KCG Working Paper

**Estimation of empirical models for margins of exports with unknown non-linear functional forms:
A Kernel-Regularized Least Squares (KRLS) approach**

Joachim Wagner

[This version: January 17, 2024]

Estimation of empirical models for margins of exports with unknown non-linear functional forms: A Kernel-Regularized Least Squares (KRLS) approach

Joachim Wagner

Abstract: Empirical models for intensive or extensive margins of trade that relate measures of exports to firm characteristics are usually estimated by variants of (generalized) linear models. Usually, the firm characteristics that explain these export margins enter the empirical model in linear form, sometimes augmented by quadratic terms or higher order polynomials, or interaction terms, to take care or test for non-linear relationships. If these non-linear relationships do matter and if they are ignored in the specification of the empirical model this leads to biased results. Researchers, however, can never be sure that all possible non-linear relationships are taken care of in their chosen specifications. This note uses for the first time the Kernel-Regularized Least Squares (KRLS) estimator to deal with this issue in empirical models for margins of exports. KRLS is a machine learning method that learns the functional form from the data. Empirical examples show that it is easy to apply and works well. Therefore, it is considered as a useful addition to the box of tools of empirical trade economists.

Joachim Wagner
Leuphana University Lueneburg and Kiel Centre for Globalization
D-21314 Lüneburg
Germany
joachim.wagner@leuphana.de

KIEL **CENTRE FOR**
GLOBALIZATION

1   Motivation

Empirical models for the intensive or extensive margins of trade that relate measures of exports to firm characteristics are usually estimated by variants of (generalized) linear models, including work-horse methods like ordinary least squares (for example, to explain the number of different countries a firm exports to), fractional logit (to take care of the fact that many firms do not export and, therefore, the share of exports in total sales is a variable with a probability mass at zero) or probit (for dichoto-mous variables like exporting or not). Usually, the firm characteristics that explain these export margins enter the empirical model in linear form, sometimes augmented by quadratic terms (like firm size and firm sized squared) or higher order polynomials, or interaction terms, to take care or test for non-linear relationships. If these non-linear relationships do matter and if they are ignored in the specification of the empirical model this leads to biased results.

Researchers, however, can never be sure that all possible non-linear relationships are taken care of in their chosen specifications, because the number of polynomials and interaction effects grows expo-nentially when the number of firm characteristics included in the empirical models for the trade mar-gins increases. One way out is the use of artificial neural networks. It is known from any textbook treatment of neural network models that they have a feature that is known as the "universal approxi-mation property". Properly designed neural networks can approximate any nonlinear relationship – and they will spot it in the data. The main disadvantage of this class of models for applications in eco-nomics is the impossibility of performing standard statistical inference for estimates of the model's parameters (see Lo (1994) for a short introductory exposition).

A second way out is the use of non-parametric regression, an appropriate alternative to standard regression models when we are unsure of the underlying functional form (see Henderson and Param-eter (2015) for a textbook treatment). One problem that makes the application of non-parametric re-gression models infeasible in the context of the estimation of empirical models for margins of exports is that they suffer from what is known as the "curse of dimensionality". Non-parametric regression models with a large number of control variables – and this includes all models with a set of dummy variables that control for industries or countries – are infeasible to estimate (see Cameron and Trivedi. 2022 p.1497).

This paper contributes to the literature by using kernel-based regularized least-squares (KRLS), in-troduced in Hainmueller and Hazlett (2014) and Ferwerda et al. (2017), and outlined in section 2 below. KRLS uses a machine learning approach to learn the functional form from the data. In doing so, it pro-tects against misspecification that leads to biased estimates. To the best of my knowledge KRLS has not been used before to estimate empirical models for margins of trade, and it has been used in the economics literature hitherto by Minviel and Ben Bouheni (2022) only in a study of the impact of re-search and development on economic growth with macro data.

To demonstrate the usefulness of the method for the estimation of intensive and extensive margins of exports this paper presents results from a study that replicates estimates reported in two papers of mine (Wagner, 2001; Wagner, 2023).

To anticipate the most important results, KRLS works fine for empirical models with continuous, fractional, and dichotomous endogenous variables and control variables that are continuous,

dichotomous, or dummy variables for industries or countries. In all three examples considered here the big picture from the original parametric models and from the models estimated by KRLS is the same. In several cases, however, the estimated average marginal effects from both models differ. These differences can be explained by the fact that the parametric model imposes a restrictive functional form in the shape of the estimated relationships, while KRLS estimated this relationship without imposing a functional form. Furthermore, KRLS reveals that the marginal effects are not constant – they are heterogeneous and tend to vary widely across the covariate space.

The rest of the paper is organized as follows. Section 2 outlines the KRLS estimator. Section 3 compares the original results from standard regression models for extensive and intensive margins of exports with the results from KRLS regressions. Section 4 concludes.

2   Kernel-Regularized Least Squares (KRLS) – A short outline

While a comprehensive discussion of the Kernel-Regularized Least Squares (KRLS) estimator is far beyond the scope of this applied note, a short outline of some of the important features and characteristics might help to understand why this estimator can be considered as an extremely helpful addition to the box of tools of empirical trade economists. For any details the reader is referred to the original papers by Hainmueller and Hazlett (2014) and Fernwerda et al. (2017).

The main contribution of the KRLS estimator is that it allows the researcher to estimate regression-type models without making any assumption regarding the functional form (or doing specification search to find the best fitting functional form). As detailed in Hainmueller and Hazlett (2014) the method constructs a flexible hypothesis space using kernels as radial basis functions and then finds the best-fitting surface in this space by minimizing a complexity-penalized least squares problem. Ferwerda et al. (2017) point out that the KRLS method can be thought of in the "similarity-based view" in two stages. In the first stage, it fits functions using kernels, based on the assumption that there is useful information embedded in how similar a given observation is to other observations in the dataset. In the second stage, it utilizes regularization, which gives preference to simpler functions (see Ferwerda et al., 2017 p.3).

The KRLS thus uses a machine learning approach to learn the functional form from the data. In doing so, it protects against misspecification that leads to biased estimates. Contrary to other methods mentioned in section 1 above KRLS allows for interpretability and inference in ways similar to the usual regression models – this is a great advantage over artificial neural networks - and it does not suffer from the curse of dimensionality, so it can deal with models that include many covariates and sets of dummy variables that control for industries or countries – a great advantage over nonparametric regression methods.

KRLS works well both with continuous outcomes and with binary outcomes. It is easy to apply in Stata using the `krls` program provided in Ferwerda et al. (2017). Instead of doing a tedious specification search that does not guarantee a successful result, users simply pass the outcome variable and the matrix of covariates to the KRLS estimator which then learns the target function from the data. As shown in Hainmueller and Hazlett (2014), the KRLS estimator has desirable statistical properties, including unbiasedness, consistency, and asymptotic normality under mild regularity conditions. An additional advantage of KRLS is that it provides closed-form estimates of the pointwise derivatives that

characterize the marginal effect of each covariate at each data point in the covariate space (see Ferwerda et al., 2017 p. 11). These estimates can be used to examine the heterogeneity of the marginal effects.

Therefore, KRLS is suitable to estimate empirical models when the correct functional form is not known for sure – which is usually the case because we do not know which polynomials or interaction terms matter for correctly modelling the relation between the covariates and the outcome variable.

3   KRLS in action: Replications of three empirical models for margins of exports

To see what we can learn from an application of the KRLS estimator we will have a close look at the results of estimates of three empirical models for different margins of exports taken from the literature that use different sets of firm-level data and three standard econometric methods, namely fractional logit (to estimate the share of exports in total sales, a fractional variable with a probability mass at zero due to a large number of non-exporting firms), probit (to estimate a model of participation in exports, a dichotomous variable that takes on the value of one or zero), and ordinary least squares (to estimate the number of firms' export destination countries, a continuous variable). While a discussion of the empirical models, the data and variables included and the theoretical hypotheses tested are beyond the scope of this short applied note and can be found in the original papers by Wagner (2001, 2023), we concentrate on a comparison of the results from the original methods used and from the alternative KRLS approach.

3.1.   Empirical model for share of exports in total sales

Table 1 reports results for an empirical model for the share of exports in total sales, defined as a fraction between zero and one. This intensive margin of exports was estimated in Wagner (2001) using the fractional logit model introduced by Papke and Wooldridge (1996) to deal with fractional variables with a probability mass at zero. Results in column 1 report the estimated average marginal effects (and its p-values) of the nine firm characteristics included in the empirical model. Note that this model includes *Firm size* (the number of employees) and *Firm size squared* to take care of the positive but decreasing effect of the number of employees on the share of exports in total sales. Furthermore, the model includes a set of 15 industry dummy variables as control variables.

[Table 1 near here]

Results for the average marginal effects estimated by KRLS (and its p-values) are reported in column 2 of table 1. A comparison of these estimates and the estimates reported in column 1 reveal that the signs are identical and the levels of significance are of a similar order of magnitude, so the big picture revealed by the two models is identical.

The estimated average marginal effects are of the same order of magnitude in five out of nine cases. KRLS estimates of average marginal effects are smaller for 3 variables and larger for one. The difference in the size of the average marginal effects can be explained by the fact that the parametric model in column 1 imposes a restrictive functional form in the shape of the estimated relationships, while KRLS estimated this relationship without imposing a functional form.

Note that KRLS was not "told" in advance to include a non-linear term (i.e. the squared number of employees). Note further that the inclusion of the 15 industry dummy variables does not pose a problem for KRLS, illustrating that this estimator is not hurt by the curse of dimensionality.

An additional advantage of KRLS compared to the parametric models used in the original estimation is that it provides closed-form estimates of the pointwise derivatives that characterize the marginal effect of each covariate at each data point in the covariate space. (see Ferwerda et al. 2017 p. 11). The last three columns of table 1 report the marginal effects estimated by KRLS at the 1st quartile, at the median, and at the 3rd quartile. We can clearly see the heterogeneity in the marginal effects. The estimated marginal effects differ widely over the quartiles and tend to increase for all variables considered here. This shows the nonlinearity and heterogeneity of the relationship between the covariates and the share of exports in total sales.

3.2.   Empirical model for export participation

Table 2 reports results for an empirical model for the participation in exports. This extensive margin of exports was estimated in Wagner (2023) using the probit model to deal with the dichotomous character of the dependent variable. Results in column 1 report the estimated average marginal effects (and its p-values) of the four firm characteristics included in the empirical model. Furthermore, the model includes a set of 26 country dummy variables as control variables.

[Table 2 near here]

Results for the average marginal effects estimated by KRLS (and its p-values) are reported in column 2 of table 2. A comparison of these estimates and the estimates reported in column 1 reveals that – like in the first example looked at above - the signs are identical and the levels of significance are of the same order of magnitude, so the big picture revealed by the two models is again identical.

The estimated average marginal effects are of the same order of magnitude in three out of four cases. KRLS estimates of average marginal effects are considerably larger for firm size, which is due to an inappropriate imposition of a linear functional form of the relationship between firm size and export participation. Again, the inclusion of a large set of (country) dummy variables does not pose a problem for KRLS.

The last three columns of table 2 report the marginal effects estimated by KRLS at the 1st quartile, at the median, and at the 3rd quartile. Again we can clearly see the heterogeneity in the marginal effects. The estimated marginal effects differ widely over the quartiles and tend to increase for all variables considered here, showing nonlinearity and heterogeneity of the relationship between the covariates and the probability of export participation.

3.3.   Empirical model for number of export destinations

Finally, table 3 reports results for an empirical model for the number of export destination countries of firms originally estimated in Wagner (2023) using ordinary least squares (OLS). Results in column 1 report the estimated regression coefficients (and its p-values) of the four firm characteristics included

in the empirical model. Furthermore, the model includes again a set of 26 country dummy variables as control variables.

[Table 3 near here]

Results for the average marginal effects estimated by KRLS (and its p-values) are reported in column 2 of table 3. A comparison of these estimates and the estimates reported in column 1 again reveals that the signs are identical and the levels of significance are the same, too, so the big picture shown by the two models is identical, and the same holds for the estimated average size of the effects here. Again, the inclusion of a large set of (country) dummy variables does not pose a problem for KRLS.

The last three columns of table 3 report the marginal effects estimated by KRLS at the 1st quartile, at the median, and at the 3rd quartile. Again we can clearly see the heterogeneity in the marginal effects. The estimated marginal effects differ widely over the quartiles and tend to increase for all variables considered here, showing nonlinearity and heterogeneity of the relationship between the covariates and the number of export destination.

3.4.   Summary of findings from three examples

The bottom line, then, is that in all three examples considered here the big picture from the original parametric models and from the models estimated by KRLS is the same. In several cases, however, the estimated average marginal effects from both models differ widely. These differences can be explained by the fact that the parametric model in column 1 imposes a restrictive functional form in the shape of the estimated relationships, while KRLS estimated this relationship without imposing a functional form. Furthermore, KRLS reveals that the marginal effects are not constant – they are heterogeneous and tend to vary widely across the covariate space.

4   Concluding remarks

The experience from the three applications of KRLS in the estimation of empirical models for various margins of exports can be summarized as follows: KRLS works fine for empirical models with continuous, fractional, and dichotomous endogenous variables and control variables that are continuous, dichotomous, or dummy variables for industries or countries. In all three examples considered here the big picture from the original parametric models and from the models estimated by KRLS is the same. In several cases, however, the estimated average marginal effects from both models differ widely because the parametric model imposes a restrictive functional form in the shape of the estimated relationships, while KRLS does not. Furthermore, KRLS reveals that the marginal effects are not constant – they are heterogeneous and tend to vary widely across the covariate space.

That said, given the ease of use thanks to the Stata program krls provided by Ferwerda et al. (2017) I suggest that KRLS should be considered as a useful addition to the box of tools of empirical trade economists. Even if the three examples considered here do not reveal that a replication using KRLS produces completely different results compared to the parametric models used in the original papers – which is good news for me as the author of these papers – it might well be the case that this will happen in future applications.

References

Cameron, A. C., and P. K. Trivedi (2022). Microeconometrics Using Stata. Volume II: Nonlinear Models and Causal Inference Methods. College Station, Texas: Stata Press.

Ferwerda, J., J. Hainmueller, and C. J. Hazlett (2017). Kernel-Based Regularized Least Squares in R (KRLS) and Stata (krls). *Journal of Statistical Software* 79 (3): 1–26.

Hainmueller, J., and C. Hazlett (2014). Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Macine Learning Approach. *Political Analysis* 22(2): 143–168.

Henderson, D. L. and C. F. Parameter (2015). Applied Nonparametric Econometrics. New York: Cambridge University Press.

Lo, A. W. (1994). Neural Networks and Other Nonparametric Techniques in Economics and Finance. In: H. R. Fogler ed., Blending quantitative and traditional equity analysis. Association for Investment Management and Research, p. 25-36.

Minviel, J.-J., and F. B. Bouheni (2022). The impact of research and development (R&D) on economic growth; new evidence from kernel-based regularized least squares. *The Journal of Risk Finance* 23(5): 583–604.

Papke, L. E., and J. M. Wooldridge (1996). Econometric Methods for Fractional Response Variables With an Application to 401(k) Plan Participation Rates. *Journal of Applied Econometrics* 11(4): 619–632.

Wagner, J. (2001). ANote on the Firm Size – Export Relationship. *Small Business Economics* 17(4): 229–237.

Wagner, J. (2023). Big Data Analytics and Exports – Evidence for Manufacturing Firms from 27 EU Countries. KCG Working Paper No. 28, Kiel Centre for Globalization.

Table 1: Empirical model for share of exports in total sales

Method	GLM Average marginal effects	KRLS Average marginal effect	P25	P50	P75
Firm size	0.0000531	0.000035	0.000025	0.000035	0.000047
(Number of employees)	(0.001)	(0.000)			
Branch plant status	0.0496	0.0490	0.0293	0.0561	0.0742
(Dummy; 1 = firm ist a branch plant)	(0.002)	(0.010)			
Craft shop	-0.093	-0.040	-0.0515	-0.0382	-0.0252
(Dummy; 1 = firm part of craft sector)	(0.000)	(0.005)			
Percentage of jobs demanding	0.0016	0.0020	0.000345	0.001679	0.00336
a university or polytech degree	(0,033)	(0.042)			
R&D/sales ratio greater zero and	0.0703	0.0412	0.0269	0.0424	0.0564
Less than 3.5 percent	(0.000)	(0,004)			
R&D/sales ratio between 3.5 and less	0.0882	0.0818	0.0579	0.0839	0.10790
han 8.5 percent	(0.000)	(0.000)			
R&D/sales ratio equal to 8.5 percent	0.0790	0.0675	0.0280	0.0839	0.1273
or more	(0.001)	(0.010)			
Patents	0.0464	0.0750	0.0498	0.0817	0.0938
(Dummy; 1 = firm registered at least one pa-tent)	(0.002)	(0.000)			
Product innovation	0.0319	0.0355	0.0195	0.0326	0.0484
(Dummy; 1 = firm introduced at least one new product)	0.016	(0.007)			
15 industry dummies	included	included			
Number of cases	768	768			

Note: GLM reports average marginal effects from a model estimated by fractional logit. KRLS reports average marginal effects and marginal effects at the 25th, 50th and 75th percentile estimated by kernel-based regularized least squares. P-values are reported in parentheses. For details, see text.

Table 2: Empirical model for export participation

Method	Probit Average marginal effects	KRLS Average marginal effect	P25	P50	P75
Big data analytics	0.112	0.111	0.0386	0.1087	0.1891
(Dummy; 1 = yes)	(0.000)	(0.003)			
Firm age	0.0015	0.0014	0.00011	0.0010	0.0025
(years)	(0.001)	(0.005)			
Firm size	0.00034	0.00082	0.00066	0.00083	0.0010
(Number of employees)	(0.000)	(0.000)			
Patent	0.212	0.186	0.1025	0.19990	0.2533
(Dummy; 1 = yes)	(0.000)	(0.000)			
26 country dummies	included	included			
Number of cases	2,355	2,355			

Note: Probit reports average marginal effects from a model estimated by ML Probit. KRLS reports average marginal effects and marginal effects at the 25th, 50th and 75th percentile estimated by kernel-based regularized least squares. P-values are reported in parentheses. For details, see text.

Table 3: Empirical model for export participation

Method	OLS Regression coefficient	KRLS Average marginal effect	P25	P50	P75
Big data analytics	0.7165	0.5116	0.3295	0.5262	0.7602
(Dummy; 1 = yes)	(0.000)	(0.000)			
Firm age	0.0110	0.0086	0.0059	0.0089	0.0119
(years)	(0.000)	(0.000)			
Firm size	0.0007	0.0011	0.00094	0.00111	0.0013
(Number of employees)	(0.003)	(0.000)			
Patent	0.9563	0.8274	0.6125	0.8796	1.0400
(Dummy; 1 = yes)	(0.000)	(0.000)			
26 country dummies	included	included			
Number of cases	1,520	1,520			

Note: OLS reports the estimated regression coefficients from a linear model. KRLS reports average marginal effects and marginal effects at the 25th, 50th and 75th percentile estimated by kernel-based regularized least squares. P-values are reported in parentheses. For details, see text.