Checking the validity of the regression model

Checking the validity of the model is done through inspection of the residuals ($e_i = Y_i - \hat{Y}_i$, the deviations of fitted from observed values). Evaluation of the model involves two issues:

1. Checking the model assumptions
2. Checking the model fit

1. Checking the model assumptions.

Recall what the model assumptions for the regression are:

i. The observations are *independent*

ii. The variance of Y is the same for any combination of the predictors X_1, X_2,…, X_k (homoskedacity).

iii. The residuals are distributed independently with each of them $\varepsilon_i \sim N(0,\sigma^2)$, $i=1,2,\ldots,n$.

We would thus expect that the residuals e_i should behave consistently with these assumptions.

2. Detecting "influential" observations

The residuals are also used to detect observations that are behaving inconsistently with the overall model. For example, very large residuals may reflect an unusual observation, or simply a poor fit. On the other hand, an observation may be worth considering even though its associated residual is small, because it exerts an inordinate "influence" on the fitted regression line.

Characteristics of the residuals

1. $\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$. This is a direct consequence of the least-squares procedure. If the fitted model is $Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$, $i=1,\ldots,n$, then minimizing $\sum \varepsilon_i^2$ involves the partial derivative, $-2\sum_{i=1}^{n}\left[Y_i - \hat{\beta}_o + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}\right] = 0$, which reduces to $\sum_{i=1}^{n}\left[Y_i - \hat{Y}_i\right] = \sum_{i=1}^{n} \varepsilon_i = 0$. This implies that the residuals are not independent!
2. $S^2 = \frac{1}{n-k-1}\sum \varepsilon_i^2 = MSE$, if the model involving $k+1$ parameters is correct.
3. If S^2 does estimate the variance of the residuals, then we would expect the *standardized residual* $z_i = \frac{e_i}{S}$ to have zero mean and unit variance (the latter since $V(z_i) = V\left(\frac{e_i}{S}\right) = \frac{V(e_i)}{S^2} \approx \frac{S^2}{S^2} = 1$). If n is large, these residuals should behave like approximate normal random variables (approximate t if n is not large). Since e_i are not independent, when n is small, these statements are suspect.

Characteristics of the residuals (continued)

4. An alternative form of standardized residuals is the *studentized residual* $r_i = \frac{e_i}{S\sqrt{1-h_i}}$, where h_i is the i[th] diagonal element (called *leverage*) of the hat matrix. This is because (see proof below) the exact variance of the residuals is $Var(\mathbf{e}) = (\mathbf{I} - \mathbf{H})\sigma^2$. The $r_i \sim t_{n-k-1}$ approximately.

5. An additional residual that we can consider is

$$r_{(-i)} = r_i\sqrt{\frac{S^2}{S^2_{(-i)}}} = \frac{e_i}{\sqrt{S^2_{(-i)}(1-h_i)}} = r_i\sqrt{\frac{(n-k-1)-1}{(n-k-1)-r_i^2}}$$

This is called the *jackknife residual*. The quantity $S^2_{(-i)}$ is the residual variance computed with the i^{th} observation deleted. Jackknife residuals are distributed exactly according to a *t* distribution with (*n*-*k*-2) degrees of freedom.

Under the usual assumptions, the standardized, studentized and jackknife residuals should behave similarly. Especially if *n*>30, they should exhibit behavior consistent to a normally distributed variable.

Optional topic: The covariance matrix of the residuals

The covariance matrix (the matrix containing variances on the diagonal, and covariances in its off-diagonal elements) of the residuals is $V(\boldsymbol{\varepsilon}) = (\mathbf{I} - \mathbf{H})\sigma^2$. This implies that the variance of the i^{th} residual is $V(e_i) = \sigma^2(1 - h_i)$ i=1,2,…,n, the i^{th} diagonal element of the covariance matrix.

Proof (Draper & Smith, p.151):

Consider that the vector of residuals $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$. Then, $\mathbf{e} - E(\mathbf{e}) = (\mathbf{I} - \mathbf{H})\mathbf{Y} - (\mathbf{I} - \mathbf{H})E(\mathbf{Y}) = (\mathbf{I} - \mathbf{H})\mathbf{Y} - (\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = (\mathbf{I} - \mathbf{H})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}$. Now the variance $V(\mathbf{e}) = E\{[\mathbf{e} - E(\mathbf{e})][\mathbf{e} - E(\mathbf{e})]'\} = (\mathbf{I} - \mathbf{H})E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')(\mathbf{I} - \mathbf{H})'$. Since $E(\varepsilon)$=0, $V(\mathbf{e}) = (\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{H})' = (\mathbf{I} - \mathbf{H})\sigma^2$.

The latter statement depends on the fact that $(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})' = (\mathbf{I} - \mathbf{H})$.

This is because $(\mathbf{I} - \mathbf{H})' = (\mathbf{I}' - \mathbf{H}') = \mathbf{I}' - \{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}' = \mathbf{I} - \{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\} = (\mathbf{I} - \mathbf{H})$, i.e., $\mathbf{H}$ is a symmetric matrix, and $\mathbf{HH} = \{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}\{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Then $(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})' = (\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = \mathbf{I} - \mathbf{H} - \mathbf{H} + \mathbf{HH} = \mathbf{I} - \mathbf{H}$. That is **I-H** is an idempotent matrix.

Graphical analysis of residuals

A very effective way to detect problems with the model or observations requiring further attention, is through plotting the residuals against the fitted values and observe the patterns that emerge.

We plot the residuals against the fitted values (instead of the observed values). This is because, the residuals are correlated with the observed values but not with the fitted values. In fact, the slope of the regression line through the residuals plotted against the fitted values is $\hat{\beta} = 1 - R^2$ (see optional proof below). Problems that can be detected by these residual plots are as follows:

1. Variance is not constant. In these cases, the plot will look like the residuals are "funneling" out with increasing $\hat{Y}_i$.
2. Data depart from linearity. There is a systematic pattern in the residuals that indicates the need for inclusion of additional (curvilinear or polynomial) terms in the regression model
3. Trends against time or general dependence patterns. The residuals should be roughly independent. When time trends or other dependencies in the data exist, the model is inadequate.

Optional topic: The correlation between residuals and observed values

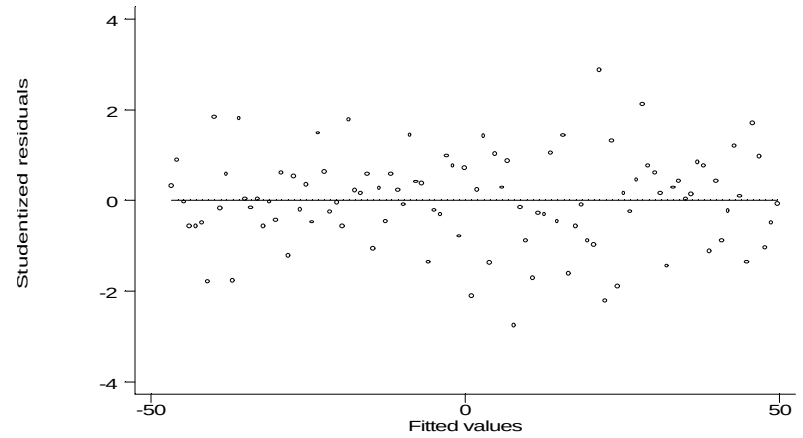
The correlation between ***e*** and **Y** is (1-R^2) and zero between ***e*** and $\hat{\mathbf{Y}}$.

Proof: (i) The correlation between ***e*** and **Y** is $r_{eY} = \frac{\sum (e_i - \bar{e})(Y_i - \bar{Y})}{\sqrt{\sum (e_i - \bar{e})^2 \sum (Y_i - \bar{Y})^2}}$. If an intercept term β_o is in the model (the sums of squares are "corrected" then $\bar{e} = 0$) then $\sum (e_i - \bar{e})(Y_i - \bar{Y}) = \sum e_i (Y_i - \bar{Y}) = \mathbf{e}'\mathbf{Y} = \mathbf{e}'\mathbf{e}$. This is since $\sum e_i \bar{Y} = \bar{Y} \sum e_i = 0$. Also, $\mathbf{e}'\mathbf{e} = \mathbf{Y}'(\mathbf{I} - \mathbf{H})'(\mathbf{I} - \mathbf{H})\mathbf{Y} = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y} = \mathbf{e}'\mathbf{Y}$ because (**I-H**) is *idempotent*.

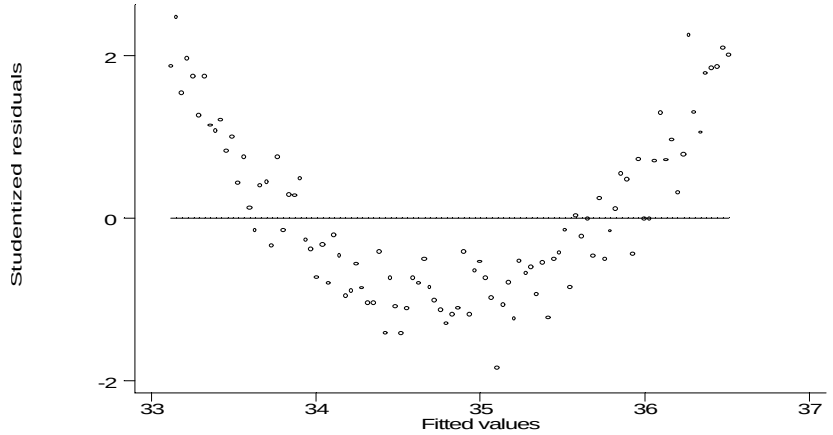
Then, $r_{eY} = \frac{\mathbf{e}'\mathbf{e}}{\sqrt{\mathbf{e}'\mathbf{e} \sum (Y_i - \bar{Y})^2}} = \left(\frac{SSE}{TCSS}\right)^{1/2} = (1 - R^2)^{1/2}$. Thus, the slope of the regression between ***e*** and **Y** will be $\beta_1 = r_{eY} \frac{S_e}{S_Y} = r_{eY} \frac{\sum (e_i - \bar{e})^2}{\sum (Y_i - \bar{Y})^2} = r_{eY} \left(\frac{SSE}{TCSS}\right)^{1/2} = (1 - R^2)^{1/2} (1 - R^2)^{1/2} = 1 - R^2$ QED.

(ii) The correlation between ***e*** and $\hat{\mathbf{Y}}$ is shown by similar methods to be zero. This is because, the numerator of the fraction $\sum (e_i - \bar{e})(\hat{Y}_i - \bar{\hat{Y}}) = \sum e_i \hat{Y}_i = \mathbf{e}'\hat{\mathbf{Y}} = \mathbf{Y}(\mathbf{I} - \mathbf{H})\hat{\mathbf{Y}} = \mathbf{Y}(\mathbf{I} - \mathbf{H})\mathbf{H}\mathbf{Y} = \mathbf{Y}(\mathbf{H} - \mathbf{H}^2)\mathbf{Y} = 0$. Here we've used the fact that $\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$, and that $\mathbf{H}\text{-}\mathbf{H}^2 = \mathbf{H}\text{-}\mathbf{H} = 0$ because **H** is *idempotent*. QED.

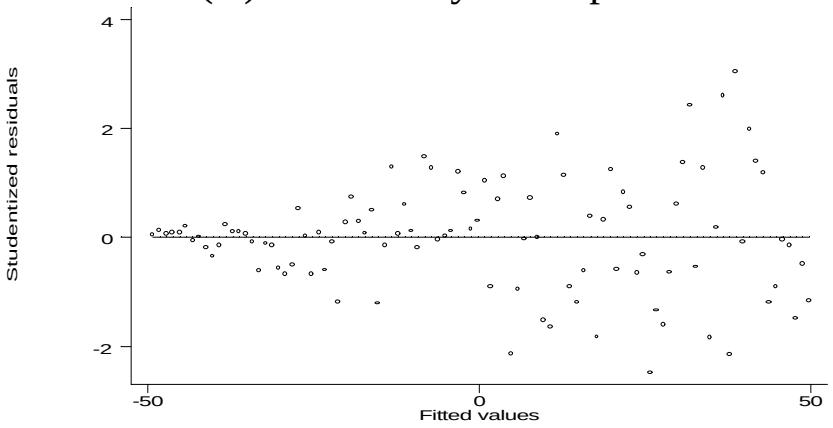
Typical plots of jackknife[1] residuals against fitted values



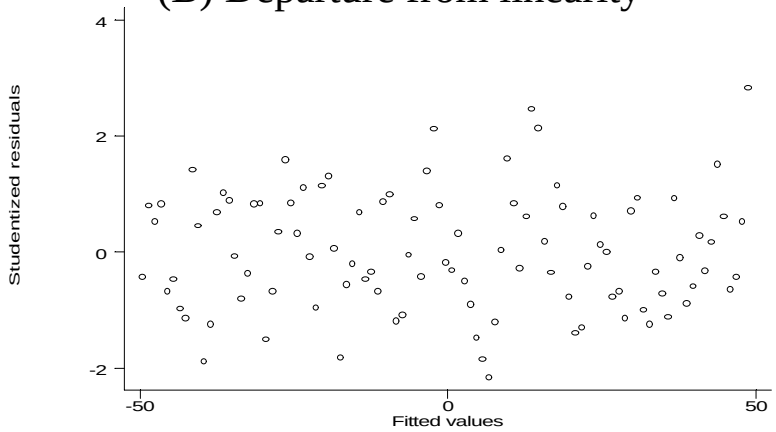
(A) Data satisfy assumptions



(B) Departure from linearity

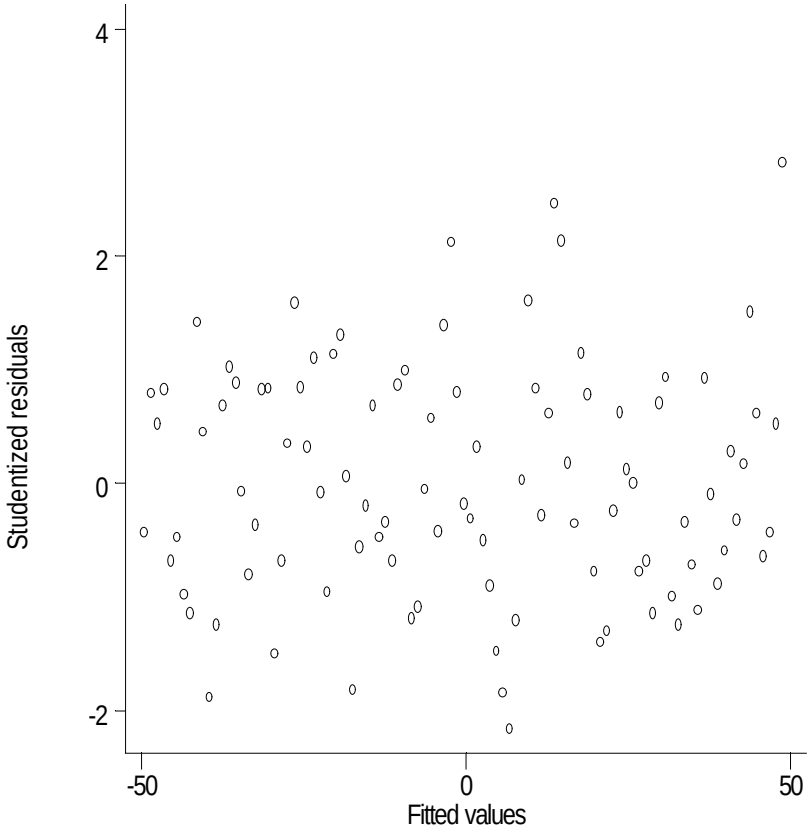


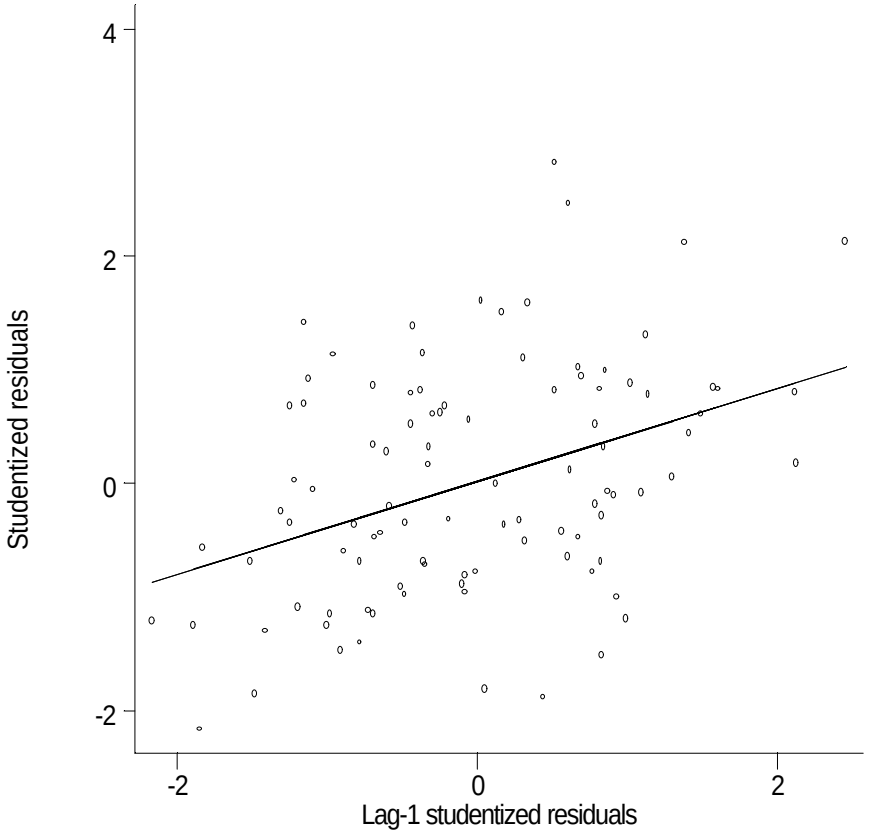
(C) Non-constant variance



(D) Dependence among data points

[1] Note that STATA calls jackknife residuals "studentized residuals"





This is the same plot of Figure (D) previously. Notice the visible correlation between the residuals and the residuals lag 1 indicating a possible “autocorrelation” structure (right).

Other graphical methods of residual analyses

Most other graphical methods of visually analyzing residuals are again trying to assess any departures from the model assumptions. With respect to the residuals all these plots will be assessing departures of the residuals from normality. Given changes in scale, it is advisable that we consider standardized residuals and out of those, the jackknife residuals due to their attractive distributional properties.

There are several methods for assessing departures from a given distribution. A number of these plots are based on comparing the observed ranks of the observations in the data to similar rankings of observations that follow a known theoretical distribution such as normal, *t*, chi-square or others.

Quantile-quantile or Q-Q plots

To create a Q-Q plot we proceed as follows:

1. The variable of interest is sorted.
2. The empirical quantiles are determined. For example, if there are 10 observations, the 5^{th} largest residual will be associated with the 50^{th} percentile.
3. The values of the inverse of the distribution of interest are then plotted against the sorted observations in the data. For example, in the case of a *normal* Q-Q plot, the median value in the data (the one associated with the 50% percentile) is plotted (on the y axis) against the value $\Phi(0.5)$ (on the x axis)[2].

[2] Actually, every i/n quantile will be plotted against $\Phi[i/(n+1)]$ or $\Phi[(i-1/2)/n]$ to avoid having to plot $\Phi[n/n] = \Phi[1] = \infty$

Standardized Probability-Probability or P-P plots

To create a P-P plot we proceed as follows:

1. The variable of interest is sorted.
2. The empirical quantiles are determined. For example, if there are 10 observations, the 5^{th} largest residual will be associated with the 50^{th} percentile.
3. The values of the inverse of the distribution of interest associated with the *standardized* data points are then plotted against their empirical quantiles. That is we plot $\Phi\left(\frac{X_i - \overline{X}}{s_x}\right)$. For example, in the case of a *standardized normal* P-P plot, the inverse-normal of the median value in the data $\Phi(m)$ is plotted (on the y axis) against the 50% percentile (on the x axis).

Symmetry plots

To create a symmetry plot we proceed as follows:

1. The variable of interest is sorted.
2. Two new variables are created. The first contains variables i (**from small to large**), $i=1,2,\ldots,n/2$ if n is even, or $i=1,2,\ldots,(n+1)/2$ if n is odd. The second contains the ordered variables $n-i+1$ (**from large to small**), where i is defined as before.
3. The distances of each variable from the median of the data are computed, and these distances are plotted. If the empirical distribution of the observations is symmetric then the plotted curve should approximately lie on a line.

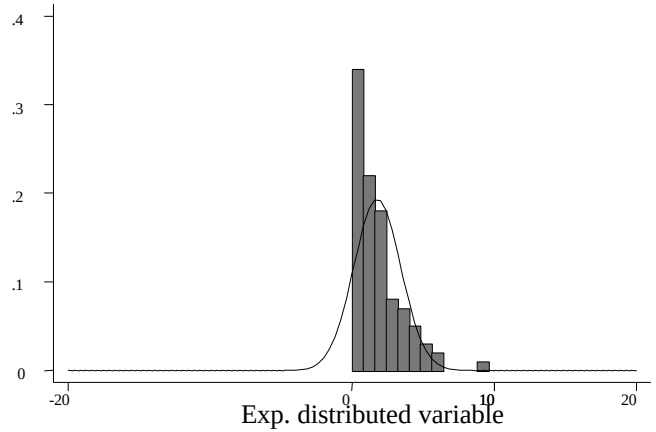
The idea behind symmetry plots is that each *pair* of ordered observation from each end of the data, should be at approximately equal distances from the median (the median is at the center of the distribution) if the distribution is symmetric. If the plotted points are above the line, then this is evidence of a distribution *skewed to the right*, while if the points are below the line, this is evidence of a distribution that is *skewed to the left.*

Example: The exponential case

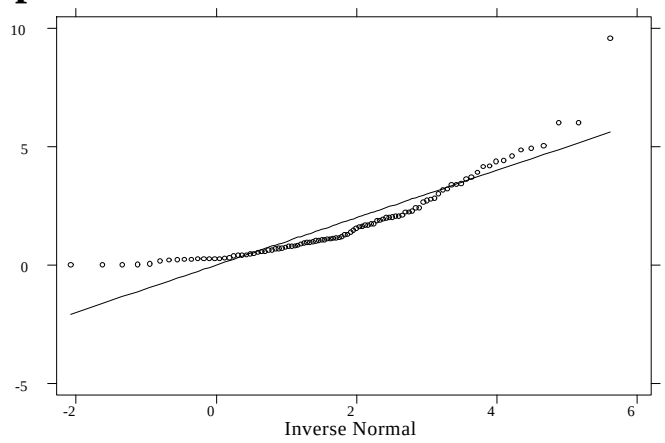
The STATA commands that will be used are as follows:

`. set obs 100`	**Sets number of observations to 100**
`. gen u=uniform()`	**Generates 100 uniform random numbers in the interval [0,1]**
`. gen c=invchi(2,(1-u))`	**Generates 100 χ_2-distributed (or equivalently exponentially distributed) random variables. Alternatively we could have written** `gen c=invchi(2,(1-uniform()))`
`. label var c "Exp. distributed variable"`	
`. graph c, normal xlab(-20,-10,0,10,20) ylab bin(50)`	**Plot a histogram of `c`, overlay a normal plot for comparison, then define the x and y axes and the number of histogram bars (50)**
`. qnorm c, xlab ylab`	**Produce a normal Q-Q plot for `c`**
`. pnorm c`	**Produce a normal P-P plot for `c`**
`. symplot c, xlab ylab`	**Produce a symmetry plot for `c`**

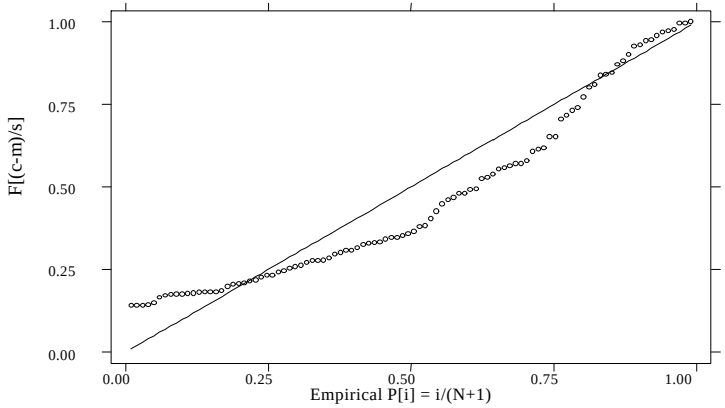
Four graphical views of an exponential variable



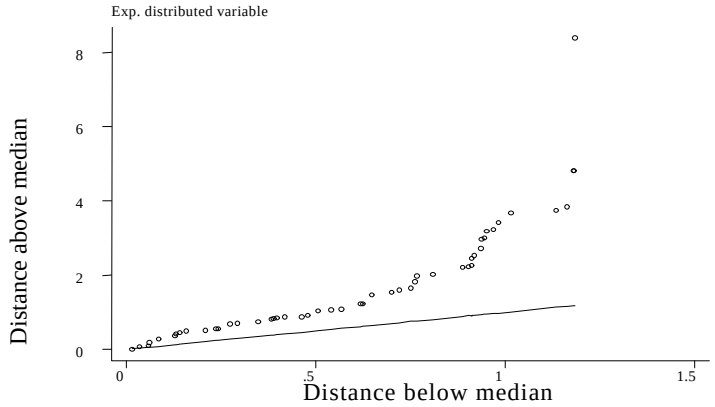
(A) Histogram of observations



(B) Normal Q-Q plot



(C) Normal P-P plot



(D) Symmetry plot

The exponential case

Comments:

(A) Note that the exponential distribution is not defined in the negative real line. The empirical distribution (histogram) of the exponentially distributed variable is compared to the (theoretical) normal distribution with the same mean and variance.

(B) Thus, the Q-Q plot of variable c against a normal distribution exhibits shorter tails among the lower (negative) values, and thicker tails among the larger values (as a result of the right-skewness of the exponential distribution; see (D))

(C) Similarly, the P-P plot shows the deviations at the tails. P-P plots also emphasize the differences in the *middle* of the distribution, while Q-Q plots are more sensitive to discrepancies in the tails of the distribution (this has to do with the arrangement of the percentiles $p_i = i/(n+1)$ versus the inverse normal percentiles $\Phi(p_i)$).

(D) The symmetry plot shows that the distribution of the variable of interest is skewed to the right (as indicated by the deviations of the plotted points *above* the line).

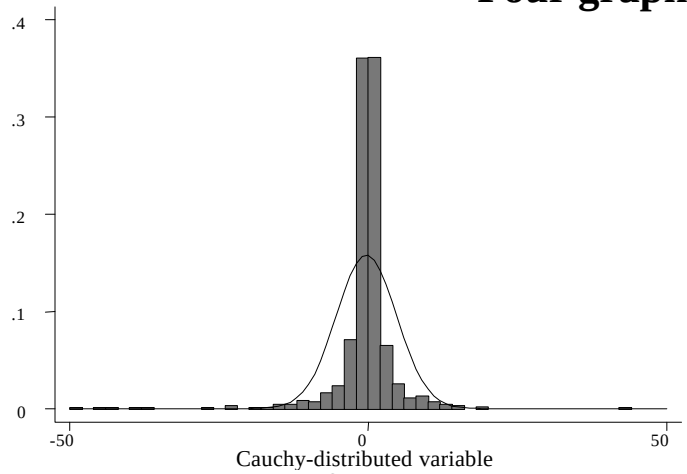
Example: The case of the Cauchy distribution

The STATA commands that will be used are as follows:

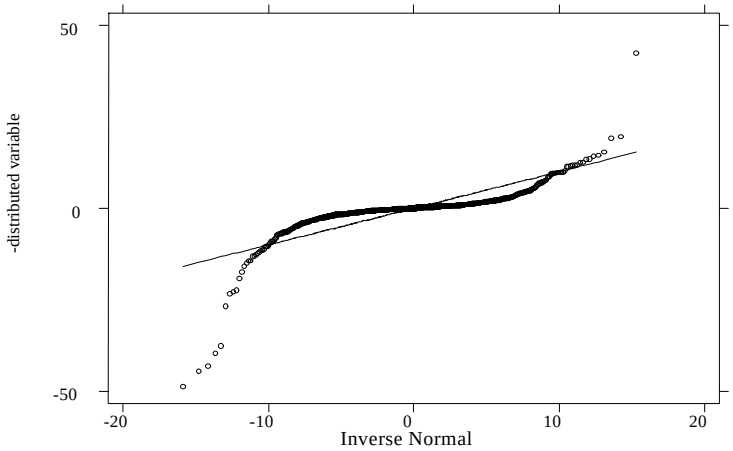
`. set obs 1000`	**Sets number of observations to 100**
`. gen u=2*uniform()-1`	**Generates 1,000 uniform random numbers in the interval [-1,1]**
`. gen t1=invt(1,u) if u>=0`	**Generates the positive half of the Cauchy-distributed random variables[3].**
`. replace t1=-invt(1,-u) if u<0`	**Generates the negative half of the variables**
`. label var t1 "Cauchy-distributed variable"`	
`. graph t1 if abs(t1)<50, normal xlab ylab bin(50)`	**Plot a histogram of `t1`, overlay a normal plot for comparison, and limit the plot to $\|t_1\|<50$ (to avoid outliers that interfere with the plot)**
`. qnorm t1, xlab ylab`	**Produce a normal Q-Q plot for `t1`**
`. pnorm t1`	**Produce a normal P-P plot for `t1`**
`. symplot t1, xlab ylab`	**Produce a symmetry plot for `t1`**

[3] We have to generate u~U[-1,1] because in STATA `invt(df,q)` gives the t value that corresponds to the quantile q=Pr(T>|t|). Thus, we can obtain the negative values by generating negative u's. Notice however, that only positive quantiles (u's) go into `invt(df,q)`, so we must use –u in the function.

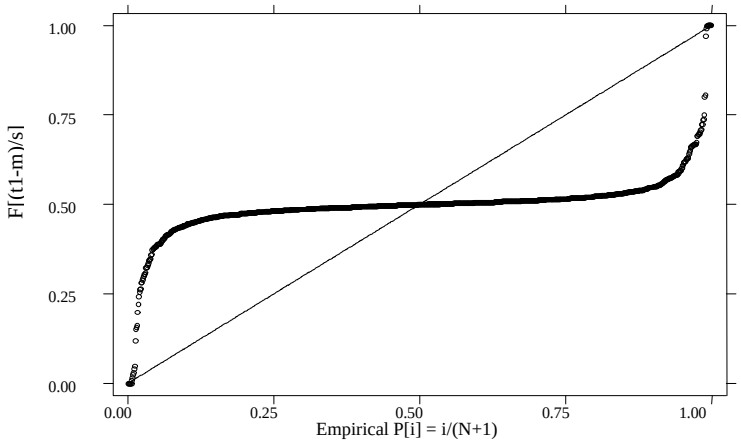
Four graphical views of a t_1 variable



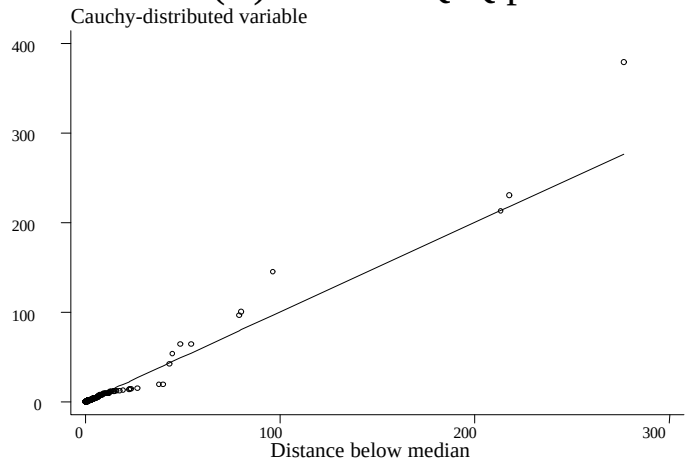
(A) Histogram of observations



(B) Normal Q-Q plot



(C) Standardized normal P-P plot



(D) Symmetry plot

The case of t_1 (Cauchy) distribution

Comments:

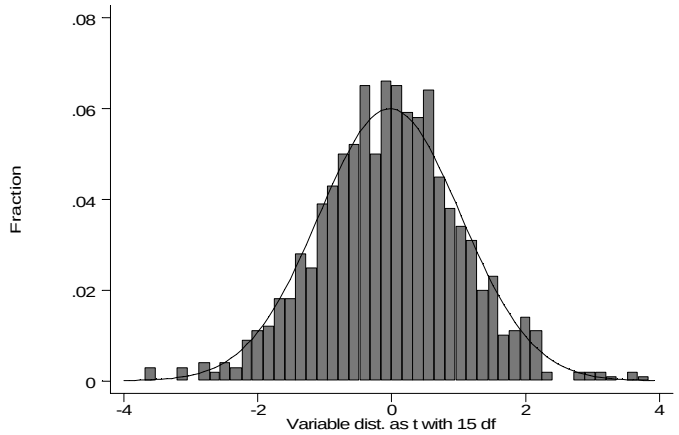
(A) Note that the empirical distribution (histogram) of the Cauchy-distributed variable has much "thicker" tails compared to the standard distribution.

(B) Thus, the Q-Q plot of this variable exhibits serious discrepancies in both tails compared to the normal distribution, as the plotted points drop below the line or rise above the line in the lower and higher ends of the distribution respectively.

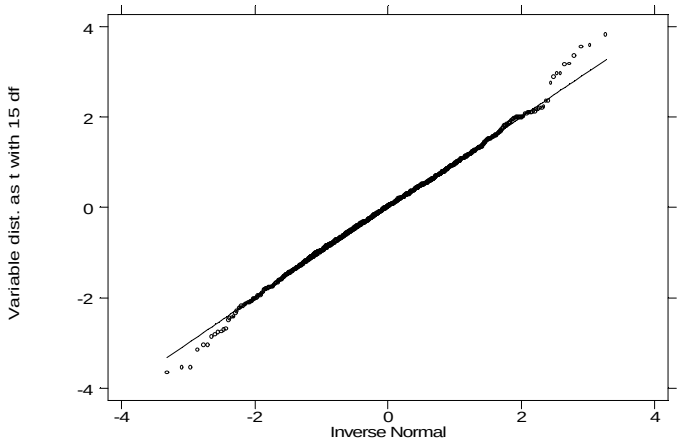
(C) Similarly, the P-P plot shows the deviations at both tails.

(D) The symmetry plot does not show marked deviations from the line, reflecting the symmetric shape of the Cauchy distribution.

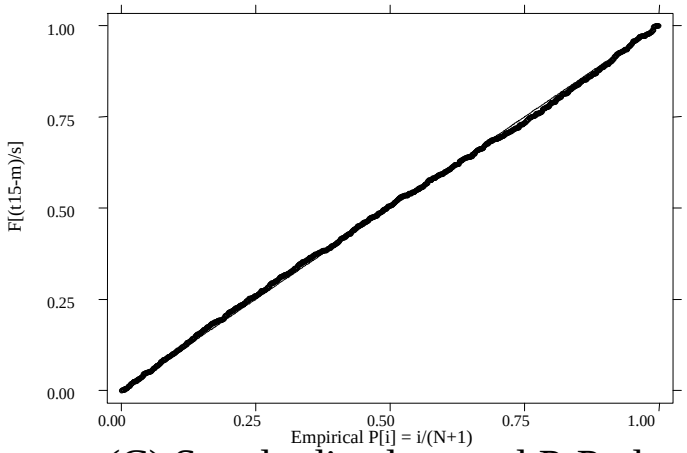
Four graphical views of a t_{15} variable



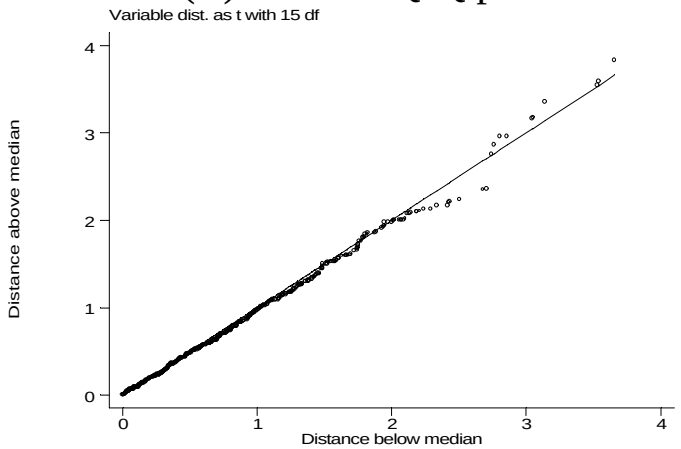
(A) Histogram of observations



(B) Normal Q-Q plot



(C) Standardized normal P-P plot



(D) Symmetry plot

The case of t_{15} distribution

Comments:

(A) Note that the empirical distribution (histogram) of the variable that is distributed as a t_{15} almost coincides with the standard normal distribution. Recall that $t_{\infty} \equiv N(0,1)$.

(B) Thus, the Q-Q plot of this variable no longer exhibits serious discrepancies in the tails compared to the normal distribution, compared to the case of the Cauchy distribution. Notice how the Q-Q plot picks up the discrepancy of the (thicker) tails better than the P-P plot

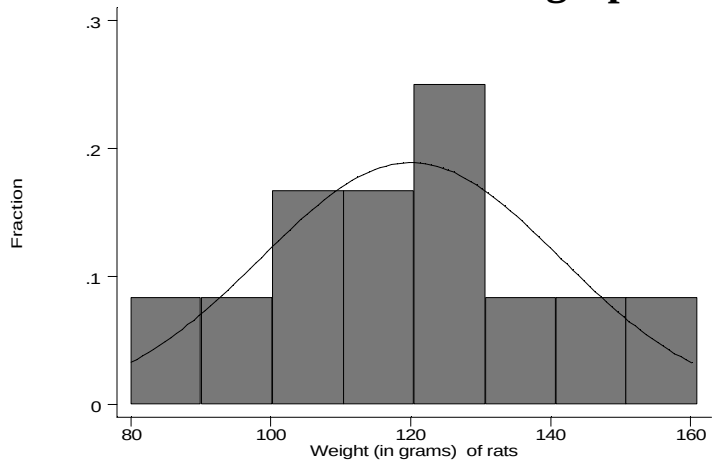
(C) Similarly, the P-P plot shows the agreement between the two distributions.

(D) The symmetry plot does not show marked deviations from the line, again reflecting the symmetric shape of the *t* distribution.

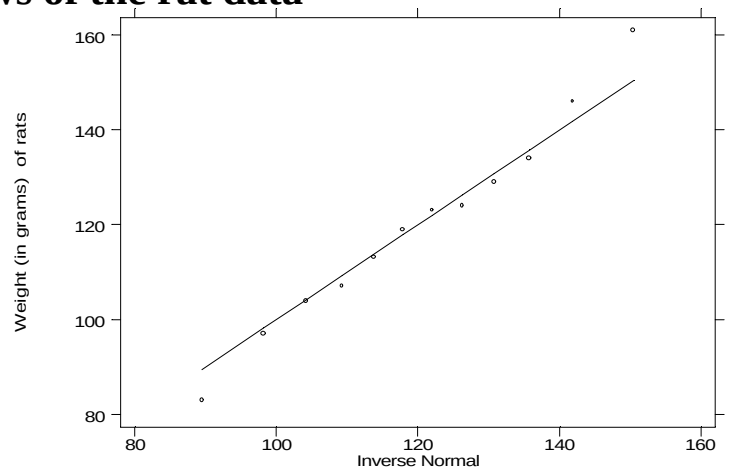
The above plots are illustrated by use of a data set on the weight of 12 female rats on a high protein diet (Snedecor & Cochran, 1980).

i	X_i	$p_i=i/(n+1)$	$\Phi^{-1}[p_i]$	$\Phi\left[\left(X_i-\overline{X}\right)/s\right]$
1	83	.077	-1.426	0.0418
2	97	.154	-1.020	0.1141
3	104	.231	-0.736	0.2272
4	107	.308	-0.502	0.2717
5	113	.385	-0.293	0.3717
6	119	.462	-0.097	0.4813
7	123	.538	0.097	0.5558
8	124	.615	0.293	0.5742
9	129	.692	0.502	0.6630
10	134	.769	0.736	0.7436
11	146	.846	1.020	0.8879
12	161	.923	1.426	0.9724

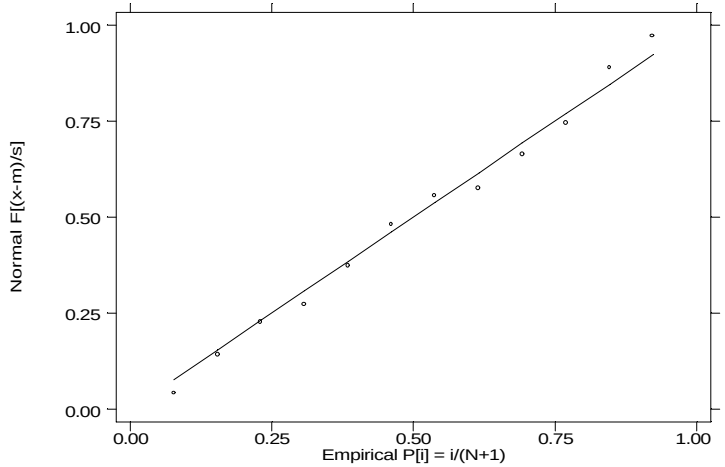
Four graphical views of the rat data



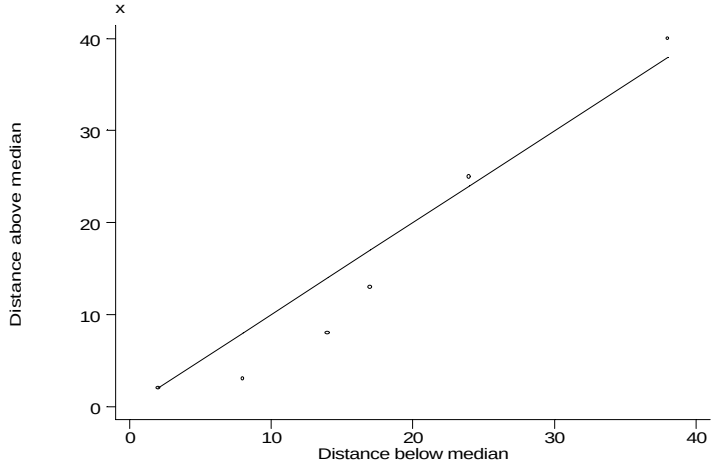
(A) Histogram of rat weights



(B) Q-Q plot of rat weights



(C) P-P plot of rat weights



(D) Symmetry plot of rat weights

Significance tests

In addition to the visual (graphical) inspection of the residuals, there are quantitative tests that are based on statistical tests that can be applied. These are:

1. *Kolmogorov-Smirnov*-type tests. These are *goodness-of-fit* tests based on the chi-square distribution (derived from aggregating deviations of observed values from their expected values).
2. A common test of normality is the *Shapiro-Wilks* test. This is very useful in cases of small sample sizes (n<50)
3. The possibility of *autocorrelation* or serial correlation among the observations is tested via the *Durbin-Watson* test. Autocorrelation is a factor sometimes when measurements are taken serially over time and refers to $r(Y_t, Y_{t-1})$, i.e., the measurement at time t versus the previous measurement.

The decision for the DW test is as follows:

i. Autocorrelation is present if $d<d_L$ or $d>4-d_L$

ii. No-autocorrelation, if $d<d_U$ and $d>4-d_U$

iii. Test inconclusive

Detecting outliers and influential observations

An *outlier* is an unusual value that does not conform to the pattern established by the rest of the data.

An *influential* observation may be an outlying observation, but it may also be one that does conform to the data pattern, but has a large contribution in the nature of the regression line.

The jackknife residual

From the definition of the jackknife residual $r_{(-i)} = \frac{e_i}{S_{(-i)}\sqrt{1-h_i}}$, we see that there are three parts that contribute in identifying outliers (large residuals) in the data.

These are:

i. The magnitude of the raw residual e_i that indicates the distance of the fitted from the observed value for that specific point

ii. The residual variance $S^2_{(-i)}$, which is the residual variance with the i^{th} observation excluded. This will be smaller than S^2 if the i^{th} point exhibits greater variability than expected, further increasing the value of the jackknife residual

iii. The *leverage* h_i is the i^{th} diagonal element of the *hat* matrix. Outliers will have large leverage values, which will tend to reduce the quantity $(1-h_i)$ and increase the size of the residual.

Usually, values of the jackknifed residual $|r_{(-i)}| > t_{n-k-2;1-\alpha/2n}$ should be scrutinized further (note that the size of the *t* tail has been adjusted to account for multiple comparisons among *n* residuals).

The leverage values $\{h_i\}$

Consider the simple linear regression case. There, the leverage of the i^{th} observation has the form, $h_i = \frac{1}{n}\frac{(X_i - \bar{X})^2}{(n-1)S_X^2} = \frac{1}{n}\frac{z_{Xi}^2}{(n-1)}$. This means that the leverage is proportionate to the standardized squared distance of the value of the i^{th} value of the predictor, from its mean.

The size of each leverage value is $0 \leq h_i \leq 1$. When there is an intercept in the model then $1/n \leq h_i \leq 1$. A value of 1 means that the regression line has been "levered" (pulled) to pass through the point.

In general, in a regression model with k parameters (k+1 including the intercept) $\sum h_i = k+1$, and thus, $\bar{h} = \frac{k+1}{n}$. It has been suggested that we should check every observation with leverage that is larger than $2(k+1)/n$. A quantitative test is based on the approximate χ^2 distribution of the leverages. Then $F_i = \frac{\left[h_i - (1/n)\right]/k}{(1-h_i)/(n-k-1)} \sim F_{k,(n-k-1)}$. Thus, any $F_i > F_{k,(n-k-i);1-\alpha/n}$ is worth examining (Appendix 9).

Cook's distance

Cook's distance measures the influence of an observation in the model. This statistic measures the change in the regression coefficients β_j when the i^{th} observation is removed from the data.

In general the statistic is proportional to a weighted average of the squared differences of β_j and $\beta_{j(-i)}$, and can be expressed in terms of leverages and studentized residuals

$$d_i = \left[\frac{1}{k+1}\right] r_i^2 \left[\frac{h_i}{1-h_i}\right] = \frac{e_i^2 h_i}{(k+1)S^2(1-h_i)}$$

The suggestion has been to check values with $d_i > 1$, but recent research has suggested that more sensitive approximations are necessary (see Table A-10 in the textbook).

Example: Calibration data

These ideas are illustrated with the following example from 17 concentrations of a pollutant (*X*) and the related readings of an instrument (*Y*).

The data are as follows:

```
. list x y

  1.         0       10.7
  2.        .5       14.2
  3.         1       16.7
  4.       1.5       19.1
  5.         2       24.9
  6.       2.5       25.4
  7.         3       32.3
  8.       3.5       30.8
  9.         4       39.6
 10.       4.5       30.3
 11.         5       37.2
 12.       5.5       37.8
 13.         6       37.5
 14.       6.5       38.6
 15.         7       42.6
 16.       7.5       44.3
 17.         8       37.2
```

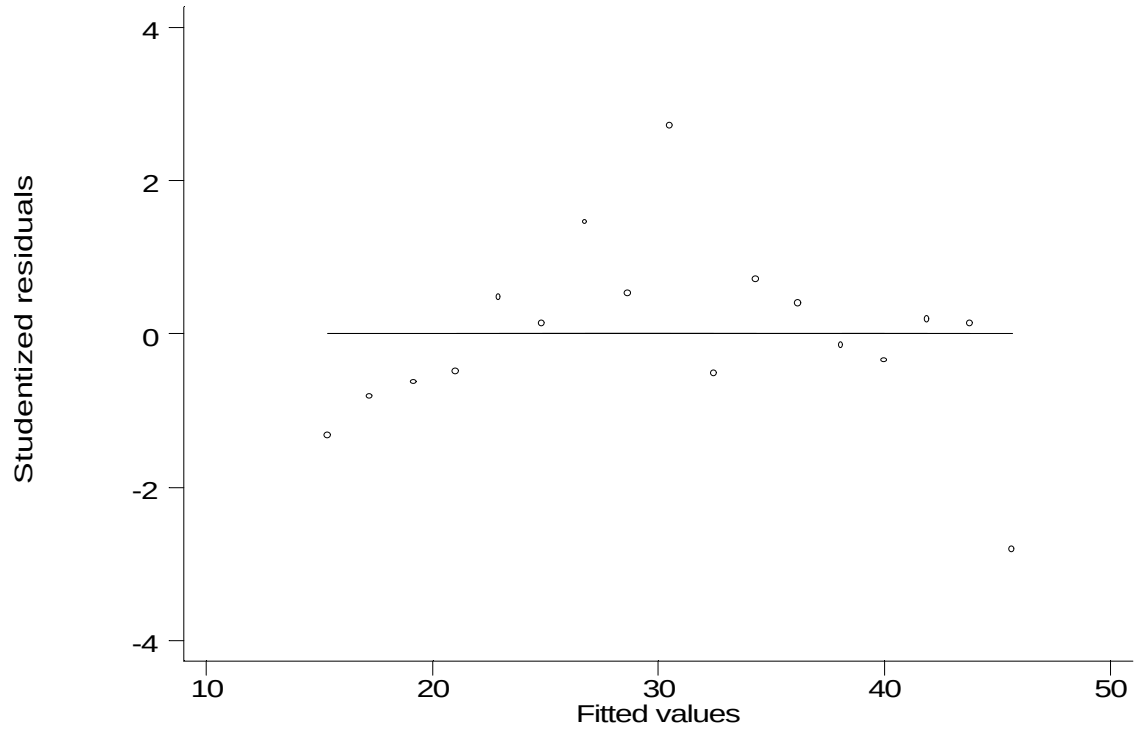
The regression model is as follows:

```
. reg y x

  Source |       SS       df       MS                  Number of obs =      17
---------+------------------------------                F(  1,    15) =   86.53
   Model |  1462.63781     1  1462.63781                Prob > F      =  0.0000
Residual |  253.543259    15  16.9028839                R-squared     =  0.8523
---------+------------------------------                Adj R-squared =  0.8424
   Total |  1716.18107    16  107.261317                Root MSE      =  4.1113

------------------------------------------------------------------------------
       y |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |   3.786765   .4070805      9.302   0.000       2.919093    4.654436
   _cons |   15.39412   1.909377      8.062   0.000       11.32438    19.46386
------------------------------------------------------------------------------
```

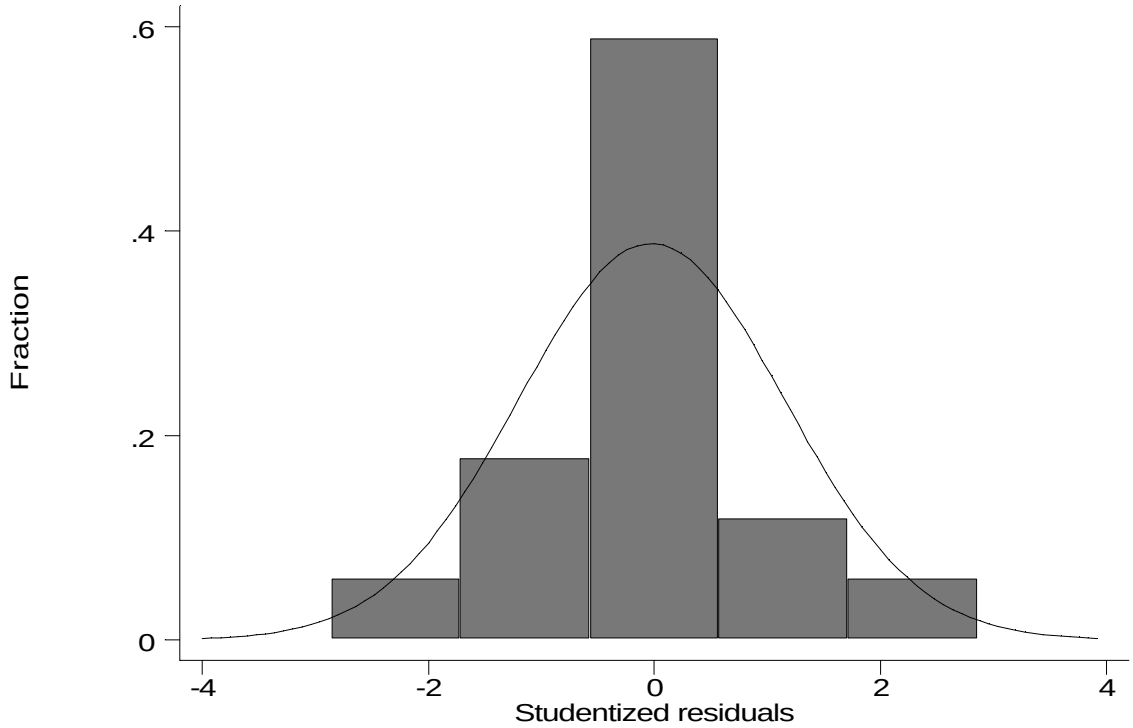
We determine the residuals and the fitted values by the commands

`. predict yhat`	**Fitted values** $\hat{Y}$
`. predict r, resid`	**Residuals** e_i
`. predict rstan, rstan`	**Standardized residuals**
`. predict rstud, rstud`	**Jackknife residuals**
`. predict d, cooksd`	**Cook's distance**
`. predict h, hat`	**Leverage (hat) values** h_i

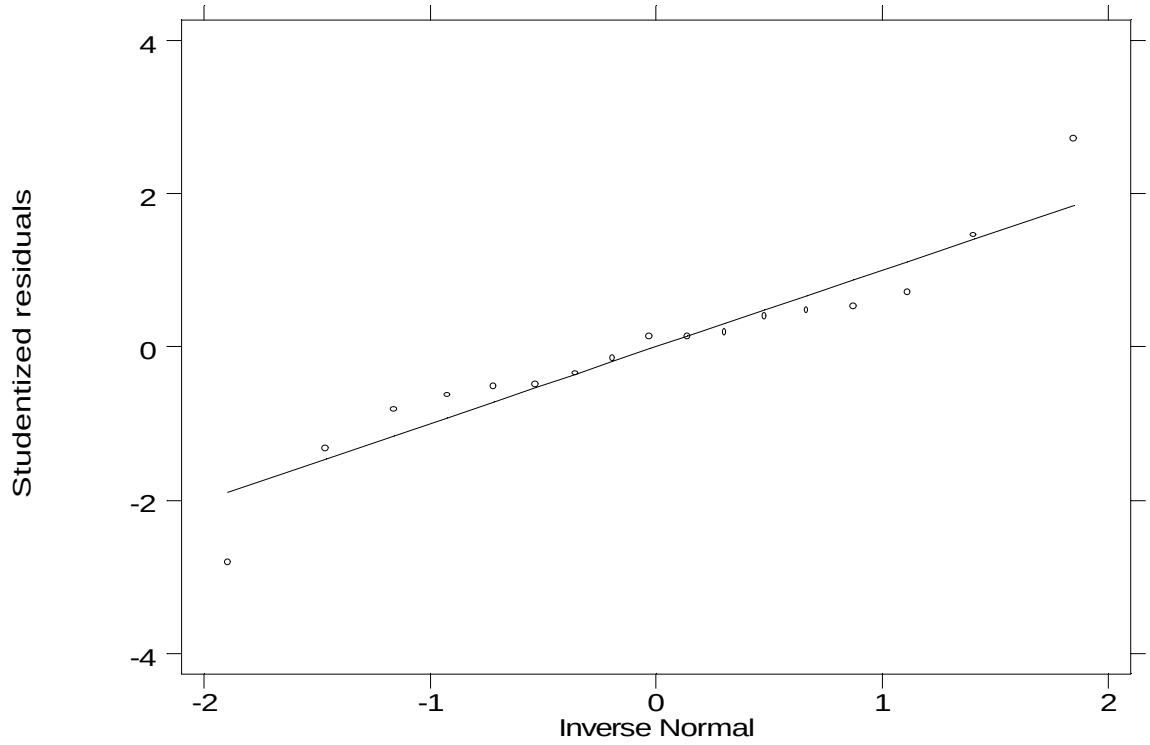
```
. gen zero=0

. graph rstud zero yhat, c(.l) s(oi) xlab ylab
```



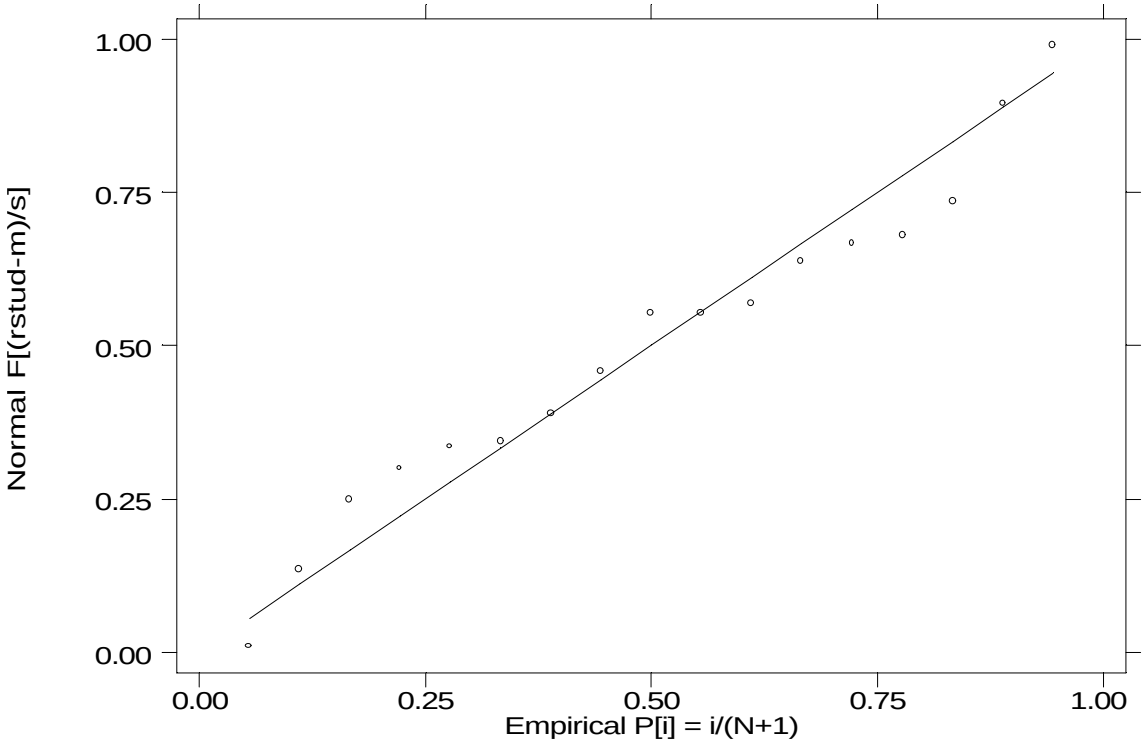
```
. graph rstud, normal bin(7) xlab ylab
```



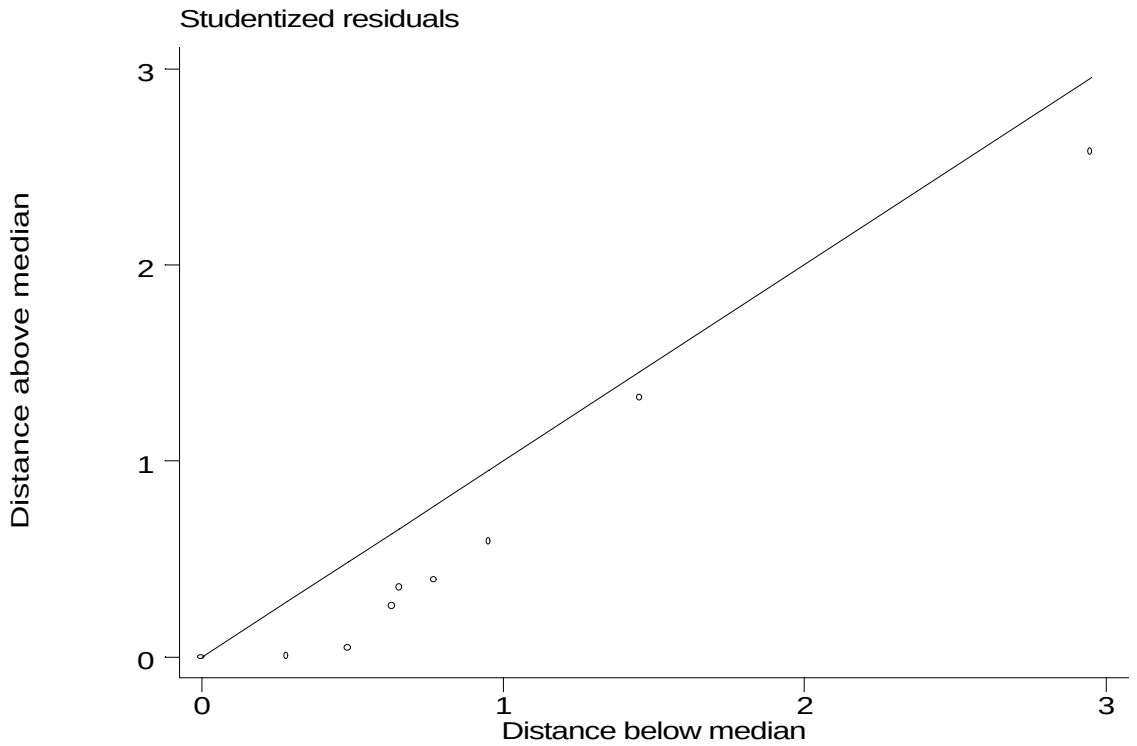
```
. qnorm rstud, xlab ylab
```



```
. pnorm rstud
```



```
. symplot rstud, xlab ylab
```



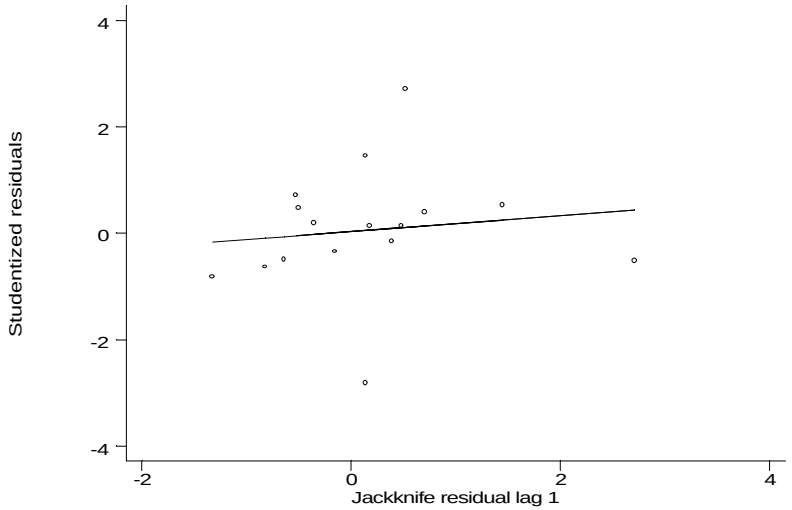
Normality tests

```
. summarize rstud

Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+--------------------------------------------------------
   rstud |      17   -.0247949   1.175543   -2.82039   2.708857


. ksmirnov rstud=normprob((rstud-(-.0247949))/ 1.175543)

One-Sample Kolomogorov-Smirnov test against theoretical distribution
           normprob((rstud-(-.0247949))/ 1.175543)

 Smaller group       D       P-value  Corrected
 ---------------------------------------------
 rstud:               0.1478    0.476
 Cumulative:         -0.1318    0.554
 Combined K-S:        0.1478    0.852      0.781

. swilk rstud
                    Shapiro-Wilk W test for normal data
 Variable |    Obs           W         V          z   Pr > z
 ---------+-----------------------------------------------------
    rstud |     17     0.94660     1.128      0.240  0.40505
```

Test for autocorrelation.

The Durbin-Watson statistic is (from the command regdw) d=1.37065. The upper and lower limits from the tables are given by d_L=1.13 and d_U=1.38. The test is inconclusive. A graphical depiction of the jackknife residual versus the residual lag-1 and the STATA commands are as follows:

```
. gen rstud_1=rstud[_n-1]
. quietly reg rstud rstud_1
. label var rstud_1 “Jackknife residual lag-1”
. predict rhat
. graph rstud rhat rstud_1, c(.l) s(oi) xlab ylab
```



Comments:

1. The regression model is highly significant ($p<0.001$).
2. There is somewhat of a dependence between residuals (with residuals in the middle being more likely above the reference line (zero line) than at the ends
3. The normality plots show good agreement with normality. This is reaffirmed by both the Kolmogorov-Smirnov test and the Shapiro-Wilks test
4. The Durbin-Watson test is inconclusive
5. The plot of the jackknife residuals versus the residuals lag-1 shows no evidence of autocorrelation.

```
. list y r rstan rstud cooksd h

             y          r      rstan      rstud          d          h

  1.      10.7  -4.694118  -1.289225  -1.320836   .2285388   .2156863
  2.      14.2    -3.0875  -.8287705  -.8196545    .074837   .1789216
  3.      16.7  -2.480881  -.6533803  -.6404042   .0368022   .1470588
  4.      19.1  -1.974264  -.5119266  -.4989459   .0178849    .120098
  5.      24.9   1.932353    .494894   .4820648   .0133109   .0980392
  6.      25.4   .5389705   .1367411   .1321869   .0008227   .0808824
  7.      32.3   5.545588   1.397673   1.447846   .0719706   .0686275
  8.      30.8   2.152205   .5402988   .5271329   .0095275   .0612745
  9.      39.6   9.058823   2.271202   2.708857   .1611987   .0588235
 10.      30.3  -2.134559  -.5358688  -.5227261   .0093719   .0612745
 11.      37.2    2.87206   .7238547   .7118535    .019304   .0686275
 12.      37.8   1.578676   .4005227   .3890275   .0070584   .0808824
 13.      37.5  -.6147053  -.1574319  -.1522195    .001347   .0980392
 14.      38.6  -1.408089  -.3651175   -.354315   .0090978    .120098
 15.      42.6   .6985285   .1839688   .1779316   .0029176   .1470588
 16.      44.3    .505147   .1355954    .131078   .0020033   .1789216
 17.      37.2  -8.488234  -2.331267   -2.82039   .7472858   .2156863
```

Comments:

1. From Table A-8(a), the critical value for the jackknife residuals and a=0.05, k=1, and n=17 is between 3.65 (n=15) and 3.54 (n=20). Since the absolute value of the largest jackknife residual is $|-2.82039|=2.82039$, there does not seem to be any problematic values in the data.
2. From Table A-8(b), the critical value for the studentized residuals and a=0.05, k=1, and n=17 is between 2.61 (n=15) and 2.77 (n=20). As the largest studentized residual is -2.331267, it does not appear to be any suspicious observation in the data.
3. No Cook's distance is larger than the critical value calculated from Table A-10. If we divide the critical values for a=0.05, k=1, and n-k-1=15, by n-k-1, we see that the distances listed in the previous dataset must not be larger than 1.037 (=15.55/15). As the largest distance is 0.747, there are no suspiciously influential observations in the data.
4. Finally, the leverage values should not be larger than 2(2)/17=0.235. Again, the largest leverage is 0.216 which is not larger than would be expected.