INSTITUT FOR MATEMATISKE FAG

AALBORG UNIVERSITET

FREDRIK BAJERS VEJ 7 G ▪ 9220 AALBORG ØST
URL: www.math.auc.dk
E-mail: jm@math.aau.dk

Tlf.: 96 35 88 63
Fax: 98 15 81 29

Model checking based on p-values

Consider first a classical statistical setting with a parametric statistical model, with density $\pi(y|\theta)$ for a random variable Y and an unknown parameter θ. In order to check this model, suppose that we have specified a *test quantity*, that is a real function $t(y)$, where for specificity we assume that large values of the test quantity are critical for the model. To be more precise, suppose we have observed $Y = y$ (the data), let θ_0 denote the "true" value of θ, and consider the p*-value* defined by obtaining something more critical than we actually observe when $\theta = \theta_0$, that is

$$p = P(t(Y) \geq t(y)|\theta_0).$$

If Y is a continuous random variable, we have

$$p = \int_{t(y)}^{\infty} \pi(x|\theta_0)\,\mathrm{d}x.$$

If Y is a discrete random variable, the integral is replaced by a sum over all x with $t(x) \geq t(y)$ and $\pi(x|\theta_0) > 0$.

A small value of p (e.g. $p \leq 0.05$) is critical for the model, since this is equivalent to that $t(y)$ is large. However, since we don't know the true value of θ, the p-value may be unknown. Therefore, one usually replace θ_0 by an estimate $\hat{\theta}$, e.g. the *maximum likelihood estimate (mle)*, that is the value of θ which maximizes $\pi(y|\theta)$ (note that in general the mle may not exists or it may not be unique; we assume here that it exists and is unique). Thereby we obtain the estimated p-value

$$\hat{p} = P(t(Y) \geq t(y)|\hat{\theta}).$$

Still it may be hard or impossible to calculate this probability. Then we may approximate $\hat{p}$ by

$$\hat{p} \approx \frac{1}{k} \sum_{i=1}^{k} \mathbf{1}[t(Y_i) \geq t(y))]$$

using a sample $Y_1, \dots, Y_k$ obtained by a simulation from the estimated density $p(\cdot|\hat{\theta})$ of the observation model.

Exercise 1

Consider an observation model with iid Bernoulli trials $Y_1, \dots, Y_n$ and parameter $\theta \in (0,1)$ of success. In other words, the observation model for $Y = (Y_1, \dots, Y_n)$ has discrete density

$$\pi(y|\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^s(1-\theta)^{n-s}$$

where $s = \sum_{i=1}^n y_i$ is the number of successes. Moreover, suppose that we have $n = 20$ trials and data

$$y = (1,1,0,0,0,0,0,1,1,1,1,1,0,0,0,0,0,0,0,0).$$

The long sequences of zeros and ones in the data indicate that the model assumption of independent Y_i's is not true; there seems to be a positive autocorrelation. To quantify this, let the test quantity $t(y) = -\mathrm{switch}(y)$ denote minus the number of switches, i.e. $\mathrm{switch}(y) = 3$ for the data. Argue why large values of $t(y) = -\mathrm{switch}(y)$ are critical for the model. Show that $\hat{\theta} = s/n = 7/20$ is the mle, and calculate $\hat{p}$ from a sample of length $k = 1000$ from the estimated observation model.

Consider next a Bayesian model where θ is replaced by a random variable Θ with prior density $\pi(\theta)$. Assuming again that we have observed $Y = y$ from the observation model $\pi(y|\theta)$, the posterior density becomes

$$\pi(\theta|y) \propto \pi(\theta)\pi(y|\theta).$$

We define the *posterior predictive distribution* as the conditional distribution of (Θ', Y') given $Y = y$, where

(i) Θ' given $Y = y$ follows the posterior density $\pi(\cdot|y)$,

(ii) Y' given $\Theta' = \theta$ follows the density $\pi(\cdot|\theta)$ of the observation model,

(iii) conditional on Θ', we have that Y' is independent of Y

(usually, by the posterior predictive distribution is meant the conditional distribution of Y' given $Y = y$, but I find it more convenient to use the present definition). Thus the posterior predictive distribution has (conditional) density

$$\pi(\theta', y'|y) = \pi(\theta'|y)\pi(y'|\theta'),$$

and we refer to the posterior predictive distribution when we write $P((\Theta', Y') \in F|Y = y)$ for events F. Note that a simulation from the posterior predictive distribution is given by first generating Θ' from the posterior density $\pi(\cdot|y)$ and second generating Y' from the density $\pi(\cdot|\Theta')$ of the observation model.

Now, in order to check the Bayesian model, suppose that we have specified a *test quantity* $t(\theta, y)$, where for specificity we assume again that large values of the test quantity are critical for the model. Note that in contrast to the classical setting, we allow $t(\theta, y)$ to depend on θ. The *Bayesian p-value* is defined by obtaining something more critical under the posterior predictive distribution than we actually observe, that is

$$p = P(t(\Theta', Y') \geq t(\Theta', y)|Y = y).$$

Then a small value of p is critical for the Bayesian model, since this means that $t(\Theta', y)$ is likely to be large. As in classical statistics, it would not make much sense if $t(\theta, y)$ does not depend on y (because otherwise $p = 1$). In Exercise 2 below, $t(\theta, y) = t(y)$ depends only on y, while in Exercise 3, $t(\theta, y)$ depends on both θ and y. Note also that we have a more clear interpretation of the p-value in a Bayesian setting than in a classical setting, since we don't need to replace θ by an estimate.

In practice, we usually approximate p from a sample $(\Theta'_1, \mathbf{Y}'_1), \ldots, (\Theta'_k, \mathbf{Y}'_k)$ of the posterior predictive distribution, calculating

$$p \approx \frac{1}{k} \sum_{i=1}^{k} \mathbf{1}[t(\Theta'_i, Y'_i) \geq t(\Theta'_i, y))].$$

As above, for each iteration $i = 1, \ldots, k$, we simply first simulate Θ'_i from $\pi(\cdot|y)$ and second simulate Y'_i from $\pi(\cdot|\Theta'_i)$.

Exercise 2

Consider again an observation model with iid Bernoulli trials $Y_1, \ldots, Y_n$ and data as in Exercise 1, but impose a uniform prior on the probability Θ of success. Thus the prior density is $\pi(\theta) = 1$ for $0 < \theta < 1$, and the posterior density is

$$\pi(\theta|y) \propto \theta^s(1-\theta)^{n-s}$$

meaning that $\Theta|Y = y$ follows the Beta-distribution with parameters $s + 1$ and $n - s + 1$. Moreover, let still the test quantity be $t(\theta, y) = -\mathrm{switch}(y)$. Generate a sample of length $k = 1000$ from the posterior predictive distribution and calculate the p-value.

Exercise 3

Let the situation be as in Exercise 2, but define the test quantity by

$$t(\theta, y) = |\mathrm{switch}(y) - E[\mathrm{switch}(Y)|\Theta = \theta]|.$$

Show that $E[\mathrm{switch}(Y)|\Theta = \theta] = 2(n-1)\theta(1-\theta)$ and calculate the p-value from a sample of the posterior predictive distribution.