

Validation of functional form in multiple regression using R

Joachim Schnurbus
(joint work with Harry Haupt and Rolf Tschernig)

Conference on Quantitative Social Science Research Using R — New York

June 18, 2009

Outline

- 1 Motivation
- 2 Computation, data set, and data exploration
- 3 Nonparametric regression with mixed data
- 4 Model validation
- 5 Conclusion

Framework

Typical setting in regression context

- data with some continuous and some discrete covariates,
- functional relationship between response variable and covariates is not known a priori.

Goal of researcher

Find most parsimonious specification that

- describes in-sample relationship well;
- has “maximal” out-of-sample prediction performance.

Framework

Typical setting in regression context

- data with some continuous and some discrete covariates,
- functional relationship between response variable and covariates is not known a priori.

Goal of researcher

Find most parsimonious specification that

- describes in-sample relationship well;
- has “maximal” out-of-sample prediction performance.

Evaluation of model performance

In-sample performance

- goodness-of-fit measures (e.g. Pseudo- R^2),
- specification tests (e.g. test of Hsiao et al., 2007).

Out-of-sample performance

- various concepts of prediction error (e.g. MSEP),
- newly collected data versus hold-out data.

Evaluation of model performance

In-sample performance

- goodness-of-fit measures (e.g. Pseudo- R^2),
- specification tests (e.g. test of Hsiao et al., 2007).

Out-of-sample performance

- various concepts of prediction error (e.g. MSEP),
- newly collected data versus hold-out data.

Regression and functional form

Parametric model: $\mathbf{y} = f(\mathbf{X}, \beta) + \mathbf{u}$

- well known properties, computationally easy,
- but results may be misleading if functional form is misspecified.

Problem: a priori parameterization.

Nonparametric model: $\mathbf{y} = g(\mathbf{X}) + \mathbf{u}$

- no assumption about exact functional form,
- but with same amount of data less precise.

Problem: “tuning” options.

Regression and functional form

Parametric model: $\mathbf{y} = f(\mathbf{X}, \beta) + \mathbf{u}$

- well known properties, computationally easy,
- but results may be misleading if functional form is misspecified.

Problem: a priori parameterization.

Nonparametric model: $\mathbf{y} = g(\mathbf{X}) + \mathbf{u}$

- no assumption about exact functional form,
- but with same amount of data less precise.

Problem: “tuning” options.

Overview of nonparametric regression

Nonparametric regression

- kernel regression
- (penalized) spline regression
- ...

Kernel regression with mixed regressors

- frequency approach (splitting data sets)
- smoothing discrete variables (Li/Racine-approach)

Overview of nonparametric regression

Nonparametric regression

- kernel regression
- (penalized) spline regression
- ...

Kernel regression with mixed regressors

- frequency approach (splitting data sets)
- smoothing discrete variables (Li/Racine-approach)

Computation

General

- R, version 2.9.0.

Nonparametric regression with mixed regressors

- `np`-package, version 0.30-2
- Authors: Tristen Hayfield, Jeffrey S. Racine.

Analyses of data and results

- `relax`-package (R Editor for Literate Analysis and lateX), version 1.2.1,
- Author: Hans Peter Wolf.

Example: Canadian housing data ($n = 546$)

- dependent variable: `lnsell` (log of sale price)
- six binary regressors:
 - `ca` (central air conditioning)
 - `drv` (driveway)
 - `ffin` (full finished basement)
 - `ghw` (gas for hot water heating)
 - `rec` (recreational room)
 - `reg` (located in preferred neighbourhood)
- four ordered categorical regressors:
 - `bdms` (number of bedrooms)
 - `fb` (number of full bathrooms)
 - `gar` (number of garage places)
 - `sty` (number of stories)
- continuous regressor: `lnlot` (log of lotsize)

Example: Canadian housing data ($I = 546$)

- dependent variable: `lnsell` (log of sale price)
- six binary regressors:
 - `ca` (central air conditioning)
 - `drv` (driveway)
 - `ffin` (full finished basement)
 - `ghw` (gas for hot water heating)
 - `rec` (recreational room)
 - `reg` (located in preferred neighbourhood)
- four ordered categorical regressors:
 - `bdms` (number of bedrooms)
 - `fb` (number of full bathrooms)
 - `gar` (number of garage places)
 - `sty` (number of stories)
- continuous regressor: `lnlot` (log of lotsize)

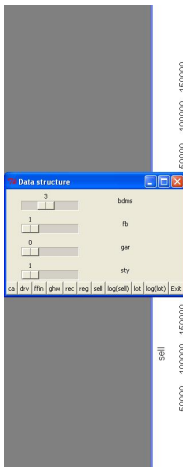
Example: Canadian housing data ($n = 546$)

- dependent variable: `lnsell` (log of sale price)
- six binary regressors:
 - `ca` (central air conditioning)
 - `drv` (driveway)
 - `ffin` (full finished basement)
 - `ghw` (gas for hot water heating)
 - `rec` (recreational room)
 - `reg` (located in preferred neighbourhood)
- four ordered categorical regressors:
 - `bdms` (number of bedrooms)
 - `fb` (number of full bathrooms)
 - `gar` (number of garage places)
 - `sty` (number of stories)
- continuous regressor: `lnlot` (log of lotsize)

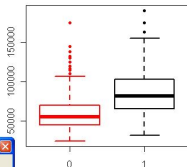
Example: Canadian housing data ($l = 546$)

- dependent variable: `lnsell` (log of sale price)
- six binary regressors:
 - `ca` (central air conditioning)
 - `drv` (driveway)
 - `ffin` (full finished basement)
 - `ghw` (gas for hot water heating)
 - `rec` (recreational room)
 - `reg` (located in preferred neighbourhood)
- four ordered categorical regressors:
 - `bdms` (number of bedrooms)
 - `fb` (number of full bathrooms)
 - `gar` (number of garage places)
 - `sty` (number of stories)
- continuous regressor: `lnlot` (log of lotsize)

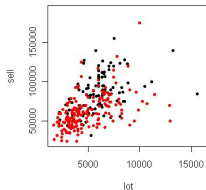
Data exploration using relax



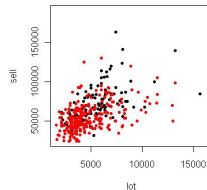
ca: 373 zeros; 173 ones.



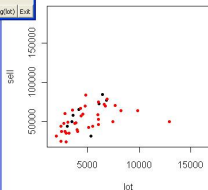
301 obs. with bdrms = 3



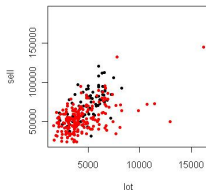
402 obs. with fb = 1



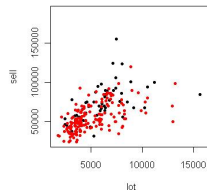
43 obs. with ordered configuration



300 obs. with gar = 0



227 obs. with sty = 1



Multiple local linear regression

Minimization calculus for a local linear regression at position $(\mathbf{x}_0, \mathbf{z}_0)$:

$$\min_{\tilde{b}_0(\mathbf{x}_0, \mathbf{z}_0), \tilde{\mathbf{b}}_1(\mathbf{x}_0, \mathbf{z}_0)} \sum_{i=1}^I \left(y_i - \tilde{b}_0(\mathbf{x}_0, \mathbf{z}_0) - \tilde{\mathbf{b}}_1(\mathbf{x}_0, \mathbf{z}_0)' \cdot (\mathbf{x}_i - \mathbf{x}_0) \right)^2 \cdot W(.).$$

Here:

- $\hat{b}_0(\mathbf{x}_0, \mathbf{z}_0)$ estimates mean effect at position $(\mathbf{x}_0, \mathbf{z}_0)$.
- $\hat{\mathbf{b}}_1(\mathbf{x}_0, \mathbf{z}_0)$ captures partial effects of continuous regressors.
- Observations in the neighborhood of $(\mathbf{x}_0, \mathbf{z}_0)$ are used to calculate values $\hat{b}_0(\mathbf{x}_0, \mathbf{z}_0)$ and $\hat{\mathbf{b}}_1(\mathbf{x}_0, \mathbf{z}_0)$.
- Neighborhood is determined by $W(\cdot)$.
- $W(\cdot)$ depends on $\mathbf{x}_0, \mathbf{x}_i, \mathbf{z}_0, \mathbf{z}_i$, and \mathbf{h} .

Generalized product kernel

$$W(\cdot) = \prod_{c=1}^C W_c(x_{0c}, x_{ic}, h_c) \cdot \prod_{d=C+1}^{C+D} W_d(z_{0d}, z_{id}, h_d),$$

with

- $W_c(\cdot)$ weighting function for continuous regressors,
- $W_d(\cdot)$ weighting function for discrete regressors,
- x_{0c}, x_{ic} values of continuous regressors,
- z_{0d}, z_{id} values of discrete regressors,
- h_c, h_d smoothing parameter of a continuous/discrete variable.

Unordered discrete regressors — Li/Racine-Kernel

Weighting function

$$W_d(z_{0d}, z_{id}, h_d) = \begin{cases} 1 & \text{for } z_{id} = z_{0d}, \\ h_d & \text{for } z_{id} \neq z_{0d}. \end{cases}$$

Smoothing parameter

$$h_d \in [0; 1]$$

Kernel behavior

- $h_d = 0 \Rightarrow$ complete separation (frequency approach),
- $h_d = 1 \Rightarrow$ complete smoothing (irrelevant regressors).

Smoothing discrete covariates with `relax`

Nonparametric with binary

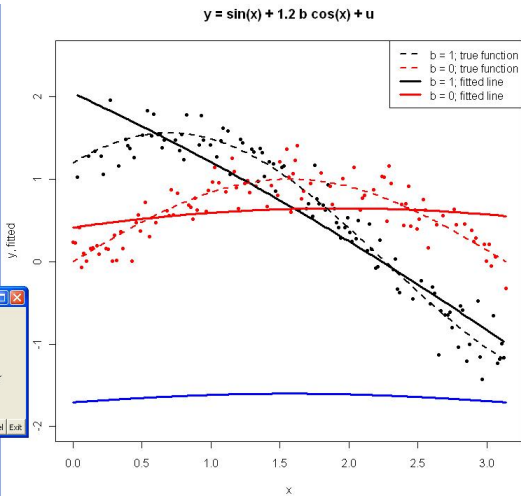
1.20 Mix - Sinus/Cosinus

50 Percentage of ones

2.00 BW - continuous regressor

0.00 BW - binary regressor

Uniform Kernel | Gaussian Kernel | Epanechnikov Kernel | Exit



Nonparametric estimation — bandwidths

Regressor	Type	Weighting Function	Estimated h	Maximal h
ln(lot)	cont.	Gaussian (2. order)	1,021,545	∞
ca	bin.	Aitchison/Aitken	0.1409	0.5
drv	bin.	Aitchison/Aitken	0.1206	0.5
ffin	bin.	Aitchison/Aitken	0.1296	0.5
ghw	bin.	Aitchison/Aitken	0.0302	0.5
rec	bin.	Aitchison/Aitken	≈ 0.5	0.5
reg	bin.	Aitchison/Aitken	0.2253	0.5
bdms	ord.	Wang/van Ryzin	0.6153	< 1
fb	ord.	Wang/van Ryzin	0.2667	< 1
gar	ord.	Wang/van Ryzin	≈ 1	< 1
sty	ord.	Wang/van Ryzin	0.5063	< 1

In-sample fit

Pseudo- R^2

$$R^2 = (\text{Corr}(\mathbf{y}, \hat{\mathbf{y}}))^2$$

Linear parametric specification

$$R^2 = 0.6865$$

Nonparametric specification

$$R^2 = 0.7606$$

Hsiao/Li/Racine-Test

Hypotheses

$$H_0 : P(E[y_i | \mathbf{x}_i, \mathbf{z}_i] = f(\mathbf{x}_i, \mathbf{z}_i, \mathbf{b})) = 1 \quad \text{for some } \mathbf{b},$$

$$H_1 : P(E[y_i | \mathbf{x}_i, \mathbf{z}_i] = f(\mathbf{x}_i, \mathbf{z}_i, \mathbf{b})) < 1 \quad \text{for all } \mathbf{b}.$$

Idea of test statistic

$$T = E \left[(E[u_i | \mathbf{x}_i, \mathbf{z}_i])^2 g(\cdot) \right] = E [u_i E[u_i | \mathbf{x}_i, \mathbf{z}_i] g(\cdot)]$$

Distribution of test statistic

- \hat{T} is asymptotically normal (under general conditions).
- Alternative: bootstrap (iid, wild, ...).

Hsiao/Li/Racine-Test — p-values

	LSCV		AIC_c	
	LC	LL	LC	LL
Asymptotic				
lnlot	0.3682	0.4400	0.4294	0.2791
lot	0.5133	0.2957	0.4566	0.2830
IID				
lnlot	0.1278	0.1805	0.0802	0.0226
lot	0.1955	0.1429	0.0702	0.0251
Wild				
lnlot	0.1529	0.2180	0.1103	0.0351
lot	0.2607	0.2030	0.1128	0.0401

A simulation approach

Monte-Carlo simulation with 10,000 replications:

In each replication:

- Estimate bandwidth for the full sample.
- Split sample randomly (e.g. 90% - 10%).
- Estimate regressions (rescaled bandwidths) with the estimation sample (90%).
- Calculate fitted values for the prediction sample (10%).
- Calculate \widehat{MSEP} for prediction sample (P observations):

$$\widehat{MSEP} = \frac{1}{P} \sum_{p=1}^P (y_p - \hat{y}_p)^2.$$

Testing the MSEP

Test-Input

- 10,000 replications result in 10,000 \widehat{MSEP} s for each specification.
- Note: predictions use partly same information, implying dependence.

Paired t-test

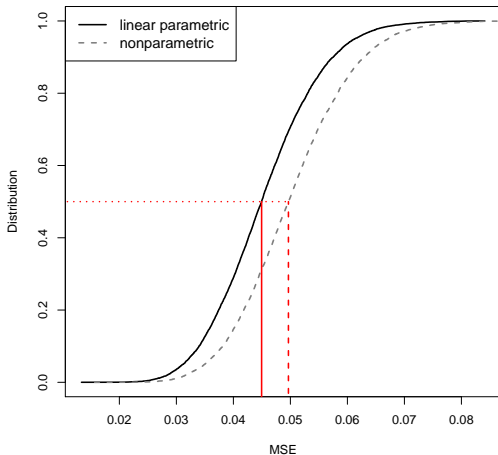
$$H_0 : MSEP(\text{specification 1}) \geq MSEP(\text{specification 2}),$$

$$H_1 : MSEP(\text{specification 1}) < MSEP(\text{specification 2}).$$

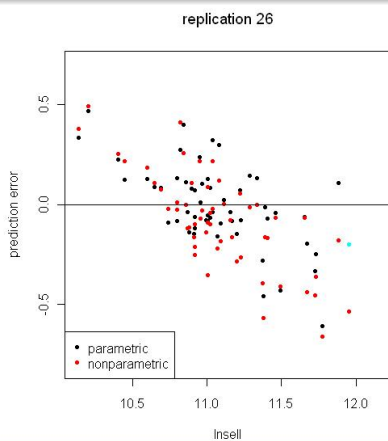
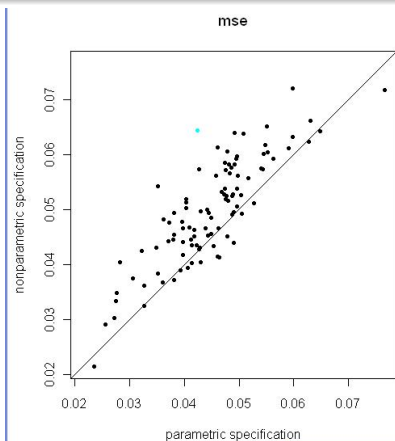
Simulation results

Split-proportion	share parametric superior	p-value (paired t-test)
70:30	0.9621	< 0.0001
80:20	0.9278	< 0.0001
90:10	0.8382	< 0.0001

Simulation results (cont.)



Analysis of the simulation results using `relax`



slider control widget

MAE	MAPE	MSE	MSPE	select replication	select observation	Exit
-----	------	-----	------	--------------------	--------------------	------

R Console

```
sell lot bdcms fb sty drv rec ffin ghv ca gar reg lnseII fitted.s1 fitted.s3  
338 155000 7500 3 3 1 1 0 1 0 1 2 1 11.95118 11.75102 11.41350
```

Conclusion

Multiple nonparametric regression (for mixed data)

- allows for capturing nonlinearities without ad hoc assumptions on functional form,
- is applicable to a wide range of data due to the approach of Li and Racine.

R in combination with the packages `np` and `relax`

- allows an extensive validation of a given specification, with respect to in- and out-of-sample-performance,
- covers the whole modeling cycle from data exploration to model estimation and validation.

Conclusion

Multiple nonparametric regression (for mixed data)

- allows for capturing nonlinearities without ad hoc assumptions on functional form,
- is applicable to a wide range of data due to the approach of Li and Racine.

R in combination with the packages `np` and `relax`

- allows an extensive validation of a given specification, with respect to in- and out-of-sample-performance,
- covers the whole modeling cycle from data exploration to model estimation and validation.

Conclusion

Multiple nonparametric regression (for mixed data)

- allows for capturing nonlinearities without ad hoc assumptions on functional form,
- is applicable to a wide range of data due to the approach of Li and Racine.

R in combination with the packages `np` and `relax`

- allows an extensive validation of a given specification, with respect to in- and out-of-sample-performance,
- covers the whole modeling cycle from data exploration to model estimation and validation.

Conclusion

Multiple nonparametric regression (for mixed data)

- allows for capturing nonlinearities without ad hoc assumptions on functional form,
- is applicable to a wide range of data due to the approach of Li and Racine.

R in combination with the packages `np` and `relax`

- allows an extensive validation of a given specification, with respect to in- and out-of-sample-performance,
- covers the whole modeling cycle from data exploration to model estimation and validation.

Work in progress

- Simulations concerning scope and pitfalls of specification tests and the measures of in- and out-of-sample performance.
- Extend the analyses using additional data sets and data configurations (including panel data and a simulated data example).

Work in progress

- Simulations concerning scope and pitfalls of specification tests and the measures of in- and out-of-sample performance.
- Extend the analyses using additional data sets and data configurations (including panel data and a simulated data example).

References

- Aitchison J. and C.G.G. Aitken (1976), Multivariate Binary Discrimination by the Kernel Method. *Biometrika* 63, 413-420.
- Haupt H., J. Schnurbus, and R. Tschernig (2009), On Nonparametric Estimation of a Hedonic Price Function. Under Revision at *JAE*.
- Hsiao, C., Q. Li, and J. S. Racine (2007), A consistent model specification test with mixed discrete and continuous data. *J Econometrics* 140, 802-826.
- Hurvich, C. M., J. S. Simonoff, and C. L. Tsai (1998), Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *JRSS B* 60, 271-293.
- Li, Q. and J. S. Racine (2007), *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.
- Parmeter C.F., D.J. Henderson, and S.C. Kumbhakar (2007), Nonparametric Estimation of a Hedonic Price Function. *JAE* 22, 695-699.
- Racine, J. S. and Q. Li (2004), Nonparametric estimation of regression functions with both, categorical and continuous data. *J Econometrics* 119, 99-130.
- Wang, M.C. and J. van Ryzin (1981), A class of smooth estimators for discrete distributions. *Biometrika* 68, 301-309.

Thanks for your attention!

Notation

- i index for I observations with $i = 1, \dots, I$.
- c index for C continuous regressors $c = 1, \dots, C$.
- d index for D discrete regressors $d = C + 1, \dots, C + D$.
- x_{ic} value of continuous regressor c for observation i .
- z_{id} value of discrete regressor d for observation i .
- y_i value of response for observation i .
- h smoothing parameter.

Continuous regressors

$$W_c(x_{0c}, x_{ic}, h_c) = K\left(\frac{x_{ic} - x_{0c}}{h_c}\right)$$

Here:

- $K(\cdot)$ is a second order Gaussian kernel
- Weight of an observation with value x_{ic} for a regression at x_{0c} decreases with increasing absolute value of $x_{ic} - x_{0c}$.
- A larger smoothing parameter h_c provides weights that are more similar (and small) for all observations.

Ordered discrete regressors — Li/Racine-Kernel

Weighting function

$$W_d(z_{0d}, z_{id}, h_d) = h_d^{|z_{0d} - z_{id}|}$$

Smoothing parameter

$$h_d \in [0; 1]$$

Kernel behavior

- $h_d = 0 \Rightarrow$ complete separation (frequency approach),
- $h_d = 1 \Rightarrow$ complete smoothing (irrelevant regressors).

Unordered discrete — Aitchison/Aitken kernel

Weighting function

$$W_d(z_{0d}, z_{id}, h_d) = \begin{cases} 1 - h_d & \text{for } z_{id} = z_{0d}, \\ \frac{h_d}{n-1} & \text{for } z_{id} \neq z_{0d}. \end{cases}$$

with n as number of categories of the discrete regressor.

Smoothing parameter

$$h_d \in \left[0; \frac{n-1}{n} \right]$$

Ordered discrete — Wang/van Ryzin kernel

Weighting function

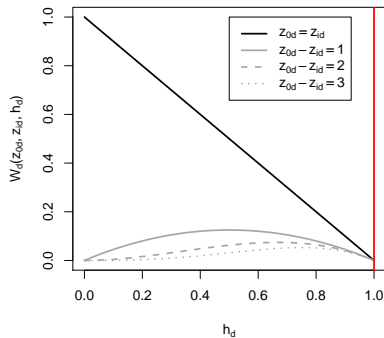
$$W_d(z_{0d}, z_{id}, h_d) = \begin{cases} 1 - h_d & \text{for } z_{id} = z_{0d}, \\ \frac{1}{2}(1 - h_d)h_d^{|z_{id} - z_{0d}|} & \text{for } z_{id} \neq z_{0d}. \end{cases}$$

Smoothing parameter

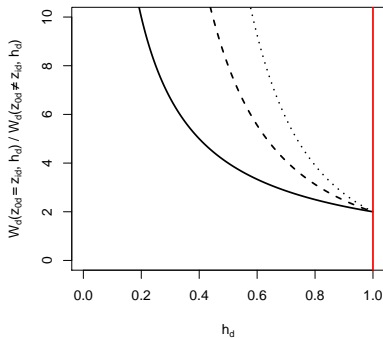
$$h_d \in [0; 1]$$

Ordered discrete — Wang/van Ryzin kernel (cont.)

Weights



Weights (relative)



Bandwidth — least-squares cross-validation (LSCV)

$$\min_{\tilde{\mathbf{h}}} CV(\tilde{\mathbf{h}}),$$

with

$$CV(\tilde{\mathbf{h}}) = \sum_{i=1}^l \left(y_i - \tilde{g}_{-i}(\mathbf{x}_i, \tilde{\mathbf{h}}) \right)^2 \cdot M(\mathbf{x}_i).$$

and

- $\tilde{g}_{-i}(\mathbf{x}_i, \tilde{\mathbf{h}})$ as the leave-one-out estimator at position \mathbf{x}_i .
- $M(\mathbf{x}_i)$ is another weighting function to ensure computability.

Bandwidth — corrected Kullback-Leibler (AIC_c)

$$\min_{\tilde{\mathbf{h}}} AIC_{corrected}(\tilde{\mathbf{h}}),$$

with

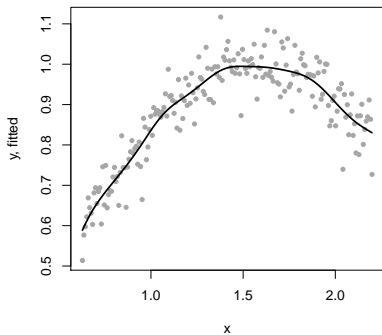
$$AIC_{corrected}(\tilde{\mathbf{h}}) = \ln \left(\frac{1}{l} \sum_{i=1}^l (y_i - \tilde{g}(\mathbf{x}_i, \tilde{\mathbf{h}}))^2 \right) + P_{corrected}$$

and

$P_{corrected}$ as a penalty function to avoid a bandwidth of zero.

Scaling the bandwidths

$h = 0.083$



$h = 0.125$

