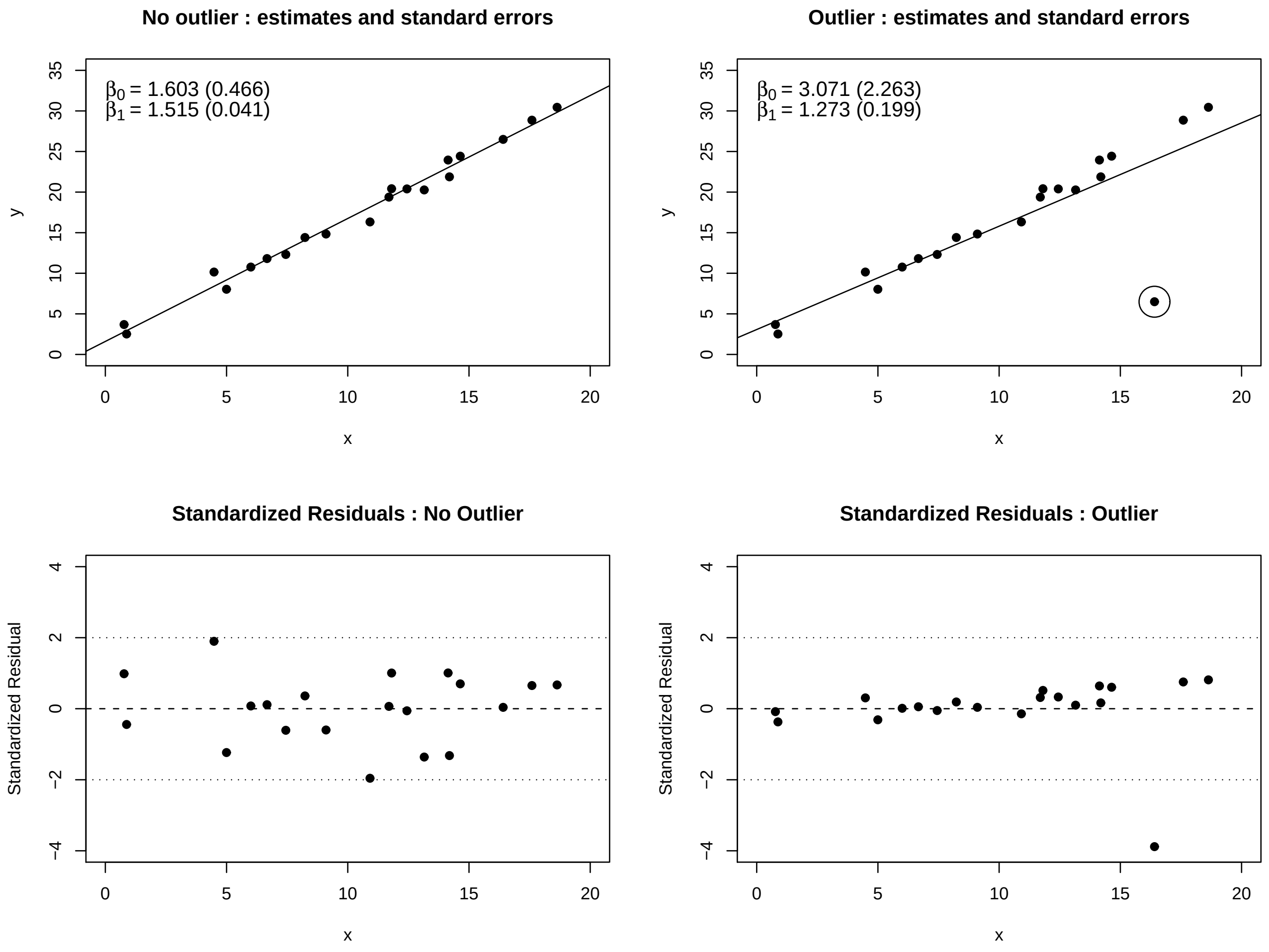
   (i) *Normality*

   (ii) *Dependence* on the covariates

   (iii) *Constant variance*

   (iv) *Outliers*: An *outlier* is an response value that gives rise to a residual which is large in magnitude, indicating that the fit of the model is poor for that data point.

   Outliers can significantly alter the fit of a model, and the parameter estimates. If an outlier is suspected, then careful consideration should be given to omitting that data point from the analysis (see below; estimates (standard errors) change in the presence of an outlier).



No outlier : estimates and standard errors

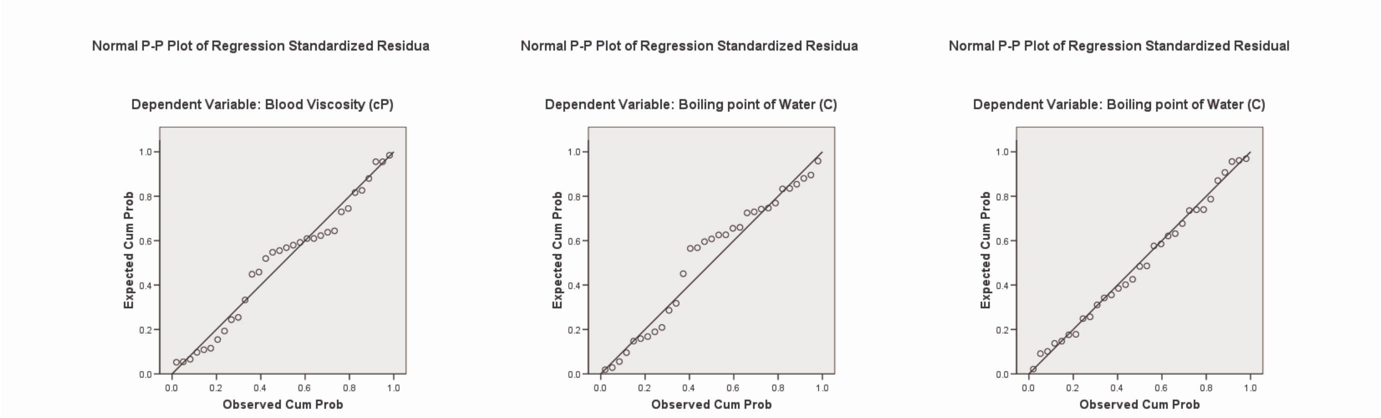
Outlier : estimates and standard errors

Standardized Residuals : No Outlier

Standardized Residuals : Outlier

To check the normality of the residuals, a *histogram* or *probability plot* can be used. A probability plot is a plot constructed using the **observed** (standardized) residuals and their **theoretical** counterparts **assuming a normal model**. Points in such a plot should lie on a straight-line with slope one; any deviation from this may indicate deviation from normality. These plots are available on the *Linear Regression* menu, after clicking the *Plots* button.

The examples below are from (a)the Viscosity vs PCV data, and (b) the straight-line and (c) the quadratic model analysis of the Hooker data. In the (a), the probability plot indicates that the residual variance is larger in the middle compared to the of the residual range. In (b), the points in the probability plot do not lie on the straight-line, so again a deviation from normality is indicated. In (c), the plot indicates normality of the residuals.



(a) Viscosity vs PCV

(b) Hooker Data: Linear Model

(c) Hooker Data: Quadratic Model

2. **Data Transformations :** The response variable and continuous covariates can be **transformed** (using log or square-root transformation say) to improve the fit of the model, or to make the model assumptions more appropriate.

3. **Model Selection :** Model selection by means of stepwise selection and sequential ANOVA-F testing can be an effective way of finding the important explanatory variables and interactions. However it must be carried out with care.

   In general, we aim to select the **simplest model** that provides an adequate fit to the data.

   The goodness of fit measures R^2 and adjusted R^2 statistics can provide a final assessment of model adequacy.

4. **Multicollinearity :** *Multicollinearity* is the term to describe dependence between the covariates used in a regression model. If the covariates are highly correlated, then the estimated coefficients for those covariates in a multiple regression need careful interpretation.

   If two covariates are highly correlated, then if one is a useful predictor of the response, the other will likely appear to be a useful predictor as well, that is if one estimated coefficient is significantly different from zero, then the other will be also. However, in a multiple regression model with both covariates included, it might be that neither coefficient is significantly different from zero.

5. **Predicting outside the Range of the Covariates :** In a regression model, the fitted parameters reflect relationships and dependencies in the *observed* data. The model can be used for prediction, but is only likely to be reliable if the prediction is carried out at x values within the range of the observed xs.

   For example, in a simple linear regression, if x takes values on the range $(0, 100)$, predictions at new x values within this range will be reliable, but predictions at, say, $x = 200$ will be much less reliable.