Goodness of Fit Tests

1 Introduction

Chapters 8, 9, 10 have some miscellaneous material about model checking. Also students have used some of these informal methods in earlier courses or earlier in this semester. Goodness of fit problems are generally of the form to test or check that data come from a particular statistical model or from any other model, perhaps with some restriction.

Often we assume that $X_i, i = 1, \ldots, n$ are iid from some normal distribution, but can we check or test this assumption. We might also wish to check if the r.v.s are consistent with an assumption that they are uncorrelated (this is not the same as testing they are independent), or perhaps some functions of the r.v.s are uncorrelated.

One may wish to check an assumption that r.v.s $X_i, i = 1, \ldots, n$ are iid from some Poisson distribution, versus an assumption that they are iid from any other form of distribution.

These notes discuss this idea, but not in a complete fashion.

Generally we test that X_i are iid with distribution F (or perhaps omf or pdf f) from a parametric family of distributions $\mathcal{F}$ versus a general alternative $F \in \mathcal{G} \backslash \mathcal{F}$, where $\mathcal{G}$ is the set of all possible distributions (of the right dimension and continuous or discrete form). Thus we are testing $H_0 : F \in \mathcal{F}$ versus$H_A : F \in \mathcal{G} \backslash \mathcal{F}$. This is a hypothesis test, but not of a parametric form. It is not easy to put this into a GLR form.

2 Pearson's chi-square Goodness of Fit Statistic

Earlier we studied the GLR for a multinomial model, usually against a general alternative, that is of the form $H_0 : \underline{p} \in \Theta_0$ versus $H_A : \underline{p} \in \Theta_A = \Theta_0^c$. We have a Theorem (see earlier notes) under the two assumptions of a regular statistical model, and for iid sampling, the r.v. $-2\log(\Lambda(\underline{X})$ converges in distribution to χ^2 as the sample size $n \to \infty$, and when this distribution is evaluated in the null hypothesis case. In some settings this is called the Pearson's chi-square goodness of fit statistic.

This terminology is used in the following sense. Amongst all possible multinomial distributions, a test is performed for a specific null hypothesis. If we accept the null hypothesis, then we decide that a special reduced form of a multinomial fits the data well. If we reject the null we decide this special form of the multinomial does not fit the data well. In other words we test if the special form of the multinomial as given by the null hypothesis is a good fit for the data, or in other words test the *goodness of fit* of the special form of the multinomial.

The Hardy Weinberg example is such a case. In this problem every trial of the experiment results in an outcome of a trinomial r.v., $M = 3$, with probability vector $\underline{p} = (p_1, p_2, p_3) \in \Theta = \mathcal{S}_3$, the simplex of order 3. A special form of this trinomial has $\underline{p} = ((1-\theta)^2, 2\theta(1-\theta), \theta^2)$ where $\theta \in [0, 1]$. This gives the null hypothesis

$$H_0 : \underline{p} \in \Theta_0 = \left\{((1-\theta)^2, 2\theta(1-\theta), \theta^2) : \theta \in [0, 1]\right\}$$

and the alternative

$$H_A : \underline{p} \in \Theta_A = \Theta_0^c \ .$$

Extension of Multinomial Test

We can use the same idea to test for the goodness of fit of a Poisson model. Here we have an experimental setting with iid r.v. $X_1, X_2, \ldots, X_n$ which come from a distribution with support N_0, the set of non negative integers. The parameter space is

$$\Theta = \left\{ \underline{p} = (p_0, p_1, p_2, \ldots) : p_i \geq 0, \sum_{i=0}^{\infty} p_i = 1 \right\} .$$

We have a null hypothesis which is that $\underline{p}$ is belongs to the set of Poisson distributions, that is it belongs to the set

$$\Theta_0 = \left\{ \underline{p} = (p_0, p_1, p_2, \ldots) : p_i = \frac{\lambda^i}{i!} e^{-\lambda}, \lambda > 0 \right\} .$$

Let $\hat{\lambda}$ be the MLE.

In this case we need to modify the application of the Pearson's chi-square multinomial goodness of fit test, since of course dimension$(\Theta) = \infty$. To do this we group the observations into M bins, $0, 1, \ldots, M-2, \geq M-1$. We then compare the *fitted* multinomial

$$\underline{p} = \left(p(0, \hat{\lambda}), p(1, \hat{\lambda}), \ldots, p(\geq M-1, \hat{\lambda}) \right)$$

against the *general multinomial* with

$$\underline{p} = (p_0, p_1, \ldots, p_{M-1}) \in S_M$$

the simplex of order M. This gives rise to the Pearson's chi-square, which under the null hypothesis has a limit $\chi^2_{(M-1-1)}$ distribution. Also note that

$$p(\geq M-1, \hat{\lambda}) = 1 - \sum_{k=1}^{M-2} p(k, \hat{\lambda}) .$$

How does one choose M? In a study outside this course, it has been found that the null distribution approximation works well provided that

$$\hat{E}_0(N_k) = np(k, \hat{\lambda}) \geq 5$$

for all the categories. Also one typically wants to use several categories, maybe 5 to 10 or 20, but not a large number of categories. These are only guidelines, obtained by simulation studies and numerical analysis.

In some applications one uses categories that might not correspond to bins $0, 1, \ldots, M-1$. Perhaps one might decide to include 0, 1 into a single bin.

This procedure generally works well in practice.

3 Test for Normality

Consider the setting where $X_1, \ldots, X_n$ are iid with some distribution F, assumed to be continuous, so that F has pdf f. We consider

$$H_0 : f \in \left\{ f = N(\mu, \sigma^2) : \mu \in \mathcal{R}, \sigma^2 > 0 \right\} \equiv \Theta_0$$

versus

$$H_A : F \in \Theta_0^c$$

where $\Theta =$ the set of all continuous distributions on $\mathcal{R} = (-\infty, \infty)$.

Formal or informal tests are used to test this type of hypothesis.

The informal or graphical test that is commonly used is a normal quantile-quantile plot, or qqnorm plot (in R terminology). These plots can be made formal, but putting confidence bands about the qqline, but this involves a topic of Brownian motion and stochastic process. This is rarely done in practice as it is not implemented in most software.

Another test that is used is the Jarque-Bera test, which is a sum of the squares the sample skewness and sample kurtosis. It has a limiting $\chi^2_{(2)}$ distribution.

$$JB = n \left(\frac{1}{6} \left(\frac{\hat{\mu}_3}{\hat{\sigma}^3} \right)^2 + \frac{1}{24} \left(\frac{\hat{\mu}_4}{\hat{\sigma}^4} - 3 \right)^2 \right)$$

which is a linear combination of squares of the sample skewness and centred sample kurtosis. If X_i are iid normal (from any normal distribution) then the population skewness is 0 and the population kurtosis is 3.

4 Empirical Distribution Function

For r.v.s $X_i, i = 1, \ldots, n$ the empirical distribution function is

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}(X_i \leq x) \ .$$

Notice that F_n has the properties of a cdf.

Various tests are constructed to examine how close F or $F(\cdot, \hat{\theta}_n)$ is to F_n. The Kolmogorov-Smirnov test statistic is the most commonly used. It is

$$KS = \sqrt{n} \max_{x \in R} |F(x) - F_n(x)| \tag{1}$$

if the null is a simple hypothesis (no parameters to be estimated), or is

$$KS = \sqrt{n} \max_{x \in R} \left| F(x, \hat{\theta}_n) - F_n(x) \right| \tag{2}$$

if parameters θ need to be estimated.

The distribution of the KS in (1) has a limit distribution form that is tabulated in the this *simple null hypothesis* case. It is also implemented in R and many other statistical languages and packages, and hence KS is calculated and critical values are provided. In the case of (2) packages do not generally implement this, and so one might need to write code to calculate KS and some other method, such as Monte Carlo or simulation methods, to obtain the appropriate critical value.