A data-driven lack-of-fit test of regression functions

Erkki P. Liski

University of Tampere
Department of Mathematics and Statistics

Outline

- The Model
- Hypothesis
- Orthogonal series representation

Outline

- The Model
- Hypothesis
- Orthogonal series representation
- Distributional properties of sample coefficients
- NML model

Outline

- The Model
- Hypothesis
- Orthogonal series representation
- Distributional properties of sample coefficients
- NML model
- Stochastic complexity
- MDL criterion
- Model probabilities

Nonparametric regression model

Observe $Y_1, \ldots, Y_n$ from

Nonparametric regression model

Observe $Y_1, \ldots, Y_n$ from

the model

$$Y_i = \mu(t_i) + \varepsilon_i, \qquad t_i = (i - 1/2)/n, \ i = 1, \ldots, n,$$
$$\varepsilon_1, \ldots \varepsilon_n \text{ are i.i.d.} \quad N(0,1),$$

where the unknown function μ is defined on $[0, 1]$.

Hypothesis

We wish to test the **no-effect hypothesis**

$$H_0: \ \mu(t) = \beta_0 \text{ for all } t \in [0, 1],$$

where $\beta_0 \in \mathbb{R}$ is an unknown constant.

Hypothesis

We wish to test the **no-effect hypothesis**

$$H_0: \ \mu(t) = \beta_0 \text{ for all } t \in [0,1],$$

where $\beta_0 \in \mathbb{R}$ is an unknown constant.

The test will rely on an orthogonal series representation for μ:

$$\mu(t) = \beta_0 + \sum_{j=1}^{\infty} \beta_j x_j(t), \quad x \in [0,1], \tag{1}$$

Hypothesis

We wish to test the **no-effect hypothesis**

$$H_0: \ \mu(t) = \beta_0 \text{ for all } t \in [0,1],$$

where $\beta_0 \in \mathbb{R}$ is an unknown constant.

The test will rely on an orthogonal series representation for μ:

$$\mu(t) = \beta_0 + \sum_{j=1}^{\infty} \beta_j x_j(t), \quad x \in [0,1], \tag{1}$$

where

$$\beta_j = \int_0^1 x_j(t)\mu(t)\,\mathrm{d}t, \quad j = 0, 1, \ldots$$

and

$\{1, x_1, x_2, \ldots\}$ is an orthonormal basis for $\mu \in L_2[0,1]$.

Example: An example of such a basis is

$$x_j(t) = \sqrt{2}\cos(\pi j t), \quad j = 1, 2, \ldots.$$

Example: An example of such a basis is

$$x_j(t) = \sqrt{2}\cos(\pi j t), \quad j = 1, 2, \ldots .$$

Wavelets and orthogonal polynomials are other examples.

Example: An example of such a basis is

$$x_j(t) = \sqrt{2}\cos(\pi j t), \quad j = 1, 2, \ldots.$$

Wavelets and orthogonal polynomials are other examples.

The basis functions are also assumed to be orthonormal with respect to the design:

$$\sum_{i=1}^{n} x_j(t_i) x_k(t_i) = \begin{cases} 0, & j \neq k \\ n, & j = k \end{cases}$$

for all $j, k \in \{0, 1, \ldots\}$ and $x_0 \equiv 1$.

Example: An example of such a basis is

$$x_j(t) = \sqrt{2}\cos(\pi j t), \quad j = 1, 2, \ldots.$$

Wavelets and orthogonal polynomials are other examples.

The basis functions are also assumed to be orthonormal with respect to the design:

$$\sum_{i=1}^{n} x_j(t_i) x_k(t_i) = \begin{cases} 0, & j \neq k \\ n, & j = k \end{cases}$$

for all $j, k \in \{0, 1, \ldots\}$ and $x_0 \equiv 1$.

Note 1 The cosine basis is orthonormal with respect to the design.

Note 2 The representation (1) has infinitely many parameters $\beta_0, \beta_1, \ldots$

Alternative models $\mathcal{A} = \{M_1, \ldots, M_K\}$ are of the form

$$\mu_j(t) = \beta_0 + \sum_{k \in \mathcal{K}_j} \beta_k x_k(t),$$

where $\mathcal{K}_j$ is a subset of $\{1, \ldots, j\}$ and $K < n$.

Alternative models $\mathcal{A} = \{M_1, \ldots, M_K\}$ are of the form

$$\mu_j(t) = \beta_0 + \sum_{k \in \mathcal{K}_j} \beta_k x_k(t),$$

where $\mathcal{K}_j$ is a subset of $\{1, \ldots, j\}$ and $K < n$.

A common version of such a model is

$$\mu_j(t) = \beta_0 + \sum_{k=1}^{j} \beta_k x_k(t).$$

Alternative models $\mathcal{A} = \{M_1, \ldots, M_K\}$ are of the form

$$\mu_j(t) = \beta_0 + \sum_{k \in \mathcal{K}_j} \beta_k x_k(t),$$

where $\mathcal{K}_j$ is a subset of $\{1, \ldots, j\}$ and $K < n$.

A common version of such a model is

$$\mu_j(t) = \beta_0 + \sum_{k=1}^{j} \beta_k x_k(t).$$

The MLE of β_k is

$$\hat{\beta}_k = \frac{1}{n} \sum_{i=1}^{n} x_k(t_i) Y_i, \quad k = 0, 1, \ldots, n-1.$$

Distributional properties Sample coefficients satisfy:

1. $\hat{\beta}_0, \ldots, \hat{\beta}_{n-1}$ are mutually independent.
2. $\hat{\beta}_k \sim \mathrm{N}(\frac{1}{n}\sum_{i=1}^{n}\mu(t_i)x_k(t_i), \frac{1}{n}), \quad k = 0, 1, \ldots, n-1.$

Distributional properties Sample coefficients satisfy:

1. $\hat{\beta}_0, \ldots, \hat{\beta}_{n-1}$ are mutually independent.
2. $\hat{\beta}_k \sim \mathrm{N}(\frac{1}{n}\sum_{i=1}^{n}\mu(t_i)x_k(t_i), \frac{1}{n}), \quad k = 0, 1, \ldots, n-1.$

Under H_0

$$\sqrt{n}\hat{\beta}_k \sim \mathrm{N}(0, 1), \quad k = 1, \ldots, n-1.$$

If we assume merely independence of errors
1. and 2. continue to hold asymptotically.

Normalized Maximum Likelihood (NML) model

Denote the class of normal densities for M_j, j=0,1,. . . ,K

$$\mathcal{M}_j = \{ f(\boldsymbol{y}; \boldsymbol{\beta}_j); \boldsymbol{\beta}_j \in \Theta_j \subset \mathbb{R}^{k_j} \}$$

where

$$\boldsymbol{\beta}_j = (\beta_0, \beta_1, \ldots, \beta_{k_j})', \quad \boldsymbol{y} = (y_1, \ldots, y_n)'.$$

Normalized Maximum Likelihood (NML) model

Denote the class of normal densities for M_j, j=0,1,. . . ,K

$$\mathcal{M}_j = \{ f(\boldsymbol{y}; \boldsymbol{\beta}_j); \boldsymbol{\beta}_j \in \Theta_j \subset \mathbb{R}^{k_j} \}$$

where

$$\boldsymbol{\beta}_j = (\beta_0, \beta_1, \ldots, \beta_{k_j})', \quad \boldsymbol{y} = (y_1, \ldots, y_n)'.$$

The shortest "codelength" of data $\boldsymbol{y}$ with $\mathcal{M}_j$ is

$$\log \frac{1}{f\left(\boldsymbol{y}; \hat{\boldsymbol{\beta}}_j(\boldsymbol{y})\right)}, \qquad \hat{\boldsymbol{\beta}}_j(\boldsymbol{y}) \quad \text{is the MLE of } \boldsymbol{\beta}_j.$$

Normalized Maximum Likelihood (NML) model

Denote the class of normal densities for M_j, j=0,1,. . . ,K

$$\mathcal{M}_j = \{ f(\boldsymbol{y}; \boldsymbol{\beta}_j); \boldsymbol{\beta}_j \in \Theta_j \subset \mathbb{R}^{k_j} \}$$

where

$$\boldsymbol{\beta}_j = (\beta_0, \beta_1, \ldots, \beta_{k_j})', \quad \boldsymbol{y} = (y_1, \ldots, y_n)'.$$

The shortest "codelength" of data $\boldsymbol{y}$ with $\mathcal{M}_j$ is

$$\log \frac{1}{f(\boldsymbol{y}; \hat{\boldsymbol{\beta}}_j(\boldsymbol{y}))}, \qquad \hat{\boldsymbol{\beta}}_j(\boldsymbol{y}) \quad \text{is the MLE of } \boldsymbol{\beta}_j.$$

If we use $q(\boldsymbol{y})$, the excess code length is

$$\log \frac{1}{q(\boldsymbol{y})} - \log \frac{1}{f(\boldsymbol{y}; \hat{\boldsymbol{\beta}}_j(\boldsymbol{y}))} = \log \frac{f(\boldsymbol{y}; \hat{\boldsymbol{\beta}}_j(\boldsymbol{y}))}{q(\boldsymbol{y})}.$$

Minimax solution

The NML density function for $\mathcal{M}_j$ is

$$\hat{f}(\boldsymbol{y};j) = \frac{f\left(\boldsymbol{y};\hat{\boldsymbol{\beta}}_j(\boldsymbol{y})\right)}{\int_{\hat{\boldsymbol{\beta}}_j(\boldsymbol{x})\in\Theta_j} f\left(\boldsymbol{x};\hat{\boldsymbol{\beta}}_j(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}}.$$

Minimax solution

The NML density function for $\mathcal{M}_j$ is

$$\hat{f}(\boldsymbol{y}; j) = \frac{f\left(\boldsymbol{y}; \hat{\boldsymbol{\beta}}_j(\boldsymbol{y})\right)}{\int_{\hat{\boldsymbol{\beta}}_j(\boldsymbol{x}) \in \Theta_j} f\left(\boldsymbol{x}; \hat{\boldsymbol{\beta}}_j(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}}.$$

It solves the minimax problem

$$\min_q \max_{\boldsymbol{y}} \log \frac{f(\boldsymbol{y}; \hat{\boldsymbol{\beta}}_j(\boldsymbol{y}))}{q(\boldsymbol{y})}$$

with the unique solution $\hat{q} = \hat{f}(; j)$.
(Shtarkov 1987, Rissanen 1996).

Maxmin solution

The NML density $\hat{f}(;j)$ is also the solution of the maxmin problem

$$\max_g \min_q E_g \log \frac{f(\boldsymbol{y}; \hat{\boldsymbol{\beta}}_j(\boldsymbol{y}))}{q(\boldsymbol{y})}$$

(Rissanen 2001 and 2007).

NML Density for $\mathcal{M}_j$

$$\hat{f}(\boldsymbol{y}; j) = \frac{f\left(\boldsymbol{y}; \hat{\boldsymbol{\beta}}_j(\boldsymbol{y})\right)}{C(j)},$$

where the numerator is the LF evaluated at the MLE:

$$f\left(\boldsymbol{y}; \hat{\boldsymbol{\beta}}_j(\boldsymbol{y})\right) = (2\pi)^{-n/2} \exp[-\frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0)^2 + \frac{n}{2}\sum_{k\in\mathcal{K}_j}\hat{\beta}_k^2]$$

NML Density for $\mathcal{M}_j$

$$\hat{f}(\boldsymbol{y}; j) = \frac{f\left(\boldsymbol{y}; \hat{\boldsymbol{\beta}}_j(\boldsymbol{y})\right)}{C(j)},$$

where the numerator is the LF evaluated at the MLE:

$$f\left(\boldsymbol{y}; \hat{\boldsymbol{\beta}}_j(\boldsymbol{y})\right) = (2\pi)^{-n/2} \exp[-\frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0)^2 + \frac{n}{2}\sum_{k \in \mathcal{K}_j} \hat{\beta}_k^2]$$

and the denominator

$$C(j) = \int_{\hat{\boldsymbol{\beta}}_j(\boldsymbol{x}) \in \Theta_j} f\left(\boldsymbol{x}; \hat{\boldsymbol{\beta}}_j(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}$$

is the parametric complexity of the model $\mathcal{M}_j$.

The MDL codelength for $\boldsymbol{y}$ with $\mathcal{M}_j$:

$$\log \frac{1}{\hat{f}(\boldsymbol{y};j)} = -\log \hat{f}(\boldsymbol{y};j) = -\log f\left(\boldsymbol{y};\hat{\boldsymbol{\beta}}_j(\boldsymbol{y})\right) + \log C(\mathcal{M}_j).$$

The MDL codelength for $\boldsymbol{y}$ with $\mathcal{M}_j$:

$$\log \frac{1}{\hat{f}(\boldsymbol{y};j)} = -\log \hat{f}(\boldsymbol{y};j) = -\log f\left(\boldsymbol{y};\hat{\boldsymbol{\beta}}_j(\boldsymbol{y})\right) + \log C(\mathcal{M}_j).$$

is the *stochastic complexity* of data $\boldsymbol{y}$ relative to $\mathcal{M}_j$.

The MDL codelength for $\boldsymbol{y}$ with $\mathcal{M}_j$:

$$\log \frac{1}{\hat{f}(\boldsymbol{y};j)} = -\log \hat{f}(\boldsymbol{y};j) = -\log f\left(\boldsymbol{y};\hat{\boldsymbol{\beta}}_j(\boldsymbol{y})\right) + \log C(\mathcal{M}_j).$$

is the *stochastic complexity* of data $\boldsymbol{y}$ relative to $\mathcal{M}_j$.

It can be shown that

$$-2\log \hat{f}(\boldsymbol{y};j) = k_j\left(\log \frac{n\|\hat{\boldsymbol{\beta}}_j\|^2}{k_j\,\sigma_n^2} - \frac{n\|\hat{\boldsymbol{\beta}}_j\|^2}{k_j\,\sigma_n^2} + 1\right) + \log k_j + \alpha(\boldsymbol{y}),$$

where $\sigma_n^2 = \frac{1}{n}$ and $\alpha(\boldsymbol{y})$ is common to all $\mathcal{M}_j$.
Denote

$$MDL_j = -2\log \hat{f}(\boldsymbol{y};j).$$

The MDL codelength for $\boldsymbol{y}$ with $\mathcal{M}_j$:

$$\log \frac{1}{\hat{f}(\boldsymbol{y};j)} = -\log \hat{f}(\boldsymbol{y};j) = -\log f\left(\boldsymbol{y};\hat{\boldsymbol{\beta}}_j(\boldsymbol{y})\right) + \log C(\mathcal{M}_j).$$

is the *stochastic complexity* of data $\boldsymbol{y}$ relative to $\mathcal{M}_j$.

It can be shown that

$$-2\log \hat{f}(\boldsymbol{y};j) = k_j\left(\log \frac{n\|\hat{\boldsymbol{\beta}}_j\|^2}{k_j\,\sigma_n^2} - \frac{n\|\hat{\boldsymbol{\beta}}_j\|^2}{k_j\,\sigma_n^2} + 1\right) + \log k_j + \alpha(\boldsymbol{y}),$$

where $\sigma_n^2 = \frac{1}{n}$ and $\alpha(\boldsymbol{y})$ is common to all $\mathcal{M}_j$.
Denote

$$MDL_j = -2\log \hat{f}(\boldsymbol{y};j).$$

Model probabilities

Given the data $\boldsymbol{y}$, $\hat{f}(\boldsymbol{y}; j)$ can be interpreted as the likelihood of the model $\mathcal{M}_j$, $j = 1, 2, \ldots, K$

Model probabilities

Given the data $\boldsymbol{y}$, $\hat{f}(\boldsymbol{y}; j)$ can be interpreted as the likelihood of the model $\mathcal{M}_j$, $j = 1, 2, \ldots, K$
This leads to the *NML distribution* for models

$$\hat{p}(j; \boldsymbol{y}) = \frac{\hat{f}(\boldsymbol{y}; j)}{\sum_{i=1}^{K} \hat{f}(\boldsymbol{y}; i)} \qquad = \frac{\exp(-MDL_j/2)}{\sum_{i=1}^{K} \exp(-MDL_i/2)}.$$

Model probabilities

Given the data $\boldsymbol{y}$, $\hat{f}(\boldsymbol{y}; j)$ can be interpreted as the likelihood of the model $\mathcal{M}_j$, $j = 1, 2, \ldots, K$

This leads to the *NML distribution* for models

$$\hat{p}(j; \boldsymbol{y}) = \frac{\hat{f}(\boldsymbol{y}; j)}{\sum_{i=1}^{K} \hat{f}(\boldsymbol{y}; i)} \qquad = \frac{\exp(-MDL_j/2)}{\sum_{i=1}^{K} \exp(-MDL_i/2)}.$$

Given the data $\boldsymbol{y}$, we may compute the probability

$$p_0(\boldsymbol{y}) = \hat{p}(\mathcal{M}_0; \boldsymbol{y}).$$

Model probabilities

Given the data $\boldsymbol{y}$, $\hat{f}(\boldsymbol{y}; j)$ can be interpreted as the likelihood of the model $\mathcal{M}_j,\ j = 1, 2, \ldots, K$
This leads to the *NML distribution* for models

$$\hat{p}(j; \boldsymbol{y}) = \frac{\hat{f}(\boldsymbol{y}; j)}{\sum_{i=1}^{K} \hat{f}(\boldsymbol{y}; i)} \qquad = \frac{\exp(-MDL_j/2)}{\sum_{i=1}^{K} \exp(-MDL_i/2)}.$$

Given the data $\boldsymbol{y}$, we may compute the probability

$$p_0(\boldsymbol{y}) = \hat{p}(\mathcal{M}_0; \boldsymbol{y}).$$

The idea of the test:

reject $\mathcal{M}_0$ at level α if $p_0(\boldsymbol{y})$ less than the α-quantile of $p_0(\boldsymbol{y})$'s null distribution.

Distribution of $p_0(\boldsymbol{y})$

A related test based on BIC. (Aerts & Claeskens 2004).
Results on the asymptotic distribution of $p_0(\boldsymbol{y})$ under

- the null hypothesis and
- local alternatives.

Distribution of $p_0(\boldsymbol{y})$

A related test based on BIC. (Aerts & Claeskens 2004).
Results on the asymptotic distribution of $p_0(\boldsymbol{y})$ under

- the null hypothesis and
- local alternatives.

Here the distribution of $p_0(\boldsymbol{y})$ seems to be more complicated.
A topic for further research.

References

Aerts, Claeskens & Hart (2004), *Bayesian-motivated tests of function fit and their asymptotis frequentists properties*, *Annals of Statistics* 32, pp. 2580–2615.

Burnham & Anderson (2002), *Model Selection and Multi-model Inference*, Springer

Liski, E. P. (2006), Normalized ML and the MDL Principle for Variable Selection in Linear Regression In: *Festschrift for Tarmo Pukkila on His 60th Birthday*, 159-172.

Rissanen, J. (1996). Fisher Information and Stochastic Complexity. *IEEE Trans. Information Theory*, IT-42, pp. 40–47.

Rissanen, J. (2000). MDL Denoising. *IEEE Trans. Information Theory*, IT-46, pp. 2537–2543.

Rissanen, J. (2001). Strong Optimality of the NML Models as Universal Codes and Information in Data. *IEEE Trans. Information Theory*, IT-47, pp. 1712–1717.

Rissanen, J. (2007), *Information and Complexity in Statistical Modeling*, Springer

Shtarkov, Yu. M. (1987). Universal Sequential Coding of Single Messages. *Problems of Information Transmission*, 23, pp. 3–17.