Statistical Methods: Model Checking

http://www.stats.ox.ac.uk/~snijders/sm.htm

Tom A.B. Snijders

University of Oxford

November 15, 2011

Literature:

F.L. Ramsey & D.W. Schafer, *The Statistical Sleuth*, 2nd edition.
Pacific Grove: Duxbury, 2002.
Chapter 11.

W.N. Venables and B.R. Ripley, *Modern Applied Statistics with S*, 4th Edition.
New York: Springer, 2002.
Section 6.3.

If you wish further background material:
A.C. Atkinson, *Plots, Transformations and Regression* (Oxford, 1985).
J. Fox, *An R and S-Plus Companion to Applied Regression* (Thousand Oaks, Sage, 2002) Chapter 6; or its successor,
J. Fox and S. Weisberg, *An R Companion to Applied Regression*,
2nd Edition. Newbury Park, CA: Sage, 2011.

The best concise overview is the chapter by Fox, or Fox & Weisberg.

Outliers

Outliers are, loosely speaking, data points with very large absolute deviations from the rest of the data.

Often, this concept is used in relation to a model: outliers then are data points that are at far away from what would be expected under the model; often, this means they have large absolute residuals.

Outliers are important because, when undetected, they may wreck the statistical analysis.

The occurrence of outliers (or too many of them) often is a sign of violations of the model.

Question.

For an i.i.d. sample X of size n from $\mathcal{N}(\mu, \sigma^2)$,
how many data points X_i do you expect approximately with

$$\frac{|\, X_i - \bar{X} \,|}{S} > 3 \text{ , or } > 4 \text{ ,}$$

(where $\bar{X}$ is the sample mean and S the sample s.d.)
do you expect when $n = 50$?
And when $n = 1000$?

In linear and generalized linear models, we must distinguish between

1. outliers in the space of explanatory variables
2. outliers in the space of dependent variables.

Often there is not a particular statistical model for the distribution of the explanatory variables, so that the first type mostly does not refer to a model.

It is always advisable to check for the occurrence of outliers.

$\Rightarrow$ Perhaps they are erroneous values.
If this is evident, they may be corrected or discarded.

It is always advisable to check for the occurrence of outliers.

$\Rightarrow$ Perhaps they are erroneous values.
If this is evident, they may be corrected or discarded.

$\Rightarrow$ Perhaps they point to incorrectness of the model,
and under a better model
they may not be outliers any more.

It is always advisable to check for the occurrence of outliers.

$\Rightarrow$ Perhaps they are erroneous values.
If this is evident, they may be corrected or discarded.

$\Rightarrow$ Perhaps they point to incorrectness of the model,
and under a better model
they may not be outliers any more.
(*What kind of better model?*)

It is always advisable to check for the occurrence of outliers.

- $\Rightarrow$ Perhaps they are erroneous values.
  If this is evident, they may be corrected or discarded.
- $\Rightarrow$ Perhaps they point to incorrectness of the model,
  and under a better model
  they may not be outliers any more.
  (*What kind of better model?*)
- $\Rightarrow$ If they remain unexplained outliers,
  it can be important to assess the sensitivity of the results
  for these outliers.

The Hat Matrix

In the case of the basic linear model,
outliers in the space of explanatory variables
can be found using the *hat matrix*. For the model

$$Y = X\beta + E \tag{1}$$

where the E_i are i.i.d. $\mathcal{N}(0, \sigma^2)$,
the OLS (and also ML) estimator for β is

$$\hat{\beta} = (X'X)^{-1}X'Y\ , \tag{2}$$

The Hat Matrix

In the case of the basic linear model,
outliers in the space of explanatory variables
can be found using the *hat matrix*. For the model

$$Y = X\beta + E \tag{1}$$

where the E_i are i.i.d. $\mathcal{N}(0, \sigma^2)$,
the OLS (and also ML) estimator for β is

$$\hat{\beta} = (X'X)^{-1}X'Y\,, \tag{2}$$

so that the predicted values are

$$\hat{Y} = X\hat{\beta} = HY\,, \text{ where } H = X(X'X)^{-1}X'\,, \tag{3}$$

the *hat matrix*. (It puts the hat on Y.)

The Hat Matrix

In the case of the basic linear model,
outliers in the space of explanatory variables
can be found using the *hat matrix*. For the model

$$Y = X\beta + E \tag{1}$$

where the E_i are i.i.d. $\mathcal{N}(0, \sigma^2)$,
the OLS (and also ML) estimator for β is

$$\hat{\beta} = (X'X)^{-1}X'Y\,, \tag{2}$$

so that the predicted values are

$$\hat{Y} = X\hat{\beta} = HY\,, \text{ where } H = X(X'X)^{-1}X'\,, \tag{3}$$

the *hat matrix*. (It puts the hat on Y.) Further,

$$\text{cov}(Y - \hat{Y}) = \sigma^2(I - H)\,. \tag{4}$$

Define $h_i = H_{ii}$, the diagonal of the hat matrix;
this is also called the *leverage* of data point (X_i, Y_i).

Then, we see:

1. h_i expresses how strongly the data point (X_i, Y_i) itself determines the fitted value $\hat{Y}_i$.

Define $h_i = H_{ii}$, the diagonal of the hat matrix;
this is also called the *leverage* of data point (X_i, Y_i).

Then, we see:

1. h_i expresses how strongly the data point (X_i, Y_i) itself determines the fitted value $\hat{Y}_i$.
2. $\text{var}(Y_i - \hat{Y}_i) = \sigma^2(1 - h_i)$.

Define $h_i = H_{ii}$, the diagonal of the hat matrix;
this is also called the *leverage* of data point (X_i, Y_i).

Then, we see:

1. h_i expresses how strongly the data point (X_i, Y_i) itself determines the fitted value $\hat{Y}_i$.
2. $\text{var}(Y_i - \hat{Y}_i) = \sigma^2(1 - h_i)$.
3. It is not hard to prove that $\sum_i h_i = p$,
   where p is the rank of X; so their average is p/n.

Define $h_i = H_{ii}$, the diagonal of the hat matrix;
this is also called the *leverage* of data point (X_i, Y_i).

Then, we see:

1. h_i expresses how strongly the data point (X_i, Y_i) itself determines the fitted value $\hat{Y}_i$.
2. $\text{var}(Y_i - \hat{Y}_i) = \sigma^2(1 - h_i)$.
3. It is not hard to prove that $\sum_i h_i = p$,
   where p is the rank of X; so their average is p/n.
4. If c_i is the number of times that row X_i
   is being replicated in the design matrix,
   then $h_i \leq 1/c_i$.

Define $h_i = H_{ii}$, the diagonal of the hat matrix;
this is also called the *leverage* of data point (X_i, Y_i).

Then, we see:

1. h_i expresses how strongly the data point (X_i, Y_i) itself determines the fitted value $\hat{Y}_i$.
2. $\text{var}(Y_i - \hat{Y}_i) = \sigma^2(1 - h_i)$.
3. It is not hard to prove that $\sum_i h_i = p$,
   where p is the rank of X; so their average is p/n.
4. If c_i is the number of times that row X_i
   is being replicated in the design matrix,
   then $h_i \leq 1/c_i$.

Large values of h_i are called *leverage points*.
Such points may lead to difficulties in any data analysis,
because they would require extrapolation of the model.

Residuals

The raw residual is

$$\hat{E}_i = Y_i - \hat{Y}_i \sim \mathcal{N}(\beta, \sigma^2(1 - h_i)) \tag{5}$$

if the model is correct
(with homoscedastic normally distributed E_i).

However, we do not know σ^2.
For standardization we need to estimate σ^2 .

Residuals

The raw residual is

$$\hat{E}_i = Y_i - \hat{Y}_i \sim \mathcal{N}(\beta, \sigma^2(1 - h_i)) \tag{5}$$

if the model is correct
(with homoscedastic normally distributed E_i).

However, we do not know σ^2.
For standardization we need to estimate σ^2 .

With the usual estimate $\hat{\sigma}^2 = RSS/(n - p)$, the value

$$r_i = \frac{Y_i - \hat{Y}_i}{\hat{\sigma}\sqrt{1 - h_i}} \tag{6}$$

is called variously the
standardized or *internally studentized* residual.

The deletion principle

It is more useful to work with $\hat{\sigma}^2_{(i)}$, defined as the estimator for σ^2 calculated from all data points excluding (X_i, Y_i). The value

$$t_i = \frac{Y_i - \hat{Y}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_i}} \tag{7}$$

is called variously the
studentized or *externally studentized* or *jackknifed* residual.
If the model is correct, $t_i \sim t(n - p - 1)$.

Studentized residuals are useful for indicating
outliers in the space of dependent variables.
For assessing heteroscedasticity, it is useful to work with
smoothed plots of t_i^2.

The deletion principle (continued)

This illustrates the *deletion* principle:

For detection of aberrant cases (or sets of cases),
it is better to estimate parameters with these cases deleted.

Cook's distance

The extent to which a data point (X_i, Y_i) is an outlier
with respect to the explanatory as well as dependent variables
an be expressed by *Cook's distance* (R.D. Cook, 1977).
This expresses how strongly (X_i, Y_i)
influences the OLS estimate $\hat{\beta}$.

Cook's distance

The extent to which a data point (X_i, Y_i) is an outlier
with respect to the explanatory as well as dependent variables
an be expressed by *Cook's distance* (R.D. Cook, 1977).
This expresses how strongly (X_i, Y_i)
influences the OLS estimate $\hat{\beta}$.

Recall that $\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2(X'X)^{-1})$.
Define $\hat{\beta}_{(i)}$ as the OLS estimate based on the data
from which case (X_i, Y_i) was deleted.
Then Cook's distance is defined by

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' X' X (\hat{\beta}_{(i)} - \hat{\beta})}{p\,\hat{\sigma}^2} = \frac{1}{p}\, r_i^2\, \frac{h_i}{1 - h_i}\,, \qquad (8)$$

where r_i is the standardized residual.

An even better method is to use external studentization:

$$\text{DFFITS}_i^2 = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' X' X (\hat{\beta}_{(i)} - \hat{\beta})}{\hat{\sigma}^2_{(i)}} = t_i^2 \frac{h_i}{1 - h_i}, \qquad (9)$$

where t_i is the studentized residual.

The more usual formula does not have the square:

$$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\hat{\sigma}_{(i)} \sqrt{h_i}} = t_i \sqrt{\frac{h_i}{1 - h_i}}, \qquad (10)$$

where $\hat{Y}_{i(i)} = H\hat{\beta}_{(i)}$ is the 'deletion fitted value' for Y_i.
This indicates how much using (X_i, Y_i) in the estimation increases the fitted value.

An even better method is to use external studentization:

$$\text{DFFITS}_i^2 = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' X' X (\hat{\beta}_{(i)} - \hat{\beta})}{\hat{\sigma}^2_{(i)}} = t_i^2 \frac{h_i}{1 - h_i}, \qquad (9)$$

where t_i is the studentized residual.

The more usual formula does not have the square:

$$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\hat{\sigma}_{(i)} \sqrt{h_i}} = t_i \sqrt{\frac{h_i}{1 - h_i}}, \qquad (10)$$

where $\hat{Y}_{i(i)} = H\hat{\beta}_{(i)}$ is the 'deletion fitted value' for Y_i.
This indicates how much using (X_i, Y_i) in the estimation increases the fitted value.

Note that these measures combine the magnitudes of the leverage and the residual.

Masking

A limitation of residuals and single case deletion
is that aberrant sets of cases may remain undetected,
because they mask each other's lack of fit.

Masking

A limitation of residuals and single case deletion
is that aberrant sets of cases may remain undetected,
because they mask each other's lack of fit.

Sometimes it can be meaningful to omit groups of cases,
or analyze meaningful subsets of the data separately,
to try and become aware of such groups.

Masking

A limitation of residuals and single case deletion
is that aberrant sets of cases may remain undetected,
because they mask each other's lack of fit.

Sometimes it can be meaningful to omit groups of cases,
or analyze meaningful subsets of the data separately,
to try and become aware of such groups.

Another procedure is to use *robust estimation methods*,
which produces results which are less sensitive
for even a larger fraction of outliers;
comparing residuals from an OLS and a robust fit
can help spotting aberrant sets of cases.

Partial Residual plots

Denote the variables comprised in X by X_j $(j = 1 \ldots, p)$.

The *partial residual plot* can be useful to assess
whether it is useful to add an extra variable W to the model,
while also assessing
the possible non-linearity of the dependence of Y on W.

For an explanatory variable W, not included among the X_j,
let E_i be the residual for case i
in the model with explanatory variables X (not W).

The plot of E_i against W_i is the *partial residual plot*
and will be a straight line + scatter
if there is a linear effect of W on Y, controlling for X.

Partial Residual plots

Denote the variables comprised in X by X_j ($j = 1 \ldots, p$).

The *partial residual plot* can be useful to assess
whether it is useful to add an extra variable W to the model,
while also assessing
the possible non-linearity of the dependence of Y on W.

For an explanatory variable W, not included among the X_j,
let E_i be the residual for case i
in the model with explanatory variables X (not W).

The plot of E_i against W_i is the *partial residual plot*
and will be a straight line + scatter
if there is a linear effect of W on Y, controlling for X.

Non-linearity also will show up (smooth the residuals!),
and the same procedure is useful for a W that is included
among the X_j to assess possible non-linearity of its effect.

Added Variable plots

An alternative method for the same purpose, but somewhat more computationally demanding, is the *added variable plot*.

Denoting by H the hat matrix for explanatory variables X, this is the plot of $E = (I - H)Y$ against $(I - H)W$: for W also its linear prediction based on X is subtracted.

This makes sense because the part of Y that covaries with HW is a linear function of X and therefore does not require W.

On the other hand, non-linear transformations of W will not show up as directly as in the partial residual plot.

Transformations of the dependent variable

Non-linear transformations of the dependent variable
may lead to simpler or more easily manageable models,
or may improve the quality of approximation of a linear model.

They may serve a variety of functions, e.g.,

- improve symmetry of distribution of residuals;
- improve homoscedasticity;
- improve linearity of the dependence on X.

A well-known example is the logarithmic transform,
often used for dependent variables representing money.
This transformation may correct high skewness
(but watch out for overcorrection!), and
changes multiplicative dependence into additive dependence.

Transformations and dispersion

To understand how a transformation affects dispersion, it is helpful to use the *delta method* :

Suppose that $\mathrm{E}(Y) = \mu$, $\mathrm{var}(Y) = \sigma^2$, and let f be a differentiable function. A first-order Taylor series gives

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

and therefore

$$\mathrm{var}\{f(Y)\} \approx \left(f'(\mu)\right)^2 \sigma^2 . \tag{11}$$

This approximation is reasonable if σ is small compared to the curvature of the function f.

Variance-stabilizing transformations

In many practical phenomena, there is a relation between the variance and the mean of random variables.

Examples for well-known families of distributions

- For the Poisson(λ) distribution, $\mathrm{var}(Y) = \lambda = \mathrm{E}(Y)$.
- For the binomial(n, p) distribution, $\mathrm{var}(Y) = np(1-p)$, $\mathrm{E}(Y) = np$ (n constant)
- For the Gamma(α, β) distribution, $\mathrm{var}(Y) = \alpha\,\beta^2$, $\mathrm{E}(Y) = \alpha\beta$ (often α constant) (recall that $\sigma^2\chi^2(n) = \mathrm{Gamma}(n/2, 2\sigma^2)$).

If we write $\mathrm{var}(Y) = \sigma^2(\mu)$ where $\mu = \mathrm{E}(Y)$, and f is a function with $f'(\mu) = c\,\sigma^{-1}(\mu)$, then f is a *variance-stabilizing transformation*:

$$\mathrm{var}\big(f(Y)\big) \approx c^2\ , \text{ a constant.}$$

Variance-stabilizing transformations (continued)

For some positive random variables Y (e.g., quantities),
the dispersion increases with the mean.
Here the variance may be stabilized by a concave increasing transformation, which has a decreasing derivative offsetting the increasing variance.

For example, for counts where the Poisson distribution is reasonable, the square root transform may be useful:

$$\frac{d\sqrt{\mu}}{d\mu} = \frac{1}{2\sqrt{\mu}}.$$

If variance $\sim$ mean^2 (scale parameters!),
the logarithmic transform may be useful:

$$\frac{d\log\mu}{d\mu} = \frac{1}{\mu}.$$

Box-Cox transformation

A family of transformations of positive variables
that connects various often-used transformations such as
$\log(Y)$, $1/Y$, $\sqrt{Y}$, etc., is the *Box-Cox family*

$$Y^{(\lambda)} = \begin{cases} \frac{Y^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \log(Y) & \lambda = 0 \end{cases} \tag{12}$$

The parameter λ can be estimated, e.g., by ML,
and this will be an attempt to obtain approximately symmetric
and homoscedastic residuals.

It may be helpful for interpretation to use a value of λ
that is a 'pretty number' rather than the precise ML estimate;
for the further inference about regression coefficients,
it is better not to take the estimated nature of λ into account.