Goodness of fit tests in semiparametric transformation models

Benjamin Colling Ingrid Van Keilegom

Institut de Statistique, Biostatistique et sciences Actuarielles - Université catholique de Louvain

Young Researchers Day - 1st February 2013

Introduction - Importance of the data transformation

- Let (X_i, Y_i), $i = 1, \ldots, n$, be a data sample and consider the following simple linear regression model :

$$Y_i = a + bX_i + \epsilon_i$$

where ϵ_i are *iid*, ϵ_i are normally distributed, $E[\epsilon_i] = 0$ and $\text{Var}[\epsilon_i] = \sigma^2$ is constant (homogeneous variance).

- Assumptions made on the linear regression model :
  - Additive model
  - Variance of ϵ and Y are constant
  - Error term ϵ and response variable Y are normally distributed

Introduction - Importance of the data transformation

- Imagine we have the following relation between X and Y (which is not an additive model) :

$$Y = \lambda(1+\rho)^X U$$

where U is the error term, U is independent of X, λ and ρ are unknown parameters.

- Taking logarithm :

$$Z = \log Y = a + bX + \epsilon$$

where $a = \log \lambda$, $b = \log(1+\rho)$, $\epsilon = \log U$, ϵ is independent of X and $\epsilon \sim N(0, \sigma^2)$

- After transformation : estimation of a and b by classical least squares estimation method.

$\hookrightarrow$ Conclusion : Multiplicative model $\longrightarrow$ Additive model

Introduction - Importance of the data transformation

- Suppose that the variance of ϵ is not constant :
- $\longrightarrow$ Stabilization of the variance taking log or $\sqrt{.}$ for example

$\hookrightarrow$ Conclusion : Heterogeneous variance $\longrightarrow$ Homogeneous variance

Introduction - Importance of the data transformation

- If $Y \nsim N(\mu, \sigma^2)$, we have to transform Y
- For example : Box-Cox transformation defined by

$$\Lambda_\theta(Y) = \begin{cases} \frac{Y^\theta - 1}{\theta} & \text{if } \theta \neq 0 \\ \ln Y & \text{if } \theta = 0 \end{cases}$$

- After transformation :

$$W = \begin{cases} \frac{Y^\theta - 1}{\theta} = a + bX + \epsilon & \text{if } \theta \neq 0 \\ \ln Y = a + bX + \epsilon & \text{if } \theta = 0 \end{cases}$$

- NB : to choose the optimal value of θ, we maximize the normal log-likelihood function of W with respect to θ.

$\hookrightarrow$ Conclusion : Response variable not normally distributed $\longrightarrow$ Response variable normally distributed

Introduction - Importance of the data transformation

- If we know that the link between X and Y is not linear but for example :
  - quadratic : replace $a + bX$ by $a + bX + cX^2$
  - exponential : replace $a + bX$ by $\exp(aX)$
  - ...

- If the link between X and Y is unknown or very difficult to describe :

$$a + bX \longrightarrow m(X)$$

  where $m(X)$ is a certain unknown function

- NB : the model $Y = m(X) + \epsilon$ leads to nonparametric regression.

Introduction - Semiparametric transformation model

- A semiparametric transformation model is defined by :

$$\Lambda_{\theta_0}(Y) = m(X) + \epsilon$$

where $\{\Lambda_\theta : \theta \in \Theta\}$ is a parametric family of strictly increasing functions and the function m is of unknown form. Suppose that we have a randomly drawn sample (X_i, Y_i), $i = 1, \ldots, n$ from this model.

- Assumptions :
  - ϵ is independent of X and $E(\epsilon) = 0$
  - $m(x, \theta_0) = E[\Lambda_{\theta_0}(Y)|X = x] := m(x)$ and $V[\Lambda_{\theta_0}(Y)|X = x] = \sigma^2$ constant.
  - $\theta_0 \in \Theta \subset \mathbb{R}^k$ where Θ is a compact subset of $\mathbb{R}^k$.

- Notations :
  - We denote θ_0 as the true unknown parameter.
  - f_X and f_ϵ are the density functions of X and ϵ respectively.
  - F_X and F_ϵ are the distribution functions of X and ϵ respectively.

Introduction - Semiparametric transformation model

- Goal : we like to test the hypothesis

$$H_0 : m \in \mathcal{M} \qquad\qquad H_1 : m \notin \mathcal{M}$$

where $\mathcal{M} = \{m_\beta : \beta \in \mathcal{B}\}$ is some parametric class of regression functions and $\mathcal{B} \subset \mathbb{R}^p$

- For example

$$\begin{cases} H_0 : \exists \beta \ s.t. \ \ m(X) = \exp(\beta X) \\ H_1 : \nexists \beta \ s.t. \ m(X) = \exp(\beta X) \end{cases}$$

- Example for $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$, $p = 4$

$$\begin{cases} H_0 : \exists \beta \ s.t. \ \ m(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 \\ H_1 : \nexists \beta \ s.t. \ m(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 \end{cases}$$

Outline

1. Proposed test
2. Main results of asymptotic theory (about the test statistics)
3. Simulations and conclusions

1. Nadaraya-Watson estimator of the function $m(x)$

- Remind : in nonparametric regression, $Y = m(X) + \epsilon$, the Nadaraya-Watson estimator of the function m is defined by

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} K(\frac{X_i - x}{h}) \cdot Y_i}{\sum_{i=1}^{n} K(\frac{X_i - x}{h})}$$

- In semiparametric transformation model, $\Lambda_{\theta_0}(Y) = m(X) + \epsilon$, the Nadaraya-Watson estimator of the function m is defined by

$$\widehat{m}(x, \theta) = \frac{\sum_{i=1}^{n} K(\frac{X_i - x}{h}) \Lambda_\theta(Y_i)}{\sum_{i=1}^{n} K(\frac{X_i - x}{h})}$$

  where K is a kernel and h a bandwidth.

- We need an estimator $\widehat{\theta}$ for θ (next slide) and we'll note $\widehat{m}(x) = \widehat{m}(x, \widehat{\theta})$.

1. Profile likelihood estimator of θ

- The idea of the profile likelihood method is to calculate the log-likelihood function of Y given X and to replace unknown expressions by their nonparametric estimators.
- The profile likelihood estimator of θ is the value of θ that maximizes this log-likelihood function.
- The log-likelihood function of Y given X is given by :

$$\sum_{i=1}^{n} \left\{ \log f_{\epsilon(\theta_0)}(\Lambda_{\theta_0}(Y_i) - m(X_i, \theta_0)) + \log \Lambda'_{\theta_0}(Y_i) \right\}$$

1. Profile likelihood estimator of θ

- The profile likelihood estimator is defined by :

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \left\{ \log \widehat{f}_{\epsilon(\theta)}(\Lambda_\theta(Y_i) - \widehat{m}(X_i, \theta)) + \log \Lambda'_\theta(Y_i) \right\}$$

where

$$\widehat{f}_{\epsilon(\theta)}(e) = \frac{1}{ng} \sum_{i=1}^{n} K_2\left(\frac{e - \widehat{\epsilon}_i(\theta)}{g} \right)$$

is the nonparametric estimator of the error density function and where

$$\widehat{\epsilon}_i(\theta) = \Lambda_\theta(Y_i) - \widehat{m}(X_i, \theta)$$

are the nonparametric residuals.

- See Linton, O., Sperlich, S., Van Keilegom, I. (2008), *Estimation of a Semiparametric Transformation Model*, Annals of Statistics, Volume 36, Number 2, 686–718.

1. Proposed test

- Principal question : how can we define the test ?
- One way to compare the different approaches proposed in the litterature is to compare the power of the tests. We consider :

$$\begin{cases} H_0 : & \exists \beta \text{ such that } m(x) = m_\beta(x) \\ H_1 : & \exists \beta \text{ such that } m(x) = m_\beta(x) + c_n r(x) \end{cases}$$

for some function r and where $c_n \to 0$.

- For example, in the case of nonparametric regression, Härdle, W. and Mammen, E. (1993) construct a test based on the following test statistic :

$$\int (\widehat{m}(x) - \widehat{m}_{\widehat{\beta}}(x))^2 \, d\widehat{F}_X(x)$$

where $\widehat{m}_{\widehat{\beta}}(x) = \frac{\sum_{i=1}^n K_h(x - X_i) m_{\widehat{\beta}}(X_i)}{\sum_{i=1}^n K_h(x - X_i)}$ and $\widehat{\beta}$ is estimated by least squares method.

- This test can detect alternatives at the rate $c_n = n^{-1/2} h^{-d/4}$.

1. Proposed test

- In the article of Van Keilegom, I., Gonzalez Manteiga, W., Sanchez Sellero, C. (2008), *Goodness of fit tests in parametric regression based on the estimation of the error distribution*, a similar test for the following heteroscedastic regression model is developped :
  $Y = m(X) + \sigma(X)\epsilon$
- They proposed a new type of test statistic where the idea is to compare $\widehat{F}_{\epsilon}(y)$ and $\widehat{F}_{\epsilon_0}(y)$.
- This test can detect alternatives at the rate $c_n = n^{-1/2} \Longrightarrow H_1$ converges faster to H_0 than if we use the rate $c_n = n^{-1/2}h^{-d/4}$ of Härdle and Mammen (1993).
- Conclusion : we use here the approach of Van Keilegom, I., Gonzalez Manteiga, W., Sanchez Sellero, C. (2008)

1. Proposed test

- Estimation of the distribution function of ϵ in a nonparametric way :

$$\widehat{F}_{\epsilon}(y) = n^{-1} \sum_{i=1}^{n} I(\widehat{\epsilon}_i \leq y)$$

where $\widehat{\epsilon}_i = \Lambda_{\widehat{\theta}}(Y_i) - \widehat{m}(X_i)$ are the nonparametric residuals.

- Estimator of the distribution function of ϵ under H_0 :

$$\widehat{F}_{\epsilon_0}(y) = n^{-1} \sum_{i=1}^{n} I(\widehat{\epsilon}_{i0} \leq y)$$

where $\widehat{\epsilon}_{i0}$ are the residuals estimated under H_0 :

$$\widehat{\epsilon}_{i0} = \Lambda_{\widehat{\theta}}(Y_i) - m_{\widehat{\beta}}(X_i, \widehat{\theta})$$

and β is estimated by the least squares method :

$$\widehat{\beta} = \arg\min_{\beta \in \mathcal{B}} n^{-1} \sum_{i=1}^{n} (\Lambda_{\widehat{\theta}}(Y_i) - m_{\beta}(X_i))^2$$

1. Proposed test

- The test statistics that we will use are the Kolmogorov-Smirnov type statistic

$$T_{KS} = n^{1/2} \sup_{y \in \mathbb{R}} |\widehat{F}_{\epsilon}(y) - \widehat{F}_{\epsilon_0}(y)|$$

- and the Cramer-von Mises type statistic

$$T_{CM} = n \int (\widehat{F}_{\epsilon}(y) - \widehat{F}_{\epsilon_0}(y))^2 \, d\widehat{F}_{\epsilon}(y)$$

- Main reference/source : Van Keilegom, I., Gonzalez Manteiga, W., Sanchez Sellero, C. (2008), *Goodness of fit tests in parametric regression based on the estimation of the error distribution*.

2. Asymptotic theory : some additional notations

- Let $\beta = (\beta_1, \ldots, \beta_p)$ and

$$\frac{\partial m_\beta(x)}{\partial \beta} = \left(\frac{\partial m_\beta(x)}{\partial \beta_r} \right)_{r=1,\ldots,p}$$

- Notations :

$$\Omega = \left\{ E\left[\frac{\partial m_{\beta_0}(X)}{\partial \beta_r} \left(\frac{\partial m_{\beta_0}(X)}{\partial \beta_s} \right)^t \right] \right\}_{r,s=1,\ldots,p}$$

- and

$$\eta_{\theta,\beta}(x,y) := \eta_\beta(x,y) = \Omega^{-1} \frac{\partial m_\beta(x)}{\partial \beta} (\Lambda_\theta(y) - m_\beta(x))$$

Theorem

Under H_0,

$$\widehat{F}_{\epsilon}(y) - \widehat{F}_{\epsilon_{\mathbf{0}}}(y) = f_{\epsilon}(y) n^{-1} \sum_{i=1}^{n} \left(\Lambda_{\theta_{\mathbf{0}}}(Y_i) - m(X_i) - \int \left(\frac{\partial m_{\beta_{\mathbf{0}}}(x)}{\partial \beta} \right)' dF_X(x) \cdot \left[-\Omega^{-1} g(X_i, Y_i) E\left[\frac{\partial \Lambda_{\theta_{\mathbf{0}}}(Y)}{\partial \theta} \frac{\partial m_{\beta_{\mathbf{0}}}(X)}{\partial \beta} \right] + \eta_{\beta_{\mathbf{0}}}(X_i, Y_i) \right] \right) + R_n(y)$$

where $\sup_{y \in \mathbb{R}} |R_n(y)| = o_P(n^{-1/2})$ *and*

$$g(X,Y) = \Gamma_1^{-1} \left[\frac{1}{f_{\epsilon}(\epsilon)} [f'_{\epsilon}(\epsilon)(\dot{\Lambda}_{\theta_{\mathbf{0}}}(Y) - \dot{m}_{\theta_{\mathbf{0}}}(X)) + \dot{f}_{\epsilon}(\epsilon)] + \frac{\dot{\Lambda}'_{\theta_{\mathbf{0}}}(Y)}{\Lambda'_{\theta_{\mathbf{0}}}(Y)} \right] \stackrel{not}{=} \Gamma_1^{-1} G_{1PL}(\theta_0, X, Y)$$

and

$$\Gamma_1 = \frac{\partial}{\partial \theta} E[G_{1PL}(\theta_0, X, Y)]$$

Corollary

Under H_0, the process $n^{1/2}(\widehat{F}_\epsilon(y) - \widehat{F}_{\epsilon_0}(y))$ converges weakly to $f_\epsilon(y)W$ where W is a zero mean normal random variable with variance

$$V(W) = E\left[\left(\Lambda_{\theta_0}(Y) - m(X) - \int \left(\frac{\partial m_{\beta_0}(x)}{\partial \beta}\right)' dF_X(x) \cdot \eta_{\beta_0}(X,Y)\right.\right.$$

$$\left.\left. + \int \left(\frac{\partial m_{\beta_0}(x)}{\partial \beta}\right)' dF_X(x) \cdot \Omega^{-1} g(X,Y) E\left[\frac{\partial \Lambda_{\theta_0}(Y)}{\partial \theta}\frac{\partial m_{\beta_0}(X)}{\partial \beta}\right]\right)^2\right]$$

2. Asymptotic theory : main results

- Remind : Definition of the test statistics

$$T_{KS} = n^{1/2} \sup_{y \in \mathbb{R}} |\widehat{F}_{\epsilon}(y) - \widehat{F}_{\epsilon_0}(y)|$$

$$T_{CM} = n \int (\widehat{F}_{\epsilon}(y) - \widehat{F}_{\epsilon_0}(y))^2 \, d\widehat{F}_{\epsilon}(y)$$

Theorem

Under H_0 *:*

$$T_{KS} \xrightarrow{d} \sup_{y \in \mathbb{R}} |f_{\epsilon}(y)| \cdot |W|$$

and

$$T_{CM} \xrightarrow{d} \int f_{\epsilon}^2(y) \, dF_{\epsilon}(y) \cdot W^2$$

Theorem

Under H_1 *:*

$$T_{KS} \xrightarrow{d} \sup_{y \in \mathbb{R}} |f_\epsilon(y)| \cdot |W + b|$$

and

$$T_{CM} \xrightarrow{d} \int f_\epsilon^2(y)\, dF_\epsilon(y) \cdot (W + b)^2$$

where

$$b = -\int \left(\frac{\partial m_{\beta_0}(x)}{\partial \beta} \right)' dF_X(x) \cdot \Omega^{-1} \cdot \int r(x) \cdot \frac{\partial m_{\beta_0}(x)}{\partial \beta}\, dF_X(x) + \int r(x)\, dF_X(x)$$

2. Asymptotic theory : main results

- The previous limiting distribution depend on $f_\epsilon(y)$, which is unknown –> we have to do bootstrap to estimate the critical values of our test.
- Then, we introduce two other test statistics defined by

$$T_{KS2} = \sup_{y \in \mathbb{R}} |\widehat{C}(y)|$$

and

$$T_{CM2} = 3 \int_{-\infty}^{+\infty} |\widehat{C}(y)|^2 \, d\widehat{F}_\epsilon(y)$$

where

$$\widehat{C}(y) = \int_{-\infty}^{y} n^{1/2} (\widehat{F}_\epsilon(s) - \widehat{F}_{\epsilon_0}(s)) \, ds$$

- See Escanciano, J.C., Pardo-Fernandez, J.C., Van Keilegom, I. (2012), *A nonparametric test for risk-return relationships*

2. Asymptotic theory : main results

Theorem

Under H_0,

$$T_{KS2} \xrightarrow{d} |W|$$

and

$$T_{CM2} \xrightarrow{d} W^2$$

Theorem

Under H_1,

$$T_{KS2} \xrightarrow{d} |W + b|$$

and

$$T_{CM2} \xrightarrow{d} (W + b)^2$$

3. Simulations

- Problem : the variance of W depends on f_ϵ and its derivatives with respect to y and θ and all these quantities are unknown –> bootstrap in order to estimate the critical values of our test.

- We realize a very simple simulation in order to illustrate the power of a test.

- We consider the following test :

$$\begin{cases} H_0 : \exists \beta \ s.t. \ \ m(X) = \exp(\beta X) \\ H_1 : \exists \beta \ s.t. \ m(X) = \exp(\beta X) + c \end{cases}$$

  where c is a certain constant.

3. Simulations

- $n = 200$, $\theta_0 = 0.5$, $\beta_0 = 2$, $X \sim U[0,1]$ and $\epsilon \sim N(0,1)$ truncated on $[-3,3]$
- Epanechnikov kernels :

$$K(u) = K_2(u) = \frac{3}{4}(1-u^2) \cdot 1_{\{-1 \leq u \leq 1\}}$$

- Bandwidth h chosen by cross-validation method :

$$\widehat{h}(\theta) = \arg\min_h n^{-1} \sum_{i=1}^{n} (\Lambda_\theta(Y_i) - \widetilde{m}_{h,-i}(X_i,\theta))^2$$

where

$$\widetilde{m}_{h,-i}(X_i,\theta) = \frac{\sum_{j=1,j\neq i}^{n} K(\frac{X_j - X_i}{h}) \Lambda_\theta(Y_i)}{\sum_{j=1,j\neq i}^{n} K(\frac{X_j - X_i}{h})}$$

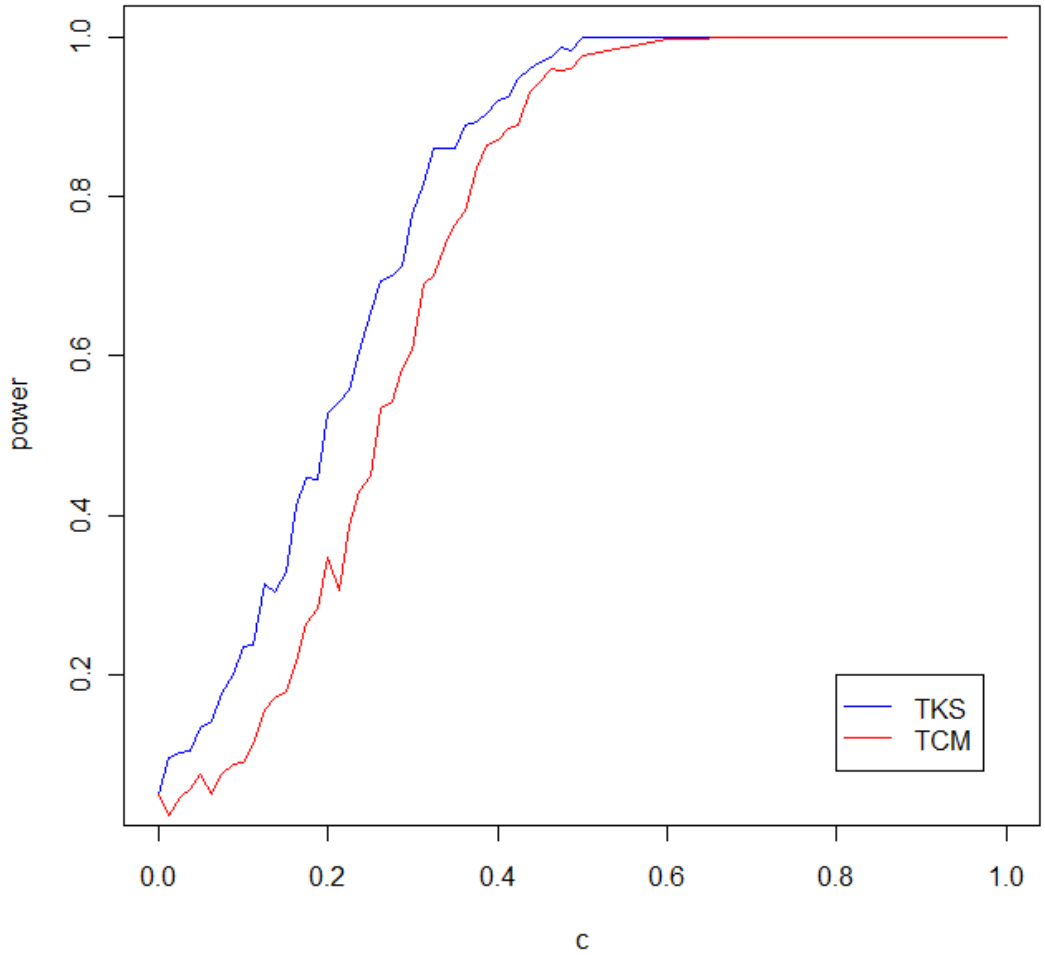
- Bandwidth g chosen by normal reference rule :
$\widehat{g}(\theta,h) = 2.34 \cdot \widehat{\sigma}_\epsilon(\theta,h) \cdot n^{-1/5}$

3. Simulations

- To determine critical values of our test :
  1. We realize 500 Monte-Carlo simulations where (X_i, Y_i) is determined under H_0
  2. Hence we obtain 500 values of T_{KS} and T_{CM}
  3. The critical values are the 475th greatest values of T_{KS} and T_{CM}
  4. NB : by construction $\alpha = 0.05$

- Goal : under H_1, for 500 simulations and for different values of c, we estimate the power of the test by the number of test statistics greater than critical values divided by 500

3. Simulations



Main references

- Akritas, M.G., Van Keilegom, I. (2001), *Non-parametric Estimation of the Residual Distribution*, Scandinavian Journal of Statistics, Volume 28, Number 3, 549–568.
- Bickel, P.J. and Doksum, K. (1981), *An analysis of transformations revisited*, Journal of the American Statistical Association, Volume 76, 296–311.
- Box, G.E.P. and Cox, D.R. (1964), *An analysis of transformations*, Journal of the Royal Statistical Society - Series B, Volume 26, Number 2, 211–252.
- Carroll, R.J. and Ruppert, D. (1988), *Transformation and Weighting in Regression*, Chapman and Hall, New-York.
- Chen, X., Linton, O., Van Keilegom, I. (2003), *Estimation of Semiparametric Models when the Criterion Function is not Smooth*, Econometrica, Volume 71, Number 5, 1591–1608.
- Escanciano, J.C., Pardo-Fernandez, J.C., Van Keilegom, I. (2012), *A nonparametric test for risk-return relationships*.

Main references

- Härdle, W. and Mammen, E. (1993), *Comparing nonparametric versus parametric regression fits*, The Annals of Statistics, Volume 21, Number 4, 1926–1947.
- Linton, O., Sperlich, S., Van Keilegom, I. (2008), *Estimation of a Semiparametric Transformation Model*, Annals of Statistics, Volume 36, Number 2, 686–718.
- Sakia, R.M. (1992), *The Box-Cox transformation technique : a review*, The Statistician, Volume 41, 169–178.
- Samb, R., Heuchenne, C., Van Keilegom, I. (2012), *Estimating the Residual Distribution in Semiparametric Transformation Models*.
- Van Keilegom, I., Gonzalez Manteiga, W., Sanchez Sellero, C. (2008), *Goodness of fit tests in parametric regression based on the estimation of the error distribution*, TEST, Volume 17, 401–415.
- Zellner, A. and Revankar, N.S. (1969), *Generalized production functions*, The Review of Economic Studies, Volume 36, Number 2, 241–250.