AXEL GANDY

Approaches and Limitations of Model Checks

Models are used in almost all branches of science. Sometimes they are very well established, used over and over again, sometimes they are just *ad hoc* choices. Especially if the particular setting has never been investigated previously, a standard regression model is often used and it is clear that the model can only be a crude approximation of reality. However, especially in medical statistics, a wrong decision based on using an incorrect model can have serious real-life implications.

The purpose of this article is to give a brief non-technical introduction to how models can be checked using statistical methods. We shall consider what type of problems occur and point out the possibilities for direct model checks to detect discrepancies of particular interest.

Dr. Axel Gandy
Lecturer in Statistics,
Imperial College London, UK
E-mail: a.gandy@imperial.ac.uk
CAS Fellow 2005/2006

Throughout the article, we consider the following simple example: Consider repeatedly tossing a die with faces numbered from 1 to 6. Suppose one observes the following sample:

1,1,5,3,1,4,1,6,1,3,1,2

Does this sample come from independent tosses of a fair die? How can one test this? The term *fair* means that all numbers are equally likely, and *independent* means that the tosses do not influence one another. We shall only consider classical hypothesis testing. Other approaches, e.g. Bayesian approaches, will not be considered.

Statistical hypothesis testing

Before proceeding with this example, let us recall the basis for statistical tests. Typically, one wants to decide between two hypotheses: the null hypothesis (denoted by H_0) and the alternative hypothesis (denoted by H_1). Because the hypotheses are treated asymmetrically, we will see that the hypothesis that needs to be 'proven' should always be inserted into alternative H_1.

As an example, consider the case where one wants to decide between two treatments (OLD, NEW) for a certain disease. Here, the usual null hypothesis H_0 is that NEW is not better than OLD and alternative hypothesis H_1 is that NEW is better than OLD. This setting is called a non-inferiority study.

After observing a sample, one has to make a decision, either rejecting H_0 or not. The following table illustrates the possible situations:

	not reject H_0	reject H_0
H_0 correct	Correct	type I error
H_1 correct	type II error	Correct

A statistical test to a certain level α (typically α=5%) is a decision rule which states that the probability of rejecting H_0 if H_0 is true is not greater than α. More to the point, the probability of a type I error is bounded by α. Rejecting H_0 guarantees the low error bound α. Accordingly, rejecting H_0 is evidence in favour of H_1.

In contrast, there is no specific requirement attached to the type II error. Indeed, the probability of a type II error can generally only be bounded by 1 - α. This is because H_0 and H_1 need not be separated, as is the case in our example: It can be the case that NEW is only slightly better than OLD. Using the typical level α=5% it can be the case that even though H_1 holds true, the null hypothesis H_0 is only rejected with roughly 95% probability. Consequently, not rejecting H_0 is not 'proof' that H_0 is true.

Hence, the only hypothesis for which 'evidence' can be gathered is H_1. It should therefore contain the hypothesis one wants to 'prove'. The above example was intended to show that NEW is better than OLD and therefore this is chosen as alternative hypothesis H_1.

Goodness-of-fit testing

Goodness-of-fit testing is concerned with two hypotheses: a given model is either correct or it is wrong. By the above, we would like to use the hypothesis that the model is correct as alternative hypothesis H_1. However, this setup implies that the probability of accepting the model if it was true would only be α. This is because the probability of not rejecting the model if it were only slightly wrong would have to be bounded by α and thus, by continuity, the probability of accepting the model if it were true would only be α. As a result, the following setup has to be used:

H_0 : Model is correct, H_1 : Model is wrong.

Hence, the only information to be gained from a goodness-of-fit test is that the model is wrong. The model cannot be 'proven' to be correct.

The first example of a goodness-fit-test is the χ^2-test ascribed to Pearson (1900): It rejects the model if the squared difference between the observed frequencies and the expected frequencies is 'too large'.

In our example of tossing a die there were 12 tosses. Thus each number is expected to appear twice. The test statistic is therefore:

$$T=(O_1\text{-}2)^2+(O_2\text{-}2)^2+\ldots+(O_6\text{-}2)^2=20$$

where, for i=1,…,6, we let O_i denote how often the number i has been observed.

Is this value of T not consistent with the model? Assuming the model is true, one can compute the so-called p-value, which in this case is the probability of observing a value at least as large as 20 if the model is true. This can be done either by simulations or by large sample results. If the p-value does not exceed the level α of the test, the model is rejected. For the example, the p-value is 0.084 and thus the model is not rejected at the 5% level, but it is rejected at the 10% level.

Directed goodness-of-fit

The χ^2-test is only one of many tests that may be used to check the model of independent tosses of a fair die. Suppose you are playing a game in which high numbers are advantageous and your opponent always uses

one specific die. You suspect that this is not a fair die. Depending on the situation, other tests than the χ^2-test may be more suitable. Consider the following two situations:

1. Your opponent is a professional gambler: You are interested in finding out whether the die gives excessively high results.

2. In the attic, you find the die your grandparents always used when playing against you. Did they let you win by using a die that is biased towards low numbers?

Now we can use different test statistics. Again we reject for large values of those statistics:

1. Test statistic for the professional gambler: $(O_4\text{-}2)^+ + 2(O_5\text{-}2)^+ + 3(O_6\text{-}2)^+$

2. Test statistic for your grandparents: $3(O_1\text{-}2)^+ + 2(O_2\text{-}2)^+ + (O_3\text{-}2)^+$

where $x^+ = x$ if $x \geq 0$ and $x\text{^}+ = 0$ otherwise. In the first case, we reject if there are too many high numbers, whereas in the second case we reject if there are too many low numbers.

If we apply these tests to our example, we get the following result: If the test statistic for the professional gambler is used, we do not reject both at the 5% and at the 10% level. If the test statistic for the grandparents is used, we reject at the 5% level (p-value 0.011).

The χ^2-test could also be used in both these situations. The advantage of using the other tests is that they will result in less type II error for the alternatives of particular interest.

Limitations of model checks

Does

1,1,2,2,3,3,4,4,5,5,6,6

look too regular? It seems to violate the assumption that tosses are independent! However, the χ^2-test will not detect this, i.e. it does not check this assumption.

This is by no means a coincidence. There are theoretical results that basically say the following: For a given fixed sample size, any goodness-of-fit test only has high power against a finite-dimensional subspace, see e.g. Janssen (2003). In other words, there will always be certain model discrepancies which one specific test will fail to detect well.

Statistical model

Typically, models are not simple since they include unknown parameters. Accordingly, it is necessary to check whether a particular one of a certain set of models is correct. The basic idea is to pick one of the models first, typically the one that is 'most likely', and then compare the observed counts with the expected counts in the model selected.

Fisher (1924) realized that the way in which one picks one of the models has to be taken into consideration as it alters the distribution of the test statistic.

Model checking for regression models in event history analysis

One of the goals of event history analysis is to study the effects on survival of a number of variables (covariates). Consider the melanoma example also considered by Borgan in his contribution to this book. The covariates in the melanoma example are sex, tumour thickness, etc. Several regression models can be used, e.g. the proportional hazards model attributed to Cox or the additive model attributed to Aalen. Furthermore, a set of covariates has to be chosen for each of those models.

More than 20 model checks have been suggested for the Cox Model. The most popular idea is to start with the differences between the observed and the expected number of events per individual, and then to aggregate these differences.

Directed tests for regression models have recently been proposed. For example, it is possible to devise tests that are good at rejecting when another regression model is true, see Gandy (2006), Gandy and Jensen (2005a), Gandy and Jensen (2005b).

Alternative approach

There are other tests besides the χ^2-test, especially for continuous observations. Instead of mentioning them, reference is made to the classical book by D'Agostino and Stephens (1986).

There is one fundamentally different approach which tries to circumvent the problem entailed by the fact that models can only be proven wrong. It makes it possible to show that the chosen model is close to the true model: The basic idea is to define a certain distance M on the space of all models. The null hypothesis H_0 is that the true model is more than a given distance away from the model being checked. In this setup, a rejection means that the model is not further away than a given distance from the true model. However, this approach is more complicated to execute, and therefore rarely used in practice.

Remarks

All models are wrong, some models are useful. (G. E. P. Box)

In the example of a die, we are relatively certain about the model. The same applies to certain models in physics. In medical statistics, however, although it is safe to assume that the model is wrong, one still hopes that it approximates reality. Model checks are a useful and important tool for detecting whether a model is 'very' wrong.

References

D'Agostino, R. B. and Stephens, M. A. *Goodness-of-fit Techniques*. Marcel Dekker Inc., 1986.

Fisher, R. A. "The conditions under which χ^2 measures the discrepancy between observation and hypothesis". *Philosophical Magazine*, 87, 1924: 442–449.

Gandy, A. *Directed Model Checks for Regression Models from Survival Analysis*. Logos Verlag, Berlin. Dissertation, Universität Ulm, 2006.

Gandy, A. and Jensen, U. "Checking a semiparametric additive risk model". *Lifetime Data Analysis*, 11(4), 2005a: 451–472.

Gandy, A. and Jensen, U. "On goodness of fit tests for Aalen's additive risk model". *Scandinavian Journal of Statistics*, 32, 2005b: 425–445.

Janssen, A. "Which power of goodness of fit tests can really be expected: Intermediate versus contiguous alternatives". *Statistics & Decisions*, 21(4), 2003: 301–325.

Pearson, K. "On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". *Philosophical Magazine*, 50, 1900: 157–175.