Residual Analysis

Hong Kong Baptist University

Fall 2021

Residual Analysis

The validity of many of the inferences associated with a regression analysis depends on the assumption that the error terms ε_i satisfy

$$
\begin{aligned}
&\text{(i)} \quad \mathbb{E}(\varepsilon_i) = 0, \\
&\text{(ii)} \quad \mathbf{var}(\varepsilon_i) = \sigma^2 \quad \text{(a constant)}, \\
&\text{(iii)} \quad \mathbf{corr}(\varepsilon_i, \varepsilon_j) = 0, \quad \text{whenever } i \neq j, \text{ and} \\
&\text{(iv)} \quad \varepsilon_i \sim \text{Normal}.
\end{aligned}
$$

The first three conditions, known as the *Gauss–Markov condition* (or Gauss–Markov assumption), while the last one is the normality assumption. These four conditions can be expressed simply by

$$\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}).$$

We have the following famous result:
Gauss–Markov Theorem: *Under the Gauss–Markov condition, the least squares estimators are best linear unbiased estimators (BLUEs), where "best" implies minimum variance.*

When the normality condition (iv) also holds, then the least squares estimators not only are BLUEs but also are uniformly minimum-variance unbiased (UMVU) estimators.

Residual Plots

Note that in this course we always assume the normal distribution, which allows us to use t-test and F-test.

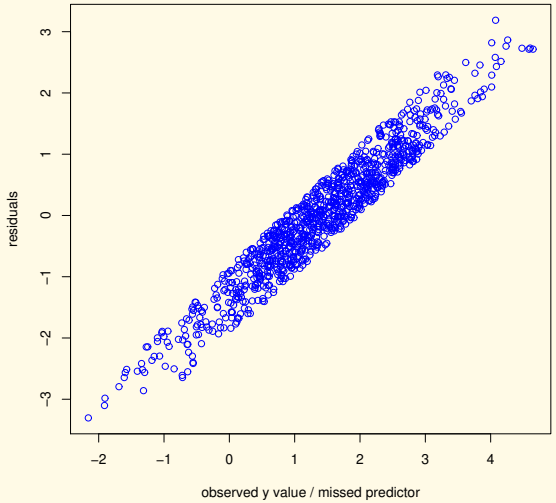
Thus, it is important for us to check whether the error terms really satisfy the Gauss–Markov condition and the normality assumption.

The exact values of ε, however, cannot be observed or calculated. Instead, we estimate them by residuals:

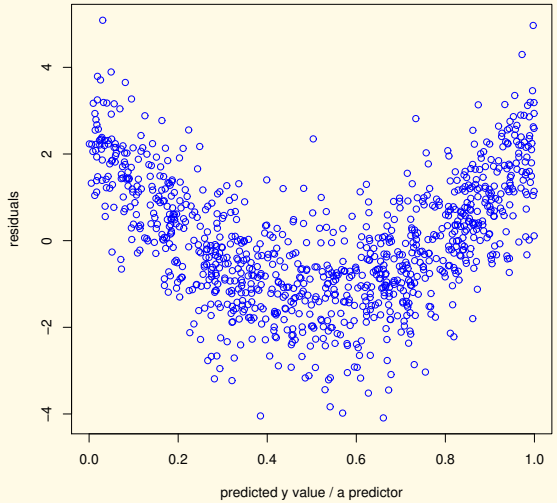
$$e_i = Y_i - \hat{Y}_i.$$

The residuals e_i can be plotted against each of the independent variables x_j ($j = 1, \ldots, k$), against the observed response Y_i and against the predicted value $\hat{Y}_i$, for $i = 1, \ldots, n$.

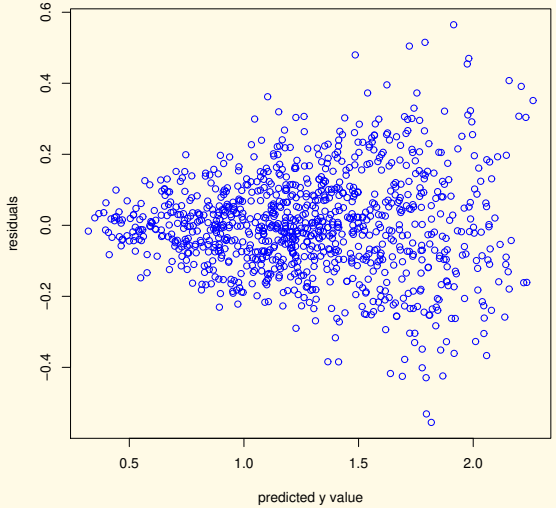
Such a plot, called a **residual plot**, should have no trends, no dramatic increases or decreases in variability and only a few residuals (about 5%) more than two standard deviations above or below zero.



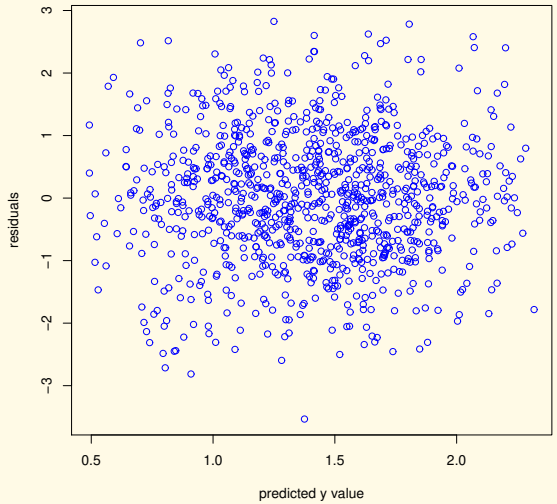
A linear pattern, indicating a wrong model; a linear term is missing.



A quadratic pattern, indicating a wrong model; a quadratic term is missing.



An outward-opening funnel pattern, indicating unequal variance.



No pattern, indicating no problem.

Partial Regression Plots

In addition to residual plots, we can also consider the marginal role of an independent variable x_j, given that the other independent variables under consideration are already in the model.

In such a plot, both Y and x_j are regressed against the other independent variables and the residuals are obtained for each and are plotted as **partial residuals** of Y versus x_j or versus residuals of x_j. The so-called partial residuals of Y are

$$e_{i[j]} = Y_i - \hat{Y}_{i[j]},$$

where $\hat{Y}_{i[j]}$ is the estimate of Y_i obtained by regressing Y against all other independent variables except x_j.

Similarly, the residuals of x_j are the residuals in the model where x_j is regressed against all the other independent variables.

Thus, the variation in the partial residuals $e_{i[j]}$ has two components: (i) the variation in x_j, which can be explained if we add x_j back to the model, and (ii) the variation in the random error ε, which is also present in the full model.

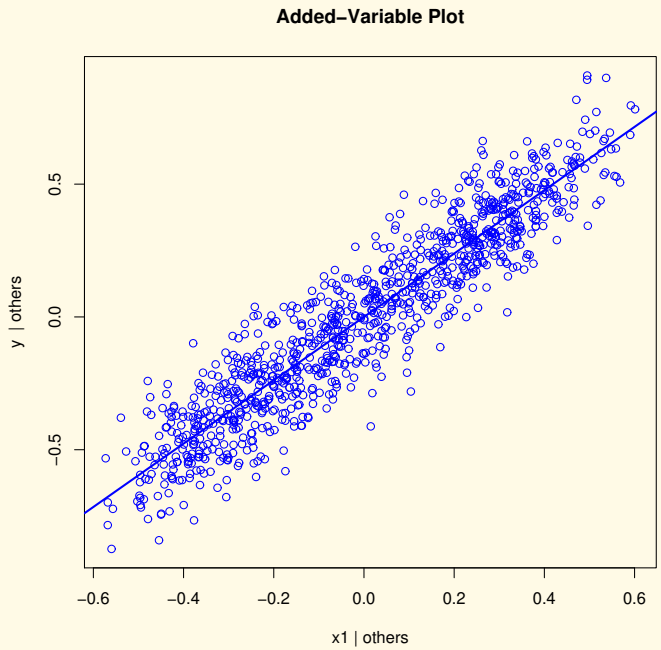
The variation in the residuals of x_j is the variation in x_j that cannot be explained by the variation in other independent variables.

This **partial regression plot**, also known as **added-variable plot**, is plotting $e_{i[j]}$ against the residuals of x_j, so that the effects of all other independent variables except x_j are filtered out.

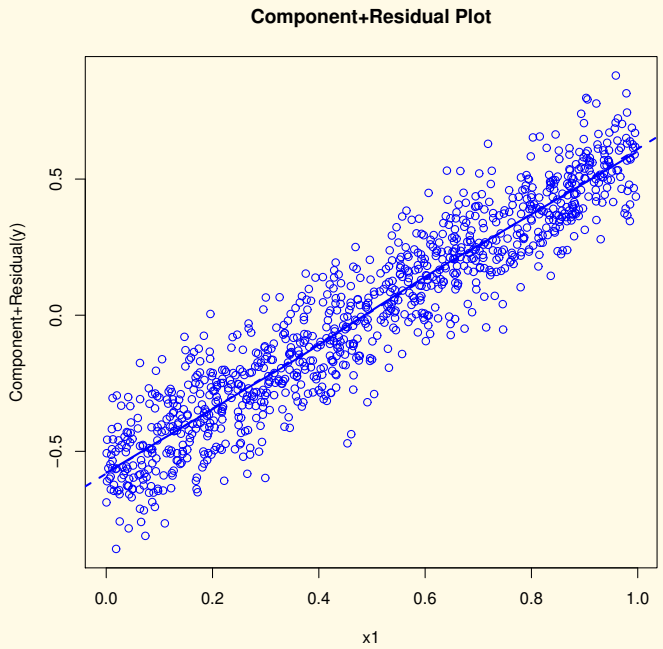
This plot shows the marginal importance of the variable x_j in reducing the residual variability and may provide information about the marginal regression relation for x_j for possible inclusion in the model.

Partial regression plots are related to, but distinct from, **partial residual plots**, also known as **component plus residual plots**, which plot partial residuals $e_{i[j]}$ against the independent variable x_j (the variable itself, not the residuals). However, the partial residual plots are not available in SAS.

Figure 2 shows an example that the two partial plots are not the same; in this example the full model is $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ and x_1 and x_2 are correlated, and we can see that the two plots are very similar but not identical because x_1, when is regressed against the correlated x_2, gives residuals that are different from x_1 itself. If the data of x_1 and of x_2 formed two orthogonal vectors, then the two plots would be identical.



(a)



(b)

Figure: Partial plots: (a) partial regression plot and (b) partial residual plot

If we think carefully, we can see two features of the partial regression plot:

(i) Fitting by the least squares the partial residuals to a straight line passing through the origin will give a regression line of slope exactly $\hat{\beta}_j$ (i.e. the LSE of the coefficient of x_j when you fit the data to the original full model), and the residuals from this fitting are identical to the residuals from the least squares fit of the original full model (Y against all the independent variables including x_j).

(ii) In the partial regression plot, the value of each partial residual is the distance from it to the horizontal line through the origin; the sum of all squared partial residuals is the $SSE_{[j]}$ of the model in which all independent variables except x_j are included, and the vertical distance from each partial residual to the line through the origin with slope $\hat{\beta}_j$ is the residual of the original full model and hence the sum of the squares of such distances is the SSE_{full} of the original full model. The difference of these two SSEs is the extra sum of squares, which can be used to form the partial F-test to test the significance of the x_j-term.
Thus, if the scatter of the points around the line through the origin with slope $\hat{\beta}_j$ is much less than the scatter of the points around the horizontal line through the origin, inclusion of the variable x_j in the model will provide a substantially further reduction in the error sum of squares.

However, these two features are not true for partial residual plots. Then, what is the purpose of partial residual plots? They are most commonly used to identify the nature of the relationship between Y and x_j (given the effect of the other independent variables in the model). This job is better done by partial residual plots, rather than by partial regression plots, because for the latter, the x-axis is not x_j and so is less helpful than the former in determining the nature of the relationship between Y and x_j.

A partial regression plot only suggests if there is a relationship but does not provide an analytic expression of the relationship. Such a plot may not show the proper form of the marginal effect of an independent variable if the functional relations for some independent variables already in the model are misspecified.

Hence, a variety of transformations of the independent variable or curvature effect terms may need to be investigated and additional residual plots utilised to identify the best transformation or curvature effect terms.

Variance-Stabilising Transformation

If the residual plots or partial plots suggest that there are missing terms, we may add extra terms to the model. How about if the plots suggest violation of other assumptions? Consider the equal variance, known as **homoscedasticity**, assumption. If we consider e.g. Y has a Poisson distribution or a binomial distribution, or if we have a multiplicative model $Y = [\mathbb{E}(Y)]\varepsilon$, then we have unequal variances, i.e. *heteroscedasticity*. If we know how the variance depends on the mean, it is possible to stabilise the variance by some **variance-stabilising transformation**:

Variance σ_i^2 proportional to	Transformation Y^*
constant	$Y^* = Y$
$\mathbb{E}(Y_i)$	$Y^* = \sqrt{Y}$
$\mathbb{E}(Y_i)\{1 - \mathbb{E}(Y_i)\}$	$Y^* = \sin^{-1}\sqrt{Y}$
$\{\mathbb{E}(Y_i)\}^2$	$Y^* = \log Y$
$\{\mathbb{E}(Y_i)\}^3$	$Y^* = 1/\sqrt{Y}$
$\{\mathbb{E}(Y_i)\}^4$	$Y^* = 1/Y$

Denote by f the required transformation so that $Y^* = f(Y)$ and by μ the mean of Y. The above table results from the first order Taylor's approximation:

$$f(Y) \approx f(\mu) + (Y - \mu) f'(\mu),$$

which provides an approximation to the variance of the transformed Y:

$$\mathrm{var}(f(Y)) \approx \mathbb{E}\{[f(Y) - f(\mu)]^2\} \approx \mathbb{E}\{(Y - \mu)^2\} f'(\mu)^2 = \sigma^2(\mu) f'(\mu)^2.$$

(The second expression above is not exactly the variance of $f(Y)$ because $\mathbb{E}[f(Y)] \neq f(\mathbb{E}[Y])$ in general.) For example, if $\sigma^2(\mu)$ is proportion to μ, then taking $f(Y) = \sqrt{Y}$ will make the last expression a constant (i.e. a term without μ).

Box–Cox Transformation

However, in many instances transformations are, or have to be, selected empirically, and so more formal, objective techniques should be applied to help specify an appropriate transformation. This can be done by the celebrated **Box–Cox transformation**:

$$Y_i^{(\lambda)} = \begin{cases} \dfrac{Y_i^{\lambda} - 1}{\lambda}, & \lambda \neq 0, \\ \log Y_i, & \lambda = 0. \end{cases}$$

The purpose of Box–Cox transformation is to have a data-driven way to correct deviations from homoscedasticity, normality and linearity.

A procedure to find the best λ is to consider the maximum likelihood estimation, which corresponds to the λ-value for which the $SSE(\lambda)$ is a minimum.

However, we may not want to use the best λ. Instead, we may prefer simple choices for λ. For example, the practical difference in $\lambda = 0.5$ and $\lambda = 0.512$ is likely to be small but $\lambda = 0.5$ is much easier to interpret.

Once a value of λ is selected, we are now free to fit the model using simply Y_i^{λ} or $\log Y_i$, rather than $Y_i^{(\lambda)}$.

SAS procedure `transreg` will report a plot log-likelihood against λ, see the figure below, from which we can see that the best value is $\lambda = 1.25$, but the 95% confidence interval is $(0, 2.75)$, and so it is more convenient to use $\lambda = 1$.

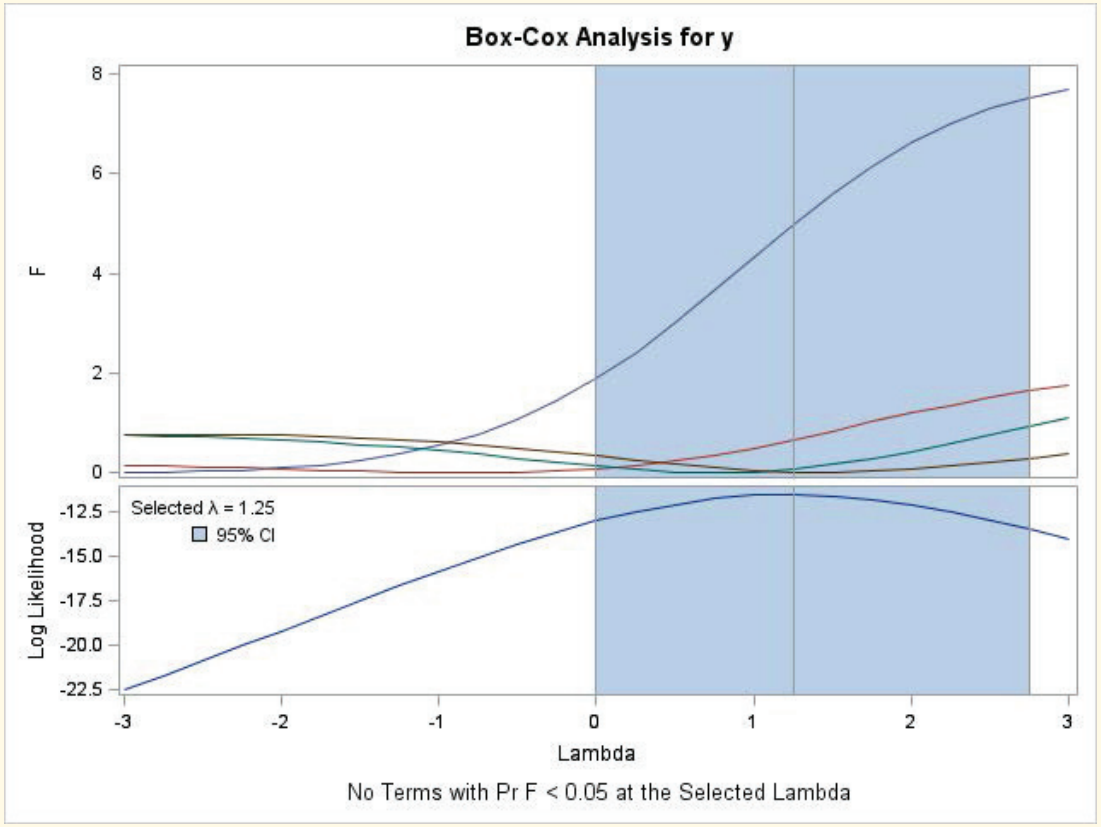


Figure: Box–Cox analysis for the Hald data with all four independent variables

However, there is a technical issue in the transformation formula, namely, as λ varies, the size of $Y_i^{(\lambda)}$ can change enormously so that the SSE for models of $Y_i^{(\lambda)}$ with different values of λ are in fact measured on different scales and so are not comparable, leading to a minor problem in data analysis.

A more preferable transformation is to include a scale factor (in fact, the Jacobian of the transformation) related to the geometric mean

$$\dot{Y} = \sqrt[n]{Y_1 Y_2 \cdots Y_n}$$

such that

$$Y_i^{(\lambda)} = \begin{cases} \dfrac{Y_i^{\lambda} - 1}{\lambda \dot{Y}^{\lambda-1}}, & \lambda \neq 0, \\ \\ \dot{Y} \log Y_i, & \lambda = 0, \end{cases}$$

so that SSE for models with different λ are comparable. If we do not include the factor related to the geometric mean $\dot{Y}$, then $SSE(\lambda)$ for $Y_i^{(\lambda)}$ are measured on different scales and so are not comparable.

Detecting Outliers, Leverage Points, and Influential Observations

The Box–Cox transformation aims at correcting not only deviations from homoscedasticity but also deviations from normality, which we will also check. But before we talk about normality, let us discuss how to detect **outliers**, **leverage points**, and **influential observations**. Outliers are data points that contain unusual response Y values. Usually, if the i^{th} observation has a residual e_i that is larger than two standard deviations (in absolute value), it is considered to be an outlier. Thus, it is natural to consider the **standardised residual**

$$z_i = \frac{e_i}{\hat{\sigma}}$$

so that whenever the standardised residual is greater than 2 (some authors use 3 instead of 2, but here we use 2), it is an outlier.

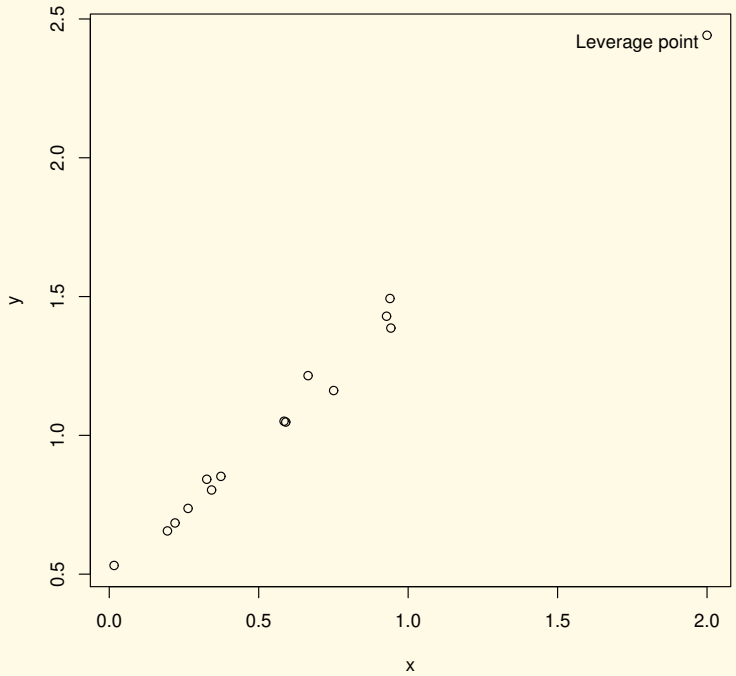
Outlier cannot always be explained by data entry or recording errors. Extremely large or small residuals may be attributable to skewness (nonnormality) of the distribution of ε, chance or unassignable causes.

Although some analysts advocate elimination of outliers, regardless of whether cause can be assigned, others encourage the correction of only those outliers that can be traced to specific causes. Anyway, before deciding the fate of an outlier, you may want to determine how much influence it has on the regression analysis.

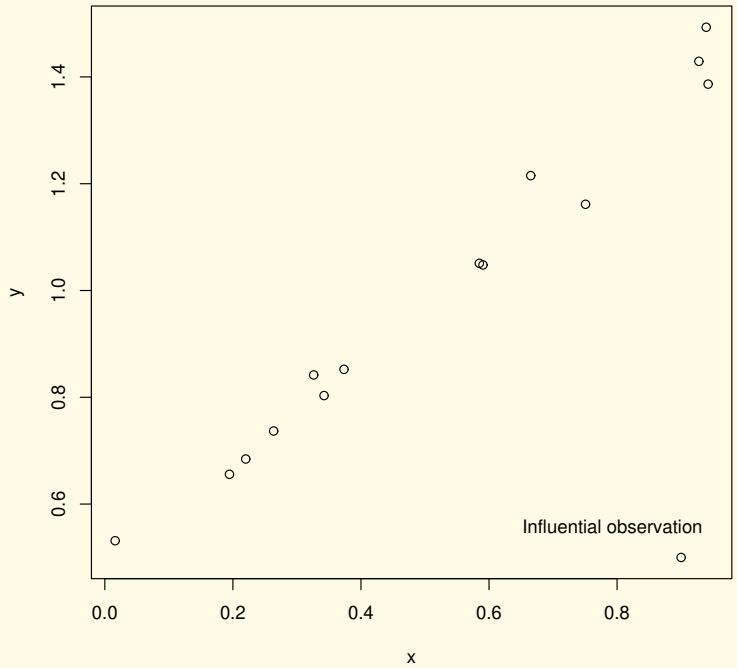
When an accurate outlier (i.e. an outlier that is not due to recording or measurement error) is found to have a dramatic effect on the regression analysis, it may be the model and not the outlier that is to be suspected.

Omission of important independent variables or high order terms could be the reason why the model is not predicting well for the outlying observation.

A data point with an unusual x-value is called a leverage point, and a leverage point is likely, but not necessarily, influential, whilst an influential observation is likely, but not necessarily, a leverage point.



(a) A leverage point that is not an outlier or influential obs.



(b) An influential observation that is an outlier but not a leverage point

Figure: Examples of leverage points and influential observations

Now, to make detection of leverage points more quantitative, we introduce the notion of *hat matrix*, which is denoted by $\boldsymbol{H}$ and defined by

$$\boldsymbol{H} := \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

so that

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y},$$

that is to say, the hat matrix converts $\boldsymbol{Y}$ to $\hat{\boldsymbol{Y}}$.
Consider the covariance matrix of the residuals. Noting that $\boldsymbol{H}' = \boldsymbol{H}$ (symmetric) and $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$ (idempotent), we obtain

$$\begin{aligned} h_{ii} &= \sum_j h_{ij}^2 = h_{ii}^2 + \sum_{j\neq i} h_{ij}^2, \qquad (1) \\ \mathrm{cov}(\boldsymbol{e}) &= \mathrm{cov}(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \mathrm{cov}(\{\boldsymbol{I} - \boldsymbol{H}\}\boldsymbol{Y}) \\ &= (\boldsymbol{I} - \boldsymbol{H})\mathrm{cov}(\boldsymbol{Y})(\boldsymbol{I} - \boldsymbol{H})' = \sigma^2(\boldsymbol{I} - \boldsymbol{H}), \qquad (2) \end{aligned}$$

where h_{ij} is the i-j element of $\boldsymbol{H}$. The most important is of course h_{ii}, which is the i^{th} diagonal element of $\boldsymbol{H}$.

Equations (1) and (2) imply that

$$0 \leq h_{ii} \leq 1,$$

and equation (2) means that the variance of e_i is not σ^2 but $\sigma^2(1 - h_{ii})$. This is understandable because the parameters of the fitted model are estimated by LSE, i.e. by minimising $\sum e_i^2$, and so if the model is correct, the variance of e_i will be smaller than the true variance of ε_i. Therefore, it is more appropriate to normalise the residual to

$$z_i^* = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}},$$

which is called the (**internally**) **studentised residual** (labelled by `Student residual` in SAS), because it is related to the *Student t-distribution*.

Technically, because $\hat{\sigma}$ in the denominator involves e_i, it is not independent of e_i and so we do not have the t-distribution. In order to get the t-distribution, we delete the i^{th} observation in the calculation of the mean squared error (MSE) and denote it by $\hat{\sigma}^2_{(i)}$. Then

$$z_i^{**} = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}},$$

called the **externally studentised residual** (labelled by `RStudent` in SAS), will follow the t-distribution with degrees of freedom $(n-1)-p$, where p is the number of parameters estimated in the regression model and $n-1$ is the number of data points used (because the i^{th} one is excluded). Anyway, often we have a large n and so this technical change will make little difference and if an observation has an internally or externally studentised residual of absolute value larger than 2, it is considered an outlier.

How to identify leverage points (data of unusual x-values)? Because $\hat{\boldsymbol{Y}} = \boldsymbol{HY}$, we have

$$\hat{Y}_i = h_{i1}Y_1 + h_{i2}Y_2 + \cdots + h_{ii}Y_i + \cdots + h_{in}Y_n.$$

Thus, h_{ii} (known as a *hat matrix diagonal element* or sometimes [when no confusion is possible] a *hat-value*) measures the influence of Y_i on $\hat{Y}_i$. On one hand, a few algebraic steps will reveal that in the simple linear regression, h_{ii} measures how far x_i is from the sample mean $\overline{x}$:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{C_{xx}},$$

and in the multiple regression model, h_{ii} is a standardised measure of the distance of the i^{th} observation from the centroid of the x-space. On the other hand, $\hat{Y}_i$ will be dominated by $h_{ii}Y_i$ if h_{ii} is large. Therefore, h_{ii} is the amount of leverage exerted by Y_i on $\hat{Y}_i$ and is hence simply called the *leverage* of the i^{th} observation. An observation with large leverage is a leverage point and is potentially influential.

Since $\boldsymbol{H}$ is idempotent, which means $\boldsymbol{H} = \boldsymbol{H}^k \Rightarrow \boldsymbol{Q\Lambda Q}^{-1} = \boldsymbol{Q\Lambda}^{\boldsymbol{k}}\boldsymbol{Q}^{-\boldsymbol{1}}$, implying that its eigenvalues are either 0 or 1, and consequently,

$$\sum h_{ii} =: \operatorname{tr}(\boldsymbol{H}) = \operatorname{rank}(\boldsymbol{H}).$$

A few more steps can show that $\operatorname{rank}(\boldsymbol{H}) = \operatorname{rank}(\boldsymbol{X}) = p$.

Let us define the mean leverage $\overline{h} = p/n$, and the rule of thumb is that if a point has a leverage more than a double of the mean leverage

$$\boxed{h_{ii} > 2\overline{h} = \frac{2p}{n}},$$

then its x-value is remote enough from the rest of the data to be called a leverage point.

However, if h_{ii} is large, it means that the fitted regression model tries to fit well Y_i, but this point is not necessarily influential. Why? Consider the case that if we delete the i^{th} observation in the estimation procedure and if the fitted regression model still, though unintentionally, fits well Y_i, then Y_i is not influential. That is, not all leverage points are going to be influential on the regression coefficients.

Typically a leverage point is influential if it is also an outlier.

To check quantitatively whether a point is influential, we apply the statistical technique we encountered before, namely **jackknife**, which means that we estimate the parameters by using all except the i^{th} observation and obtain $\hat{\boldsymbol{Y}}^{(i)}$, and then compare it with $\hat{\boldsymbol{Y}}$. If Y_i is influential, then these two vectors of fitted values will be very different.

To measure the difference between these two vectors, we introduce **Cook's distance** D_i, defined by

$$\begin{aligned} D_i &= \frac{(\hat{\boldsymbol{Y}}^{(i)} - \hat{\boldsymbol{Y}})'(\hat{\boldsymbol{Y}}^{(i)} - \hat{\boldsymbol{Y}})}{p \cdot \hat{\sigma}^2} = \frac{(\hat{\boldsymbol{\beta}}^{(i)} - \hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(i)} - \hat{\boldsymbol{\beta}})}{p \cdot \hat{\sigma}^2} \\ &= \cdots = \frac{e_i^2}{p \cdot \hat{\sigma}^2} \frac{h_{ii}}{(1-h_{ii})^2} = \frac{z_i^{*2}}{p} \frac{h_{ii}}{1-h_{ii}}. \end{aligned}$$

We can see that Cook's distance can be interpreted as a summary measure of distances between $\hat{\beta}_0$ and $\hat{\beta}_0^{(i)}$, $\hat{\beta}_1$ and $\hat{\beta}_1^{(i)}$, etc. And, it can also be viewed as a measure of the i^{th} observation's influence, incorporating both its studentised residual z_i^* and leverage h_{ii}.

A large Cook's distance D_i indicates that the i^{th} observation has strong influence on the fitted values and also on the parameter estimates; examples are outliers with large leverages or leverage points with large studentised residuals.

For small samples, the magnitude of D_i is usually assessed by comparing with $F_{0.5,p,n-p}$. It is because if $D_i = F_{0.5,p,n-p}$, then deleting point i would move the parameter estimates to the boundary of an approximately 50% confidence region for $\boldsymbol{\beta}$ based on the complete data set. This is a huge displacement and indicates that $\hat{\boldsymbol{\beta}}$ is sensitive to the i^{th} data point. Hence, if

$$\boxed{D_i > F_{0.5,p,n-p} \approx 1},$$

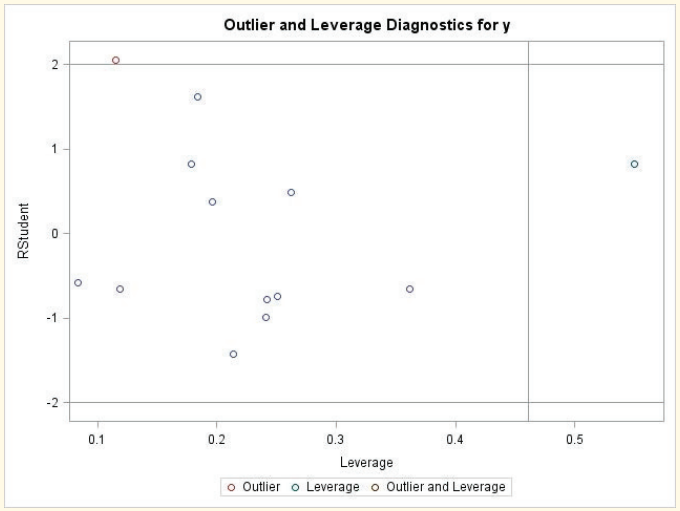
then the i^{th} observation is influential. Note that (i) the value $F_{0.5,p,n-p}$ is often (but not always) close to 1 and (ii) the distance D_i is in fact not an F-statistic, but (iii) the cutoff value 1 works pretty well in practice for small samples.

When n is large, another cutoff criterion for Cook's distance has been suggested:

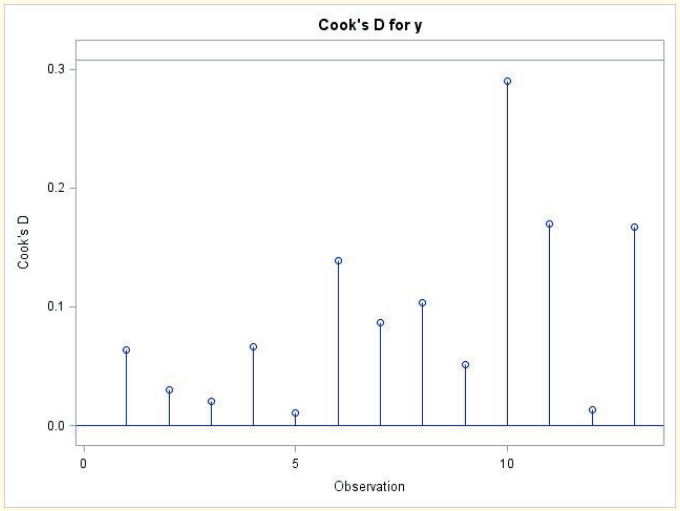
$$\boxed{D_i > \frac{4}{n-p} \approx \frac{4}{n}},$$

which results from setting $z_i^* = 2$ and $h_{ii} = \overline{h} = p/n$.

In SAS, the externally studentised residuals will be plotted against the leverages, with the cutoff values shown as a vertical (for $h_i > 2p/n$) and two horizontal reference lines (for $|z_i^{**}| > 2$), while in another plot each Cook's distance will be represented by a vertical bar, with a horizontal reference line (for $D_i > 4/n$), see the figure below.



(a)



(b)

Figure: Plots for fitting $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ to the Hald data: (a) Externally studentised residuals vs leverages in SAS, (b) Cook's distance plot in SAS

There are other measures of influence. The most well-known three criteria are

$$|\texttt{DFBETAS}_{j,i}| = \left| \frac{\hat{\beta}_j - \hat{\beta}_j^{(i)}}{\sqrt{\hat{\sigma}^2_{(i)} c_{j+1,j+1}}} \right| > \frac{2}{\sqrt{n}},$$

where $c_{j+1,j+1}$ is the $(j+1)^{\text{st}}$ diagonal element of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ [note that $j = 0, \ldots, k$],

$$|\texttt{DFFITS}_i| = \left| \frac{\hat{Y}_i - \hat{Y}_i^{(i)}}{\sqrt{\hat{\sigma}^2_{(i)} h_{ii}}} \right| > 2\sqrt{\frac{p}{n}},$$

$$|\texttt{CovRatio}_i - 1| = \left| \frac{\det\left((\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)})^{-1}\right) \hat{\sigma}^2_{(i)}}{\det\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\right) \hat{\sigma}^2} - 1 \right| > \frac{3p}{n},$$

where the suggested cutoff points are for large samples.

The cutoff value for $\texttt{DFFITS}_i$ comes from deriving $\texttt{DFFITS}_i = \cdots = z_i^{**}\sqrt{h_{ii}/(1-h_{ii})}$, then substituting $z_i^{**} = 2$ and $h_{ii} = \overline{h} = p/n$ into it and finally approximating $\sqrt{p/(n-p)}$ by $\sqrt{p/n}$.

The cutoff value for $\texttt{DFBETAS}_{j,i}$ is obtained in a similar spirit (but involving more laborious arguments).

For $\texttt{CovRatio}_i$, it can be shown that

$$\texttt{CovRatio}_i = \frac{1}{\left(\frac{n-p-1}{n-p} + \frac{{z_i^{**}}^2}{n-p}\right)^p (1-h_{ii})},$$

and substituting two extreme cases: $z_i^{**} \geq 2$ with h_{ii} at its minimum $1/n$ and $z_i^{**} = 0$ with $h_{ii} \geq 2p/n$ and using first order Taylor's approximation, we will get the above cutoff value $3p/n$. These measures of influence and their cutoff values are suggested in the classical book Belsley, Kuh and Welsch (1980, *Regression Diagnostics: Identify Influential Data and Sources of Collinearity*, Wiley).

The interpretation of them is as follows.

The first two are obviously measuring, when we fit a model and estimate the parameters with and without the i^{th} observation, the difference (DF) in the β_j's estimate (BETA) and in the i^{th} fitted value (FIT), both of them scaled (S).

Whilst these $\texttt{DFBETAS}_{j,i}$ and $\texttt{DFFITS}_i$ provide insight about the influence of the i^{th} observations on $\hat{\beta}_j$ and on $\hat{Y}_i$, respectively, they do not provide any information about overall precision of estimation.

It is a common practice to use the determinant of the covariance matrix as a convenient scalar measure of precision, called the **generalised variance**:

$$\mathrm{GV}(\hat{\boldsymbol{\beta}}) = \det\big(\mathrm{var}(\hat{\boldsymbol{\beta}})\big) = \det\big((\boldsymbol{X}'\boldsymbol{X})^{-1}\big)\,\sigma^2.$$

To express the role of the i^{th} observation on the precision of estimation, we define $\texttt{CovRatio}_i$ as the ratio of the generalised variance of jackknifed version $\hat{\boldsymbol{\beta}}^{(i)}$ to the generalised variance of the least squares estimates $\hat{\boldsymbol{\beta}}$ based on the complete data set. The precision may be better or worse by deleting the i^{th} observation and so a large deviation from 1 in absolute value indicates an influential observation.
These three measures, as well as the leverages (i.e. the hat matrix diagonals) and the jackknifed version of the studentised residuals z_i^{**}, denoted by `RStudent` in SAS, can be obtained by the following code:

```
proc reg;
model y = x1 x2 x3 x4 / influence;
```

What should we do when we have problematic data?

First of all, they should not be ignored nor deleted automatically and without reflection. Keep this in mind: many of the great discoveries in science were made by paying attention to the outliers. We definitely cannot erase an outlier just because its deletion makes the result of the regression model more beautiful. Unless the observation is clearly an error such as recording error, most that can be done is to report the results both with and without the outlier.

The exception to this is the case of extreme x values. It is possible to reduce the range over which our predictions will be valid. For instance, it is fine to build a relationship between weight (Y) and height (x) only for those of height between, say, 5 feet and 6.5 feet, even if you have an observation of, say, $x = 7.5$ feet.

To limit the influence an outlying observation has, we may consider, instead of the least squares estimation (LSE) , the so-called **robust estimation**.

One example is the least absolute deviations (LAD) estimation, which minimises $\sum |Y_i - \hat{Y}_i|$. For the LSE, large deviations are prohibited because a large deviation will have a large weight (which equals itself), whilst for LAD, all deviations have the same weight and so large deviations will not be more important than small deviations.

Checking Normality

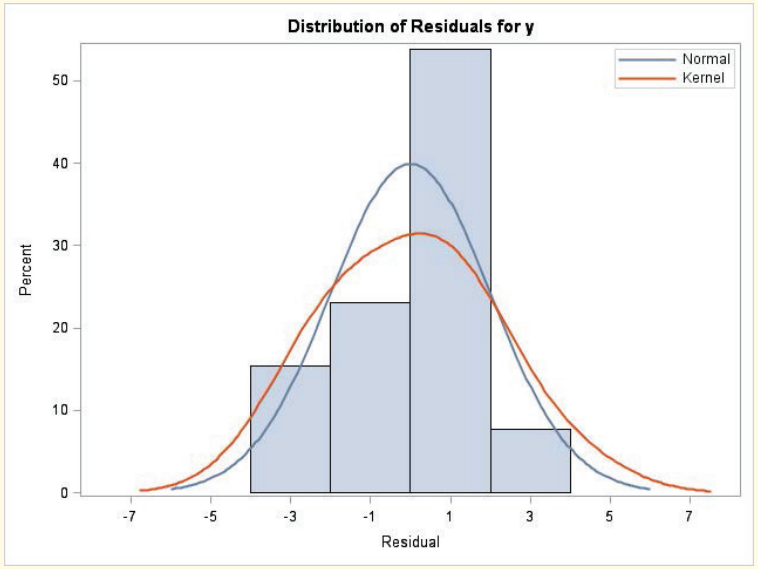
Normal probability paper is a specially constructed type of graph paper.

Although the unnumbered horizontal axis is marked by equal division in the usual way, the vertical axis is marked by a nonlinear scale such that the cumulative distribution function (cdf) of the normal distribution will be straightened to a straight line.

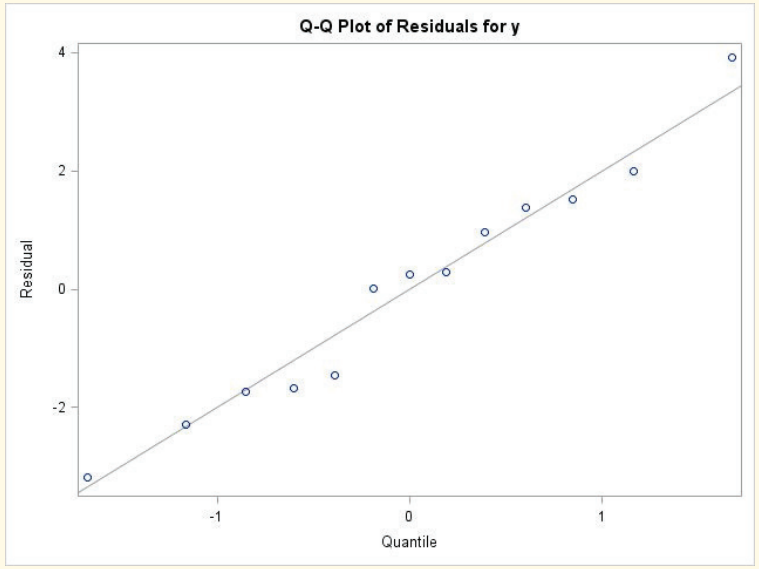
Nowadays, most normal probably plots are done on the computer, and there the vertical axis is often converted to a normal score such that 2.5% is converted to -1.96, 50% to 0, 95% to 1.645, and so on.

The normal probability plot is a special Q-Q plot (where Q stands for quantile).

The normal probability plot, together with a histogram of the residuals, is also a part of the SAS default output of the `model` statement in `proc reg`.



(a)



(b)

Figure: Plots in SAS for checking the normality assumption of the residuals: (a) histogram and (b) normal probability plot.

If the residuals are close to the reference line, then the normality assumption is fine.

However, if the residuals deviate from the straight line, then the normality assumption may be violated.

Examples of nonnormal distribution can be found in the following figure, in which the heavy-tailed distribution used is the t-distribution, the light-tailed the uniform, the positively skewed the χ^2 and the negatively skewed is minus one times χ^2.

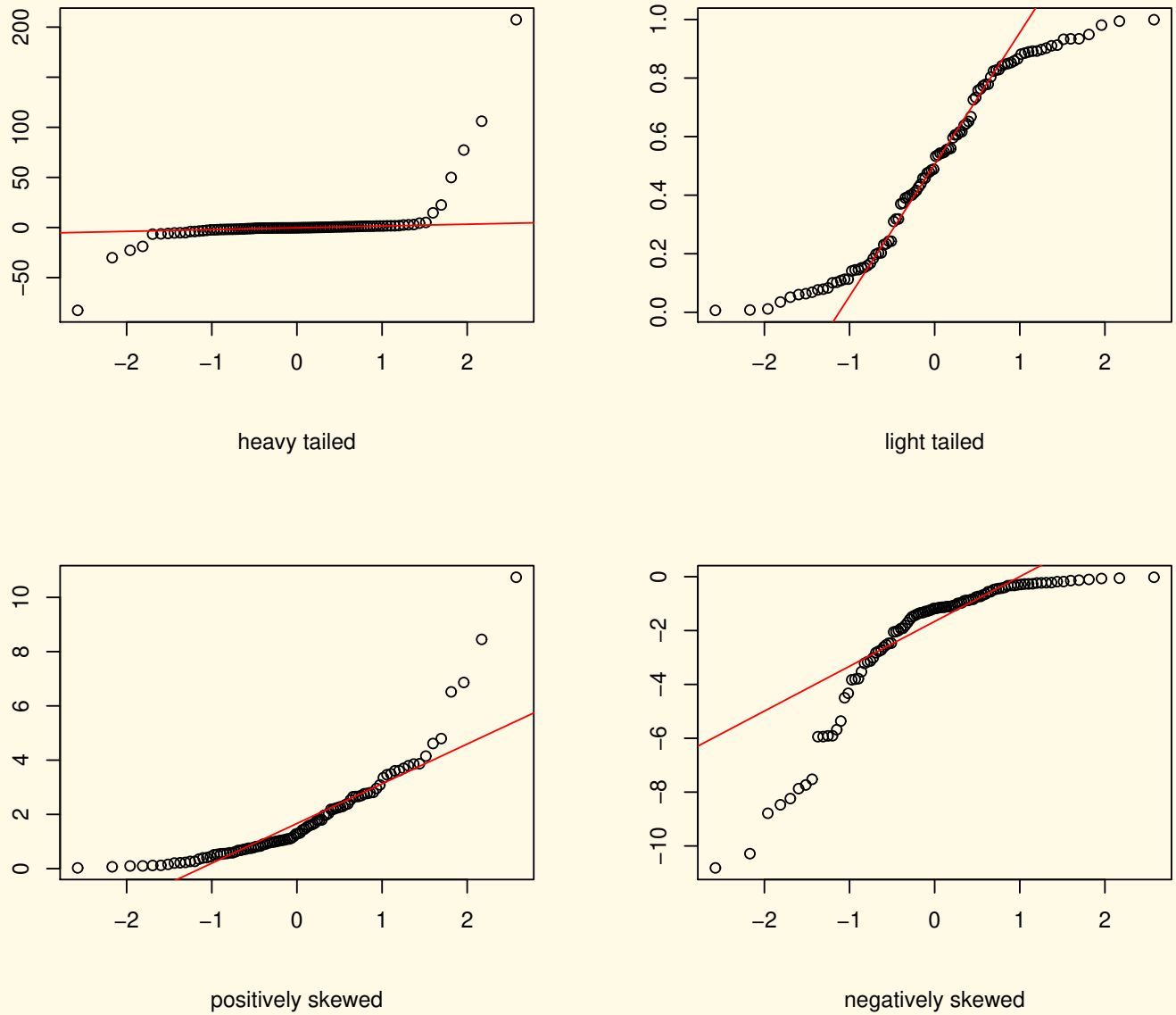
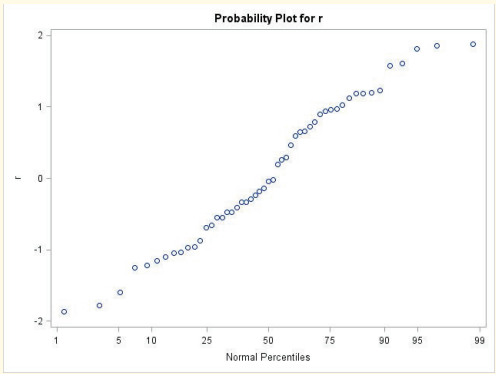
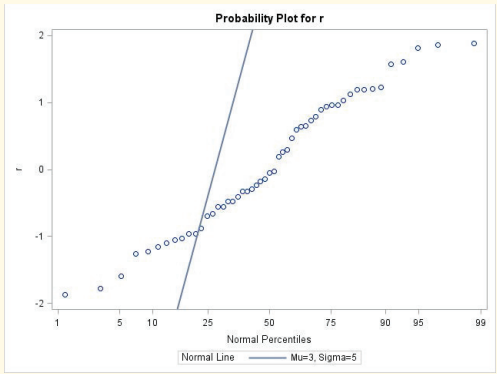


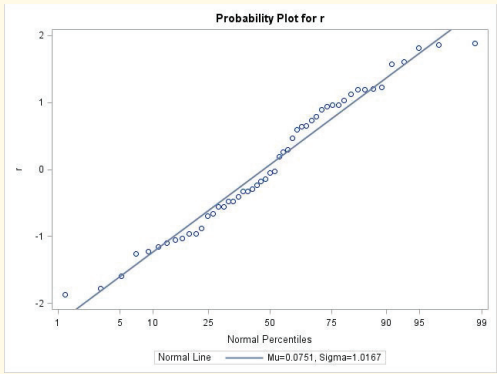
Figure: Normal Probability Plots of nonnormal samples.

The statistics for a univariate variable reported by `proc univariate` are very straightforward. What is important to us is the statement `probplot r/normal`, which produces a Q-Q plot for the normal distribution, i.e. the normal probability plot. See the figure below. We can add a reference line for the normal distribution with either (i) specified mean and specified standard deviation by inputting the specified values to the SAS code directly, e.g. `normal(mu=3 sigma=5)`, or (ii) estimated mean and estimated standard deviation by `normal(mu=est sigma=est)`:

```
proc univariate;
var r;
probplot r/normal;
probplot r/normal(mu=3 sigma=5);
probplot r/normal(mu=est sigma=est);
```







`normal;` `normal(mu=3 sigma=5);` `normal(mu=est sigma=est);`

Figure: Normal probability plots by `proc univariate` using `probplot r/normal`

There are of course formal statistical testing procedures to test the normality assumption (I am aware of more than 50 different tests) but we do not consider such tests here. This concludes Chapter 8.

Nonlinear Regression

There are many situations where a linear regression model may not be appropriate.

Like linear regression, a nonlinear regression model tries to relate a response Y to a vector of independent variables $\boldsymbol{x}$ through a known nonlinear function.

For example, the engineer or scientist may have direct knowledge of the form of the relationship between Y and $\boldsymbol{x}$, often from the theory underlying the phenomena. It may be a differential equation or the solution to a differential equation, e.g.

$$\mathbb{E}(Y) = \theta_1 \mathrm{e}^{\theta_2 x}, \tag{3}$$

which is a solution to

$$\frac{\mathrm{d}\mathbb{E}(Y)}{\mathrm{d}x} = \lambda \mathbb{E}(Y).$$

The expression on the right-hand side of equation (3) is a nonlinear function of the parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)$.

In general, a nonlinear regression model is expressed as

$$Y_i = f(\boldsymbol{x}_i, \boldsymbol{\theta}) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is the vector of p parameters and $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{ki})$ is the i^{th} observed vector of the k independent variables, where unlike in linear regression, p is not necessarily $k + 1$.

To estimate the parameters, we still employ least squares estimation, i.e. find the $\boldsymbol{\theta}$ that solves the minimisation problem:

$$\text{minimise} \quad \sum_{i=1}^{n} \{Y_i - f(\boldsymbol{x}_i, \boldsymbol{\theta})\}^2.$$

The normal equations are

$$\sum_{i=1}^{n} \{Y_i - f(\boldsymbol{x}_i, \boldsymbol{\theta})\} \left. \frac{\partial f(\boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \theta_j} \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} = 0 \quad \text{for } j = 1, \ldots, p,$$

which may be very difficult to solve analytically when f is a nonlinear function. Such a least squares estimation is often called nonlinear least squares estimation.

In particular, if ε_i are i.i.d. normal, then the nonlinear least squares estimator $\hat{\boldsymbol{\theta}}$ is also the maximum likelihood estimator. This feature makes the least squares attractive because then we may apply the asymptotic theory of MLE for statistical inference.

Another possible way to estimate the parameters is to transform the nonlinear model to a linear model. For example, consider again model in (3), we may express it as

$$\log \mathbb{E}(Y) = \log \theta_1 + \theta_2 x.$$

Therefore, it is tempting to consider the model

$$\log Y = \log \theta_1 + \theta_2 x + \varepsilon' = \beta_0 + \beta_1 x + \varepsilon', \tag{4}$$

and using linear least squares to estimate β_0 and β_1. However, the linear least squares of (4) will not in general be equivalent to the nonlinear least squares of the original model because the former minimises the SSE on $\log Y$. Consequently, the estimation is just based on a geometrical argument and there is a lack of statistical theory for it. Also, the distribution of the error term ε' is different from that of the original random error ε. That is, if ε follows the normal distribution, then ε' does not.

Thus, usually the nonlinear least squares estimates are obtained by iterative procedures such that we begin at some initial guess $\hat{\boldsymbol{\theta}}_0$ and calculate the corresponding $SSE(\hat{\boldsymbol{\theta}}_0)$. Then, using some mathematical method we get a new estimate from the old one

$$\hat{\boldsymbol{\theta}}_{i+1} = \hat{\boldsymbol{\theta}}_i + \boldsymbol{\Delta}_i,$$

and if the reduction in the error sum of squares $SSE(\hat{\boldsymbol{\theta}}_i) - SSE(\hat{\boldsymbol{\theta}}_{i+1})$ is large for some $\boldsymbol{\Delta}_i$ (the determination of $\boldsymbol{\Delta}_i$ depends on what numerical method we use), then we take $\hat{\boldsymbol{\theta}}_{i+1}$ as our new estimates and repeat the procedure until the reduction is smaller than a pre-specified threshold value.

If we have a linear regression model, then the contours of the SSE will be ellipsoidal and we have a unique global minimum, which is easy to locate.

If we have a nonlinear regression model, then the contours will often be of the banana-shape, in which the contours near the optimal solution are typically quite elongated so that many quite different values of $\hat{\theta}$ will produce values of the error sum of squares that are close to the global minimum. In some situations the contours may be so irregular that there are several local minima. Thus, a good initial value $\hat{\theta}_0$ is very important.

There is, however, no standard "crank the handle" mechanism for getting initial estimate $\hat{\theta}_0$. The typical general way is, to consider the behaviour of the response function, such as the limit as $x \to \pm\infty$ or the intercept at $x = 0$, and substitute in for observations that most nearly represent those conditions; then solve the resulting equations.

Before we illustrate the idea of getting an initial value, let us see one of the best known category of nonlinear models, namely the growth models, which describe how something grows with changes in an independent variable, often the time. Typical applications are in biology and ecology, where organisms and plants grow with time, but there are also applications in economics and engineering.

1. Exponential growth/decay:

$$Y = \theta_1 \mathrm{e}^{\theta_2 x} + \varepsilon,$$

   where the *hazard rate* of f, defined by f'/f, is a constant.

2. Gompertz's model:

$$Y = \theta_1 \mathrm{e}^{\theta_2 \mathrm{e}^{\theta_3 x}} + \varepsilon,$$

   where the hazard rate is exponential; such a model is used to model e.g. with negative θ_2 and θ_3, the population size in a confined space, as birth rates first increase and then slow as resource limits are reached.

3. Weibull's model:

$$Y = \theta_1 \mathrm{e}^{\theta_2 x^{\theta_3}} + \varepsilon,$$

where the hazard rate follows the power law; such a model is used to model failures of technical devices.

4. Logistic growth model:

$$Y = \frac{\theta_1}{1 + \theta_2 \mathrm{e}^{\theta_3 x}} + \varepsilon,$$

which will give us an S-shaped curve for, e.g., the population, in which the initial stage of growth is approximately exponential; then, as saturation begins, the growth slows, and at maturity, the growth stops.

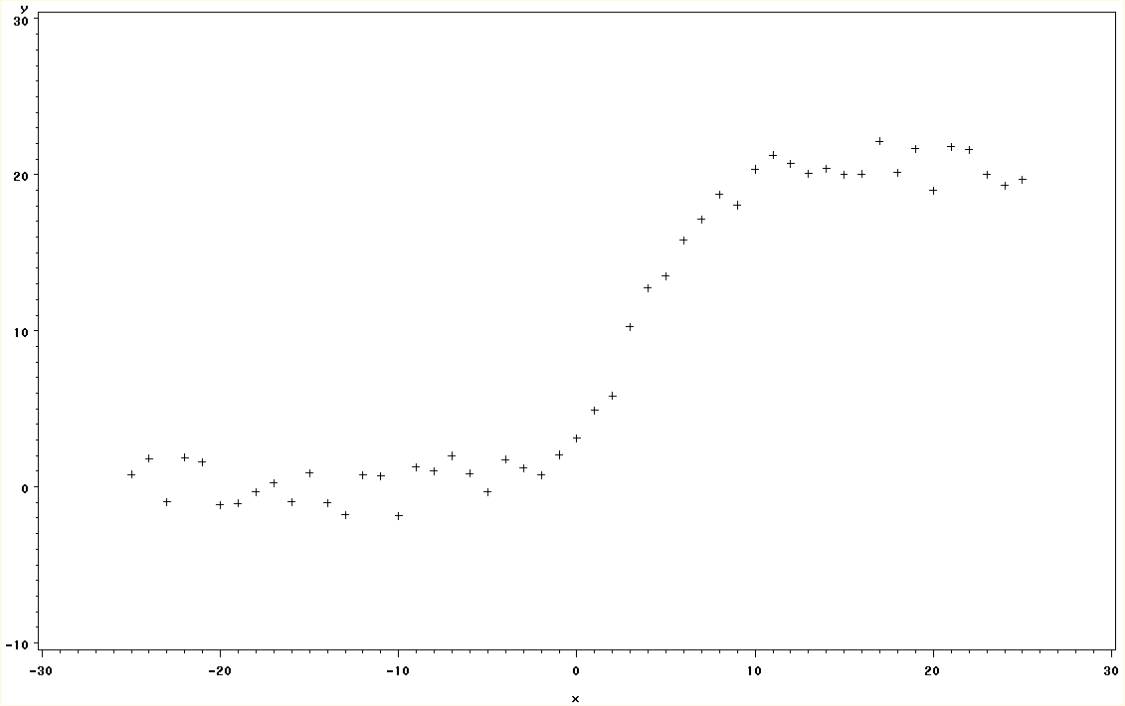


Figure: A scatterplot of data generated by a logistic growth model.

Consider the logistic model. To get the initial estimates of θ_1, θ_2 and θ_3, we first check the scatterplot. For example, in the above figure, we can see that θ_3 should be negative, and hence as $x \to \infty$, the mean of Y, denote by $f(x)$, goes to θ_1. Thus, our initial guess of θ_1 is the maximum of Y (or e.g. 110% of the maximum Y). We may take e.g. $\hat{\theta}_1 = 25$. When $x = 0$, $f(0)$ equals $\theta_1/(1 + \theta_2)$ and hence

$$\hat{\theta}_2 = \frac{\hat{\theta}_1}{Y\text{-intercept}} - 1 \approx \frac{25}{3} - 1 = 7.3.$$

Finally, noting that $f(x)/\theta_1 = 1/(1+\theta_2 \mathrm{e}^{\theta_3 x})$, we have

$$\log\Big(\frac{\theta_1}{f(x)} - 1\Big) = \log\theta_2 + \theta_3 x.$$

Taking the derivative, we have

$$\theta_3 = \frac{1}{\frac{\theta_1}{f(x)} - 1}\Big(-\frac{\theta_1}{f(x)^2}\Big) f'(x) = -\frac{\theta_1}{\theta_1 f(x) - f(x)^2}\, f'(x).$$

Hence, estimating the slope of f at $x = 0$ by the two data points at $x = -1$ and $x = 1$, which are $(-1, 2.06)$ and $(1, 4.89)$ in this scatterplot:

$$\hat{f}'(0) = \frac{4.89 - 2.06}{2} = 1.415,$$

we get our initial value for θ_3:

$$\hat{\theta}_3 = -1.415 \times \frac{25}{25 \cdot 3 - 3^2} \approx -0.5.$$

The SAS code for fitting a nonlinear model, plotting the residuals versus the x-values and versus the predicted $\hat{Y}$, and plotting the normal probability plot, is

```
data logistic;
input x y;
datalines;
-25 0.7927979
-24 1.8153270
. .
. .
25 19.6970758
;
proc gplot;
plot y*x;
proc nlin;
parms b1=25 b2=7.3 b3=-0.5;
model y=b1/(1+ b2*exp(b3*x));
output out=growth predicted=p residual=r;
proc print data=growth;
```

```
proc gplot data=growth;
plot r*x;
plot r*p;
proc univariate;
var r;
probplot r/normal(mu=est sigma=est);
run;
```

In the above code we introduced two more new procedures, namely `proc nlin` for fitting a nonlinear model and `proc univariate` for reporting many statistics for a univariate variable (specified by the `var` statement), including the possibility of the normal probability plot.

The procedure `proc nlin` produces an ANOVA table, testing the null hypothesis that all parameters are zero.

Note that in procedure `reg` the ANOVA tests the null hypothesis that all parameters except the intercept are zero and hence the value of the degrees of freedom for a linear regression model is $p - 1$.

However, in nonlinear regression we test all parameters and so the value of the degrees of freedom is exactly equal to the number of parameters.

The p-value given in the ANOVA table, as well as the confidence intervals for the parameters, are all approximates only. It is because in nonlinear regression the least squares (or maximum likelihood) estimators of the parameters do not enjoy any of the attractive properties that their counterparts do in linear regression.

Statistical inference in nonlinear regression depends on large-sample (asymptotic) results. The large-sample theory generally applies for both normally and nonnormally distributed errors and the key asymptotic properties are (i) the bias goes to zero as $n \to \infty$, and (ii) approximately $\hat{\boldsymbol{\theta}} \sim \mathrm{Normal}$. The second property allows us to get an approximated p-value and confidence limits.

How good are these approximates? If the nonlinear regression estimation algorithm converges in only a few iterations, say 4 iterations in this example, then it is likely that the asymptotic results will apply nicely. Convergence requiring many iterations is a symptom that the asymptotic results may not apply.