An updated review on Goodness–of–fit tests for regression models with some recent results

Wenceslao González Manteiga and Rosa M. Crujeiras

Statistics and Operations Research
Universidade de Santiago de Compostela

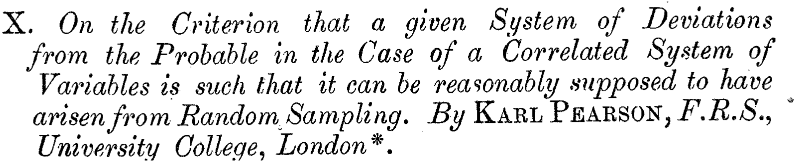
Motivation: the χ^2 test

Goodness–of–fit tests for regression

Tests based on the estimation of the regression function

Calibration, size and power

Some extensions (models, data, contexts)



X. *On the Criterion that a given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from Random Sampling. By* KARL PEARSON, *F.R.S., University College, London**.

Pearson, K. (1900)
On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling.
Philosophical Magazine, Vol. L, 157-175.

Parametric tests

Let X denote the variable of interest with distribution F_θ, $\theta \in \Theta$. We may be interested in testing:

- $H_0 : \theta = \theta_0$, vs. $H_a : \theta \neq \theta_0$
- $H_0 : \theta \in \Theta_0$, vs. $H_a : \theta \in \Theta_1$
- $\ldots$

Nonparametric tests (Specification tests)

NP tests are concerned with structural hypothesis about X, without imposing a parametric underlying distribution:

- $H_0 : F = F_0$, vs. $H_a : F \neq F_0$
- $H_0 : F \in \mathcal{F}_\theta$, vs. $H_a : F \notin \mathcal{F}_\theta$
- $\ldots$

Goodness–of–fit tests

The term *goodness–of–fit* (GoF) was introduced by Pearson at the beginning of the 20th century and refers to statistical tests that check the quality of a distribution's fit to a set of data.

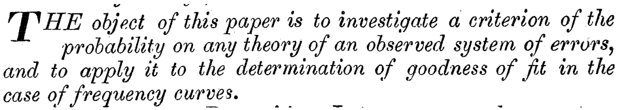
Nowadays, GoF is not only used for distribution problems, but in more general contexts such as regression.

The χ^2 idea (χ^2 tests)

The general idea of χ^2 test consists of classifying the possible values of the theoretical distribution in groups/bin/classes and compare, in each of them, the actual observed number of data points with the expected quantity under the null hypothesis.

$$T_n = \sum_k \frac{(O_k - E_k)^2}{E_k}$$

- Goodness–of–fit, homogeneity, independence.



THE object of this paper is to investigate a criterion of the probability on any theory of an observed system of errors, and to apply it to the determination of goodness of fit in the case of frequency curves.

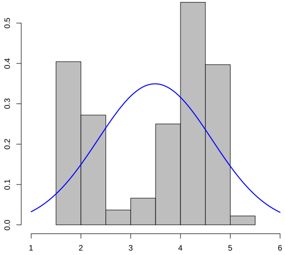


Figura: Histogram and normal density (f_0) for the eruptions time in Faithful Geyser dataset.

Density based χ^2 test

Denote by $I_k = [x_k, x_{k+1})$ and by $\hat{f}_{n,H}$ the histogram. For the testing problem $H_0 : \; f = f_0$ vs. $H_a : \; f \neq f_0$, the χ^2 test can be written in terms of the histogram $\hat{f}_{n,H}$ as follows:

$$\begin{aligned} T_n &= \sum_k \frac{(O_k - E_k)^2}{E_k} \\ &= n \sum_k \frac{\left[\int_{I_k} \left(\hat{f}_{n,H}(x) - f_0(x)\right) dx\right]^2}{\int_{I_k} f_0} \\ &\simeq n \left(\int \frac{\hat{f}^2_{n,H}(x)}{f_0(x)} dx - 1 \right) \end{aligned}$$

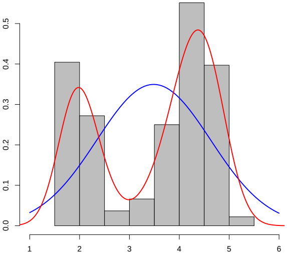


Figura: Histogram, kernel density estimator and normal density (f_0) for the eruptions time in Faithful Geyser dataset.

Other density based tests

Distance based tests:

$$T_n = d\left(\hat{f}, f_0\right)$$

$\hat{f}$ is a nonparametric estimator of f and d is a functional distance: $d(f,g) = \int |f-g|$ or $d(f,g) = \left[\int (f-g)^2\right]^{1/2}$, for instance.

Just in a while...

Testing problem (simple null hypothesis)

Let X denote the random variable of interest, with distribution F.

$$H_0 : \ F = F_0, \quad \text{vs.} \quad H_a : \ F \neq F_0$$

Empirical cumulative distribution function

From a random sample $X_1, \ldots, X_n$ of X, the empirical cumulative distribution function (ECDF) is defined as:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x) = \begin{cases} 0 & \text{if } x \in (-\infty, X_{(1)}) \\ \frac{k}{n} & \text{if } x \in [X_{(k)}, X_{(k+1)}) \\ 1 & \text{if } x \in [X_{(n)}, \infty) \end{cases}$$

Based on the empirical process...

$$\alpha_n(x) = n^{1/2}\left(F_n(x) - F_0(x)\right) = n^{-1/2}\sum_{i=1}^{n}\left(\mathbb{I}\left(X_i \leq x\right) - F_0(x)\right)$$

- KS test: $T_n = \sup_x |\alpha_n(x)| = \|\alpha_n(\cdot)\|_\infty$
- CvM test: $T_n = \int \alpha_n^2(x)dF_0(x)$

Asymptotic behaviour is determined by the continuous functional operating on a Gaussian limit process α.

Durbin, J. (1973)
Weak convergence of the sample distribution function when parameters are estimated.
Annals of Statistics, Vol. 1, 279–290.

Distribution based tests (composite H_0)

Empirical process with estimated parameters:

$$\alpha_n(x) = n^{1/2}\left(F_n(x) - F_{\hat{\theta}}(x)\right),$$

leading to a Kolmogorov-Smirnov test:

$$T_n = \sup_{x\in\mathbb{R}} n^{1/2}\left|F_n(x) - F_{\hat{\theta}}(x)\right|,$$

or to a Cramer-von Mises test:

$$T_n = \int \left[n^{1/2}(F_n(x) - F_{\hat{\theta}}(x))\right]^2 dF_{\hat{\theta}}(x).$$

The use of kernel density estimator for testing

Nonparametric kernel density estimators can be used as *pilot* estimations for parametric models. Let's start with a simple null hypothesis:

$$H_0: \ f = f_0, \quad \text{vs.} \quad H_a: \ f \neq f_0$$

$$T_n = \int \left[\sqrt{nh} \left(\hat{f}_{n,K}(x) - \mathbb{E}_{H_0}(\hat{f}_{n,K}(x)) \right) \right]^2 \omega(x) dx$$

where $\mathbb{E}_{H_0}$ denotes the expected value of $\hat{f}_{n,K}$ (a kernel density estimator) under the null hypothesis and ω is a weight function.

Bickel, P.J, and Rosenblatt, M. (1973)
On some global measures of the deviations of density function estimates.
Annals of Statistics, Vol. 1, 1071-1095.

Kernel estimator

$$\hat{f}_{n,K}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i),$$

with rescaled kernel

$$K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right)$$

(the density of a r.v. hX_K, being X_K a r.v. with density K).

Rosenblatt, M. (1956)
Remarks on some nonparametric estimates of a density function.
Annals of Statistics, Vol. 27, 832–837.

Parzen, E. (1962)
On estimation of a probability density function and mode.
Annals of Statistics, Vol. 44, 1065–1076.

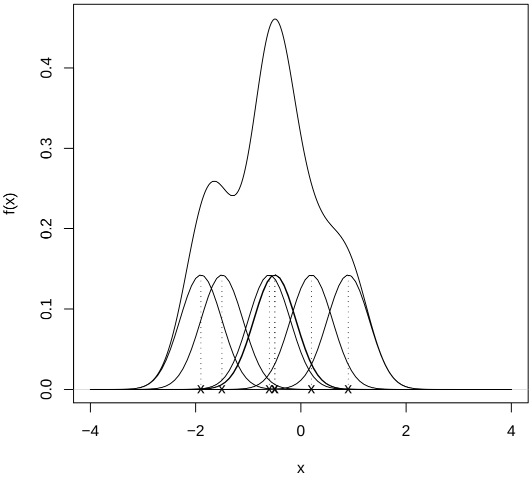


Figura: Construction of the KDE.

Asymptotic distribution (under H_0)

$$h^{-1/2}(T_n - \mu(K,\omega)) \to N(0, \sigma^2(K,\omega))$$

- Asymptotic mean:

$$\mu(K,\omega) = \left(\int f_0(x)\omega(x)dx\right)\left(\int K^2(x)dx\right)$$

- Asymptotic variance:

$$\sigma^2(K,\omega) = 2\left(\int (K*K)^2(x)dx\right)\left(\int \omega^2(x)f_0^2(x)dx\right)$$

The use of kernel density estimator for testing (II)

$$H_0: \ f \in \mathcal{F}_\theta, \quad H_a: \ f \notin \mathcal{F}_\theta$$

we may construct a test statistic measuring the *distance* between $\hat{f}_{n,K}$ and $f_{\hat{\theta}}$ (being $\hat{\theta}$ a suitable estimator). For instance:

$$T_n = nh \int \left(\hat{f}_{n,K}(x) - f_{\hat{\theta}}(x) \right)^2 \omega(x) dx$$

Bickel, P.J, and Rosenblatt, M. (1973)
On some global measures of the deviations of density function estimates.
Annals of Statistics, Vol. 1, 1071-1095.

General asymptotic structure

$$h^{-1/2}(T_n - \mu(K,\omega)) \to N\left(0, \sigma^2(K,\omega)\right)$$

- Asymptotic mean:

$$\mu(K,\omega) = \left(\int f_{\theta_0}(x)\omega(x)dx\right)\left(\int K^2(x)dx\right)$$

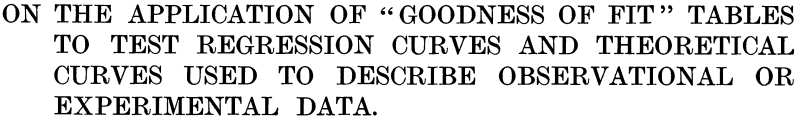
- Asymptotic variance:

$$\sigma^2(K,\omega) = 2\left(\int (K*K)^2(x)dx\right)\left(\int \omega^2(x)f_{\theta_0}^2(x)dx\right)$$

Bachmann, D. and Dette. H. (2005)
A note on the Bickel–Rosenblatt test in autoregressive time series
Statistics and Probability Letters, Vol. 74, 221-234.

Cao, R. and Lugosi, G. (2005)
Goodness–of–fit tests based on the kernel density estimates
Scandinavian Journal of Statistics, Vol. 32, 559–616.

Tenreiro, C. (2009)
On the choice of the smoothing parameter for the BHEP goodness–of–fit test
Computational Statistics and Data Analysis, Vol. 53, 1038–1053.



ON THE APPLICATION OF "GOODNESS OF FIT" TABLES TO TEST REGRESSION CURVES AND THEORETICAL CURVES USED TO DESCRIBE OBSERVATIONAL OR EXPERIMENTAL DATA.

Pearson, K. (1916)
On the application of goodness–of–fit tables to test regression curves and theoretical curves used to describe observational or experimental data.
Biometrika, Vol. 11(3), 239–261.

Regression model

Regression is usually formalized as the expected value of Y conditionally on the values of X:

$$m(x) = \mathbb{E}(Y|X = x), \quad \text{for each } x \in Supp(X)$$

Then, the response variable can be decomposed as:

$$Y = m(X) + \varepsilon$$

with ε an error variable with $\mathbb{E}(\varepsilon|X = x) = 0$.

González–Manteiga, W. and Crujeiras, R.M.(2013)
An updated review of Goodness-of-Fit tests for regression models.
Test, Vol. 22, 361-411.

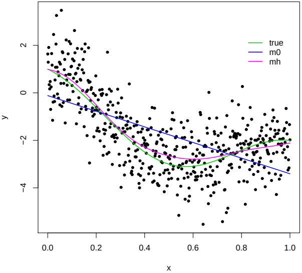
The kernel estimator

For each x, the kernel regression estimator could be constructed as a weighted average of observations Y_i, taking into account the distance of X_i to x:

$$m_{nh}(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i, \quad \sum_{i=1}^{n} W_{ni}(x) = 1$$

Weights $W_{ni}(x)$ will be constructed taking a kernel K and a bandwidth h.

- Priestley–Chao (fixed design)
- Nadaraya–Watson
- Local polynomial



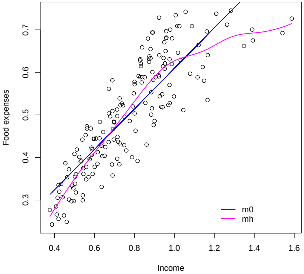
Model:

$$Y = 2x^2 - 5x + \cos(2\pi x) + \varepsilon$$

with $n = 500$ and $\varepsilon \sim N(0,1)$.

Test:

$$H_0 : \ m \ \text{ is linear}$$

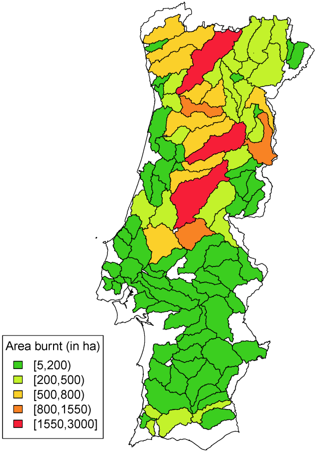


Test:

$$H_0 : \; m \text{ is linear}$$

Or generally:

$$H_0 : \; m \in \mathcal{M}_\theta = \{m_\theta, \; \theta \in \Theta\}$$

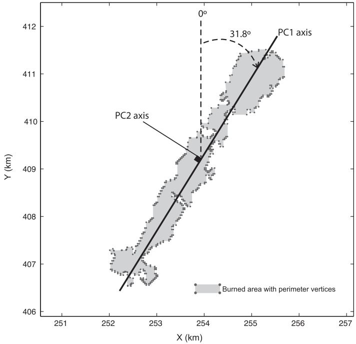


Regression with complex data

In forest fire modelling, there is an interest in relating fire orientation and size. Before stating a regression model (with fire orientation as explanatory variable), an independence test has been carried out.

E. García–Portugués et al. (2014)
A test for directional-linear independence, with applications to wildfire orientation and size.
SERRA, Vol. 28, 1261–1275.

H. Xu and F.P. Schoenberg (2011)
Point process modeling of wilfire hazard in LA county, CA.
Ann. Appl. Stat., Vol. 5, 684–704.



Fire orientation construction

- Data: orientations $(\mathbf{X})$ and log–burnt areas (Y) of 26870 wildfires in Portugal during 1985–2005.
- What is the relationship m between $\mathbf{X}$ and Y?

$$Y = m(\mathbf{X}) + \sigma(\mathbf{X})\varepsilon$$

- About m:
  - Estimate m nonparametrically.
  - Check if m can be specified as a certain parametric model.

Empirical processes

The initial empirical process, for the p-dimensional case in the explanatory variable, is given by:

$$\begin{aligned}
\overline{\alpha_n}(x) &= \sqrt{nh^p}\left(m_{nh}(x) - \mathbb{E}_{\hat{\theta}}(m_{nh}(x))\right) \\
&= \sqrt{nh^p}\sum_{i=1}^{n} W_{ni}(x)\left(Y_i - m_{\hat{\theta}}(X_i)\right) \\
&= \sqrt{nh^p}\sum_{i=1}^{n} W_{ni}(x)\hat{\varepsilon}_i
\end{aligned}$$

where $\mathbb{E}_{\hat{\theta}}$ is the estimation of $\mathbb{E}_{\theta_0}$ (with θ_0 theoretical parameter under H_0) and $\hat{\theta}$ is a $\sqrt{n}$-consistent estimator of θ_0 (for instance, least squares, maximum likelihood, ...).

Empirical processes (an example)

For instance, in order to test a polynomial regression model:

$$H_0 : \ m(x) = \sum_{j=1}^{q} \theta_j x^{j-1}, \quad \text{vs.} \quad H_a : \ m(x) \neq \sum_{j=1}^{q} \theta_j x^{j-1}$$

we may consider $T_n = \int \overline{\alpha_n}^2(x)\omega(x)dx$. It holds that:

$$h^{-1/2}(T_n - c_1) \xrightarrow{d} N(0, c_2)$$

where

$$c_1 = \int \widetilde{K}^2(x)dx \int \frac{\sigma^2(x)\omega(x)}{f(x)}dx, \quad c_2 = 2\int (\widetilde{K} * \widetilde{K})^2(x)dx \int \frac{\sigma^4(x)\omega^2(x)}{f^2(x)}dx$$

Smooth–based tests (I)

Consider $H_0 : m \in \mathcal{M}_\theta$. A first proposal:

$$d_1(m, H_0) = \int \left(\hat{m}(x) - \hat{m}_{\hat{\theta}}(x)\right)^2 \omega(x) dx,$$

where $\hat{m}_{\hat{\theta}}(x)$ is the local polynomial regression function from $\{(X_i, m_{\hat{\theta}}(X_i))\}_{i=1}^n$.

Härdle, W. and Mammen, E. (1993)
Comparing nonparametric versus parametric regression fits.
Annals of Statistics, Vol. 21, 1926–1947.

Smooth–based tests (II)

A second proposal:

$$d_2(m, H_0) = \frac{1}{n} \sum_{i \neq j} K_h(X_i - X_j)(Y_i - m_{\hat{\theta}}(X_i))(Y_j - m_{\hat{\theta}}(X_j))\omega(X_i).$$

Zheng, J.X. (1998)
A consistent test of functional form via nonparametric estimation techniques.
Journal of Econometrics, Vol. 75, 263–289.

Smooth–based tests (III)

A third proposal:

$$d_3(m, H_0) = \sum_{i=1}^{n} \left(Y_i - m_{\hat{\theta}}(X_i)\right)^2 \omega(X_i) - \sum_{i=1}^{n} \left(Y_i - \hat{m}_h(X_i)\right)^2 \omega(X_i).$$

Dette, H. (1999)
A consistent test for the functional form of a regression based on a difference of variance estimators.
Annals of Statistics, Vol. 27, 1012–1040.

Drawbacks of the smoothing approach

1. Bandwidth choice.
2. Slow rate of convergence of T_n to its normal limit.
3. Unknown curves involved in the test statistic requires estimation.

Model

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1, \ldots, n$$

where $\{\varepsilon_i\}$ is a sequence of i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables and the X_i have density support in $[0,1]$. Assume that:

$$\mathcal{M} = \{m \in L^2[0,1]; \int (m^{(k)}(x))^2 dx \leq c\}$$

Testing problem

$$H_0 : m(x) = \theta_0 + \theta_1 x \quad \text{vs.} \quad H_a : m(x) \neq \theta_0 + \theta_1 x$$

The loglikelihood associated with the previous model is given by:

$$l(m, \sigma) = -n \log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - m(X_i))^2$$

Log–likelihood, revisited

Denote by $\hat{\theta}_0$ and $\hat{\theta}_1$ the maximum likelihood estimators (MLE) under H_0 and $\hat{m}_{MLE}$ the MLE under $\mathcal{M}$. This estimator is the one that minimizes:

$$\sum_{i=1}^{n}(Y_i - m(X_i))^2 \quad \text{subject to} \int (m^{(k)}(x))^2 dx \leq c.$$

Then, $\hat{m}_{MLE}$ is the smoothing spline with smoothing parameter such that $\|\hat{m}_{MLE}^{(k)}\|^2 = c$. Therefore,

$$RSS_0 = \sum_{i=1}^{n}(Y_i - \hat{\theta}_0 - \hat{\theta}_1 X_i)^2, \; RSS_1 = \sum_{i=1}^{n}(Y_i - \hat{m}_{MLE}(X_i))^2.$$

Likelihood ratio test

$$\lambda_n = l(\hat{m}_{MLE}, \hat{\sigma}) - l(\hat{m}_0, \hat{\sigma}_0) = \frac{n}{2} \log \frac{RSS_0}{RSS_1},$$

with $\hat{\sigma}^2 = RSS_1/n$, $\hat{\sigma}^2 = RSS_0/n$ and $\hat{m}_0(x) = \hat{\theta}_0 + \hat{\theta}_1 x$.

A note of caution...

Although in this particular situation, the MLE exists under $\mathcal{M}$ models, the constant c is unknown and in many situations, $\hat{m}_{MLE}$ may not exist...

Generalized likelihood ratio test (GLRT)

The generalized likelihood ratio tests (GLRT) considers an estimator under $\mathcal{M}$ which may not coincide with the MLE, for instance, the local linear fit $\hat{m}_h$. In this way:

$$l(\hat{m}_h, \hat{\sigma}) = -\frac{n}{2}\log(RSS_1) - \frac{n}{2}\left(1 + \log\frac{2\pi}{n}\right),$$

$$l(\hat{m}_0, \hat{\sigma}_0) = -\frac{n}{2}\log(RSS_0) - \frac{n}{2}\left(1 + \log\frac{2\pi}{n}\right),$$

and the GLRT statistic is given by:

$$\lambda_n = l(\hat{m}_h, \hat{\sigma}) - l(\hat{m}_0, \hat{\sigma}_0) = \frac{n}{2}\log\frac{RSS_0}{RSS_1},$$

where

$$RSS_1 = \sum_{i=1}^{n}(Y_i - \hat{m}_h(X_i))^2.$$

Generalized likelihood ratio test (GLRT)

Under some regularity conditions, in Fan *et al.* (2001) it is proved that:

$$r_k \lambda_n \sim \chi^2_{\nu_n}, \quad \nu_n = \frac{r_k c_k |\Omega|}{h}$$

where $|\Omega|$ is the measure of the support of X, $r_k = c_k/d_k$, $c_k = K(0) - \frac{1}{2}\|K\|^2$ and $d_k = \|K - \frac{1}{2}K * K\|^2$.

- Fan, J., Zhang, C. and Zhang, J. (2001)
Generalized likelihood ratio statistics and Wilks phenomenon.
Annals of Statistics, Vol. 29, 153–193.

- Fan, J. and Jiang, J. (2007)
Nonparametric inference with generalized likelihood ratio tests.
Test, Vol. 16, 409–444.

Location–scale regression model

Assume that the regression model can be written in a location–scale form as

$$Y = m(X) + \sigma(X)\varepsilon,$$

with ε independent of X and with error distribution $F_\varepsilon(y) = \mathbb{P}(\varepsilon \leq y) = \mathbb{P}\left((Y - m(X))|\sigma(X) \leq y\right)$.

Location–scale regression model

If $\tilde{\theta}_0$ denotes the argument that minimizes $\mathbb{E}((m(X) - m_\theta(X))^2)$ over the parameter set $\Theta \subset \mathbb{R}^q$, then $m_{\tilde{\theta}_0}$ is the parametric model with minimum distance to m, and the error distribution under this model is built as

$$F_{\varepsilon_0}(y) = \mathbb{P}(\varepsilon_0 \leq y) = \mathbb{P}\left((Y - m_{\tilde{\theta}_0}(X))|\sigma(X) \leq y\right).$$

Hence, the null hypothesis $H_0 : m \in \mathcal{M}_\theta$ is true if and only if the error distributions F_ε and F_{ε_0} are the same.

The process

This result opens a way for GoF considering continuous functionals of the process $\{\hat{F}_{\varepsilon}(\cdot) - \hat{F}_{\varepsilon_0}(\cdot)\}$, where the estimators of the error distribution can be given by

$$\hat{F}_{\varepsilon}(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(\frac{Y_i - m_{nh}(X_i)}{\widehat{\sigma}(X_i)} \leq y\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\widehat{\varepsilon}_i \leq y)$$

and

$$\hat{F}_{\varepsilon_0}(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(\frac{Y_i - m_{\widehat{\theta}}(X_i)}{\widehat{\sigma}(X_i)} \leq y\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\widehat{\varepsilon}_{i0} \leq y)$$

The variance

And the variance estimator is given by

$$\widehat{\sigma}^2(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i^2 - m_{nh}^2(x)$$

being $\{W_{ni}\}_{i=1}^{n}$ a sequence of Nadaraya–Watson weights and $\widehat{\theta}$ a least squares estimator.

Van Keilegom, I., WGM and Sánchez–Sellero, C. (2008)
Goodness–of–fit tests in parametric regression based on the estimation of the error distribution.
Test, Vol. 17, 401–415.

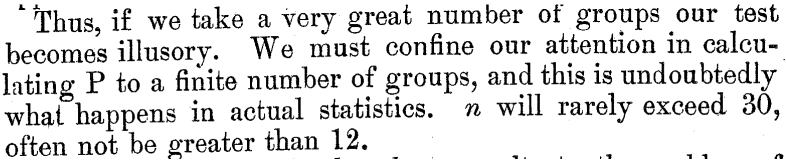
Neumeyer, N. and Van Keilegom, I. (2010)
Estimating the error distribution in nonparametric multiple regression with applications to model testing.
Journal of Multivariate Analysis, Vol. 101, 1067–1078.

The test

Based on the empirical distribution of the residuals, the Kolmogorov–Smirnov and Cramér–von–Mises tests are given by:

$$T_{nKS} = n^{1/2} \sup_{y \in \mathbb{R}} |\hat{F}_\varepsilon(y) - \hat{F}_{\varepsilon_0}(y)|, \quad \text{and} \quad T_{nCM} = n \int (\hat{F}_\varepsilon(y) - \hat{F}_{\varepsilon_0}(y))^2 d\hat{F}_{\varepsilon_0}(y).$$

From this methodology, a test for the error distribution can be also constructed, without further assumptions on m and σ, just comparing the empirical distribution of the residuals $\{\widehat{\varepsilon}_i\}_{i=1}^n$ with the one estimated under $H_0 : F_\varepsilon \in \mathcal{F}_\theta$.



Thus, if we take a very great number of groups our test becomes illusory. We must confine our attention in calculating P to a finite number of groups, and this is undoubtedly what happens in actual statistics. n will rarely exceed 30, often not be greater than 12.

Pearson, K. (1900)
On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling.
Philosophical Magazine, Vol. L, 157-175.

Motivation

A great deal of the theory developed during the nineties, considers tests statistics constructed from the comparison of a nonparametric estimator of the regression model and an estimator under the null hypothesis (that is, based on the $\overline{\alpha_n}$ process).

In this case, the curse of dimensionality as p increases, being p the dimension of the explanatory variable, can be appreciated.

Lavergne, P. and Patilea, V. (2008)
Breaking the curse of dimensionality in nonparametric testing.
Journal of Econometrics, Vol. 143, 103–122.

Xia, Y. (2009)
Model checking in regression via dimension reduction.
Biometrika, Vol. 96, 133–148.

Dimension reduction

Inspired on the projection pursuit ideas, the null hypothesis $H_0 : m \in \mathcal{M}_\theta$ is true if and only if $m = m_{\theta_0} \in \mathcal{M}_\theta$, and this is also equivalent to $\mathbb{E}(\varepsilon|X) = \mathbb{E}(\varepsilon_0|X) = \mathbb{E}(Y - m_{\theta_0}(X)|X) = 0$. In addition, this is also equivalent to:

$$\sup_{\beta,\, \|\beta\|=1} \sup_{\nu} |\mathbb{E}(\varepsilon|\beta^t X = \nu)| = 0 \Leftrightarrow \sup_{\beta,\, \|\beta\|=1} \mathbb{E}(\varepsilon \mathbb{E}(\varepsilon|\beta^t X)) = 0$$

under some regularity conditions, and this allows for the construction of sthe following test:

$$T_n = \sup_{\beta,\, \|\beta\|=1} \sum_{i<j} K_h(\beta^t(X_i - X_j))(Y_i - m_{\widehat{\theta}}(X_i))(Y_j - m_{\widehat{\theta}}(X_j)).$$

Dimension reduction

Another interesting idea consists in projecting the covariate X in the direction of $\beta = \beta_0$ such that this β_0 (with $\|\beta_0\| = 1$) minimizes

$$\mathbb{E}^2(\varepsilon - \mathbb{E}(\varepsilon|\beta^t X)) = \mathbb{E}^2(\varepsilon - m_\beta(X)),$$

the single–indexing procedure obtained through the corresponding empirical counterparts.

Dimension reduction

This enables to construct test statistics such as

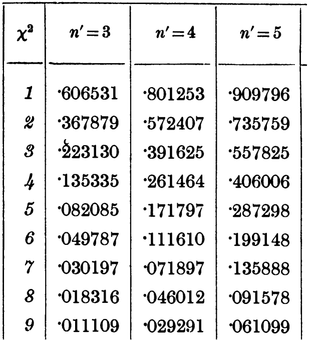
$$T_n = \frac{1}{n}\sum_{i=1}^{n} \omega(X_j)\left(\widehat{\varepsilon}_{j0} - \widehat{m}_{\widehat{\beta}_j}(\widehat{\beta}_j^t X_j)\right)^2$$

where

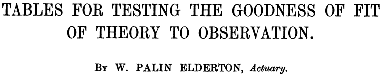
$$\hat{\beta}_j = \arg \underset{\beta,\,\|\beta\|=1}{\text{mín}} \sum_{i\neq j}\left(\widehat{\varepsilon}_{i0} - \widehat{m}_{\beta}^{j}(X_i)\right)^2, \quad j = 1,\ldots,n$$

being

$$\widehat{m}_{\beta}^{j}(x) = \frac{1}{n\hat{f}_{\beta}^{j}(X_j)}\sum_{i\neq j} K_h(\beta^t(x - X_i))\widehat{\varepsilon}_{i0},\ \hat{f}_{\beta}^{j}(x) = \frac{1}{n}\sum_{i\neq j} K_h(\beta^t(x - X_i)).$$



χ^2	$n'=3$	$n'=4$	$n'=5$
1	·606531	·801253	·909796
2	·367879	·572407	·735759
3	·223130	·391625	·557825
4	·135335	·261464	·406006
5	·082085	·171797	·287298
6	·049787	·111610	·199148
7	·030197	·071897	·135888
8	·018316	·046012	·091578
9	·011109	·029291	·061099



TABLES FOR TESTING THE GOODNESS OF FIT
OF THEORY TO OBSERVATION.

By W. PALIN ELDERTON, *Actuary.*

Elderton, W.P. (1902)
Tables for testing the goodness–of–fit of theory to observation.
Biometrika, Vol. 1, 155–163.

In the general testing problem:

$$H_0 : \ g \in \mathcal{G} = \{g_\theta\}_{\theta\in\Theta}, \quad \text{vs.} \quad H_a : \ g \notin \mathcal{G} = \{g_\theta\}_{\theta\in\Theta}$$

with test statistic $T_n = T(g_n, g_{\hat{\theta}})$ where g may be F_θ, f_θ or m_θ, calibration of critical points is crucial.

Critical point calibration

Estimate c_α such that:

$$\mathbb{P}_{H_0}(T_n \geq c_\alpha) = \alpha$$

How can we estimate c_α?

- Using the asymptotic normality (Case $g_n = f_{nh}$ or $g_n = m_{nh}$).
- Approximating the distribution of the empirical process α_n.
- Using Bootstrap: naive, wild,...

Partially linear model

$$H_{0PL}:\ \mathbb{E}(X|Y) = \theta_0' X_1 + m_2(X_2), \quad X = (X_1, X_2)$$

Test proposal:

$$T_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{i \neq j} K_h(X_i - X_j)\widehat{\varepsilon}_{i0}\widehat{\varepsilon}_{j0}\hat{f}_2(X_{2i})\hat{f}_2(X_{2j})$$

with $\varepsilon_{i0} = Y_i - \theta_0' X_{1i} - m_2(X_{2i})$.

Fan, Y. and Li, Q. (1996)
Consistent model specification tests: omitted variables and semiparametric functional forms.
Econometrica, 64, 865-890.

Simplified model

$$H_{0SM}: \mathbb{E}(X|Y) = m_1(X_1)$$

Test proposal:

$$T_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{i \neq j} K_h(X_i - X_j) \widehat{\varepsilon}_{i0} \widehat{\varepsilon}_{j0} \hat{f}_1(X_{1i}) \hat{f}_1(X_{1j})$$

with $\widehat{\varepsilon}_{i0} = Y_i - \widehat{m}_1(X_{1i})$.

Fan, Y. and Li, Q. (1996)
Consistent model specification tests: omitted variables and semiparametric functional forms.
Econometrica, 64, 865-890.

Single index model

$$H_{0SIM}:\ \mathbb{E}(X|Y)=\mathcal{H}(\theta_0'X)$$

with $\mathcal{H}$ unknown link. Test proposal:

$$T_n=\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{i\neq j}K_h(X_i-X_j)\widehat{\varepsilon}_{i0}\widehat{\varepsilon}_{j0}\hat{f}_{\hat{\theta}}(\hat{\theta}'X_i)\hat{f}_{\hat{\theta}}(\hat{\theta}'X_j)$$

with $\varepsilon_{i0}=Y_i-\mathcal{H}(\theta_0'X_i)$.

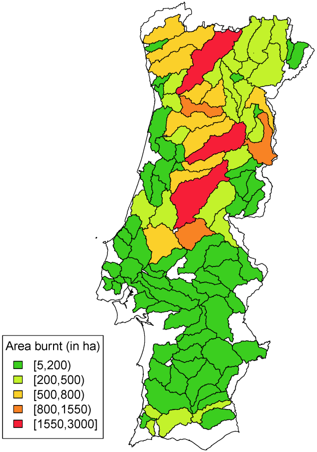
Futher models and methods

- Also for testing (generalized) additive models
- Tests for the variance function:
- Comparison of regression curves

Handling regression models with complex data

In some cases, the sample from the regression model may not provide complete information about the underlying population or may exhibit some complexities that should be captured in the modelling processes. GoF tests on regression models may take into account:

- Dependence (spatial and/or temporal)
- Censoreship and/or truncation
- Missing data
- Measurement errors
- Biased data



- Data: orientations $(\mathbf{X})$ and log–burnt areas (Y) of 26870 wildfires in Portugal during 1985–2005.
- What is the relationship m between $\mathbf{X}$ and Y?

$$Y = m(\mathbf{X}) + \sigma(\mathbf{X})\varepsilon$$

- About m:
  - Estimate m nonparametrically.
  - Check if m can be specified as a certain parametric model.

- Assessing a parametric model: $H_0 : m \in \mathcal{M}_\Theta = \{m_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$.
- The proposed test statistic is a smoothed weighted $\mathcal{L}^2$–distance between $\widehat{m}_{h,p}$ and $m_{\widehat{\boldsymbol{\theta}}}$:

Test statistic

$$T_n = \int_{\Omega_q} \left(\widehat{m}_{h,p}(\mathbf{x}) - \mathcal{L}_{h,p} m_{\widehat{\boldsymbol{\theta}}}(\mathbf{x})\right)^2 \widehat{f}_{n,L}(\mathbf{x}) w(\mathbf{x}) \omega_q(d\mathbf{x})$$

where $\mathcal{L}_{h,p} m(\mathbf{x}) = \sum_{i=1}^n W_{ni}(\mathbf{x})\, m(\mathbf{X}_i)$.

W. Härdle and E. Mammen (1993)
Comparing nonparametric versus parametric regression fits.
Ann. Statist., 21(4):1926–1947.

W. González-Manteiga and R.M. Crujeiras (2013)
An updated review of goodness-of-fit tests for regression models.
TEST, 22(3):361–411.

Theorem (Limit distribution of T_n)

Under $H_0 : m \in \mathcal{M}_\Theta$

$$h^{-q/2}\left(T_n - C(L,q)\int_{\Omega_q} \sigma^2_{\boldsymbol{\theta}_0}(\mathbf{x})w(\mathbf{x})\,\omega_q(d\mathbf{x})\right) \xrightarrow{d} \mathcal{N}\left(0, 2\nu^2_{\boldsymbol{\theta}_0}\right),$$

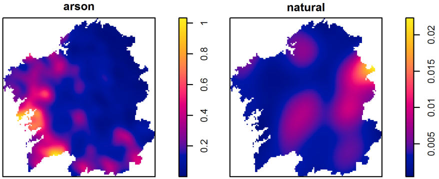
where $\sigma^2_{\boldsymbol{\theta}_0}(\mathbf{x}) = \mathbb{E}\left[(Y - m_{\boldsymbol{\theta}_0}(\mathbf{X}))^2 | \mathbf{X} = \mathbf{x}\right]$*.*

- If L is the von Mises kernel,

$$\nu^2_{\boldsymbol{\theta}_0} = (8\pi)^{-\frac{q}{2}} \int_{\Omega_q} \sigma^4_{\boldsymbol{\theta}_0}(\mathbf{x})w(\mathbf{x})^2\,\omega_q(d\mathbf{x}).$$

- Results under local alternatives.
- Calibration in practice by a consistent bootstrap procedure.

García-Portugués, E., Van Keilegom, I., Crujeiras, R.M. and González-Manteiga, W. (2016)
Testing parametric models in linear-directional regression.
Scandinavian Journal of Statistics, Vol. 43, 1178–1191.



Wildfire patterns

- Do arson and natural wildfires have the same spatial distribution?
- If two point processes, $\mathbf{X_1} = \{\mathbf{x_i}\}_{i=1}^{N_1}$ and $\mathbf{X_2} = \{\mathbf{x_j}\}_{j=N_1+1}^{N}$, have the same spatial structure ⇒ their densities λ_{01} and λ_{02} of event locations are equal.
- Hypothesis test:

$$H_0 : \lambda_{01}(x) = \lambda_{02}(x) \quad \text{vs.} \quad H_a : \lambda_{01}(x) \neq \lambda_{02}(x)$$

- Conditional on $N_1 = n_1$ and $N_2 = N - N_1 = n_2$, $\mathbf{X_1}$ and $\mathbf{X_2}$ are random samples of the bivariate distributions with densities $\lambda_{01}(x)$ and $\lambda_{02}(x)$.
- Test statistics: a squared discrepancy measure

$$\begin{aligned} T &= \int_W (\lambda_{01}(x) - \lambda_{02}(x))^2 \, dx \\ &= \psi_{0,1} + \psi_{0,2} - (\psi_{0,12} + \psi_{0,21}) \end{aligned}$$

where $\psi_{0,j} = \int_W \lambda_{0j}(x)^2 \, dx$ and $\psi_{0,ij} = \int_W \lambda_{0i}(x) \lambda_{0j}(x) \, dx$, for $j = 1, 2$.

Duong, T., Goud, B. and Schauer, K. (2012)
Closed-form density-based framework for automatic detection of cellular morphology changes.
Proceedings of the National Academy of Sciences of the United States of America, Vol. 109, 8382-8387.

Fuentes-Santos, I., González-Manteiga, W., Mateu, J. (2017)
A nonparametric test for the comparison of first-order structures of spatial point processes.
Spatial Statistics

- Our test statistic is:

$$\hat{T} = \hat{\psi}_{0,1} + \hat{\psi}_{0,2} - \left(\hat{\psi}_{0,12} + \hat{\psi}_{0,21}\right)$$

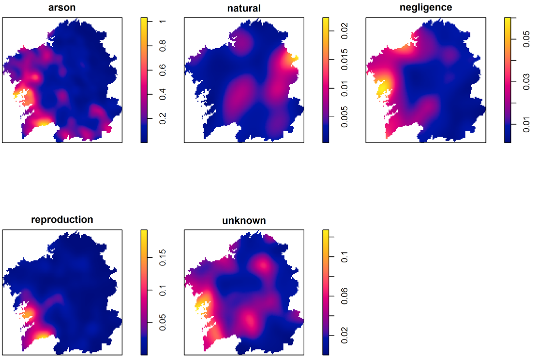
where

$$\hat{\psi}_{0,1} = \frac{1}{n_1^2} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_1} k_{G_1}\left(\mathbf{x_{i_1}} - \mathbf{x_{i_2}}\right), \ \hat{\psi}_{0,2} = \frac{1}{n_2^2} \sum_{j_1=n_1+1}^{n} \sum_{j_2=n_1+1}^{n} k_{G_2}\left(\mathbf{x_{j_1}} - \mathbf{x_{j_2}}\right)$$

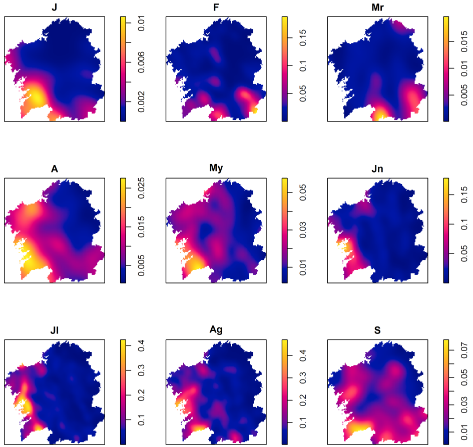
$$\hat{\psi}_{0,12} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=n_1+1}^{n} k_{G_1}\left(\mathbf{x_i} - \mathbf{x_j}\right), \ \hat{\psi}_{0,21} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=n_1+1}^{n} k_{G_2}\left(\mathbf{x_i} - \mathbf{x_j}\right)$$

- In the multivariate density framework Duong *et al.* (2012) showed that, under $\mathcal{H}_0$, $\hat{T} \to N(\mu_T, \sigma_T)$ and proposed nonparameric estimators of μ_T and σ_T.
- This property can be directly extended to the Poisson point process framework.
- **Problem**: the null distribution of $\hat{T}$ for small datasets or non-Poisson point processes may not be normal.
- **Solution**: use nonparametric bootstrap to estimate the distribution of $\hat{T}$ under $\mathcal{H}_0$.

- Dataset: 6851 wildfires registered in Galicia (NW-Spain) from January to September 2006.
- Negligible number of wildfires from October onwards.
- Null hypotheses:
  - The spatial distribution of wildfires does not depend on their cause.
  - The spatial distribution of wildfires remains constant over time.
- Kernel intensity estimators with 2-stages plug-in bandwidth matrices.
- Bootstrap calibration ($B = 1000$)

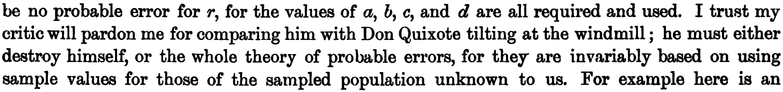


- $p - value < 0{,}01$ in all the pairwise comparisons.

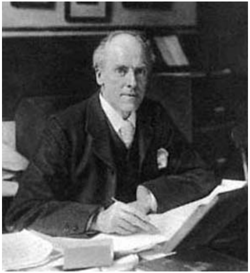

J
0.002
0.006
0.01
F
0.05
0.1
0.15
Mr
0.005
0.01
0.015
A
0.005
0.015
0.025
My
0.01
0.03
0.05
Jn
0.05
0.1
0.15
Jl
0.1
0.2
0.3
0.4
Ag
0.1
0.2
0.3
0.4
S
0.01
0.03
0.05
0.07

	n	J	F	Mr	A	My	Jn	Jl	Ag
J	60								
F	659	< 0.01							
Mr	89	< 0.01	< 0.01						
A	268	< 0.01	< 0.01	< 0.01					
My	441	< 0.01	< 0.01	< 0.01	0.12				
Jn	638	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01			
Jl	1789	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01		
Ag	2238	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	
S	669	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01

- **Application to real data**
  - The spatial distribution of wildfires during 2006 depended on their cause.
  - The spatial distribution of wildfires during 2006 varied over months.



be no probable error for r, for the values of a, b, c, and d are all required and used. I trust my critic will pardon me for comparing him with Don Quixote tilting at the windmill; he must either destroy himself, or the whole theory of probable errors, for they are invariably based on using sample values for those of the sampled population unknown to us. For example here is an



Thanks for your attention!

Pearson, K. (1900)
On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling.
Philosophical Magazine, Vol. L, 157-175.

Pearson, K. (1916)
On the application of goodness–of–fit tables to test regression curves and theoretical curves used to describe observational or experimental data.
Biometrika, Vol. 11(3), 239–261.

Elderton, W.P. (1902)
Tables for testing the goodness–of–fit of theory to observation.
Biometrika, Vol. 1, 155–163.