Jensen, Cathrine Ulla; Panduro, Toke Emil

Working Paper

PanJen: A test for functional form with continuous variables

IFRO Working Paper, No. 2016/08

Provided in Cooperation with:
Department of Food and Resource Economics (IFRO), University of Copenhagen

Suggested Citation: Jensen, Cathrine Ulla; Panduro, Toke Emil (2016) : PanJen: A test for functional form with continuous variables, IFRO Working Paper, No. 2016/08, University of Copenhagen, Department of Food and Resource Economics (IFRO), Copenhagen

This Version is available at:
https://hdl.handle.net/10419/204400

DEPARTMENT OF FOOD AND RESOURCE ECONOMICS
UNIVERSITY OF COPENHAGEN

IFRO Working Paper

PanJen:
A test for functional form with continuous variables

Cathrine Ulla Jensen
Toke Emil Panduro

2016 / 08

IFRO Working Paper 2016 / 08

PanJen: A test for functional form with continuous variables

Authors: Cathrine Ulla Jensen, Toke Emil Panduro

JEL Classification: C01, C52

Published: October 2016

See the full series IFRO Working Paper here:
www.ifro.ku.dk/english/publications/foi_series/working_papers/

Department of Food and Resource Economics (IFRO)
University of Copenhagen
Rolighedsvej 25
DK 1958 Frederiksberg DENMARK
www.ifro.ku.dk/english/

PanJen: A test for functional form with continuous variables

Corresponding author
Cathrine Ulla Jensen

Contact details
Email: cuj@ifro.ku.dk
Phone +45 3533 3677

Authors:
Cathrine Ulla Jensen and Toke Emil Panduro

Affiliation:
Department of Food and Resource Economics, Faculty of Science, University of Copenhagen, Rolighedsvej 23, 1958 Frederiksberg Copenhagen, Denmark

Abstract

PanJen provides users the opportunity to explore the relationship between a dependent variable and its covariates with minimal restrictions. The package offers an easy and data-driven way to choose a functional form in multiple linear regression models by comparing a range of parametric transformations. The parametric functional forms are benchmarked with a non-parametric smoothed relationship. The package allows users to generate plots that show the functional form relationship between the explanatory variable and the dependent variable. Furthermore, PanJen allows users to specify specific functional transformations, driven by an a priori and theory-based hypothesis. The plots and model fit metrics enable users to make an informed choice of how to specify the functional form the regression. We show that the PanJen ranking outperforms the Box-Tidwell transformation, especially in the presence of inefficiency, heteroscedasticity or endogeneity.

Keywords

linear regression, functional form, semi-parametric

1 Introduction

The functional form in a regression model describes the relationship between a dependent variable and its covariates. There are numerous examples of researchers who have neglected to reflect on functional form relationships and applied the default linear relationship between variables in their regression models (8,9,3,2). From a superficial point of view, these models may provide efficient parameter estimates with narrow standard errors, high t-values and significance. A strong non-linear relationship between a dependent variable and an independent variable will often provide reasonable test statistics with a default linear functional form specification. However, positive test statistics are clearly not the same as proof of a linear relationship. Even more important is the fact that a misspecified functional form can lead to a wrong interpretation and prediction of the relationship between the dependent variable and a given covariate. We believe that researchers should give the functional form in their regression model the attention it deserves. The specification should be driven by theory with an a priori hypothesis of the relationship between the dependent and independent variables. However, a transformation should not be forced through, and there are circumstances in which theory provides little or no guidance regarding the question of functional form. A priori hypotheses should be tested, and this calls for an initial flexible form. There are situations in which a data-driven search for the best fit is justifiable (26).

A model that fits data well but is unrelated to theory is limited to describing correlations, whereas a model with both a good fit and a theoretically sound foundation can give insights to hypotheses on causality. The PanJen package was developed over several years of applied research on property value models. Here, the sales price is estimated as a function of it characteristics, such as the size of the living space, the number of rooms, and access to shopping. However, the package is applicable in most cases in which the relationship between a continuous dependent and its covariates is explored. In the applied econometric literature on house prices, this task has previously been solved using power transformations in initial analyses (23). Power transformations such as Box-Cox and Box-Tidwell are used to transform dependent and independent variables with the objective to obtain the highest possible model fit (7, 8). These transformations can be difficult to interpret, do not necessarily relate back to a theory-driven hypothesis and do not detect whether the relationship is changing across the distribution. Another approach to the functional form issue is to abandon the parametric model altogether and approach the challenge from a non- or semi-parametric angle. Non- or semi-parametric models provide a data-driven approach to establish the relationship between dependent and independent variables. In many situations, a non-parametric model approach is more attractive as the functional form is given by data and does not need to be predefined. However, the non-parametric approach comes at a cost. Non-parametric models do not provide one parameter estimate of the relationship between the dependent and independent variables; instead it provides local ones where the relationship often is depicted in a plot. In non-parametric models, the relationship is fitted to the sample to the extent that the estimated relationship is at risk of being over-fitted. That is, the estimated effect captures random error or noise in combination with the underlying relationship in the population (27). A main criticism of the non-parametric method is that the model results are case specific and difficult to generalize and extend outside the sample (22). The main objective of the PanJen package is to bridge the gap between parametric and non-parametric methods by providing users a ranked set of transformations that indicate which parametric transformation captures the most variance of the dependent variable. The ranked specification includes a non-parametric specification that uses a generalized additive model (GAM) and thin plate-smoothing regression splines. This non-parametric relationship can be used as a benchmark to compare a number of parametric specifications. In

the case that the relationship is ill-captured by a parametric form, the flexible non-parametric model will outperform less flexible forms.

The idea of the PanJen package was incubated in a research environment mainly focused on property value models, but it is readily applicable in situations that require the definition of a relationship between a continuous dependent and independent variable in a linear regression model. The literature on property value models has a long tradition of estimating property prices as a function of property characteristics such as the size of the living space, the number of rooms, and distance to parks. Many years of combined experience in the literature on valuation models have converged to provide guidance on the parametric specification with respect to major characteristics such as living space. However, for numerous minor variables, such guidance is not available. We believe that the PanJen packages will help establish a common understanding of the relationship between the dependent variable and a wide range of covariates within the literature on property valuation models and in other fields concerned with continuous variables.

Power transformations such as Box-Cox and Box-Tidwell were suggested in the 1960s (5,6). The shortcomings of power transformations are well described in the current literature (19, 28). While power transformations perform well in many circumstances, they struggle with omitted or variables with a high variance. Nonetheless, the ability of power transformations to detect functional forms resulted in extensive applications in the academic literature, e.g., (18, 10, 13), and are still used in applied studies (14, 15, 20). In the academic literature, a number of alternatives have been proposed and used, such as non-parametric or semi-parametric methods (1,16,12,4,17). The gain in flexibility comes at the cost of interpretation, which is perhaps why parametric models are often used in applied work.

In property valuation, the focal point is not just the functional form but the control of relationships across space that spill over through other variables in the pricing function. One way to do this without imposing assumptions on the functional form is through smoothing splines that control for space and time. This has been performed using a GAM (29, 24). We use the same approach to show the true functional form for a given covariate. PanJen targets practitioners who want the best from both the parametric and non-parametric world. It allows both a data-driven functional form and the simplicity of well-known parametric models.

In the next section, we briefly describe how semiparametric models work in a GAM context. In section 3, we explain how the PanJen ranking works. In section 4, we show how to use the package with a real example from our own research. In section 5, we compare the PanJen ranking with the Box-Tidwell transformation. We simulated 10,000 datasets and recovered the functional form of one variable in a model with different impediments to show the merits of PanJen relative to the conventional approach. In section 6, we conclude the paper with a short discussion of when the package is relevant.

2 A semi-parametric model for benchmark

Generalized additive models (GAM) are generalized linear models that contain one or more smoothing functions of continuous independent variables. The GAM can be written as follows:

$$Y_i = X_i\beta + f_1(x_{1i}) + \epsilon_i \quad \textbf{(1)}$$

Y_i is the dependent variable of observation i that consists of any type of exponential family distribution. X_i is a matrix of independent variables that are parametrically related to the dependent variable. β is the

corresponding vector of the parameter estimate, and f_1 is a smoothing function of independent variable x_{1i}. The model described in equation 1↑ is similar to the smoothing function applied in the PanJen ranking.

The GAM allows a flexible specification of independent variables by only specifying as a smooth function, i.e., it is possible to avoid specifying a specific functional relationship to the dependent variable and obtain the smoothing function to derive a data-driven relationship. The smooth function comprises the sum of k thin plate regression spline bases $b_h(\bullet)$ multiplied by their coefficients to be estimated: $f = \sum_{h=1}^{k} \beta_h b_h(x_1)$. The non-parametric component of the model $f(x_1;k)$ is fitted using thin plate regression splines with a penalty on "wiggliness". The penalty, θ, is determined from the data using generalized cross-validation or related techniques. The penalty enters the objective function directly through an additional term capturing "wiggliness" in the smooth function, i.e.,

$$\left\| Y_i - \hat{Y}_i \right\|^2 + \theta \int f''^{(x_1)2} dx_1 \quad \textbf{(2)}$$

Here, $\hat{Y}$ is the fitted dependent variable, and the second derivatives of the smooth function describe its "wiggliness". The objective function explicitly contains the trade-off between bias and variance (27). People who apply the GAM must choose the flexibility of the model by setting the number of basis functions, k. This is a balancing act between accurately capturing the attribute without overfitting the model, although the penalty term also reduces the probability of overfitting. We estimate the generalized additive model using the mgcv R-package (25).

3 The main idea of the PanJen Ranking

The PanJen package is built on the idea that the choice of a functional form can be extrapolated from model fit measures. In the PanJen ranking, a given number of similar models are estimated, except for a specific independent variable, which enters into the models with a different transformation. The models are then ranked according to the Akaika information criterion (AIC). The AIC provides the relative goodness-of-fit measure while accounting for the complexity of the model. More formally, the PanJen ranking estimates a model $Y = X\beta + g(x) + \varepsilon$, where Y is the dependent variable, X is a matrix of independent variables that are parametrically related to the dependent variable, β is the corresponding vector of the parameter estimate, and $g(x)$ represents a set of functional form transformations:

$$g(x) = \left\{ \frac{1}{x^2}, \frac{1}{x}, \frac{1}{\sqrt{x}}, \log(x), \sqrt{x}, x, x^2, f(x), 0 \right\} \quad \textbf{(3)}$$

The ranked values AIC shows how each transformation performs relative to the others. By including a semi-parametric transformation, it is possible to assess how well parametric transformations perform relative to a flexible non-parametric function. In most cases, the expectation is that the smoothing will capture the relationship better than a parametric transformation. The AIC scores are supplemented with the closely related Bayesian information criterion (BIC). Both the AIC and the BIC penalize the model complexity, although the penalty term in BIC is larger (11). In cases where the "true" functional form relationship is closely related to a specific functional form, the model complexity introduced by the non-parametric smooth term may result in better parametric performance.

The PanJen ranking is supported by a plot function that graphically outlines the relationship between the dependent variable and independent variable x. The plot is created by predicting the dependent variable using the median for all covariates other than the one in question. The covariate in question varies across

a scale from the 5th quantile to the 95th quantile of the actual distribution in the dataset. The PanJen plot shows the user how each transformation captures the relationship across the distribution of the dependent variable. If the smoothing spline far outperforms all parametric transformations, the reason may be that the relationship changes across the distribution. The plot will reveal whether this is the case. When a parametric transformation fits the relationship only in the tails or around the median, the results should be interpreted with caution. Vice versa, if a parametric transformation performs well across the whole distribution, it is a close approximation of the "true" relationship within the data. The PanJen packages also allow users to define other transformations that they deem more appropriate to their specific model challenge and aligned to their hypothesis of the relationship.

4 Using the package

We illustrate the use of the PanJen package by estimating a house pricing function (23). For this example, you only need to know that we model the price of the home as a function of its qualities, measured by a range of variables. We do not know a priori how the characteristics of the house are related to the price. One example is that we expect the price to increase with size and decrease with age, but we do not know whether it is a linear relationship or there are marginal increasing or decreasing effects to take into account. It is this type of question that PanJen was developed to answer, i.e., to find the functional form relationship between the dependent and independent variables.

4.1 An example: the cost of more living area

The packages feature a dataset called "hvidovre". It includes 901 single detached homes sold between 2007 and 2010 with a Danish municipality called "Hvidovre". The dataset was compiled from different Danish databases as a part of a larger hedonic study on households' willingness to pay for different urban and recreational services (21).

We have 9 continuous and 7 dummy variables for quality at our disposal. In addition, the dataset includes 3 year dummies to control for price trends. The variables are listed in table 1.

Table 1: Variables

Continuous		Dummy	
Variable	**Description**	**Variable**	**Description**
lprice	sales price, log price in 1000 EUR	rebuild70	home rebuilt in 1970s
area	living area in square meters	rebuild80	home rebuilt in 1980s
age	year built	rebuild90	home rebuilt in 1990s
bathrooms	number of bathrooms	rebuild00	home rebuilt in 2000s
lake_SLD	distance to nearest lake in meters	brick	Construction made out of brick =1
highways	distance to nearest highway in meters	roof_tile	roof made out of tiles =1
big_roads	distance to nearest large road in meters	roof_cement	roof made out of cement =1
railways	railways distance to nearest railway in	y7,y8,y9	home sold in 2007, 2007, or

	meters	2009
nature_SLD	SLDdistance to nearest nature area in meters	

First, we load the package and the dataset:

```
library(PanJen)

## Loading required package: mgcv
## Loading required package: nlme
## This is mgcv 1.8-14. For overview type 'help("mgcv-package")'.
## Loading required package: RColorBrewer
## Loading required package: Formula
## Loading required package: lasso2
## R Package to solve regression problems while imposing
##   an L1 constraint on the parameters. Based on S-plus Release 2.1
## Copyright (C) 1998, 1999
## Justin Lokhorst   <jlokhors@stats.adelaide.edu.au>
## Berwin A. Turlach <bturlach@stats.adelaide.edu.au>
## Bill Venables     <wvenable@stats.adelaide.edu.au>
##
## Copyright (C) 2002
## Martin Maechler maechler@stat.math.ethz.ch

data("hvidovre")
```

Then, we set up a formula object. Ten of the variables are dummies for which transformations are irrelevant. We include these only in the first regressions:

```
formBase<-formula(lprice ~brick+roof_tile+roof_cemen+ rebuild70+rebuild80
+rebuild90+rebuild00+y7+y8+y9)
summary(lm(formBase, data=hvidovre))

## Call:
## lm(formula = formBase, data = hvidovre)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2358 -0.1282  0.0217  0.1635  1.0298
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.61058    0.02909 192.902  < 2e-16 ***
## brick        0.11530    0.02660   4.334 1.63e-05 ***
## roof_tile    0.06238    0.02328   2.679 0.007511 **
## roof_cemen   0.08845    0.02969   2.979 0.002969 **
## rebuild70    0.08357    0.03158   2.646 0.008285 **
## rebuild80    0.14506    0.04382   3.310 0.000970 ***
## rebuild90    0.14718    0.05356   2.748 0.006122 **
## rebuild00    0.21120    0.04275   4.940 9.33e-07 ***
## y7           0.14188    0.02653   5.347 1.14e-07 ***
## y8           0.09193    0.02847   3.229 0.001286 **
## y9          -0.09633    0.02784  -3.460 0.000566 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2965 on 890 degrees of freedom
## Multiple R-squared:  0.1555, Adjusted R-squared: 0.146
## F-statistic: 16.39 on 10 and 890 DF, p-value: < 2.2e-16
```

The initial model explains just above 15% of the variation in price. The coefficients are significant and exhibit signs in line with expectations. The Danish housing market experienced a steep decline starting around ultimo 2007 and the trend did not reverse until the end of 2009 or later. This is a trend we also discover in our preliminary results. A home traded in 2009 sold for nearly 9% less than in 2010, and a home in 2007 sold for just above 14% more than in 2010.

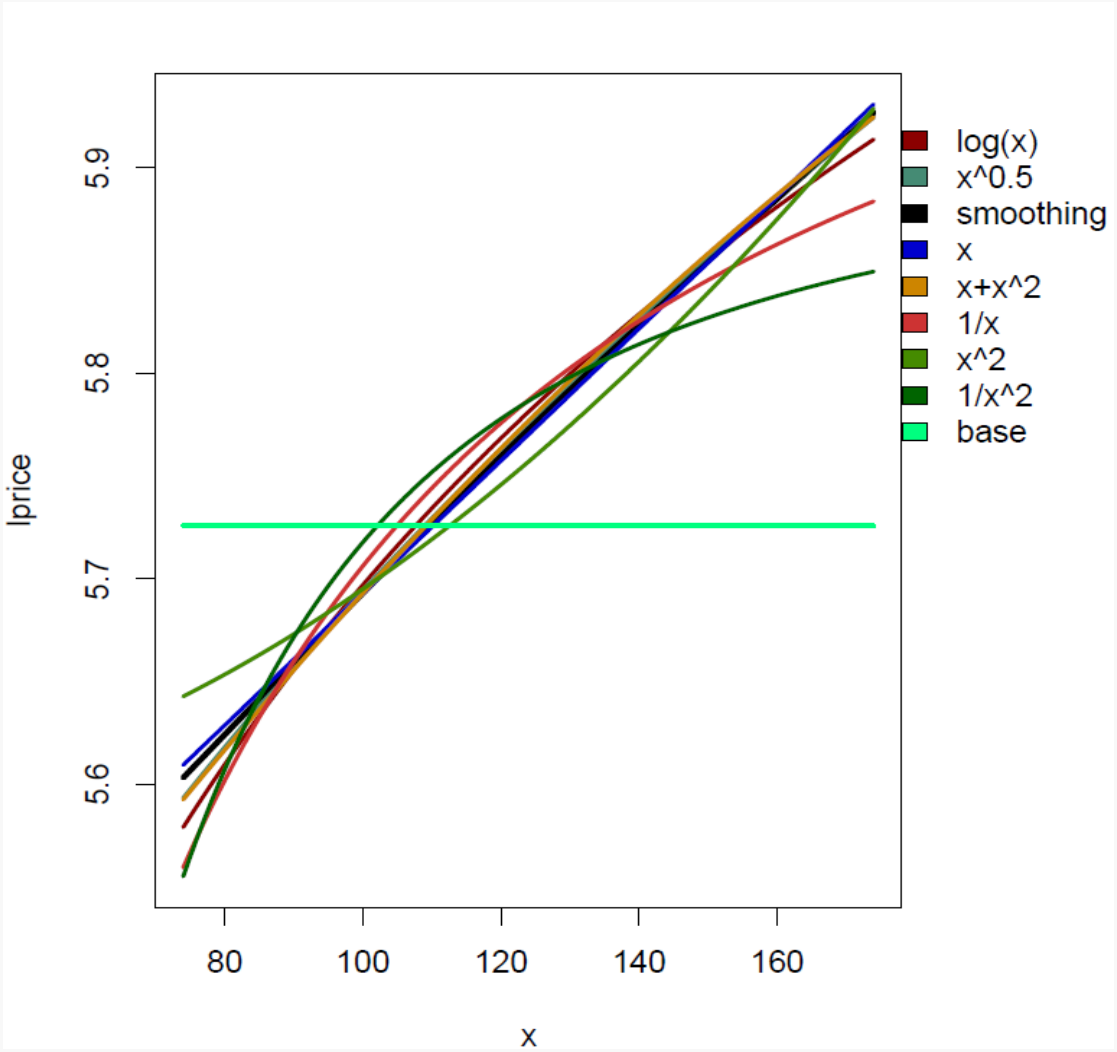
The first attribute we want to add is the size of the home. The living area in square meters is stored under "*area*".

We start out by using the default transformations supplied by the PanJen function *fform().* This function ranks the fit of nine predefined transformations and a smoothing. The mandatory inputs are the name of the dataset, the model formula and the new variable we wish to test using the PanJen ranking:

```
PanJenArea<-fform(hvidovre,"area",formBase)

##                AIC      BIC ranking (AIC)
## log(x)    290.0449 352.4905             1
## x^0.5     290.0884 352.5340             2
## smoothing 291.5307 355.8986             3
## x         291.7010 354.1465             4
## x+x^2     292.4278 359.6769             5
## 1/x       294.4382 356.8838             6
## x^2       299.1506 361.5962             7
## 1/x^2     303.3169 365.7625             8
## base      379.1957 436.8377             9
```

The results are ranked according to their Akaike information criteria (AIC). Strictly according to this ranking, we should log-transform the area. This implies that a % change in living area results in a % change in price. The differences in score for the four lowest AIC are small, and it might be a matter of differences in the tails of the distribution. This can be checked by plotting the predicted price against the area. *plot.PanJen()* generates a plot with the predicted price against the area from the 5th to the 95th percentiles with all other covariates variables at their median value:

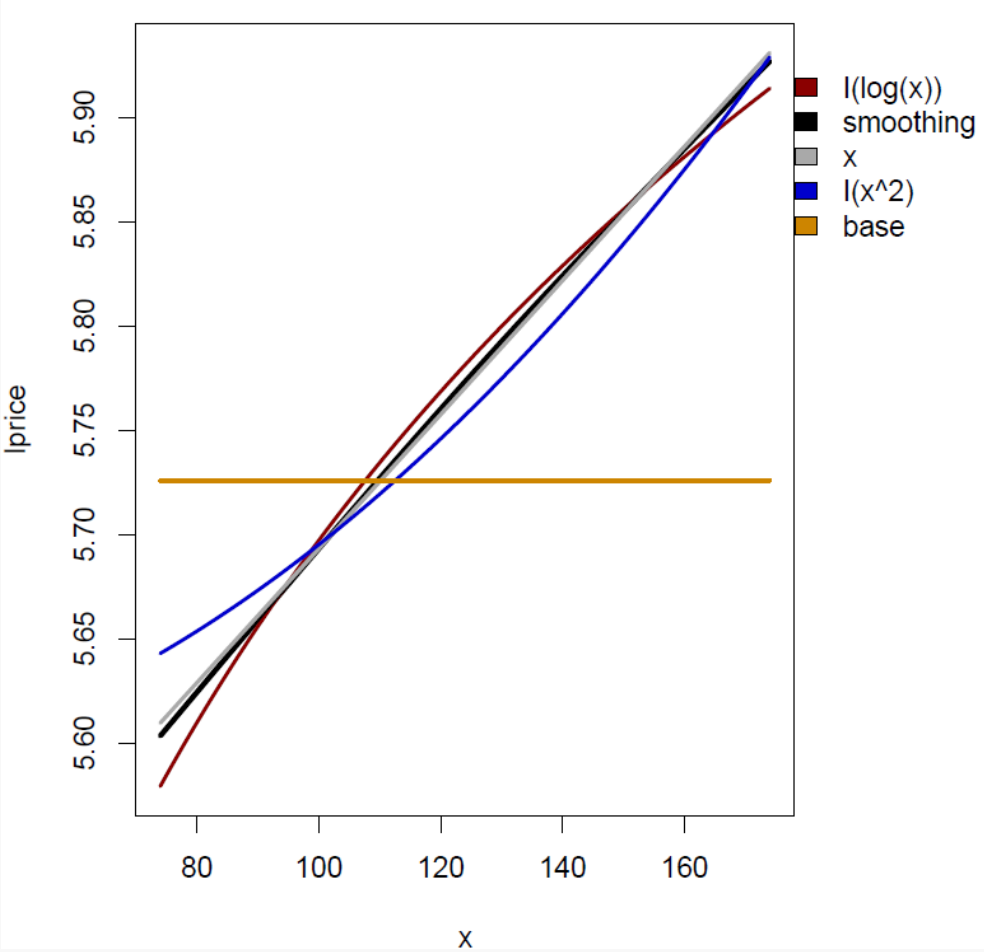
```
plotff(PanJenArea)
```



The black line is the smoothing function. The three with the lowest AIC differ mainly in the tails. Either way, the conclusion is that the marginal willingness to pay for living area is positive and slightly marginally declining. We can focus on a subset of these transformations by using *choose.fform()* and then plot them.

You can specify your own transformations by adding them to the formula object and then using the *choose.form()* function. In the following, we test three transformations: "*area*", "*log(area)*" and "$area^2$":

```
formsArea<-formula(lprice ~brick+roof_tile+roof_cemen+ rebuild70+rebuild80
                     +rebuild90+rebuild00+y7+y8+y9|x|I(log(x))|I(x^2))
PanJenAreaC<-choose.fform(hvidovre,"area",formsArea)

##                 AIC      BIC ranking (AIC)
## I(log(x)) 290.0449 352.4905             1
## smoothing 291.5307 355.8986             2
## x         291.7010 354.1465             3
## I(x^2)    299.1506 361.5962             4
## base      379.1957 436.8377             5

plotff(PanJenAreaC)
```



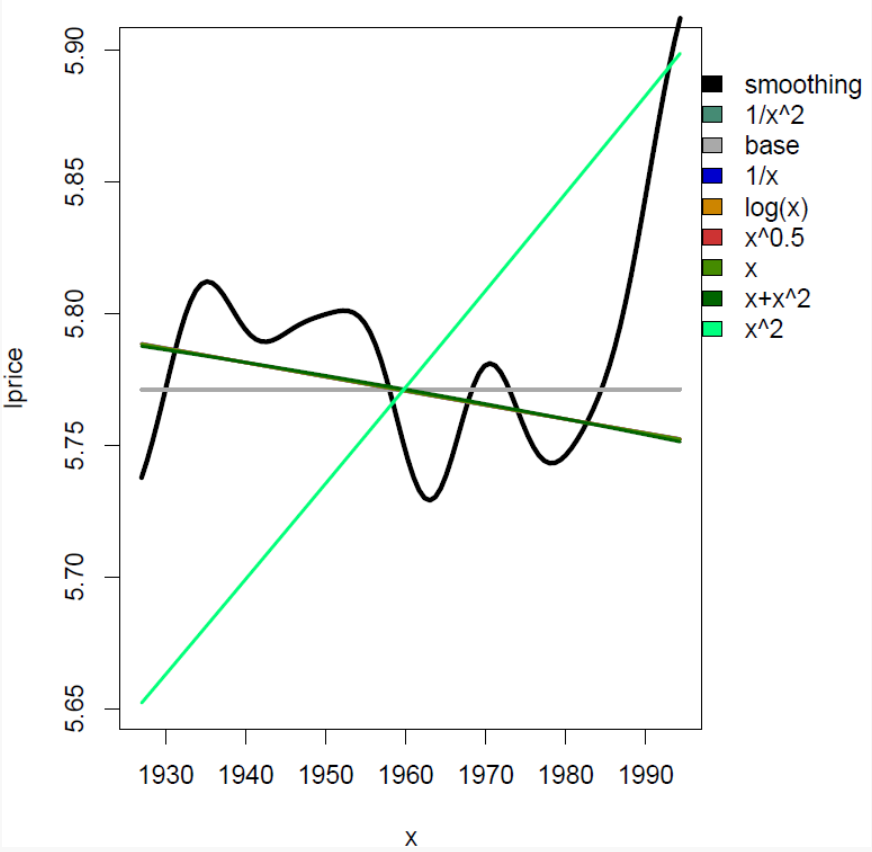
We log-transform the area, clean up the regression and are now able to explain nearly 23% of the variation in price:

```
formArea<-formula(lprice ~brick+rebuild80+rebuild90+rebuild00+y7+y8+y9+log(area))
summary(lm(formArea, data=hvidovre))

##
## Call:
## lm(formula = formArea, data = hvidovre)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2109 -0.1000  0.0194  0.1479  0.9255
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  3.76482    0.17863  21.076 < 0.0000000000000002 ***
## brick        0.08695    0.02477   3.510             0.000471 ***
## rebuild80    0.07485    0.04224   1.772             0.076714 .
## rebuild90    0.11285    0.05084   2.220             0.026690 *
## rebuild00    0.12213    0.04126   2.960             0.003159 **
## y7           0.14800    0.02517   5.880        0.00000000579 ***
## y8           0.09012    0.02704   3.333             0.000894 ***
## y9          -0.10620    0.02645  -4.015        0.00006453604 ***
## log(area)    0.40308    0.03793  10.627 < 0.0000000000000002 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2818 on 892 degrees of freedom
## Multiple R-squared: 0.2352, Adjusted R-squared: 0.2284
## F-statistic: 34.3 on 8 and 892 DF, p-value: < 0.00000000000000022
```

4.2 A changing relationship

We would expect the age of the home to matter for the price, but it is not given how. A newly built home is up to date, an old house can be charming and authentic, and homes built during the building boom in Denmark in the 1960s may seem cheap due to poor materials. We define a formula object without the age of the home and run *fform*():

```
  PanJenAge<-fform(hvidovre,"age",formArea)

##                AIC      BIC ranking (AIC)
## smoothing 272.2769 379.2664              1
## 1/x^2     285.8198 333.8549              2
## base      285.8198 333.8549              3
## 1/x       286.7114 339.5500              4
## log(x)    286.7154 339.5539              5
## x^0.5     286.7173 339.5558              6
## x         286.7191 339.5577              7
## x+x^2     286.7567 339.5953              8
## x^2       359.7224 407.7575              9

plotff(PanJenAge)
```



Here, the smoothing far outperforms all seven transformations. The best parametric transformation is $\frac{1}{x^2}$, but it is only slightly better than not just controlling for age. In conclusion, none of the tested parametric transformations captures the relation. Given the plot, it is difficult to think of a parametric relation that will. If the age of the home is somehow related to the research question, the best solution might be to actually use the smoothing function. If age is nothing other than a control variable, one could perhaps resolve to interval dummies similar to the year dummies in the model. As a part of the final model testing, it would be worthwhile to test to what degree the variable of interest is robust to the way age enters the pricing function. In our setting, what should be noted is that the complexity of the relation between age and price would have gone unnoticed if we had compared only the parametric transformations.

4.3 Interacting with other packages

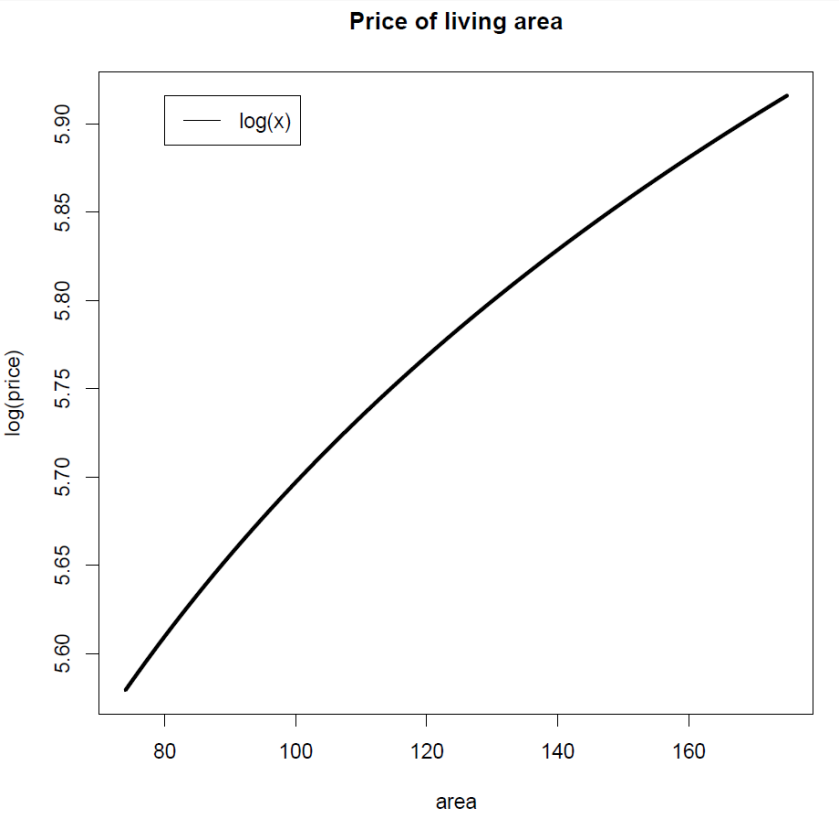
When you run *choose.form()* or *fform(),* all generated models and datasets are stored in a new "list of list" object. Within the list "models", all estimated models are stored as "*gam*", "*glm*" and "*lm*" objects. This means that all procedures used for those objects, such as plots, predictions or other diagnostics, are easily available. Here, we show how to use this to make good plots. The plotting function in PanJen is simple, and it works for a search for functional form but perhaps not for producing plots for a third party.

For example, you can create a new plot of just one transformation using *"predict"* from mgcv and a base R plot:

```
library(PanJen)
data("hvidovre")

# setting the formular
formBase<-formula(lprice ~brick+roof_tile+roof_cemen
                  + rebuild70+rebuild80+rebuild90+rebuild00+y7+y8+y9)

# running the PanJen ranking
PanJenArea<-fform(hvidovre,"area",formBase)

##                AIC      BIC ranking (AIC)
## log(x)    290.0449 352.4905             1
## x^0.5     290.0884 352.5340             2
## smoothing 291.5307 355.8986             3
## x         291.7010 354.1465             4
## x+x^2     292.4278 359.6769             5
## 1/x       294.4382 356.8838             6
## x^2       299.1506 361.5962             7
## 1/x^2     303.3169 365.7625             8
## base      379.1957 436.8377             9

# names of models
names(PanJenArea$models)
```

```
## (1) "model_log(x)"    "model_x^0.5"     "model_smoothing" "model_x"
## (5) "model_x+x^2"     "model_1/x"       "model_x^2"       "model_1/x^2"
## (9) "model_base"

# getting the variable names used in the log model transformation
nV<-all.vars(formula(PanJenArea$models(("model_log(x)"))))(1:11)

# creating a prediction dataframe with median values
pred_frame<-data.frame(matrix(rep(sapply(hvidovre(nV),median),each=100),nrow=100))
# giving the prediction dataframe variable names
names(pred_frame)<-nV

# Finding the 0.05 quantile and the 0.95 quantile of the area variable
min05<-as.numeric(quantile(hvidovre$area,0.05))
max95<-as.numeric(quantile(hvidovre$area,0.95))


# Changing the area variable in the prediction dataframe to a scale that goes from
the 0.05 quantile to the 0.95 quantile of the area variable
pred_frame$variable<-seq(min05,max95,length.out=100)

# predicting the lprice using the prediction dataframe
pred_frame$lprice=predict(PanJenArea$models(("model_log(x)"))
                          ,newdata=pred_frame, type="response")

# Defining limits for plot
limx=c(min(pred_frame$variable),max(pred_frame$variable))
limy=c(min(pred_frame$lprice),max(pred_frame$lprice))
```

```r
# ploting the predicted lprice as afunction of the scaled area variable
plot(pred_frame$lprice~pred_frame$variable,
     data=pred_frame, type="l",sub="",xlab="area",ylab="log(price)",
     lwd=3, col="black", xlim=limx, ylim=limy, main="Price of living area")


# Adding a legend
legend(80,limy(2),cex=1,lty=1,"log(x)", horiz=FALSE)
```



5 A test of performance

We tested the performance of the PanJen ranking against the well-known Box-Tidwell transformation (7). The basic idea behind the Box-Tidwell transformation is to find the transformation that minimizes non-normality in the error term and linearizes the relationship between the dependent variable and the covariate. It tests a range of power transformations by a maximum likelihood function. The transformation enables the researcher to determine a functional form relationship between a dependent variable and one covariate. From a superficial point of view, the Box-Tidwell transformation is very

similar to the PanJen ranking. However, the approach to obtaining the result is dissimilar, and this is what drives the differences in their performance.

In the table below, we present nine Monte Carlo simulations of 10,000 runs that compare the ability of PanJen and Box-Tidwell to recover the underlying functional form. The basic idea is that we create three random variables and define how they are related to a fourth dependent variable. In this way, we define a data-generating process and know the true functional form. We then establish a base model:

$$Y = x_1\beta_1 + x_2\beta_2 + f(x_3) + \varepsilon$$

where x_3 is the variable of interest, and x_1 and x_2 are two covariates.

The functional relationship between $\hat{Y}$ and $f(x_3)$ is tested by PanJen and the Box-Tidwell transformation. The table shows the share (in percent) of the 10.000 simulations that PanJen and Box-Tidwell reported the true functional form. In the Box-Tidwell, case the transformation parameter was allow to vary by 0.2 from the correct specification while still being identified as correct.

Table 2: Simulation results

Simulation	Simulation description	PanJen	Box-Tidwell
Identification	Relationship squared $f(x_3) = x_3^2$	100	96
Identification	Relationship linear $f(x_3) = x_3$	99	90
Identification	Relationship root $f(x_3) = x_3^{0.5}$	99	99
Efficiency	Relationship squared $f(x_3) = x_3^2$ and with a high variance	99	64
Collinearity	Relationship squared $f(x_3) = x_3^2$ while being highly correlated with x_2	100	95
Omitted variables	Relationship squared $f(x_3) = x_3^2$ while being correlated with an omitted variable	100	92
Heteroscedasticity	Relationship squared $f(x_3) = x_3^2$, and x_3 suffers from heteroscedasticity	99	59
Endogeneity	Relationship squared $f(x_3) = x_3^2$, while x_3 is endogenous	99	3
Misspecification	Relationship squared $f(x_3) = x_3^2$, while x_2 is misspecified	100	95

Each of the nine simulations tested the robustness of the methods in relation to different well-known econometric challenges. Overall, the PanJen ranking performed acceptably, obtaining the correct functional form relationship from 99 to 100% of the times. The Box-Tidwell transformation performed less well and could only match the PanJen ranking in the simulation, where the underlying relationship was square root. In the simulations concerned with parameter efficiency, heteroscedasticity and endogeneity, the Box-Tidwell transformation performed poorly. The poor performance of Box-Tidwell is the product of how the method works and thus is not a surprise. High variance, heteroskedasticity and endogeneity pose serious problems for the Box-Tidwell method.

6 Conclusion

In this paper, we present the PanJen package. We provide a simple and intuitive description of the PanJen ranking. Based on a house price dataset, we show how the functions in the package can be applied to determine the relationship between a dependent variable and its covariates. Furthermore, we compare the PanJen ranking method to the Box-Tidwell transformation and show, using Monte Carlo simulation, that the PanJen ranking outperforms the Box-Tidwell transformation, especially in situations where the independent variable suffers from efficiency, heteroscedasticity or endogeneity.

In some circumstances, theory provides little or no guidance on the functional relationship between the dependent and independent variables in a multiple regression model. In such circumstances, the PanJen package can support users in their decision on the functional form of the independent variables. If the functional form relationship is more complex than a simple parametric transformation, we suggest considering a semi- or non-parametric model. The package has deliberately been restricted to test one independent variable at a time based on the recognition that each variable included in a multiple regression model is hypothesized to be tested. It is our sincere hope that people will use the PanJen package and improve their models by specifying relationships that more accurately fit their data. In doing so, users should still consider PanJen as a guide and a helping hand in their initial data analysis. It is not a substitute for a priori hypotheses.

7 Bibliography

(1) Paul M Anglin, Ramazan Gençay: "Semiparametric estimation of a hedonic price function", *Journal of Applied Econometrics*, pp. 633—648, 1996.

(2) Joshua D Angrist, Jorn-Steffen Pischke: "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics", *The Journal of Economic Perspectives*, pp. 3-30, 2010.

(3) Richard A Berk: *Regression analysis: A constructive critique*. Sage, 2004.

(4) Okmyung Bin: "A prediction comparison of housing sales prices by parametric versus semi-parametric regressions", *Journal of Housing Economics*, pp. 68—84, 2004.

(5) G. E. P. Box, D. R. Cox: "An analysis of transformations", *Journal of the Royal Statistical Society. Series B (Methodological*, pp. 211—252, 1964.

(6) GEP Box, PW Tidwell: "Transformation of the independent variables", *Technometrics*, 1962. URL http://amstat.tandfonline.com/doi/abs/10.1080/00401706.1962.10490038.

(7) GEP Box, PW Tidwell: "Transformation of the independent variables", *Technometrics*, 1962. URL http://amstat.tandfonline.com/doi/abs/10.1080/00401706.1962.10490038.

(8) George EP Box: “Science and statistics”, *Journal of the American Statistical Association*, pp. 791—799, 1976.

(9) Leo Breiman, others: “Statistical modeling: The two cultures (with comments and a rejoinder by the author)”, *Statistical Science*, pp. 199—231, 2001.

(10) Thomas P Brennan, Roger E Cannaday, Peter F Colwell: “Office rent in the Chicago CBD”, *Real Estate Economics*, pp. 243—260, 1984.

(11) Kenneth P Burnham, David R Anderson: “Multimodel inference understanding AIC and BIC in model selection”, *Sociological methods & research*, pp. 261—304, 2004.

(12) J Clapp, C Giaccotto: “Evaluating house price forecasts”, *Journal of Real Estate Research*, pp. 26, 2002. URL http://ideas.repec.org/a/jre/issued/v24n12002p1-26.html.

(13) Jeffrey A Clark: “Estimation of economies of scale in banking using a generalized functional form”, *Journal of Money, Credit and Banking*, pp. 53—68, 1984.

(14) Jacob Cohen, Patricia Cohen, Stephen G West, Leona S Aiken: *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge, 2013.

(15) Bilal Farooq, Eric Miller, Murtaza Haider: “Hedonic analysis of office space rent”, *Transportation Research Record: Journal of the Transportation Research Board*, pp. 118—127, 2010.

(16) Ramazan Gençay: “A statistical framework for testing chaotic dynamics via Lyapunov exponents”, *Physica D: Nonlinear Phenomena*, pp. 261—266, 1996.

(17) Ghislain Geniaux, Claude Napoleone: *Semi-parametric tools for spatial hedonic models: An introduction to mixed geographically weighted regression and geoadditive models* in *Hedonic Methods in Housing Markets: Pricing Environmental Amenities and Segregation*. 2008.

(18) Joseph G Kowalski, Peter F Colwell: “Market versus assessed values of industrial land”, *Real Estate Economics*, pp. 361—373, 1986.

(19) Andrew Levin, Russell Davidson, James G. MacKinnon: “Estimation and Inference in Econometrics.”, *Journal of the American Statistical Association*, pp. 1143, 1993.

(20) Heike Link: "A cost function approach for measuring the marginal cost of road maintenance", *Journal of Transport Economics and Policy (JTEP)*, pp. 15—33, 2014.

(21) Thomas Lundhede, Toke Emil Panduro, Linda Kummel, Alexander Staahle, Axel Heyman, Bo Jellesmark Thorsen: *Værdisætning af bykvaliteter-fra hovedstad til provins: Appendiks*. Institut for Fødevare-og Ressourceøkonomi, Københavns Universitet, 2013.

(22) Daniel P. McMillen, Christian L. Redfearn: "Estimation and hypothesis testing for nonparametric hedonic house price functions", *Journal of Regional Science*, pp. 712—733, 2010.

(23) Raymond B Palmquist: "Property value models", *Handbook of environmental economics*, pp. 763—819, 2005.

(24) T E Panduro, B J Thorsen: "Evaluating two model reduction approaches for large scale hedonic models sensitive to omitted variables and multicollinearity", *Letters of spatial and resource sciences*, pp. 85—102, 2014.

(25) S. N. Wood: "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models", *Journal of the Royal Statistical Society (B)*, pp. 3-36, 2011.

(26) R M Sakia: "The Box-Cox Transformation Technique: A Review", *Journal of the Royal Statistical Society. Series D (The Statistician)*, pp. 169—178, 1992. URL http://www.jstor.org/stable/2348250.

(27) Simon Wood: *Generalized additive models: an introduction with R*. CRC press, 2006.

(28) Jeffrey M Wooldridge: "Some Alternatives to the Box-Cox Regression Model", *International Economic Review*, pp. p 935—955, 1992. URL \url{http://www.jstor.org/stable/2527151}">\url{http://www.jstor.org/stable/2527151}.

(29) Kathrine von Graevenitz, Toke Emil Panduro: "An Alternative to the Standard Spatial Econometric Approaches in Hedonic House Price Models", *Land Economics*, pp. 386—409, 2015.