Statistical Methods: Robustness

http://www.stats.ox.ac.uk/
Statistical Methods_MT2011_Snijders

Tom A.B. Snijders

University of Oxford

November 13, 2011

Literature:

W.N. Venables and B.R. Ripley, *Modern Applied Statistics with S*, 4th Edition.
New York: Springer, 2002.
Section 5.5.

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.

Assumptions

All inferential statistical methods are based on *assumptions* — but are they ever true?

Assumptions

All inferential statistical methods are based on *assumptions* — but are they ever true?

The saying

'All models are wrong; some models are useful'

is attributed to the statistician George Box.

http://en.wikiquote.org/wiki/Talk:George_E._P._Box

Assumptions

All inferential statistical methods are based on *assumptions* — but are they ever true?

The saying

'All models are wrong; some models are useful'

is attributed to the statistician George Box.

http://en.wikiquote.org/wiki/Talk:George_E._P._Box

Whether this makes sense depends on what you define as the 'truth' of a model.

A model is an image of reality, like a picture;
it is not equal to reality.

Assumptions

All inferential statistical methods are based on *assumptions* — but are they ever true?

The saying

'All models are wrong; some models are useful'

is attributed to the statistician George Box.

http://en.wikiquote.org/wiki/Talk:George_E._P._Box

Whether this makes sense depends on what you define as the 'truth' of a model.

A model is an image of reality, like a picture;
it is not equal to reality.

Only in mathematical derivations and simulation studies can we be sure of the model generating the observations.

Model assumptions such as linearity of regressions,
normal distributions, independence, all are made
to obtain simplified representations of reality
that are mathematically tractable.

Model assumptions such as linearity of regressions,
normal distributions, independence, all are made
to obtain simplified representations of reality
that are mathematically tractable.

Some assumptions are serious restrictions
of the validity of statistical inferences;
others are made for convenience,
and restrict inference only in rare cases.

Deviations from assumptions are serious
if they affect the performance of statistical procedures.

For parameter estimation and testing,
primary issues are:

1. type-I error rates of tests

For parameter estimation and testing,
primary issues are:

1. type-I error rates of tests
2. coverage rates of confidence intervals
   — which is practically the same

For parameter estimation and testing,
primary issues are:

1. type-I error rates of tests
2. coverage rates of confidence intervals
   — which is practically the same
3. reliability of standard errors
   — which also is practically the same,
   as standard errors are often used
   to make confidence intervals
   (if the estimator if approximate normally distributed).

Thus, the major question is not, whether assumptions are satisfied; but (in the first place) whether
tests for the parameters of interest
have approximately correct type-I error rates.

Thus, the major question is not, whether assumptions are satisfied; but (in the first place) whether tests for the parameters of interest have approximately correct type-I error rates.

For (generalized) linear models, the main parameters of interest are mostly (not necessarily) the regression coefficients β.

Thus, the major question is not, whether assumptions are satisfied; but (in the first place) whether
tests for the parameters of interest
have approximately correct type-I error rates.

For (generalized) linear models, the main parameters of interest are mostly (not necessarily) the regression coefficients β.

The proof of the pudding is in the approximate normality
(or *t* distribution)
of the standardized variables $(\hat{\beta}_k - \beta_k)/\text{s.e.}(\hat{\beta}_k)$.

The quality of the normal approximation is serious mainly for tail probabilities like 0.001, 0.01, 0.10, 0.90, 0.99, 0.999.

Example: t-test

Consider two samples $X_i,\ i = 1, \dots n_X$ and $Y_i,\ i = 1, \dots n_Y$ and denote $n = n_X + n_Y,\ \lambda = n_x/n$.
It is assumed that these are independent i.i.d. samples from distributions with finite variances.

The two-sample t-test statistic is

$$t = \frac{\overline{X} - \overline{Y}}{s.e.\left(\overline{X} - \overline{Y}\right)}$$

where for the classical t-test the standard error is the pooled standard error, assuming equal variances, but the default in R is the standard error for unequal variances combined with the Welch/Satterthwaite approximation for d.f.
(See keyword `var.equal` for `t.test`.)

The Law of Large Numbers implies

$$\mathcal{L}\Big(\sqrt{n}(\overline{X} - \overline{Y} - (\mu_X - \mu_Y))\Big) \to \mathcal{N}\Big(0, \frac{\sigma_X^2}{\lambda} + \frac{\sigma_Y^2}{1-\lambda}\Big) .$$

The delta method implies that for large n_X and n_Y,
$\sqrt{n} \times$ *s.e.* may be replaced by its probability limit.
This value is always asymptotically correct
for the unequal-variances standard error,
but for the pooled standard error only if $\sigma_X^2 = \sigma_Y^2$.

Therefore, t has an asymptotic standard normal distribution under $\mu_X = \mu_Y$, except when using the pooled standard error if the variances are different.

Therefore, t has an asymptotic standard normal distribution under $\mu_X = \mu_Y$, except when using the pooled standard error if the variances are different.

This is the mathematical reasoning showing that the t-test with the pooled s.e. is *robust* against deviations from normality; the t-test with the unequal-variances s.e. is also *robust* against unequal variances.
This comes at the price of a small loss of power for the case that actually the variances are equal.

Simulations can be used to show the same, but with more questionable generality.

The assumption of normal distributions of the populations is made only to be able to derive the *t*-test as an *optimal* test with an *exact t* distribution (if $\sigma_X^2 = \sigma_Y^2$);

for the *practical* validity of the *t*-test, one may say that the choice between the two variants depends on whether the null hypothesis is restricted or unrestricted,

$$H_0^{(r)} : \mu_X = \mu_Y,\ \sigma_X^2 = \sigma_Y^2 \qquad \text{or} \qquad H_0^{(u)} \mu_X = \mu_Y;$$

the *t*-test is robust against non-normality;
this test is in doubt only when there can be serious outliers (long-tailed distributions – note the finite variance assumption);
or when sample sizes are small
and distributions are far from normal.

. . . exercise . . .

Make a simulation study of the robustness of the *t*-test under various assumptions, and find

1. specifications with far from normal distributions where the *t*-test performs well;
2. specifications where the *t*-test performs poorly.

(Hint: for a more complicated simulation study, see `LM_Robustness.r`.)

Alternatives to the *t*-test

If the *t*-test is in doubt
(outliers or (small samples & serious non-normality)),
which two-sample tests can be used instead?

Alternatives to the t-test

If the t-test is in doubt
(outliers or (small samples & serious non-normality)),
which two-sample tests can be used instead?

1. Nonparametric test: Wilcoxon-Mann-Whitney
   `wilcox.test`
2. Permutation test: e.g., package `coin`

The null hypothesis for these tests is that
the two distributions are equal (not just their means),
but this distribution may have any shape.

Assumptions of Linear Models

The assumptions of the linear model:

1. $Y = X\beta + E$; X is fixed or conditioned upon.
2. The E_i are independent.
3. The E_i have normal distributions with expected values 0.
4. The E_i have constant variance.

How important are these?

Assumptions of Linear Models

The assumptions of the linear model:

1. $Y = X\beta + E$; X is fixed or conditioned upon.
2. The E_i are independent.
3. The E_i have normal distributions with expected values 0.
4. The E_i have constant variance.

How important are these?

1. Linear model: crucial, $\Rightarrow$ model specification.
   Minor deviations from linearity are tolerable.

Assumptions of Linear Models

The assumptions of the linear model:

1. $Y = X\beta + E$; X is fixed or conditioned upon.
2. The E_i are independent.
3. The E_i have normal distributions with expected values 0.
4. The E_i have constant variance.

How important are these?

1. Linear model: crucial, $\Rightarrow$ model specification.
   Minor deviations from linearity are tolerable.
2. Independent residuals: crucial;
   violations $\Rightarrow$ e.g., time series, multilevel models.

Assumptions of Linear Models

The assumptions of the linear model:

1. $Y = X\beta + E$; X is fixed or conditioned upon.
2. The E_i are independent.
3. The E_i have normal distributions with expected values 0.
4. The E_i have constant variance.

How important are these?

1. Linear model: crucial, $\Rightarrow$ model specification.
   Minor deviations from linearity are tolerable.
2. Independent residuals: crucial;
   violations $\Rightarrow$ e.g., time series, multilevel models.
3. Normally distributed residuals: not important as such;
   outliers can be risky, $\Rightarrow$ regression diagnostics;

Assumptions of Linear Models

The assumptions of the linear model:

1. $Y = X\beta + E$; X is fixed or conditioned upon.
2. The E_i are independent.
3. The E_i have normal distributions with expected values 0.
4. The E_i have constant variance.

How important are these?

1. Linear model: crucial, $\Rightarrow$ model specification.
   Minor deviations from linearity are tolerable.
2. Independent residuals: crucial;
   violations $\Rightarrow$ e.g., time series, multilevel models.
3. Normally distributed residuals: not important as such;
   outliers can be risky, $\Rightarrow$ regression diagnostics;
4. Deviations from constant variances (homoskedasticity) can
   be serious; however, this issue is often ignored.

'Robust' standard errors

Standard errors for regression coefficients that are robust for non-constant variances were developed by Huber and White.

These are implemented by the function `hccm` in `car` (hccm for *heteroscedasticity-corrected covariance matrix*). Also see `Anova` and `linearHypothesis`.

‘Robust’ standard errors

Standard errors for regression coefficients that are robust for non-constant variances were developed by Huber and White.

These are implemented by the function `hccm` in `car` (hccm for *heteroscedasticity-corrected covariance matrix*). Also see `Anova` and `linearHypothesis`.

Depending on the type of suspected heteroskedasticity, transformations and weighted LS may be alternatives.

See Fox (2002), Section 6.3.

The use of this ‘robust’ standard error in cases different from a well-specified linear model with heteroscedasticity is critically discussed in D.A. Freedman (2006), ‘On the so-called “Huber Sandwich estimator” and “robust standard errors”’. *The American Statistician*, 60, 299–302.

Robust estimators of location

The usual estimators of location and scale are the *mean* and the *standard deviation*.

These are highly sensitive to outliers: one observation can change the mean to anything, and change the s.d. to arbitrarily high values.

Robust estimators of location

The usual estimators of location and scale are
the *mean* and the *standard deviation*.

These are highly sensitive to outliers:
one observation can change the mean to anything,
and change the s.d. to arbitrarily high values.

The *median* is a location estimator that is much less sensitive.
Under normality, the mean is more efficient;
under long-tailed distributions, the median can be more efficient
(cf. Venables & Ripley, p. 121).

Robust estimators of location

The usual estimators of location and scale are the *mean* and the *standard deviation*.

These are highly sensitive to outliers:
one observation can change the mean to anything,
and change the s.d. to arbitrarily high values.

The *median* is a location estimator that is much less sensitive.
Under normality, the mean is more efficient;
under long-tailed distributions, the median can be more efficient (cf. Venables & Ripley, p. 121).

Another robust location estimator is the α–*trimmed mean*, where the fraction α lowest, and highest, sample elements are discarded.

(The median may be regarded as the 50% trimmed mean.)

Robust estimators of scale

One way to obtain a more robust scale estimator is to work not with squared deviations:

e.g., the *mean absolute deviation* from the mean

$$\frac{1}{n}\sum_{i=1}^{n} \mid X_i - \overline{X} \mid$$

or from the median

$$\frac{1}{n}\sum_{i=1}^{n} \mid X_i - \mathrm{median}_j(X_j) \mid \quad .$$

Robust estimators of scale

One way to obtain a more robust scale estimator is to work not with squared deviations:

e.g., the *mean absolute deviation* from the mean

$$\frac{1}{n}\sum_{i=1}^{n} | X_i - \overline{X} |$$

or from the median

$$\frac{1}{n}\sum_{i=1}^{n} | X_i - \mathrm{median}_j(X_j) | \ .$$

More usual robust estimators of scale are based on quantiles; the *interquartile range*

$$\mathrm{IQR} = X_{(3n/4)} - X_{(n/4)}$$

and the *mean absolute deviation*

$$\mathrm{MAD} = \mathrm{median}_i | X_i - \mathrm{median}_j(X_j) | \ .$$

Even if there are less than $n/4$ outliers,

the MAD will not be strongly affected.

We say that the MAD has a *breakdown point* of 25%.

Even if there are less than $n/4$ outliers,
the MAD will not be strongly affected.
We say that the MAD has a *breakdown point* of 25%.

For normal distributions, with $n \to \infty$,

$$\text{MAD} \to 0.6745\,\sigma, \qquad \text{IQR} \to 1.349\,\sigma\ .$$

Therefore often the values

$$1.4826\,\text{MAD} = \frac{1}{0.6745}\,\text{MAD} \quad \text{and} \quad 0.741\,\text{IQR} = \frac{1}{1.349}\,\text{IQR}$$

are used.

These robust estimators of location and scale illustrate two basic issues concerning robustness and sensitivity:

1. higher powers of deviations are more sensitive;
2. quantiles that are not in the tails are less sensitive.

Robust regression

Estimators for linear models (and glm) have been developed that aim to limit the influence of outliers.

These are called *robust regression methods*.

An important type of robust regression methods are M-estimators, minimizing

$$\sum_i \rho(Y_i - X_i\beta)$$

as a function of β, for a suitable function ρ.

$\rho(e) = e^2$ yields the LS estimator; robust estimators are obtained if for large e, $\rho(e)$ increases less than quadratically.

See Venables Ripley, Section 5.5.

R functions `huber`, `rlm`, `lqs`.

Concluding points

It is good to use methods that are not very sensitive to model assumptions that might be false.

However, we also wish to work with models in which we can have some confidence.
This gives greater scientific and practical insight, and potentially higher statistical efficiency (robustness is often bought at the cost of efficiency loss!).

Graphical explorations and diagnostic methods can help to improve the models being fitted.

Sometimes, robust estimates may be presented as such; in many cases, however, they are an intermediate step to diagnose sensitivity of the results to deviations from the model and thereby to diagnose the fit of the model.