

Empirical Applications of Smoothing Categorical Variables

We present some empirical applications which involve smoothing categorical explanatory variables. In most applications, the number of cells are (much) larger than the sample size which means that the conventional frequency nonparametric methods are not valid and cannot be applied. As we show below, smoothing the categorical variables and using a cross-validation method to select the smoothing parameters often leads to superior performance of the nonparametric estimator relative to many commonly used parametric estimation methods.

1 An Example with Mixed Discrete & Continuous Regressors: Modeling Count Panel Data

We consider the data used by Hausman, Hall, & Griliches (1984, *Econometrica*) in which they model the number of successful patent applications made by firms in scientific and non-scientific sectors across a seven-year period. The variables they used were as follows:

1. PATENTS: number of successful patent applications in the year,
2. CUSIP: firm identifier,
3. YEAR: year of data,
4. SCISECT: dummy for firms in scientific sector,
5. LOGR: log of R&D spending,
6. LOGK: log of R&D stock at beginning of year

This dataset is a balanced panel containing 896 observations on six variables, and the first four variables are categorical while the last two are continuous. For the categorical variables there were 227 unique values for the variable PATENTS, 128 values for CUSIP (128 firms), 7 for YEAR, and 2 for SCISECT. For this dataset, the number of discrete cells exceeds the sample size, therefore, the conventional frequency estimator cannot be used which is not uncommon in practice.

We wish to assess the dynamic predictive ability of the proposed method for this time-series count panel, and we use the first six years of data for estimation purposes and the remaining seventh year for evaluation purposes leaving $n_1 = 768$ (128 firm \times 6 years) and $n_2 = 128$ (128 firm \times 1 year).

For comparison purposes we consider three parametric models found in the literature: 1. A nonlinear OLS regression of $\log(\text{PATENTS})$ on the explanatory variables, where $\log(\text{PATENTS})$ is set to zero and a dummy variable used when $\text{PATENTS}=0$ (Hausman, Hall, & Griliches (1984, page 912)); 2. A pooled Poisson count panel model; and 3. A Poisson count panel model with firm-specific effects.

We again assess predictive ability on the independent data using $MSE = n_2^{-1} \sum_{i=1}^{n_2} (\text{PATENTS}_i - \widehat{\text{PATENTS}}_i)^2$ where $\widehat{\text{PATENTS}}_i$ denotes the predicted values generated from each model. As well, we compute the correlation coefficient between the actual and predicted values of PATENTS , $\hat{\rho}_{\hat{y},y}$. Results appear on Table 1.

Table 1:

Model	Prediction MSE	$\hat{\rho}_{\hat{y},y}$
OLS	2618.3	0.86
Poisson (pooled)	1915.9	0.87
Poisson (firm-effects)	2834.7	0.82
Unordered Kernel	403.4	0.97
Ordered Kernel	385.2	0.98

These results show that the new approach completely dominates the parametric specifications and that accounting for the natural order in the explanatory variables leads to further finite-sample efficiency gains. The squared prediction error of our nonparametric estimator (using the ordered kernel) is only 14% to 20% of those obtained by various parametric methods which have been used to model this dataset.

2 Regression Models with Discrete Regressors

In this section we present two empirical examples, both containing only categorical variables. When all the explanatory variables are categorical, the cross-validation selection of λ , $\hat{\lambda}$, has a fast rate of convergence (to zero) of $O_p(n^{-1})$. This result cannot be obtained as a corollary from the mixed discrete and continuous regressors case which appears in the enclosed paper, and a separate proof is needed to establish the result. Below we consider two empirical applications (where all explanatory variables are categorical) which demonstrate the usefulness of the proposed approach relative to commonly used parametric specifications which appear in the literature.

We shall focus on the out-of-sample predictive performance of both the proposed estimator and common parametric specifications which have been used to model each dataset. For each dataset we apply a random shuffle and then split it into a training and evaluation set having n_1 and n_2 observations respectively. We then estimate each model on the training data and generate predictions using the explanatory variables in the evaluation dataset. We then compute the squared prediction error as the square of the difference between the predicted and actual values of the dependent variable in the evaluation dataset. To avoid the criticism of our results reflecting a particular way of splitting the data, we randomly split the sample into two n_1 and n_2 sub-sample 100 times, and report mean and median values of squared prediction errors based on these 100 replications.

2.1 Count Survival Data - Veteran's, Administration Lung Cancer Trial

We consider the dataset taken from Kalbfleisch and Prentice (1980, pg 223-224) which models survival in days of cancer patients with six categorical explanatory variables being treatment type, cell type, Karnofsky score, months from diagnosis, age in years, and prior therapy.

The dataset contains 137 observations, and the number of cells greatly exceed the number of observations. Clearly, the conventional frequency nonparametric method cannot be used for this data set. The goal here is to model the expected survival in days. The estimation sample is $n_1 = 132$, and the prediction sample is $n_2 = 5$. We compute the out-of-sample $MSE = n_2^{-1} \sum_{i=1}^{n_2} (y_i - \hat{y}_i)^2$, where \hat{y}_i is the predicted value of $E(y_i|x_i)$ computed using x_i and $\{x_j, y_j\}_{j=1}^{n_1}$ (the independent estimation sample). We randomly split the sample into two sub-sample of size n_1 and n_2 100 times. Table 2 gives the out-of-sample mean and median squared prediction errors by different estimation methods. As can be seen from Table 2, the average nonparametric squared prediction error is only 60% to 62% of those obtained from various parametric methods.

Table 2: Veteran’s Lung Cancer Data

Model	Median MSE	Mean MSE	(Mean) MSE_{NONP}/MSE
NONP	9,770.5	17,614.9	100%
OLS	13,831.7	28,422.6	62%
POISSON	15,319.8	29,052.9	60%

2.2 Count Data - Fair’s 1977 Extramarital Affairs Data

We consider Fair’s (1978) dataset which models how often an individual engaged in extramarital affairs during the past year. The paper with a complete description of the data is located at <http://fairmodel.econ.yale.edu/rayfair/pdf/1978A200.PDF>. The dataset contains 601 observations and has eight categorical explanatory variables given by sex, age, number of years married, number of children, how religious, level of education, and how they rate their marriage. Following Greene (2000, pg 920-926), the parametric models include Probit, Tobit, and Poisson specifications. The estimation sample is $n_1 = 500$, and the prediction sample is $n_2 = 101$. The goal is to model the expected number of affairs. We split the sam-

ple into two sub-sample of size n_1 and n_2 to compute the out-of-sample squared prediction errors, and repeat this procedure (randomly) 100 times. Table 3 reports the median and mean out-of-sample squared prediction errors by different estimation methods. We observe that our nonparametric squared prediction error is 69% to 82% of those obtained by the various parametric methods.

Table 3: Fair’s 1977 Extramarital Affairs Data

Model	Median MSE	Mean MSE	(Mean) MSE_{NONP}/MSE
NONP	9.97	10.18	100%
OLS	11.90	12.39	82%
PROBIT	14.66	14.74	69%
TOBIT	12.81	13.04	78%
POISSON	12.88	12.99	78%

We wish to be clear that is is not our intention to suggest that our method will outperform a correctly specified parametric model. However, when parametric models are misspecified, nonparametric estimators should out-perform such parametric models if the sample sizes are sufficiently large. The two examples reported above show that, even for small and moderate sample sizes, our data-driven nonparametric estimator can yield *better* out-of-sample predictions than commonly used parametric models. In each example, the number of discrete cells is large relative to the sample size, and clearly the conventional frequency nonparametric estimator cannot be used in such cases.

3 Conditional Probability Estimation: Involving Categorical Variables

3.1 Modeling U.S. Female Labor Force Participation

Our first application uses the Mroz (1987, *Econometrica*) data file which is taken from the 1976 Panel Study of Income Dynamics, and is based on data for 1975. There are 753 observations in this dataset, the first 428 for women with positive hours worked and the remaining 325 observations for women who did not work for pay. For a complete discussion of the data see Mroz (1987) or Chapter 11 in Berndt (1991). We consider an application of the proposed method to modeling the female labor force participation decision.

The following variables from the data file were therefore used:

1. LFP: A dummy variable equal to 1 if the woman worked in 1975, 0 otherwise.
2. KL6: The number of children less than 6 years old in the household.
3. K618: The number of children between ages 6 and 18 in the household.
4. WA: The woman's age.
5. WE: The woman's educational attainment, in years.
6. CIT: A dummy variable equal to 1 if the woman lives in a large city (SMSA), 0 otherwise.
7. AX: The actual years of the woman's previous labor market experience.
8. UN: The unemployment rate in county of residence, in percentage points. This is taken from bracketed ranges.
9. LWW1: The log of the wage (woman's average hourly earnings, in 1975 dollars) for working women, the log of predicted wage for non-workers.
10. PRIN: The woman's property income computed as total family income minus the labor income earned by the woman.

The Logit and Probit approaches model Item 1 as the dependent variable and items 2 through 10 as explanatory variables in addition to a constant term, while the proposed approach does not require the use of the constant term. For the proposed method we use the fact that items 1 through 7 are categorical while items 8 through 10 are continuous. Again, bandwidths were determined via the proposed method of cross-validation. The confusion matrices and classification rates for both the proposed and Probit approaches are summarized in Table 4 (a confusion matrix is one whose diagonal elements are correctly predicted outcomes and whose off-diagonal elements are incorrectly predicted outcomes).

Table 4: Confusion matrix and classification rates for the kernel and Logit models

		Kernel		Logit	
A/P		0	1	A/P	
0		314	11	0	166
1		57	371	1	80
	% Correct	91.0%		% Correct	68.2%
	% CCR(0)	96.6%		% CCR(0)	51.0%
	% CCR(1)	86.7%		% CCR(1)	81.3%

The estimated Logit and Probit models of labor force participation correctly predict 514 (68.2%) and 512 (68.0%) of the labor force participation decisions respectively. As can be seen from Table 4, the proposed method correctly predicts 685 (91%) labor force participation choices which translates into an additional 171 choices being correctly predicted, which is a fairly dramatic improvement in terms of prediction accuracy.

3.2 One Final Empirical Application

We now consider an application of the proposed approach to modeling discrete choice. This example shows how the proposed estimator can be used to obtain superior (out-of-sample) predictive performance relative to commonly used parametric models of discrete choice.

We use the data of Gerfin (1996, J. of Applied Econometrics) who models the labor market participation of married Swiss women using a cross-section data set of size $n = 872$ having six explanatory variables. He uses a Probit model along with three semiparametric specifications, and finds that the Probit specification cannot be rejected and that the models yield similar results. He concludes that “more work is necessary on specification tests of semiparametric models and on simulations using these models”. We simply use this dataset to see whether predictions given by the Probit and semiparametric specifications can be substantially improved upon (we do not include Gerfin’s (1996) semiparametric results here as they all yielded similar results.) Data for this study can be found at <http://qed.econ.queensu.ca/jae/1996-v11.3/gerfin/>.

The variables used by the Gerfin (1996) study are

1. LFP: Labor force participation dummy.
2. LNNLINC: Log of non-labor income.
3. AGE: Age in years.
4. EDUC: Years of formal education.
5. NYC: Number of young children (younger than 7).
6. NOC: Number of older children.
7. FOREIGN: Dummy, = 1 if observation is not Swiss.

We compute the conditional distribution as the ratio of the joint distribution of variables 1 through 7 and the marginal distribution of variables 2 through 7,

$$\hat{f}_{(\text{LFP}|\text{LNNLINC, AGE, EDUC, NYC, NOC, FOREIGN})} = \frac{\hat{f}_{(\text{LFP, LNNLINC, AGE, EDUC, NYC, NOC, FOREIGN})}}{\hat{f}_1(\text{LNNLINC, AGE, EDUC, NYC, NOC, FOREIGN})} \quad (1)$$

and bandwidths are chosen via cross-validation. Finally, we predict LFP=1 if $\hat{f}_{(\text{LFP} = 1|\cdot)} > \hat{f}_{(\text{LFP} = 0|\cdot)}$ where $|\cdot)$ denotes the conditioning variables, otherwise we predict LFP=0.

We compare the results of our estimator with those from Gerfin (1996), and the confusion matrices and classification rates for both the proposed and Probit approaches are summarized in Table 5. We also report the overall correct classification rate and correct classification rates for each values assumed by the categorical dependent variable¹. As can be seen from Table 5, the proposed method correctly predicts 74.1% of all observations while a Probit model correctly predicts 66.5% which represents a marked improvement in model performance. To address potential concerns that these results might be an artifact of within-sample ‘overfitting’, we randomized the data and split it into independent estimation and evaluation samples². The predictive ability of the model as measured by performance on the independent data mirrors the within-sample results reported in Table 5 for a large number of different splits indicating that this is indeed a general improvement in predictive ability and not simply an artifact of overfitting.

Table 5: Confusion matrix and classification rates for the kernel and Probit models.

	Kernel		Probit		
A/P	0	1	A/P	0	1
0	360	111	0	358	113
1	115	286	1	179	222
% Correct	74.1%		% Correct	66.5%	
% CCR(0)	76.4%		% CCR(0)	76.0%	
% CCR(1)	71.3%		% CCR(1)	55.4%	

This application simply demonstrates how the proposed method can be used to obtain superior predictions of categorical variables relative to predictions based upon commonly used parametric specifications such as the Probit model.

¹For example, CCR(0) is the number of predicted zeros that are in fact zeros \div number of zeros in the sample $\times 100$.

²For example, we considered estimation samples of size $n_1 = 700$ and prediction samples of size $n_2 = 172$, $n_1 = 750$ and $n_2 = 122$ and so on.

References

- Aitchison, J. & Aitken, C.G.G. (1976) Multivariate binary discrimination by the kernel method. *Biometrika* 63, 413-420.
- Fair, R. (1978) "A theory of extramarital affairs," *Journal of Political Economy*," 86, 45-61.
- Gerfin, M. (1996), "Parametric and Semiparametric estimation of the binary response model of labor market participation," *Journal of Applied Econometrics* 11, 321-340.
- Greene, W.H. (2000), *Econometric Analysis*, Prentice Hall: London.
- Hausman, Jerry and Bronwyn H. Hall and Zvi Griliches (1984), "Econometric Models for Count Data with an Application to the Patents-R&D Relationship", *Econometrica*, 52, Number 4, 909–938.
- Kalbfleisch, J.D. and R.L. Prentice (1980), *The Statistical Analysis of Failure Time Data*, New York: Wiley (1980).
- Mroz, T.A. (1987) "The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions," 55, 765-799.