

# LECTURE 12: MORE ON GOODNESS OF FIT AND MODEL CHECKING

Text reference: Shalizi Chapter 9

## Overview

How does one check how well a parametric model fits the data? There are several ways of doing this formally. The general idea is to carry out *hypothesis testing* by first specifying alternative models (one corresponding to the null hypothesis and the other to a more general hypothesis). Then test statistics are calculated which measure the “goodness of fit” or “lack of fit” (e.g. MSE or deviance) of each model and these are compared. Typically the model corresponding to the null hypothesis is simpler, so if it fits the data about as well as a more complex model it is usually preferred on the grounds of parsimony (i.e., we retain the null hypothesis).

[Dobson 2002] *“For generalized linear models, the two models should have the same probability distribution and the same link function but the linear component of one model has more parameters than the other. The simpler model, corresponding to the null hypothesis  $H_0$ , must be a special case of the other more general model. [...] goodness of fit statistics may be based on the maximum value of the likelihood function, the maximum value of the log-likelihood function, the minimum value of the sum of squares criterion or a composite statistic based on the residuals.”*

Hence, the process of goodness of fit tests can be summarized as follows:

1. Specify a model  $M_0$  corresponding to  $H_0$ . Specify a more general model  $M_1$  (with  $M_0$  as a special case of  $M_1$ )  
  

---

---

---
2. Fit  $M_0$  and calculate the goodness of fit statistic  $G_0$ . Fit  $M_1$  and calculate the goodness of fit statistic  $G_1$ .
3. Calculate the improvement in fit, usually  $G_1 - G_0$ .
4. Use the sampling distribution of  $G_1 - G_0$  (or some related statistic) to test the null hypothesis that  $G_1 = G_0$  against the alternative hypothesis  $G_1 \neq G_0$ .
5. If the hypothesis that  $G_1 = G_0$  is not rejected, then  $H_0$  is not rejected and  $M_0$  is the preferred model. If the hypothesis  $G_1 = G_0$  is rejected then  $H_0$  is rejected and  $M_1$  is regarded as the better model.

Alternatively, we can specify a "lack-of-fit" statistic, such as the in-sample  $MSE$  (i.e. the residual sum of squares  $RSS$ ) or the estimated deviance ( $D$ ) in likelihood ratio tests. The corresponding test statistics  $T$  would then be defined by:

---

---

Instead of carrying out hypothesis testing with  $p$ -values, one could also (in theory) compute *confidence intervals*.

---

---

Confidence intervals are in some sense more informative than hypothesis tests as the width of a confidence interval provides a measure of the precision with which inferences can be made [see also Wasserman's "All of Statistics" p.155-156]. Here's an illustration of scientific significance versus statistical significance from "All of Statistics":

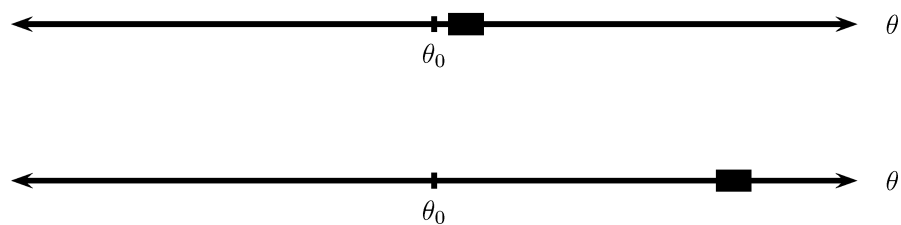


FIGURE 10.2. Scientific significance versus statistical significance. A level  $\alpha$  test rejects  $H_0 : \theta = \theta_0$  if and only if the  $1 - \alpha$  confidence interval does not include  $\theta_0$ . Here are two different confidence intervals. Both exclude  $\theta_0$  so in both cases the test would reject  $H_0$ . But in the first case, the estimated value of  $\theta$  is close to  $\theta_0$  so the finding is probably of little scientific or practical value. In the second case, the estimated value of  $\theta$  is far from  $\theta_0$  so the finding is of scientific value. This shows two things. First, statistical significance does not imply that a finding is of scientific importance. Second, confidence intervals are often more informative than tests.

Nevertheless, for both forms of inference (using confidence intervals or

hypothesis tests), **sampling distributions** are required.

---

---

---

---

So far, we have relied on standard normal theory or asymptotic results for MLE. What if the form of the sampling distribution can not be easily worked out?

*Think bootstrap.* We have used bootstrap before to assess the uncertainty of different statistics, and for computing confidence intervals; for example:

---

---

---

---

---

---

---

---

---

---

Similarly, one can also use a version of *parametric bootstrap* to test parametric regression specifications with nonparametric regression.

## Testing Parametric Regression Specifications with Nonparametric Regression

Refer to Shalizi Chapters 9 and 11 for details and examples; see e.g. Sec 9.1 and 11.6.

The basic idea is to test your parametric model against a nonparametric model (which by definition will include the parametric model), and estimate the null distribution of  $T$  by simulating data from the parametric model (i.e. it is a form of *parametric bootstrap*). As before, compute a p-value based on the estimated null distribution (from the parametric bootstrap samples) and the observed test statistic  $t$  (from the original sample).

A non-parametric model should match the training data (as measured by e.g. log-likelihood or deviance) at least as well as the parametric model, because the former is more flexible. When the parametric model is correct, however, the non-parametric model ends up approximating the parametric model, and the difference should be small and shrinking with sample size.

