Working Paper 96-65
Statistics and Econometrics Series 28
May 1997

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9849

NONPARAMETRIC CHECKS FOR COUNT DATA MODELS: AN APPLICATION TO DEMAND FOR HEALTH CARE IN SPAIN.

Begoña Alvarez and Miguel A. Delgado*

Abstract

This paper presents model specification checking procedures for count data regression models which are consistent in the direction of nonparametric alternatives. The discussion is motivated in the context of a model of demand for health care in Spain. The parameters of the regression model are estimated by maximum likelihood based on Poisson and Negative Binomial specifications as well as by ordinary least squares and semiparametric generalized least squares. However, our interest is not only centered on the estimation of the regression parameters, but also the conditional probabilities of counts. Therefore, the specification of the conditional distribution function of counts is the main focus of attention. A useful preliminary diagnosis tool consists of comparing the conditional probabilities estimates by nonparametric regression and by maximum likelihood methods based on alternative models. We present formal specification procedures based on new developed testing methods for regression model checking. The test statistics are based on marked empirical processes which are not distribution free, but their critical values are well approximated by bootstrap. Such tests are valid for testing the functional form of the conditional mean and conditional probabilities resulting from alternative distributional specifications. In our health care demand model, the linear exponential regression model with a Negative Binomial seems to be appropiate for this data set.

Keywords and phrases: consistent specification testing; nonparametric fitting; count data models; demand for health.

*Facultad de Ciencias Jurídicas y Sociales. Universidad Carlos III de Madrid. C/ Madrid, 126 28903 Getafe, Madrid. e-mail: delgado@est-econ.uc3m.es. Research supported by the Spanish Dirección General de Investigación Científica y Técnica (DGICYT), reference number PB95-0292, and by Spanish Fondo de Investigaciones Sanitarias (FIS), reference number 96/1787. We are most thankfull to Félix Lobo for his many helpful comments and by his advice on interpreting data.

1. INTRODUCTION

This paper presents model checking techniques for count regression models. The performance of the different techniques is illustrated in the context of model specification of demand for health care in Spain.

The interest in count regression models is centered on the estimation of conditional probabilities rather than the mere estimation of parameters in the conditional mean of the model. In fact, once consistent, but inefficient, parameter estimates are obtained by ordinary least squares, valid inferences are available when the estimators variances are robustly estimated. Moreover, efficient inferences can be performed using semiparametric generalized least squares as suggested by Delgado and Kniesner (1997). Unfortunately, it is not possible to estimate conditional probabilities of counts without specifying the underlying conditional distribution function or, when a parametric distribution model is not available, without resorting to nonparametric methods. Therefore, maximun likelihood methods are well motivated in practice. The Poisson regression model has been the most popular in applications. However, economic count data sets exhibit an excess of zero observations and long right tails, both relative to Poisson regression. This is why models allowing for overdispersion, like the Negative Binomial model, have been popular in recent applications (see e.g. Hausman, Hall and Griliches (1984) and Cameron and Trivedi (1986)).

In the next section we propose a linear exponential regression model for the number of visits to a doctor and the number of visits to emergency rooms using data from the 1993 Spanish Health Survey. The parameters of the model are estimated by ordinary least squares and by maximun likelihood based on Poisson and Negative Binomial specifications. The coefficient estimates from alternative procedures are not very different and there is evidence in favour of overdispersed distributions, like the Negative Binomial specification, with respect to the Poisson based on a Wald test and based on auxiliary regressions as suggested by Cameron and Trivedi (1986).

In Section 3, we study the goodness of fit of the alternative models comparing

the estimated conditional probabilities with respect to nonparametric regression estimates. The sample marginal frequencies of the dependent variable taking the value zero are fairly well approximated by the different methods. The marginal frequencies for higher values of the dependent variable are better approximated by the Negative Binomial.

In section 4, we discuss formal specification tests for regression based on ideas developed by Delgado (1992), Stute (1996) and Stute et al. (1996). These tests are consistent in the direction of nonparametric alternatives. They are based on marked empirical processes which are not distribution free. However, the critical values can be approximated by bootstrap. After applying the nonparametric tests, specifying the conditional expectation, we conclude that the linear exponential approximation is satisfactory for our data set. The Poisson specification seems unable to fit well the conditional probabilities for any given count value. However, the Negative Binomial specification cannot be rejected in order to model the conditional probabilities. The critical values are approximated by a 'wild bootstrap' when testing the functional form of the regression function, and by parametric bootstrap when testing the functional form of conditional probabilities of counts. We also apply the testing procedure to a test of the functional form of the distribution function, as suggested by Andrews (1996). The critical values are also approximated by a parametric bootstrap and the Negative Binomial model can neither be rejected.

2. DATA AND MODEL

Data is coming from the 1993 Spanish Health Survey (SHS)[1] which offers survey data on 21.061 people concernig health demand information, like number of doctor visits and emergency room visits, as well as demographic characteristics, like health habits (smoking and drinking), education, socieconomic position, geographical variables, etc.. There is not income information on this survey and family income values

[1] The data has been collected in February 1993 by the Spanish Ministry of Health and Consumer Affairs.

have been predicted from income equations based on the socioeconomic characteristics of the individual in both the SHS and the 1990-91 Spanish Household Expenditure Survey (SFES) at 1993 prices.

Sample observations $\{(Y_i, X_i), i = 1, ..., n\}$ are independent and identically distributed; Y_i are count data variables denoting the number of visits to a doctor or to the emergency rooms, and $X_i = (X_{i1}, X_{i2}, ..., X_{ik})'$ is a $k \times 1$ vector of explanatory variables. Using the Cameron et al. (1988) approach, we face the regression model

$$E(Y_i \mid X_i = x) = \exp(x'\beta_0) \ a.s. \tag{1}$$

Tables 1, 2 and 3 about here.

Table 1 presents the explanatory variables X_i used in the study and Tables 2 reports summary statistics.

Our empirical analysis includes two demand variables: the number of doctor consultations during the two weeks before the interview and the number of emergency room visits during the year before the interview. Table 3 shows the sample frecuency distributions of both variables. It is important to remark that the survey does not make any difference between private medical visits and insurance covered visits.

Following Cameron et al. (1988), we assume that demand for health care depends on health status and other demographic variables, such as age and sex. The model distinguishes two health dimensions: CHRONIC ILLNESS, if the individual suffers from an illness which is perceived as a permanent problem, and ACUTE ILLNESS, if illness is perceived as transitory and having limiting effects on the individual activity. The variable ACCID indicates if the interviewed person has suffered from an accident during the previous year.

There are other factors, besides the biological ones, that determine the demand for health care. Studies about this subject have mainly focused on income, prices and education. The Spanish National Health Service guarantees universal health care provision. Under this system, spanish citizens do not need to make any out-of-pocket expense in order to receive medical care. However, this hardly means that services

are free. Health care users face an implicit price: the time opportunity cost. The analysis of this aspect is necessarily vinculated to both employment status and time travel cost (see e.g. Wagstaff (1986)). Given that the SHS does not provide any information about the travel component, we focus only on the employment status. For this purpose, we classify individuals according to their potential cash earning looses when away from their usual activity. Thus, we consider SELF-EMPLOYED, WAGE-EARNING and NON-WORKING individuals, to control from higher to lower time opportunity cost. We also include INCOME and SCHOOL variables to test for their relevance in this framework.

Obviously, individual's health decisions are dependent on the degree in which they appreciate their own health. Unfortunately, we are not able to measure this dimension of individual preferences. Nevertheless, we can consider individual's lifestyles as proxies of these preferences. We claim that, ceteris paribus, individuals with healthy habits use more medical care than others. To examine the relationship between lifestyles and health care use, we selected smoking habit -SMOKER, NON-SMOKER and EX- SMOKER- and usual alcohol consumption -ALCOHOL- as regressors in the model.

Finally, we consider region -NORTH, CENTER and SOUTH- and population size -RURAL- in order to control for geographical differences.

Given the regression specification in (1), we have alternative conditional probability specifications,

$$P_y(x) = Pr(Y_i = y \mid X_i = x) \tag{2}$$

The function $P_y(x)$ will also typically be a function of β_0 and other 'nuisance' parameters. For instance, if Y_i are conditionally Poisson we have

$$P_y(x) = \frac{\exp(-\exp(x'\beta))\exp(yx'\beta)}{y!} \tag{3}$$

If Y_i are conditionally distributed as a Negative Binomial then

$$P_y(x) = \frac{\Gamma(y+\delta)}{\Gamma(\delta)\Gamma(y+1)}\left(\frac{\delta}{\exp(x'\beta)+\delta}\right)^{\delta}\left(\frac{\exp(x'\beta)}{\exp(x'\beta)+\delta}\right)^{y} \tag{4}$$

where the parameter δ^{-1} is the precision parameter. Specifications (3) and (4) are compatible with the regression specification in (1).

The coefficient estimates and standard errors are reported in Tables 4 and 5. Different estimation methods have been applied: nonlinear least squares (NLS), semiparametric generalized least squares (SGLS) -see Robinson (1987) and Delgado and Kniesner (1996)- and maximun likelihood based on (2) and (3).

Tables 4 and 5 about here

Focusing on the semiparametric estimations, we summarize the main results as follows. First, the aging process seems to be highly related to health care use but not always in the same direction. While old people use ordinary doctor consultations more often than young people, visits to emergency rooms decrease with age. Illness, in all its dimensions -chronic, acute and accidental-, is the main factor to explain different health care demand levels. Regarding sex, we find that there are less doctor visits for men but there are not significant differences between male and female emergency visits.

Lifestyle coefficient estimates confirm a positive relationship between unhealthy habits and low health care use. An interesting result with respect to smoking habits is that ex-smokers show higher demand for health care than non-smokers. This relationship could be explained by means of their higher concern about health, directly related to the decision of quitting smoking and maintaining it.

The coefficient estimates of employment status are consistent with our initial assumptions. This result highlights the importance of including time opportunity cost variables into health care demand models. On the contrary, family income does not appear to be significant in any case. Furthermore, education has a significant negative effect on emergency visits. Despite the fact that education implies a higher health conciousness, we could relate this result with a more rational use of this special service.

Finally, geographical variables appear to have significant effects. On the one hand,

as we move South in the country, health care demand increases. Geographical demand differences could be caused by tastes, cultural patterns and even services availability. In order to create NORTH and SOUTH variables, we have aggregated several administrative regions with heterogeneous health service endowments. Then, it is not plausible to explain our results by means of specific interregional endowment differences; it would be interesting to further investigate this empirical evidence. On the other hand, we find that people in rural areas use less emergency services. This coefficient estimate can be missinterpreted. The SHS asks for specific emergency room visits, so it is not accounting for the fact that most emergency consultations in rural areas happen at home.

The coefficient estimates obtained by the different methods are quite similar. So, it seems hard to choose between the different methods based on a mere comparison between the alternative coefficient estimates. However, there is some evidence in favour of the Negative Binomial model with respect to the Poisson alternative. On one hand the robust and non robust standard errors are quite similar for the Negative Binomial specification, but they are very different for the Poisson specification. On the other hand, the Wald or likelihood ratio test provide evidence in favour of the Negative Binomial. We also offer corrected $R^{2'}s$ for all models as suggested by Cameron and Windmeijer (1996)[2]. The Binomial Negative offers the higher R^2.

We also implemented the regression based test suggested by Cameron and Trivedi (1990). The test consists of the t-ratio on

$$[Y_i - \exp(X_i'\hat{\beta})]^2 - Y_i = \hat{\alpha}\exp(X_i'\beta) + residual.$$

Such a test provides evidence of overdispersion with a t-ratio equal to 5.287 for the doctor visits model and a t-ratio equal to 9.665 for the emergency room visits model.

The above model checking procedures are not formal specification tests and do not provide evidence on neither, the specification correctness of the regression model or

[2]For the Poisson and Negbin models, we calculated R^2 measures based on deviance residuals. In the case of the SGLS estimates we calculated a R^2 based residuals weighted by nonparametrically estimated variances.

the conditional probabilities of counts. These problems will be addressed in the next sections.

3. CONSISTENT SPECIFICATION TESTING OF REGRESSION MODELS.

In this section we present a formal test for the hyphothesis

$$H_0 \ : \ Pr\{E(Y_1 \mid X_1) = m(X_1, \beta_0)\} = 1, \ \beta_0 \in B$$

and the alternative is the negation of H_0, where $m(.,.)$ is a known function and $\beta_0 \in B$ is an unkown parameter, where $B \subset \mathbb{R}^k$ is the parameter space. In our application $m(x, \beta) = \exp(x'\beta)$.

Formal specification procedures based on the difference between nonparametric and parametric estimates are available in generous supply (see e.g Härdle and Mammen (1993), Eubank and Spiegelman (1990), Stute and Manteiga (1995) and Fan and Li (1996), to mention only a few).

The fact that for two Borel functions $m(.)$ and $g(.)$

$$Pr\{m(X_1) = g(X_1)\} = 1 \Longleftrightarrow \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} .... \int_{-\infty}^{x_k} (m(x) - g(x))dF_x(x) = 0, \qquad (5)$$

$$\forall x = (x_1, x_2, ..., x_k)' \in \mathbb{R}^k$$

where $F_x(.)$ is the distribution function of X_i, implies that

$$H_0 \ holds \iff E\{(Y_1 - \exp(X_1'\beta_0)) \prod_{j=1}^{k} 1(X_{1j} \leq x_j)\} = 0, \ \forall x \in \mathbb{R}^k. \qquad (6)$$

Then (6) suggest the statistic

$$T_n(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i - m(X_i, \hat{\beta})) \prod_{j=1}^{k} 1(X_{ij} \leq x_j), \qquad (7)$$

where $\hat{\beta}$ is any $\sqrt{n} - consistent$ estimator under H_0, e.g. the ordinary least squares estimator. Stute (1996) has obtained the limit process of $T_n(x)$ for general nonlinear regression models under very weak regularity conditions. This statistic resembles

in spirit to statistics proposed by Bierens (1982), Bierens(1990) and Bierens and Ploberger (1997). Our model holds the conditions required in Stute (1996). Then, under H_0,

$$T_n(x) \longrightarrow T(x) \text{ in distribution on } D(\mathbb{R}^k), \tag{8}$$

where $T(x)$ is a Gaussian process centered at zero and with covariance structure

$$\begin{aligned} K'(s,t) \equiv Cov(T(t),T(s)) &= K(s,t) + G(s,\beta_0)'L(\beta_0)G(t,\beta_0) \\ &\quad -G(s,\beta_0)'E\left[\sigma^2(X_1)\dot{m}(X_1,\beta_0)\prod_{j=1}^{k}1(X_{1j}\leq t_j)\right] \\ &\quad -G(t,\beta_0)'E\left[\sigma^2(X_1)\dot{m}(X_1,\beta_0)\prod_{j=1}^{k}1(X_{1j}\leq s_j)\right], \end{aligned} \tag{9}$$

where $t = (t_1, t_2, ..., t_k)'$, $s = (s_1, s_2, ..., s_k)'$, $\sigma^2(x) = Var[Y_1 \mid X_1 = x]$, $\dot{m}(x,\beta) = \partial m(x,\beta)/\partial\beta$, $G(t,\beta) = E[\dot{m}(X_1,\beta)\prod_{j=1}^{k}1(X_{1j}\leq t_j)]$, $L(\beta) = E[\dot{m}(X_1,\beta)\dot{m}(X_1,\beta)'\sigma^2(X_1)]$ and

$$\begin{aligned} K(s,t) &= E\left[\sigma^2(X_1)\prod_{j=1}^{k}1(X_{1j}\leq \min(s_j,t_j))\right] \\ &= \int_{-\infty}^{s_1\wedge t_1}\int_{-\infty}^{s_2\wedge t_2}...\int_{-\infty}^{s_k\wedge t_k}\sigma^2(x)dF_x(x). \end{aligned}$$

An asymptotic test is based on the Cràmer-von Mises statistic

$$C_n = \frac{1}{n}\sum_{i=1}^{n}T_n(X_i)^2. \tag{10}$$

By the continuous mapping theorem

$$C_n \xrightarrow{d} C \equiv \int_{\mathbb{R}^k}T(x)^2dx. \tag{11}$$

The statistic is not distribution free. Stute et al. (1996) suggested to apply 'Wild Bootstrap' in order to approximate the standard errors. We need to generate i.i.d. V_i, $i = 1, ..., n$ independent of X_i, such that $E(V_i) = 0$, $E(V_i^2) = E(V_i^3) = 1$. Then, we construct

$$Y_i^* = \exp(X_i'\hat{\beta}) + \epsilon_i^*,$$

where $\epsilon_i^* = \hat{\epsilon}_i V_i$, and $\hat{\epsilon}_i = Y_i - \exp(X_i'\hat{\beta})$. With the new bootstrap sample (Y_i^*, X_i), $i = 1, ..., n$, one computes

$$\hat{\beta}^* = \arg\min_{\beta \in B} \sum_{i=1}^{n} (Y_i^* - \exp(X_i'\beta))^2,$$

and then, the bootstrap process

$$T_n^*(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i^* - \exp(X_i'\hat{\beta}^*)) \prod_{j=1}^{k} 1(X_{ij} \leq x_j). \tag{12}$$

The bootstrap statistic is

$$T_n^* = \frac{1}{n} \sum_{i=1}^{n} T_n^*(X_i)^2. \tag{13}$$

Stute et al. (1996) showed that this bootstrap test is valid in the sense that with probability one $C_n^* \to_d C^*$, where C^* is distributed as C.

In order to implement the bootstrap test, we generate bootstrap samples $\{(Y_i^{*(j)}, X_i), i = 1, ..., n\}$, $j = 1, ..., B$, and the critical values, at the α level of significance, of T_n are approximated by $T_{n\alpha}^*$ defined as

$$\frac{1}{B} \sum_{j=1}^{B} 1(T_n^{*(j)} > T_{n\alpha}^{*B}) = \alpha. \tag{14}$$

When B is large, $T_{n\alpha}^{*B}$ is a good approximation to $T_{n\alpha}^*$ defiend as $\Pr[T_n^* > T_{n\alpha}^*] = \alpha$. If the observed T_n is greater than $T_{n\alpha}^*$, we reject H_0. To implement this test we restrict ourselves to the subsample of 764 smoker men without illnesses in the reference period, from the Center and South of Spain, between 23 and 64 years old and wage-earning. In this subsample, the p-value of the test for testing the null hypothesis $H_0 : E(Y_i \mid X_i = x) = \exp(x'\beta_0)$ *a.s.* in the model of the demand for doctor visits is 0.599. So, we are unable to reject H_0.

4. SPECIFICATION TESTING OF CONDITIONAL PROBABILITIES OF COUNTS.

In this section, we present model checking procedures for the conditional probability model $P_y(x, \hat{\beta}, \hat{\delta})$, where $\hat{\beta}$ and $\hat{\delta}$ are the maximun likelihood estimates of β_0 and δ_0 respectively, fits the conditional probability $P_y(x)$.

A preliminary specification tool consists of comparing the estimates of the marginal probabilities, $\mathbf{P}_y = Pr(Y_i = y)$, by the alternative procedures: Poisson and Negative Binomial maximun likelihood. That is $\mathbf{P}_y$ is estimated by

$$\hat{\mathbf{P}}_y = \frac{1}{n}\sum_{i=1}^{n} P_y(X_i, \hat{\beta}, \hat{\delta}). \tag{15}$$

These estimates will be compared with the sample frequencies

$$\widetilde{\mathbf{P}}_y = \frac{1}{n}\sum_{i=1}^{n} 1(Y_i = y), \tag{16}$$

which always estimate consistently $\mathbf{P}_y$ and also with estimates based on nonparametric regression

$$\bar{\mathbf{P}}_y = \frac{1}{n}\sum_{i=1}^{n} \bar{P}_y(X_i), \tag{17}$$

where

$$\bar{P}_y(x) = \frac{\sum_{i=1}^{n} 1(Y_i = y)K(\frac{X_i - x}{h})}{\sum_{i=1}^{n} K(\frac{X_i - x}{h})}, \tag{18}$$

$K(u) = \prod_{j=1}^{K} k(u_k)$, $u = (u_1, ..., u_K)'$, $k(.)$ is a kernel function and h is the bandwidth number [3]. Notice that $P_y(x)$ is estimating the regression function

$$P_y(x) = E[Y_1 = y \mid X_1 = x]. \tag{19}$$

We limit this comparisons to the same subsample used in the above section. Tables 6 reports Poisson and Negative Binomial estimates for this restricted data set.

Tables 6 and 7 about here

The different $\mathbf{P}_y$ estimates are reported in Table 7. The $\hat{\mathbf{P}}_y$ based on the Negative Binomial model is always closer to $\widetilde{\mathbf{P}}_y$ and $\bar{\mathbf{P}}_y$ than the $\hat{\mathbf{P}}_y$ based on the Poisson model. The Negative Binomial $\hat{\mathbf{P}}_y$ for $y = 0, 1$ are quite close to $\widetilde{\mathbf{P}}_y$ and $\bar{\mathbf{P}}_y$, but $\hat{\mathbf{P}}_y$ is no so closed to $\widetilde{\mathbf{P}}_y$ as $\bar{\mathbf{P}}_y$ is.

[3] We have used a Gaussian kernel in the application. An optimal bandwidth, $h = Cn^{-1/5}$, has been calculated following a "plug-in" method in wich h was selected to minimize the mean integrated squared error.

Notice that $\mathbf{P}_y$ can be accurately estimated even when $P_y(x)$ is incorrectly specified. In fact

$$\mathbf{P}_y = E[1(Y_1 = y)] = E[E[1(Y_1 = y) \mid X_1]] = E[P_y(X_1)]. \tag{20}$$

In our example, the Poisson specification can approximate reasonably well the marginal frequencies, though it is proven to be a poor approximation for the underlying conditional distribution of counts. It may be due to the fact that a high proportion of ceros are observed in this sample.

If the functional form is correctly specified, always $\mathbf{P}_y = E[P_y(X_1, \beta_0, \delta_0)]$. Under misspecification, it is also possible to find cases where $\mathbf{P}_y = E[P_y(X_1, \beta_0, \delta_0)]$. That is, different fuctional forms $P_y(x, \beta_0, \delta_0)$ can produce the same $\mathbf{P}_y$. Therefore, comparing different $\mathbf{P}_y$ estimates is not always a good method for model checking.

In order to avoid this problem we can compare the nonparametric estimates of $P_y(x)$, $\bar{P}_y(x)$, with parametric estimates based on alternative likelihood functions $\hat{P}_y(x, \hat{\beta}, \hat{\delta})$.

In Figure 1 and 2 we report nonparametric and maximun likelihood estimates of the conditional probabilities. In Figure 1, the variable ALCOHOL is fixed at its mean value, while in Figure 2 the variable AGE is fixed at its mean value. The Negative Binomial estimates are closer to the nonparametric estimates and they are always inside the nonparametric confidence bands. The Poisson estimates seem much more biased and, sometimes, it is out of the nonparametric confidence bands.

Figures 1 and 2 about here.

The mere graphical inspection is not a formal specification tool. This procedure suffers from several problems. Firstly, in multidimensions, graphical devices are difficult to implement. Secondly, we are not taking into account the variability of the nonparametric estimators, which can vary a lot for different bandwidth choices.

Formally, we are interested in testing

$$H_0 : Pr\{P_y(X_1) = P_y(X_1, \beta_0, \delta_0)\} = 1, \ \textit{some} \ (\beta_0, \delta_0)' \in \Theta, \tag{21}$$

A test can be developed in the same way as in Section 3 by noting that $P_y(x)$ is, in fact, a regression model since $P_y(x) = E[1(Y_1 = y) \mid X_1 = x]$. Hence, we can employ the same statistic as (7) where now

$$T_n(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (1(Y_i = y) - P_y(X_i, \hat{\beta}, \hat{\delta})) \prod_{j=1}^{k} 1(X_{ij} \leq x_j). \tag{22}$$

Under H_0, $\hat{\beta}$ and $\hat{\delta}$ are $\sqrt{n}$- consistent and $T_n(x)$ has a limiting Gaussian process, $T(x)$, centered at zero and with covariance structure $K'(t, s)$, where now, Y_i is substituted by $1(Y_i = y)$ and $m(X_i, \hat{\beta})$ by $P_y(X_i, \hat{\beta}, \hat{\delta})$.

As in Section 3, the test statistic is

$$C_n = \frac{1}{n} \sum_{i=1}^{n} T_n(X_i)^2 \xrightarrow{d} C. \tag{23}$$

The asymptotic null distribution is useless for performing inferences since the test statistic is not distribution free. Therefore, we have to resort to bootstrap approximations of the critical values of the test.

The bootstrap in this case is different because now the estimates $\hat{\beta}$ and $\hat{\delta}$ are not obtained by regression methods but by maximun likelihood. Since under H_0, the conditional distribution of $Y_i \mid X_i$ is perfectly known, we can apply a parametric bootstrap. That is, for each X_i, Y_i^* is generated according to the distribution $P_y(x, \hat{\beta}, \hat{\delta})$. For instance, if we are testing the Poisson hypothesis, Y_i^* is generated according to a Poisson with parameter $\exp(X_i'\beta)$. From the generated sample (Y_i^*, X_i), $i = 1, ..., n$, we obtain bootstrap estimates of β_0 and δ_0, $\hat{\beta}^*$ and $\hat{\delta}^*$ respectively. The bootstrap statistic is

$$T_n^* = \frac{1}{n} \sum_{i=1}^{n} T_n^*(X_i)^2, \tag{24}$$

where

$$T_n^*(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (1(Y_i^* = y) - P_Y(X_i, \hat{\beta}^*, \hat{\delta}^*)) \prod_{j=1}^{k} 1(X_{ij} \leq x_j). \tag{25}$$

P-values for different values of y are reported in Table 8. The Poisson model is rejected for all count values. However, the Negative Binomial model cannot be rejected in any case. This is in agreement with the nonparametric estimates reported

in figures 1 and 2. The Negative Binomial parametric estimates are closer to the nonparametric estimates and they are always inside the confidence bands. Note that the kernel estimates are very unstable for observations in the frontiers of the data set.

Table 8 about here

The testing procedure can also be applied to a test for the specification of the distribution function. The null hypothesis to be tested is

$$H_0 : Pr\{Pr(Y_1 \mid X_1) = P(Y_1, X_1, \beta_0, \delta_0)\} = 1, \tag{26}$$

where $P(y, x, \beta, \delta) \equiv P_y(x, \beta, \delta)$. Andrews (1996) considers the statistic

$$C_n = \frac{1}{n} \sum_{i=1}^{n} T_n(Y_i, X_i)^2, \tag{27}$$

where

$$T_n(y, x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{1(Y_i = y) - P(y, X_i, \hat{\beta}, \hat{\delta})\} \prod_{j=1}^{k} 1(X_{ij} \leq x_j). \tag{28}$$

Andrews (1996) finds the asymptotic process of $T_n(y, x)$, proposing a parametric bootstrap in order to approximate the critical values of the statistic C_n. The parametric bootstrap is, as before, based on the statistic

$$C_n^* = \frac{1}{n} \sum_{i=1}^{n} T_n^*(Y_i^*, X_i)^2, \tag{29}$$

where

$$T_n^*(y, x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{1(Y_i^* = y) - P(y, X_i, \hat{\beta}^*, \hat{\delta}^*)\} \prod_{j=1}^{k} 1(X_{ij} \leq x_j). \tag{30}$$

In Table 9, we apply this test to our data set, and we find that the Poisson specification is rejected and the Negative Binomial specification cannot be rejected.

Table 9 about here

REFERENCES

[1] Andrews, D.W. (1996), ' A Conditional Kolmogorov Test,' Cowles Foundation Discussion Paper 1111R, Yale University.

[2] Bierens, H.J., (1982), ' Consistent Model Specification Tests,' *Journal of Econometrics*, **10**, 105-134.

[3] Bierens, H.J., (1990), 'A Consistent Conditional Moment Test of Functional Form,' *Econometrica*, **58**, 1443-1458.

[4] Bierens, H.J. and Ploberger W., (1997), 'Asymptotic Theory of Integrated Conditional Moment Test,' *Econometrica*, forthcoming.

[5] Cameron, A.C. and P.K. Trivedi (1990), 'Regression-Based Tests for Overdispersion in the Poisson Model,' *Journal of Econometrics*, **46**, 347-364.

[6] Cameron, A.C. and P.K. Trivedi (1986), 'Econometric Models Based on Count Data: Comparisons and Applications of some Estimators and Tests,' *Journal of Applied Econometrics*, **1**, 29-53.

[7] Cameron, A.C., P.K. Trivedi, F. Milne and J. Piggot (1988), 'A Microeconomic Model of the Demand for Health Care and Health Insurance in Australia,' *Review of Economic Studies*, **55**, 85-106.

[8] Cameron, A.C. and F.A Windmeijer (1996), ' R- Squared Measures for Count Data Regression Models with Applications to Health Care Utilization,' *Journal of Business & Economic Statistics*, **14**, 209-220.

[9] Delgado, M.A. (1992), 'Semiparametric Generalized Least Squares in the Multivariate Nonlinear Regression Model,' *Econometric Theory*, **8**, 203-222.

[10] Delgado, M.A. and T.J. Kniesner (1997), 'Count Data Models with Variance of Unknown Form: An Application to a Hedonic Model of Worker Absenteeism,' *Review of Economics and Statistics*, **LXXIX**, 41-49.

[11] Eubank, R. and S. Spiegelman (1990), 'Testing the Goodness of Fit of a Linear Model Via Nonparametric Regression Techniques,' *Journal of the American Statistical Association*, **85**, 387-392.

[12] Fan, Y. and Q. Li (1996), 'Consistent Model Specification Tests: Omitted Variables and Semiparametric Functional Forms,' *Econometrica*, **64**, 865-890.

[13] Hausman, J., B.H. Hall and Z. Griliches (1984), 'Econometric Models for Count Data with an Application to the Patents R & D Relationship,' *Econometrica*, **52**, 909-938.

[14] Robinson, P.M. (1987), 'Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form,' *Econometrica*, **55**, 875-891.

[15] Stute, W (1996), 'Nonparametric Model Checks for Regression,' forthcoming in *Annals of Statistics*.

[16] Stute, W., W. González Manteiga and M. Presedo Quindimil, (1996), 'Bootstrap Approximations in Model Checks for Regression,' forthcoming in *Journal of the American Statistical Association*.

[17] Wagstaff, A. (1986), 'The Demand for Health: Some New Empirical Evidence,' *Journal of Health Economics*, **5**, 195-233.

TABLA 1: DEFINITION OF VARIABLES

NAME	DEFINITION
DOCTOR VISITS	Number of doctor visits in the last 2 weeks
EMERGENCY ROOM VISITS	Number of emergency room visits in the last year
MALE	Dichotomous variable for sex (=1 if male, =0 if female)
AGE	Age of the individual
CHRONIC ILLNESS(*)	Dichotomous variable for chronic health condition (=1 if yes, =0 otherwise)
ACUTE ILLNESS	Dichotomous variable for acute health condition in the reference period (=1 if yes, =0 otherwise)
ACCIDENT	Dichotomous variable for suffering an accident in the last year (=1 if yes, =0 otherwise)
SMOKER	Dichotomous variable for current smoker (=1 if yes, =0 otherwise)
EX-SMOKER	Dichotomous variable for ex-smoker (=1 if yes, =0 otherwise)
NON-SMOKER	Dichotomous variable for non-smoker (=1 if yes, =0 otherwise)
ALCOHOL(**)	Daily alcohol consumption (cubic centimeters)
WAGE-EARNING	Dichotomous variable for wage-earning (=1 if yes, =0 otherwise)
NON-WORKING	Dichotomous variable for non-working: student, unemployed, inactive (=1 if yes, =0 otherwise)
SELF-EMPLOYED	Dichotomous variable for self-employed (=1 if yes, =0 otherwise)
SCHOOL	Years of schooling completed
INCOME	Estimated household income
NORTH	Dichotomous variable for region (=1 if living in Asturias, Navarra Cantabria, País Vasco, Lugo, Lérida, or Castilla-León, =0 otherwise)
SOUTH	Dichotomous variable for region (=1 if living in Andalucía, Murcia or Extremadura, =0 otherwise)
CENTER	Dichotomous variable for region (=1 if not living in none of above regions, =0 otherwise)
RURAL	Dichotomous variable for residing in a rural area of population < 10.000 (=1 if yes, =0 otherwise)

(*) Illnesses in question are: heart disease, chronic bronchitis, asthma, diabetes, hypertension, allergy, high levels of cholesterol and estomach ulcer.

(**) To create this variable, we have transformed quantites of consumed alcoholic drinks and its alcoholic graduations in alcoholic cubic centimeters. For this purpose, we have used the conversion table reported by the Spanish Ministry of Health and Consumer Affairs.

TABLE 2: DESCRIPTIVE STATISTICS

	MEAN	STD. DEV	MIN	MAX
MALE	0.28	0.67	0	1
AGE	49.44	16.41	16	95
CHRONIC ILLNESS	0.33	0.47	0	1
ACUTE ILLNESS (2 weeks)	0.093	0.29	0	1
ACUTE ILLNESS (1 year)	0.184	0.387	0	1
ACCIDENT	0.071	0.257	0	1
SMOKER	0.40	0.49	0	1
EX-SMOKER	0.20	0.40	0	1
NON-SMOKER	0.36	0.48	0	1
ALCOHOL	14.56	21.72	0	125
WAGE EARNING	0.60	0.49	0	1
NON-WORKING	0.17	0.37	0	1
SELF-EMPLOYED	0.22	0.42	0	1
SCHOOL	9.05	4.33	2.5	18
INCOME $\times 10^{-6}$	3.298822	1.38	0.863778	11.985321
NORTH	0.27	0.44	0	1
SOUTH	0.17	0.38	0	1
CENTER	0.55	0.49	0	1
RURAL	0.30	0.46	0	1

TABLE 3: FREQUENCIES OF COUNTS

COUNTS	DOCTOR VISITS	EMERGENCY ROOM VISITS
0	6167	6841
1	1264	807
2	297	108
3	54	31
4	19	10
5	3	7
6	0	1
7	2	4
8+	11	8
Total	7817	7817

TABLE 4: ESTIMATES FOR THE NUMBER OF DOCTOR VISITS

	POISSON	NEGBIN	NLS	SGLS
INTERCEPT	-3.4339	-3.4800	-2.9164	-4.2689
	(0.3457)	(0.3799)	(0.3368)	(0.4143)
	[0.4106]	[0.4156]	[0.5247]	[0.4273]
log(AGE)	0.4855	0.5038	0.3517	0.6494
	(0.0800)	(0.0880)	(0.0777)	(0.0967)
	[0.0954]	[0.0968]	[0.1194]	[0.0990]
MALE	-0.2236	-0.2486	-0.1539	-0.2657
	(0.0558)	(0.0625)	(0.0490)	(0.0653)
	[0.0670]	[0.0669]	[0.0796]	[0.0729]
CHRONIC	0.4419	0.4718	0.3026	0.4398
ILLNESS	(0.0467)	(0.0518)	(0.0426)	(0.0551)
	[0.0556]	[0.0557]	[0.0673]	[0.0601]
ACUTE	1.3061	1.3622	1.2403	1.2770
ILLNESS	(0.0472)	(0.0564)	(0.0401)	(0.0551)
	[0.0575]	[0.0560]	[0.0586]	[0.0610]
log(ALCOHOL)	-0.0738	-0.0737	-0.0676	-0.0694
	(0.0148)	(0.0164)	(0.0141)	(0.0172)
	[0.0166]	[0.0168]	[0.0206]	[0.0176]
SMOKER	0.0160	0.0121	0.0541	-0.0165
	(0.0587)	(0.0646)	(0.0552)	(0.0691)
	[0.0682]	[0.0681]	[0.0828]	[0.0737]
EX-SMOKER	0.1918	0.1954	0.1579	0.2488
	(0.0593)	(0.0667)	(0.0524)	(0.0672)
	[0.0703]	[0.0699]	[0.0895]	[0.0764]
NON-WORKING	0.2309	0.2502	0.1170	0.2866
	(0.0518)	(0.0595)	(0.0450)	(0.0586)
	[0.0606]	[0.0623]	[0.0715]	[0.0624]
SELF-EMPLOYED	-0.3067	-0.3134	-0.2523	-0.3323
	(0.0665)	(0.0718)	(0.0667)	(0.0766)
	[0.0772]	[0.0772]	[0.0842]	[0.0765]
SCHOOL	-0.0041	-0.0054	-0.0020	0.0046
	(0.0065)	(0.0074)	(0.0057)	(0.0073)
	[0.0078]	[0.0080]	[0.0099]	[0.0081]
INCOME $\times 10^{-6}$	0.0319	0.0268	0.0494	0.0233
	(0.0210)	(0.0234)	(0.0183)	(0.0234)
	[0.0260]	[0.0258]	[0.0340]	[0.0267]
RURAL	0.0220	0.0072	0.0930	0.0264
	(0.0499)	(0.0559)	(0.0440)	(0.0570)
	[0.0599]	[0.0607]	[0.0724]	[0.0611]
NORTH	-0.1409	-0.1495	-0.1210	-0.1924
	(0.0532)	(0.0591)	(0.0502)	(0.0632)
	[0.0633]	[0.0647]	[0.0715]	[0.0628]
SOUTH	0.1547	0.1480	0.1782	0.1895
	(0.0552)	(0.0634)	(0.0455)	(0.0595)
	[0.0637]	[0.0655]	[0.0746]	[0.0662]
δ		1.7393		
		(0.2004)		
		[0.3064]		
log-lik	-4836.68	-4761.51		
R^2	0.18	0.19	0.14	0.10

() Standard errors. [] Eicker-White robust standard errors.

TABLE 5: ESTIMATES FOR THE NUMBER OF EMERGENCY ROOM VISITS

	POISSON	NEGBIN	NLS	SGLS
INTERCEPT	-0.4566	-0.3420	-0.6262	-0.6229
	(0.4046)	(0.4818)	(0.3656)	(0.4577)
	[0.5728]	[0.6174]	[0.6317]	[0.5106]
log(AGE)	-0.5013	-0.5507	-0.3750	-0.5183
	(0.0946)	(0.1132)	(0.0840)	(0.1081)
	[0.1329]	[0.1437]	[0.1430]	[0.1190]
MALE	0.0384	0.0300	0.0431	-0.0623
	(0.0749)	(0.0892)	(0.0668)	(0.0884)
	[0.0981]	[0.0988]	[0.1161]	[0.1046]
CHRONIC	0.4089	0.4910	0.1684	0.3124
ILLNESS	(0.0620)	(0.0736)	(0.0549)	(0.0764)
	[0.0866]	[0.0893]	[0.0955]	[0.0874]
ACUTE	1.3371	1.3876	1.2603	1.5546
ILLNESS	(0.0596)	(0.0699)	(0.0802)	(0.0761)
	[0.0770]	[0.0756]	[0.0586]	[0.0872]
ACCIDENT	1.1950	1.4238	0.8575	1.0297
	(0.0669)	(0.0903)	(0.0528)	(0.0806)
	[0.0793]	[0.0746]	[0.0914]	[0.0966]
log(ALCOHOL)	-0.0731	-0.0924	-0.0109	-0.0592
	(0.0188)	(0.0224)	(0.0163)	(0.0223)
	[0.0240]	[0.0241]	[0.0322]	[0.0239]
SMOKER	0.0041	0.0340	-0.1162	-0.0308
	(0.0722)	(0.0860)	(0.0640)	(0.0837)
	[0.0935]	[0.0928]	[0.1311]	[0.0951]
EX-SMOKER	0.1330	0.1663	0.0459	0.1850
	(0.0782)	(0.0936)	(0.0678)	(0.0936)
	[0.1032]	[0.1040]	[0.1232]	[0.1055]
NON-WORKING	0.1699	0.1859	0.1106	0.1265
	(0.0686)	(0.0837)	(0.0588)	(0.0916)
	[0.0962]	[0.0964]	[0.1136]	[0.1006]
SELF-EMPLOYED	-0.1897	-0.2316	-0.0188	-0.2752
	(0.0816)	(0.0945)	(0.0729)	(0.0968)
	[0.0992]	[0.0961]	[0.1530]	[0.1171]
SCHOOL	-0.0265	-0.0282	-0.0215	-0.0163
	(0.0088)	(0.0104)	(0.0080)	(0.0103)
	[0.0106]	[0.0108]	[0.0124]	[0.0113]
INCOME $\times 10^{-6}$	0.0625	0.0268	0.0190	0.0277
	(0.0267)	(0.0313)	(0.0245)	(0.0307)
	[0.0374]	[0.0345]	[0.0513]	[0.0356]
RURAL	-0.2767	-0.3042	-0.2186	-0.2795
	(0.0681)	(0.0799)	(0.0614)	(0.0882)
	[0.0868]	[0.0857]	[0.1135]	[0.0917]
NORTH	-0.1554	-0.1630	-0.1765	-0.2406
	(0.0700)	(0.0819)	(0.0673)	(0.0899)
	[0.0856]	[0.0868]	[0.0956]	[0.0873]
SOUTH	0.0395	0.0450	0.0030	0.1012
	(0.0740)	(0.0892)	(0.0633)	(0.0857)
	[0.0973]	[0.0967]	[0.1169]	[0.0988]
δ		0.7821		
		(0.0803)		
		[0.1127]		
log-lik	-4836.68	-4761.51		
R^2	0.20	0.22	0.10	0.07

() Standard errors. [] Eicker-White robust standard errors.

TABLA 6: ML estimates for the number of doctor visits. Subsample consists of smoker male individuals that are head of household, wage-earning, without chronic and acute illnesses, from the Center and the South of Spain and between 23 and 64 years old (N=764).

	POISSON	NEGBIN
INTERCEPT	-2.662	-2.645
	(0.301)	(0.360)
	[0.405]	[0.395]
$AGE^2 \times 10^{-4}$	0.275	0.271
	(0.125)	(0.156)
	[0.153]	[0.154]
Log(ALCOHOL)	-0.043	-0.047
	(0.067)	(0.083)
	[0.084]	[0.084]
δ		0.212
		(0.071)
		[0.081]
Log-lik	-265.954	-248.777

() Standard errors. [] Eicker-White robust standard errors.

TABLE 7: ESTIMATED MARGINAL PROBABILITIES

	$P(Y=0)$	$P(Y=1)$	$P(Y=2)$	$P(Y \geq 3)$
SAMPLE	0.9210	0.0640	0.0090	0.0060
POISSON	0.9047	0.0903	0.0024	0.0026
NEGBIN	0.9219	0.0617	0.0122	0.0042
NONPARAMETRIC	0.9221	0.0633	0.0095	0.0053

TABLE 8: P-VALUES FOR STUTE TEST

N-Bootstrap = 1000

	POISSON	BINEG
$P(Y=0 \mid X)$	0.009	0.610
$P(Y=1 \mid X)$	0.008	0.408
$P(Y \geq 2 \mid X)$	0.004	0.505

TABLE 9: P-VALUES FOR ANDREWS TEST

N-Bootstrap = 1000

POISSON	BINEG
0.000	0.558

FIGURE 1:Estimated conditional probabilities (ALCOHOL is fixed at its mean value)

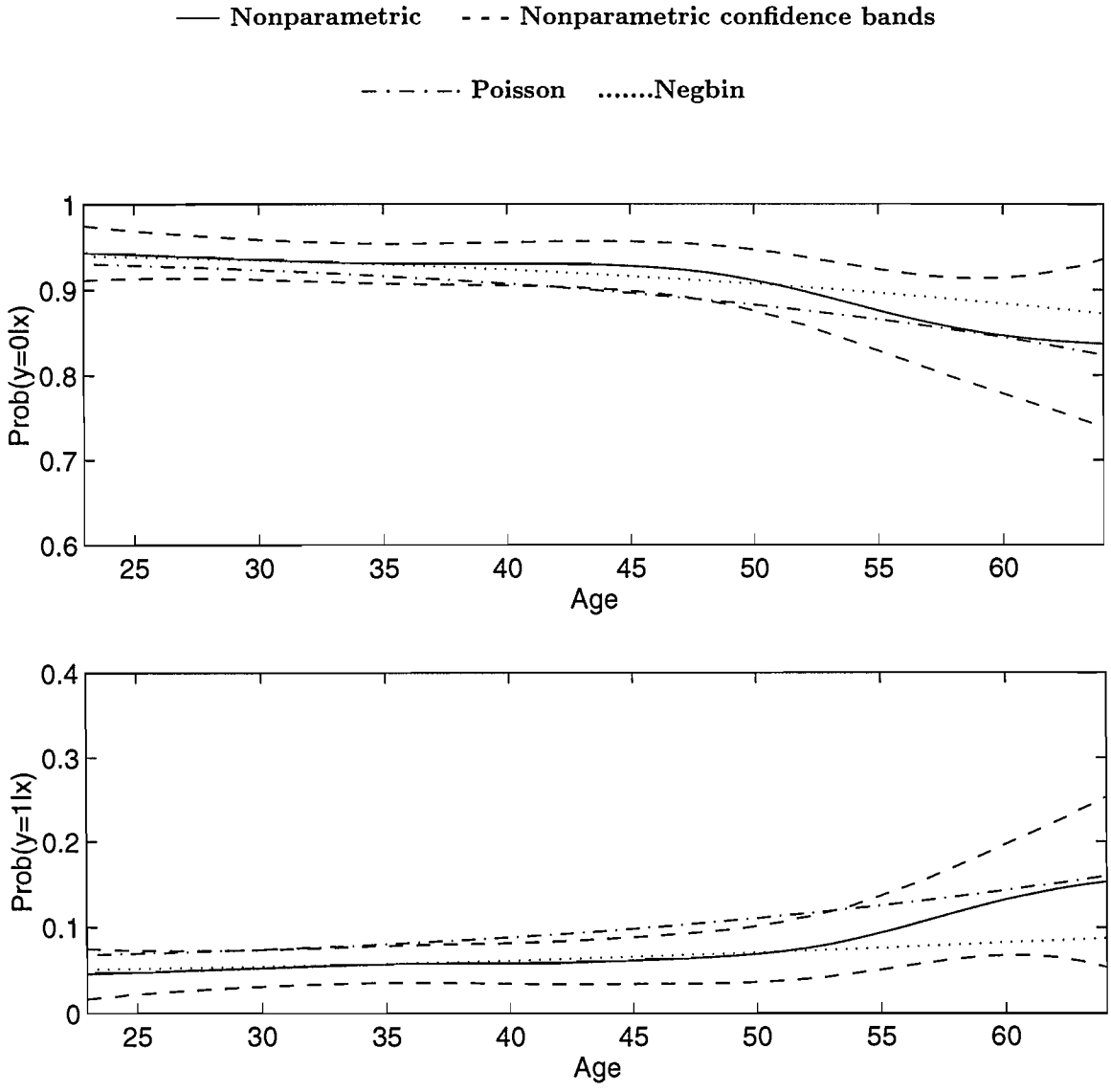

FIGURE 2:Estimated conditional probabilities (AGE is fixed at its mean value)