Socio-Economic Planning Sciences 95 (2024) 101976

Contents lists available at ScienceDirect

Socio-Economic Planning Sciences

journal homepage: www.elsevier.com/locate/seps

Semi-parametric approach for modelling overdispersed count data with application to Industry 4.0

S. Bonnini [a], M. Borghesi [a], M. Giacalone [b,*]

[a] University of Ferrara, Department of Economics and Management, Via Voltapaletto, 11, Ferrara, 44121, Italy
[b] University of Campania "Luigi Vanvitelli", Department of Economics, Corso Gran Priorato di Malta I, Capua, 81043, Italy

ARTICLE INFO

Keywords:
Generalized linear model
Count data
Overdispersion
Permutation test
Industry 4.0
Public policy

ABSTRACT

The paper deals with a test for the goodness-of-fit of a model for count data, in the framework of Generalized Linear Models. The motivating example concerns the study on the effectiveness of policy incentives for the adoption of 4.0 technologies by Small and Medium Enterprises. According to the literature, openness to Industry 4.0 should be measured in terms of the number of 4.0 technologies adopted, represented by a count variable. To investigate the effectiveness of public policy interventions to encourage the adoption of 4.0 technologies, we propose the application of a model for count data with a permutation ANOVA to test the goodness-of-fit and for the model selection. When the distribution of the response is neither Poisson nor Negative Binomial, and in the quite common situation in which the response variance is much greater than the mean, the classic Poisson regression and Negative Binomial regression are not valid. The proposed testing method is based on the combination of permutation tests on the significance of the regression coefficient estimates. The power behaviour of the proposed semi-parametric solution is investigated through a comparative Monte Carlo simulation study. The performance of such a method is compared to those of the two main parametric competitors. The application of the permutation test to an interesting case study is presented. The dataset is original, and related to a sample survey carried out in Italy, about the adoption of Industry 4.0 technologies by Italian enterprises.

1. Introduction

This research deals with the application of a regression analysis for count data, using Generalized Linear Models (GLMs). In particular, we focus on the test for the validity of the model. When the probability distribution of the dependent variable is neither Poisson nor Negative Binomial, the classic likelihood ratio test of the Poisson regression and Negative Binomial regression may not be suitable solutions. Furthermore, in the quite frequent situation of overdispersed data, i.e. when the variance is very large and, above all, much greater than the mean, the parametric inference, in particular the Poisson regression, may not be effective and performant [1]. To overcome these drawbacks, we propose a solution based on a multiple permutation test on the significance of the regression coefficient estimates.

The motivating example concerns the effectiveness of public policies in enhancing the innovative capacity of companies concerning Industry 4.0 technologies. In the literature, some studies have been conducted on the effect of policy interventions to improve the innovative capacity of companies. However, many aspects related to the implementation of Industry 4.0 technologies still need to be explored [2]. Furthermore, there are some barriers to the adoption of Industry 4.0 technologies that can be overcome through the allocation of public incentives aimed at encouraging training and skills development programs [3]. The list of 4.0 technologies includes a wide and heterogeneous set of innovative solutions. According to [4], the advent of Industry 4.0 has led to some disadvantages but also many benefits. Hence, in recent years, public incentives to support the innovative processes of private companies have been proposed by the governments [5]. The data of the motivating example concerns an Italian case study. In January 2022 a sample survey was carried out in the northern regions of Italy to assess the effect of policy incentives on the adoption of Industry 4.0 technologies by Small and Medium Enterprises (SMEs). To this aim, a suitable model to predict the number of adopted innovations of Industry 4.0 as a function of the incentives used by the companies may be defined. Hence, the dependent variable takes non-negative integer values and we deal with a model for count data.

Given that the classic linear regression analysis is not appropriate for count data because the dependent variable is not continuous [6], the GLM (Generalized Linear Model) approach is widely used in the

* Corresponding author.
E-mail address: massimiliano.giacalone@unicampania.it (M. Giacalone).

https://doi.org/10.1016/j.seps.2024.101976
Received 30 January 2024; Received in revised form 22 April 2024; Accepted 7 June 2024
Available online 8 June 2024

literature. Within the family of GLMs for count data, according to the assumed probability distribution for the dependent variable, the Poisson regression or the Negative Binomial regression are mostly considered [7–9]. In the theory of GLM, the inference on the model is usually carried out by using the likelihood approach, i.e. the maximum likelihood estimates and the likelihood ratio test for estimation and test of hypotheses respectively. In some cases, such as perfect or high collinearity between regressors, parameter estimates do not exist or cannot be calculated. Since this type of identification failure has not been widely recognized as a problem in count data models, often standard software does not check for the multicollinearity and consequently unreliable results may be obtained [10].

Furthermore, as mentioned above, overdispersion (and sometimes underdispersion) is a commonly encountered problem in the context of regression analysis for count data. Under some conditions, to handle overdispersion, alternative methods to Poisson regression may be considered, including Negative Binomial regression, mixed effects models, and Conway-Maxwell-Poisson regression [11]. These methods are parametric and therefore based on restrictive assumptions about the probability distribution of the response. The proposal presented in this paper is based on a semi-parametric method, robust and less restrictive from the point of view of the model assumptions. Specifically, the proposal is based on Combined Permutation Tests (CPTs), a family of tests for complex problems based on the permutation approach [12]. Such a solution is suitable for testing problems that can be broken down into partial tests. The partial tests, in this specific case, are based on the parametric estimators of the regression coefficients. The combined use of permutation tests and parametric estimators makes the method classifiable as semiparametric. CPTs do not require the assumption that the probability distribution of the dependent variable belongs to a certain family of distributions. The application of this method is possible when the error terms satisfy the mild condition of exchangeability [12]. The semi-parametric nature of the method implies flexibility for the conditions of applicability.

Permutation tests have been widely applied in empirical studies [13,14], with numeric variables but also categorical data [15,16], for big data problems [17], in regression analysis [18], to test directional and non-monotonic hypotheses [16,19], and in many other problems. In this work, the application concerns models for count data. The conditional inference has been proven to be suitable and effective in mixed models [20,21], and, in GLM [22,23]. Permutation goodness-of-fit tests, based on partial sums or cumulative sums of residuals, have been proposed for linear regression models [24–26]. To test the effect of covariates, [21] proposed the use of the nonparametric combination of permutation tests. The idea of considering the test on the validity of a multivariate linear model as a multiple test was presented by [27] in the framework of rotation tests.

The rest of the paper is structured as follows. Section 2 concerns the statistical problem related to the goodness-of-fit test for count data. Section 3 deals with the proposed solution based on the CPT methodology. The simulation study to compare the performance of the most typical parametric methods and the proposed permutation tests is presented in Section 4. The case study concerning the sample survey recently carried out in the northern regions of Italy, to assess the effect of policy incentives on the adoption of Industry 4.0 technologies, is described in Section 5. Finally, Section 6 includes the conclusions of the paper.

2. Statistical problem

Let us consider a regression model where Y_i, the response related to the i-th statistical unit, is a count variable (e.g. the number of 4.0 innovations) with $i = 1, \dots, n$. Furthermore, $x_{i1}, \dots, x_{iq}$ represent the sample values of q predictors observed on the i-th statistical unit. The conditional expectation of the response, given the observed values of the q predictors, for the i-th statistical unit is

$$E\left[Y_i | x_{i1}, \dots, x_{iq}\right] = \mu_i.$$

As typical of problems with count data, we consider the log-linear model, where the relationship between the conditional mean of the dependent variable and the predictors is expressed by the following link function:

$$\log\left(\mu_i\right) = \beta_0 + \sum_{k=1}^{q} \beta_k x_{ik}. \tag{1}$$

Note that the choice of such a link function mainly depends on the fact that μ_i takes only non-negative values. In fact, in case the predictions of the response returned by the right-hand side of the regression equation were negative, and the left-hand side of was the mean, we would have an inconsistency because the predicted values would be inadmissible. Since $\log(\mu_i) \in \mathbb{R}$, the log-linear specification allows us to overcome the problem. According to , the fitted values of the response mean are

$$\hat{\mu}_i = exp\left(\hat{\beta}_0 + \sum_{k=1}^{q} \hat{\beta}_k x_{ik}\right),$$

where $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_q$ are suitable estimators of the regression coefficients.

We are interested in the test on model adequacy, in other words on the goodness-of-fit. Hence, the system of hypotheses is the following:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0 \\ H_1 : \bar{H}_0. \end{cases} \tag{2}$$

Under the null hypothesis, the conditional expectation of Y_i does not depend on the observed values of the explanatory variables. Whatever the values of the predictors, the conditional mean does not change and corresponds to $exp(\beta_0)$. In other words, the log-linear model is unuseful in predicting μ_i as a function of the q considered explanatory variables. In the alternative hypothesis, the model is valid because at least one regression coefficient is not equal to zero and therefore at least one independent variable affects the response mean.

The regression analysis can be carried out with the GLM approach [7]. According to the assumed distribution of Y_i, conditional on the observed values of the explanatory variables, different solutions are possible. The most popular are the Poisson regression and the Negative Binomial regression. In the theory of GLM, to estimate the parameters $\beta_0, \beta_1, \dots, \beta_q$, the most commonly used methods are the Maximum Likelihood (ML) estimation [28] and Maximum Quasi-Likelihood (MQL) estimation [29]. Instead, to test the model adequacy, the Likelihood Ratio Test is the typically adopted solution [30]. Within the family of GLMs for count data, the Poisson model is the oldest and, still today, the most widespread. On the other hand, the Negative Binomial model is receiving growing attention and having good success, also thanks to some interesting properties which, in some cases, make it preferable to the Poisson model, overcoming some of the latter's limitations [8].

As said, a relevant limitation concerns the situation of overdispersion. Indeed, when the sample variance is much greater than the mean, the assumption that the dependent variable follows the Poisson distribution is not plausible, given that one characteristic property of the Poisson probability distribution is that the variance is equal to the mean. If the sample variance is not too much greater than the mean, then the Negative Binomial regression is appropriate and preferable to the Poisson regression. A Poisson model estimated on overdispersed data may lead to underestimated standard errors of the parameter estimators [1]. According to the literature, the Negative Binomial regression is among the possible solutions in the presence of overdispersion, because the corresponding model has a higher tolerance for extra variability [1]. Nevertheless, when the ratio between variability and central tendency in the sample data reaches particularly high levels, Negative Binomial regression also does not work [31]. A study

conducted by [1] identified the cases of moderate overdispersion, in which a simple Poisson model may be utilized, intermediate overdispersion, in which negative binomial regression is preferable, and extreme overdispersion, where neither solution is effective. In these extreme cases, an effective solution has yet to be proposed and there is therefore room for exploration to identify appropriate innovative methodological solutions.

3. Methodological solution

As said, our proposed approach is based on the CPT testing method [32,33]. The only required assumption is the exchangeability of the errors with respect to units under the null hypothesis [12,34]. The main idea is to conceive the problem as a multiple test, composed of the q partial tests on the significance of the regression coefficients. Specifically, the k-th partial test statistic is

$$T_k = |\hat{\beta}_k|, \tag{3}$$

where $\hat{\beta}_k$ is an estimator of the k-th regression coefficient. We may consider, as alternative options, the maximum likelihood estimators of the Poisson regression and of the Negative Binomial regression. For the overall testing problem defined by (2), according to the CPT theory, a suitable test statistic is based on the combination of the p-values of the q partial tests. One of the main advantages is that there is no need to know or assume either the marginal distribution of each partial test statistic or the joint distribution of the q partial test statistics (q-variate overall test statistic).

The procedure of combined permutation tests is the following:

1. Compute the vector of observed values of the partial test statistics $\mathbf{t_0} = (t_{01}, \dots, t_{0q})' = t(\mathbf{X})$, where $t_{0k} = |b_k|$, with $k = 1, \dots, q$.
2. Carry out B independent random permutations of the rows of the $\mathbf{X}$ matrix: $\mathbf{X_1^*}, \dots, \mathbf{X_B^*}$ by keeping fixed the vector of observed values of the dependent variable Y.
3. Compute the vector of q partial test statistics for each of the B dataset permutations, $\mathbf{t_r^*} = t(\mathbf{X_r^*}) = (t_{r1}^*, \dots, t_{rq}^*)'$ and the corresponding vector of p-values $\mathbf{l_r^*} = (l_{r1}^*, \dots, l_{rq}^*)'$ with $r = 1, \dots, B$. Each partial p-value is obtained through the application of the significance level function, according to the null permutation distribution. Formally, $l_{rk}^* = L_k(t_{rk}^*)$, with

$$L_k(t) = \left(\sum_{s=1}^{B} I_{(-\infty, t_{sk}^*]}(t) + 0.5 \right) / (B+1),$$

where $I_A(t) = 1$ if $t \in A$ and $I_A(t) = 0$ otherwise. The p-values, computed as indicated above, represent estimates of those obtained from the exact null permutation distribution, i.e. the distribution obtained by considering all the $n!$ possible permutations of the rows of $\mathbf{X}$. For computational reasons, instead of considering all the possible permutations, it is common practice to use a random sample from the permutation space. The approximation is good if the number of permutations randomly generated is at least 1000.
4. Compute the combined test statistic for each permutation and for the observed dataset by using a suitable function

$$\psi : [0,1]^q \to \mathbb{R}, t_{\psi r}^* = \psi\left(\mathbf{l}_r^*\right),$$

with $r = 1, \dots, B$. By assuming, without loss of generality, that the null hypothesis is rejected for large values of the test statistic, the combining function ψ must be a non-increasing function of the p-values, it tends to the supremum when one argument tends to zero, and has a critical value finite and less than the supremum. The most commonly used combining rule corresponds to the Fisher omnibus function:

$$\psi(\mathbf{l_r^*}) = -2 \sum_{k=1}^{q} \log\left(l_{rk}^*\right). \tag{4}$$

5. Compute the p-value of the combined test according to the null permutation distribution:

$$l_{\psi r}^* = L_\psi(t_{\psi r}^*). \tag{5}$$

The described solution has been implemented by the authors through original R scripts specifically created for the problem. In order to compute the parametric estimates used as test statistics for the partial tests, the R function *glm* has been used for the Poisson regression and the function *glm.nb* with regards to the Negative Binomial regression.

4. Simulation study

In this section, the results of a Monte Carlo simulation study are analysed to assess the power behaviour of goodness-of-fit tests of the GLM for count data. In particular, the two CPTs based on the maximum likelihood estimates of Poisson regression and Negative Binomial regression are compared with the parametric counterparts based on the likelihood ratio test. It is well known that, when the distribution of the response variable is Poisson or Negative Binomial, the likelihood ratio test of the Poisson and Negative Binomial regression respectively, is the best possible choice in terms of power behaviour. Hence, in our Monte Carlo simulation study, we considered a distribution other than Poisson and Negative Binomial and we focused in particular on the overdispersion case. The goal was to detect specific conditions in which the classic parametric methods are not suitable, unlike the proposed semi-parametric tests. All the simulations were carried out through R scripts created by the authors.

Let n and q denote the sample size and the number of explanatory variables of the model. The $n \times q$ matrix of the predictors $\mathbf{X}$ was simulated by randomly generating n observations from a q variate normal distribution with null mean vector and variance–covariance matrix $\mathbf{\Sigma}$, i.e. $\mathbf{X} \sim \mathcal{N}_q(\mathbf{0_q}, \mathbf{\Sigma})$. We assumed that the variance of each predictor is equal to σ_x^2 and the correlation between each couple of independent variables is ρ_x. Consequently,

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_x^2 & \sigma_x^2\rho_x & \cdots & \sigma_x^2\rho_x \\ \sigma_x^2\rho_x & \sigma_x^2 & \cdots & \sigma_x^2\rho_x \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_x^2\rho_x & \sigma_x^2\rho_x & \cdots & \sigma_x^2 \end{pmatrix},$$

which can be written in a compact way as $\mathbf{\Sigma} = \sigma_x^2\left[\rho_x \mathbf{J_q} + (1-\rho_x)\mathbf{I_q}\right]$, where $\mathbf{J_q}$ denotes the $q \times q$ all-ones matrix and $\mathbf{I_q}$ is the identity matrix of order q.

The values of the dependent variable conditional to the observed values of the predictors were generated according to a normal distribution and then transformed into non-negative integer numbers by considering the absolute value of the integer part. Formally,

$$Z_i|\mathbf{X} \sim \mathcal{N}\left(\eta_i, \sigma_z^2\right),$$

where the mean is linked to the observed values of the explanatory variables as follows

$$\eta_i = \exp\left(\beta_0 + \sum_{k=1}^{q} \beta_k x_{ik}\right).$$

The final transformation of the values to create count data is

$$Y_i = |[Z_i]|.$$

For each setting, simulations were performed by randomly generating 1000 datasets, and the null permutation distribution of the test statistics was estimated by considering 1000 permutations. Due to computational complexity, we considered the case of $q = 2$ independent variables. Firstly, simulations were carried out under the null hypothesis H_0, with $\beta_0 = 1$ and $\beta_1 = \beta_2 = 0$.

Fig. 1 shows the rejection rates of both the proposed semi-parametric tests and the corresponding parametric versions as functions of σ_z^2. We considered the cases of $\sigma_z^2 = 2$, 4, 6, and 9, with

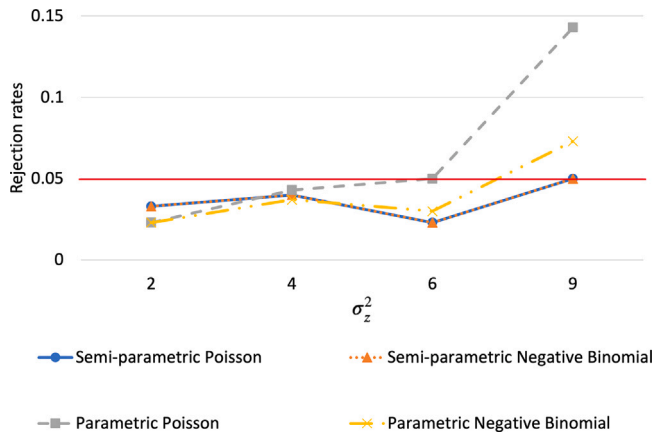


Fig. 1. Rejection rates under H_0 as a function of $\sigma_z^2 \in \{2, 4, 6, 9\}$, with $n = 600$, $q = 2$, $\rho_x = 0.2$, $\sigma_x^2 = 1$, $\beta_0 = 1$, $\alpha = 0.05$.
Source: Original data simulated by the authors.

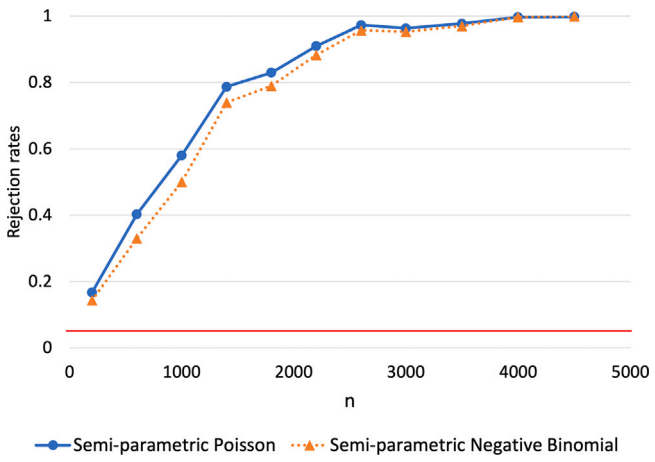


Fig. 2. Rejection rates of the semi-parametric tests under H_1 as a function of n, with $q = 2$, $\rho_x = 0.2$, $\sigma_x^2 = 1$, $\sigma_z^2 = 9$, $\beta_0 = \beta_1 = \beta_2 = 1$, $\alpha = 0.05$.
Source: Original data simulated by the authors.

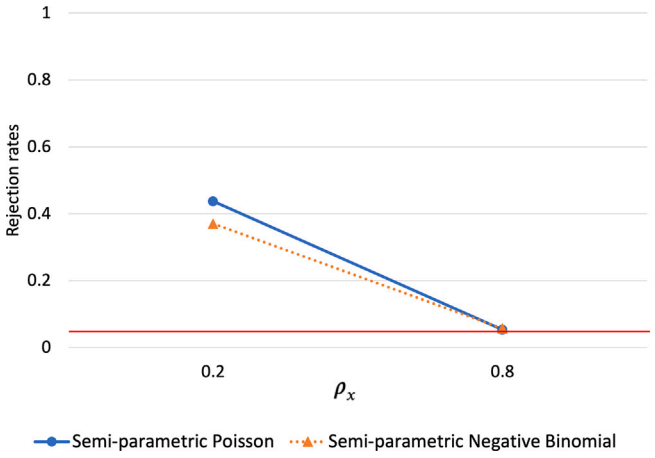


Fig. 3. Rejection rates of the semi-parametric tests under H_1 as a function of $\rho_x \in \{0.2, 0.8\}$, with $n = 600$, $q = 2$, $\sigma_x^2 = 1$, $\sigma_z^2 = 9$, $\beta_0 = \beta_1 = \beta_2 = 1$, $\alpha = 0.05$.
Source: Original data simulated by the authors.

$n = 600$, $q = 2$, $\rho_x = 0.2$, $\sigma_x^2 = 1$, $\beta_0 = 1$, $\alpha = 0.05$. Since, under H_0, $\eta_i = \exp(\beta_0) = e^1 = 2.7183$, the ratio σ_z^2/η_i in the four considered cases takes the values 0.736, 1.472, 2.207 and 3.311 respectively. It is worth noting that σ_z^2 is the variance of the normal random variable used to generate the count variable Y and not the variance of the (discrete and non-negative) response σ_Y^2, hence the mentioned ratios do not compare the variance and the mean of the dependent variable. However, they can be used as a measure of overdispersion. As can be seen, the higher the variability of the response (overdispersion) *ceteris paribus*, the greater the rejection rates of the likelihood ratio tests of the Poisson regression and Negative Binomial regression. Up to $\sigma_z^2 = 6$ and $\sigma_z^2/\eta_i = 2.207$, the rejection rates of all the tests are less than the significance level $\alpha = 0.05$. When $\sigma_z^2 = 9$ and $\sigma_z^2/\eta_i = 3.311$, the rejection rates of both the parametric tests do not respect the nominal α level. On the other hand, the tests based on the proposed semi-parametric method lead to rejection rates lower than or equal to α, whatever the considered σ_z^2 value. Thus, in case of overdispersion, the parametric tests of the Poisson and the Negative Binomial regression are anticonservative. In order to deepen the properties of the proposed semi-parametric approach with overdispersed data, when it is preferable to the classic likelihood approach, we investigate the power behaviour, i.e. the probability of rejection of the null hypothesis, of the two permutation tests under H_1 when $\sigma_z^2 = 9$.

Simulations were carried out under H_1, with the same setting parameters of Fig. 1 ($q = 2$, $\rho_x = 0.2$, $\sigma_x^2 = 1$, $\beta_0 = 1$, $\alpha = 0.05$), with $\sigma_z^2 = 9$, $\beta_1 = \beta_2 = 1$ and different n values, ranging from 200 to 4500. The estimated power of the proposed tests, under H_1, is represented as a function of the sample size, to evaluate the consistency of the test, in Fig. 2. First of all, the powers are always greater than α (horizontal red line in the graph). Then, both the tests are unbiased because the power under the alternative hypothesis is always greater than the power under the null hypothesis. Second, the power of the test based on Poisson estimators seems to be slightly greater than that of the test based on Negative Binomial estimators, but the performance is very similar. It is evident that the power of both tests increases with the sample size and tends to 1 as n diverges. The power is approximately 1 at $n = 4000$ for the semi-parametric Poisson test and at $n = 4500$ for the semi-parametric Negative Binomial test. Hence, both the tests are consistent and the power convergence rate to 1 of the former is slightly greater than that of the latter.

In Table 1, the effect of two different values of β_1 and β_2 can be seen in the two cases of $n = 200$ and $n = 600$ ($q = 2$, $\sigma_x^2 = 1$, $\sigma_z^2 = 9$, $\beta_0 = 1$, $\alpha = 0.05$). As shown above, the powers when $n = 600$ are greater than in the case where $n = 200$, due to the consistency of the tests. As expected, when $\beta_1 = \beta_2 = 2$, the rejection rates are greater than in the case where $\beta_1 = \beta_2 = 1$. Hence, the greater the true value of the regression coefficients, the further we are from the null hypothesis, and the greater the probability that the tests correctly reject H_0.

Finally, in Fig. 3, the effect of multicollinearity of the explanatory variables can be assessed. The setting parameters are: $n = 600$, $q = 2$, $\sigma_x^2 = 1$, $\sigma_z^2 = 9$, $\beta_0 = \beta_1 = \beta_2 = 1$, $\alpha = 0.05$. The rejection rates in the two cases of weak and strong correlation between the two predictors, i.e. $\rho_x = 0.2$ and $\rho_x = 0.8$ respectively, are represented. A clear drop in power, going from weak to strong collinearity, can be observed. Also in this case the result was expected because the collinearity affects the strength of the dependence between the partial tests. The greater the dependence strength, the lower the marginal informative contribution of each partial test net of the other.

5. Application to industry 4.0

We consider the application of the proposed method to an original dataset concerning a sample survey carried out in Italy in January 2022. The survey was conducted in the northern regions of Italy by the Department of Economics and Management of the University of Ferrara. A stratified random sample of manufacturing enterprises in North Italy was interviewed. We focus on the region Emilia–Romagna, one of the most developed and productive regions of the Country. In this region, more than half of the companies have embraced the 4.0 paradigm. Such

Table 1
Rejection rates of the semi-parametric tests under H_1 as a function of $\beta_1 = \beta_2 \in \{1, 2\}$, with $n = 200$ and $n = 600$, $q = 2$, $\sigma_x^2 = 1$, $\sigma_z^2 = 9$, $\beta_0 = 1$, $\alpha = 0.05$.

n	$\beta_1 = \beta_2$	Semi-parametric Poisson	Semi-parametric Negative Binomial
200	1	0.167	0.143
	2	0.273	0.170
600	1	0.403	0.330
	2	0.563	0.387

a region represents an important case study to evaluate how much the evolution of regional institutions has favoured the creation of a system capable of promoting innovative capacity [35]. The number of companies from Emilia–Romagna interviewed in this study is 613. The goal is to investigate the specific role of recent public policies, in enhancing the innovative capacity of companies regarding Industry 4.0 technologies. The response variable is a count variable representing the number of 4.0 technologies, such as those listed below, adopted by each company in the two-year period 2018–2019. The technologies taken into consideration, because they refer to Industry 4.0 innovations, are the following:

- advanced manufacturing solutions,
- additive manufacturing,
- augmented reality,
- simulation,
- horizontal or vertical integration,
- industrial internet,
- cloud computing,
- cyber-security,
- big data/analytics.

The predictors are dichotomous variables and represent policy incentives provided by the government in the mentioned period aimed at encouraging companies to adopt 4.0 technologies. Each independent variable takes 1 if a company has used the corresponding incentive and 0 otherwise. The predictors/incentives are the following:

- Hyper and super depreciation.
- New Sabatini law.
- Guarantee fund.
- R&D tax credit.
- Development contracts.
- Innovative startups and SMEs.
- Patent box.
- Training tax credit.
- Regional incentive measures for R&D and innovation.
- Other.

The classic covariates, or control variables, such as company age and company size (number of employees) were also included in the model. The test presented in Section 3, i.e. the CPT to test the significance of the estimates of the regression coefficients, was applied to the data of the problem at the significance level $\alpha = 0.05$. In the simulation study of Section 4, the semi-parametric test based on Poisson estimators was proved to be the most powerful. Hence, we considered this version of the proposed testing method. The resulting overall p-value, is equal to 0.007. Since it is less than 0.05 we reject the null hypothesis in favour of the alternative hypothesis that at least one regression coefficient is not equal to zero.

One of the advantages of CPTs is that it is a multiple test. Hence, the possible significance of the overall test can be attributed to one or more partial tests by considering the adjusted p-values. The correction of the p-values is necessary to control the familywise error rate and prevent the type I error rate from exceeding α [34,36–38]. The p-values, are adjusted with the Bonferroni–Holm method.

Table 2 shows the estimates of the regression coefficients and the corresponding adjusted p-values. According to this output, there is empirical evidence that the implementation of Industry 4.0 technologies was significantly affected by the following policy incentives: hyper and super depreciation and the New Sabatini law.

6. Conclusions

This work deals with the goodness-of-fit test of a model for count data. Given that the dependent variable takes integer non-negative values, the suitable methodological framework is that of Generalized Linear Models. The maximum likelihood tests of the Poisson and of the Negative Binomial regressions can fail when the real distribution of the dependent variable is neither Poisson nor Negative Binomial and in particular situations such as the case of overdispersed data. The proposed test is semi-parametric, and therefore more robust and less restrictive from the point of view of model assumptions. In particular, it does not require that a specific distribution of the dependent variable is assumed.

Actually, the CPT for the goodness-of-fit of the regression model for count data represents a suitable solution when the Poisson regression and the Negative Binomial regression cannot be applied, in particular in the case of overdispersion. This was proved by the simulation study, where the classic Poisson and Negative Binomial likelihood ratio test revealed their anticonservative behaviour in case of high variability of the response. On the other hand, with or without overdispersion, the proposed semi-parametric tests appeared to be always well approximated. Furthermore, such tests were proved to be powerful, unbiased and consistent under the alternative hypothesis.

Hence, the most important scientific innovation proposed in this manuscript consists of a new performant semi-parametric test suitable for overdispersed data. In fact, it is well known that, when the distribution of the response variable is Poisson or Negative Binomial, the likelihood ratio test of the Poisson and Negative Binomial regression respectively, are the best possible choices in terms of power behaviour. But, when the Poisson and Negative Binomial regressions cannot be applied (as in the case of overdispersion), the proposed semi-parametric approach represents a valid alternative.

The application of the proposed method to the original dataset concerning the sample survey about Industry 4.0 carried out in Italy in 2022, proves the practical utility of the proposal. The results provide empirical evidence in favour of the hypothesis that Italian SMEs' adoption of Industry 4.0 technologies depends on policy incentives. In particular, the public incentives that seemed to be relevant are the Hyper and super depreciation and the New Sabatini law. Indeed, the overvaluation of 250% of investments in the purchase of new capital goods, devices, and technologies functional to the 4.0 transformation of production processes (hyper depreciation), has been an important factor in stimulating companies to adopt 4.0 innovations. Super depreciation supervalues investments in newly purchased or leased capital goods by 130% and, in turn, takes on a decisive role as an incentive to innovation. Finally, the New Sabatini law provides for an interest contribution of 2.75 points ("ordinary" investments) over 5 years and 3.57 points in the case of 4.0 assets ("digital" investments) on financing or leasing aimed at purchasing new capital goods intended for business activities. Hence, it represents a relevant tool to facilitate access to corporate credit.

Future goals include extending to a fully non-parametric version of the proposed semi-parametric methodology. In this way, the approach would be completely robust with respect to the underlying distribution (Poisson, Negative Binomial, ...). Furthermore, from the application point of view, the empirical analysis could be extended by considering the possible effect of the region and the sector of activity.

Table 2
Estimates and adjusted *p*-values of the partial permutation tests on the regression coefficients of the regression model (significant estimates in bold).

	Coefficients	Adjusted *p*-values (Bonferroni–Holm)
Intercept	−2.207	
Age	0.002	1.000
Dimension	0.001	1.000
Hyper and super depreciation	1.347	**0.001**
New Sabatini law	0.708	**0.023**
Guarantee fund	0.340	1.000
R&D tax credit	0.594	0.173
Development contracts	1.351	1.000
Innovative startups and SMEs	0.866	1.000
Patent box	0.619	1.000
Training tax credit	0.716	1.000
Regional incentive measures for R&D and innovation	0.316	1.000
Other	−0.519	1.000

CRediT authorship contribution statement

S. Bonnini: Conceptualization of this study, Methodology, Software, Data processing, Writing. **M. Borghesi:** Conceptualization of this study, Methodology, Software, Data processing, Writing. **M. Giacalone:** Conceptualization of this study, Methodology, Software, Data processing, Writing.

Data availability

Data will be made available on request.

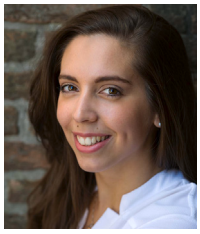
Acknowledgements

Authors thank the University of Ferrara, Italy for funding the project entitled “Public policies, 4.0 technologies and enterprise performance. Empirical analyses on a representative sample of manufacturing enterprises of northern Italy (Politiche pubbliche, tecnologie 4.0 e performance d’impresa. Analisi empiriche su un campione rappresentativo di imprese manifatturiere del Nord Italia)” for the period 2022–2024, with the Departmental Research Incentive Fund - FIRD 2022.

References

[1] Payne EH, Gebregziabher M, Hardin JW, Ramakrishnan V, Egede LE. An empirical approach to determine a threshold for assessing overdispersion in Poisson and negative binomial models for count data. Commun Stat Simul Comput 2018;47(6):1722–38.
[2] Dalenogare LS, Benitez GB, Ayala NF, Frank AG. The expected contribution of Industry 4.0 technologies for industrial performance. Int J Prod Econ 2018;204:383–94.
[3] Kumar S, Raut RD, Aktas E, Narkhede BE, Gedam VV. Barriers to adoption of industry 4.0 and sustainability: a case study with SMEs. Int J Comput Integr Manuf 2023;36(5):657–77.
[4] Cucculelli M, Dileo I, Pini M. Filling the void of family leadership: institutional support to business model changes in the Italian Industry 4.0 experience. J Technol Transf 2022;47:213–41.
[5] Cugno M, Castagnoli R, Büchi G. Openness to Industry 4.0 and performance: The impact of barriers and incentives. Technol Forecast Soc Change 2021;168:120756.
[6] Xia F. Why to use Poisson regression for count data analysis in consumer behavior research. J Mark Anal 2023;11(3):379–84.
[7] Murad NS, Abidi FAA. A comparison between some methods of analysis count data by using R-packages. In: AIP conference proceedings. Vol. 2776, (1). 2023.
[8] Zeileis A, Kleiber C, Jackman S. Regression models for count data in R. J Stat Softw, J Mark Anal 2008;27(8):1–2.
[9] Engel J. The analysis of dependent count data. Wageningen University and Research; 1987.
[10] Silva JS, Tenreyro S. On the existence of the maximum likelihood estimates in Poisson regression. Econom Lett 2010;107(2):310–2.
[11] Kurosawa T, Hui FK, Welsh AH, Shinmura K, Eshima N. On goodness-of-fit measures for Poisson regression models. Aust N Z J Stat 2020;62(3):340–66.
[12] Pesarin F. Multivariate permutation tests with applications in biostatistics. Chichester: Wiley; 2001.
[13] Alibrandi A, Giacalone M, Zirilli A. Psychological stress in nurses assisting Amyotrophic Lateral Sclerosis patients: A statistical analysis based on non-parametric combination test. Mediterr J Clin Psychol 2022;10(2).
[14] Giacalone M, Alibrandi A. A non parametric approach for the study of the controls in the production of agribusiness products. Electron J Appl Stat Anal 2011;4(2):235–44.
[15] Bonnini S. Testing for heterogeneity with categorical data: Permutation solution vs bootstrap method. MatCommun Stat Theory Methodshematics 2014;43(4):906–17.
[16] Bonnini S, Borghesi M, Giacalone M. Simultaneous marginal homogeneity versus directional alternatives for multivariate binary data with application to circular economy assessments. Appl Stoch Models Bus Ind 2023.
[17] Bonnini S, Melak Assegie G. Advances on permutation multivariate analysis of variance for big data. Stat Transit 2022;23(2):163–83.
[18] Giacalone M, Alibrandi A. Overview and main advances in permutation tests for linear regression models. Int J Math Syst Sci 2015;5(2):53–9.
[19] Bonnini S, Borghesi M, Giacalone M. Advances on multisample permutation tests for V-shaped and U-shaped alternatives with application to circular economy. Ann Oper Res 2023;1–16.
[20] Lee OE, Braun TM. Permutation tests for random effects in linear mixed models. Biometrics 2012;25:486–93.
[21] Basso D, Finos L. Exact multivariate permutation tests for fixed effects in mixed models. Commun Stat Theory 2012;41:2991–3001.
[22] Winkler A, Ridgway GR, Webster MA, Smith SM, Nicholas TE. Permutation inference for the general linear model. NeuroImage 2014;92:381–97.
[23] Goeman JJ, Van Houwelingen HC, Finos L. Testing against a high-dimensional alternative in the generalized linear model. Biometrika 2011;98:381–90.
[24] Stute W, Thies S, Zhu L. Model checks for regression: An innovation process approach. Ann Statist 1998;26:1916–34.
[25] Hattab MW, Christensen R. Lack of fit tests based on sums of ordered residuals for linear models. Aust N Z J Stat 2018;60:230–57.
[26] Blagus R, Peterlin J, Stare J. Goodness-of-fit testing in linear regression models. 2019, arXiv, arXiv:911.07522v1.
[27] Solari A, Finos L, Goeman JJ. Rotation-based multiple testing in the multivariate linear model. Biometrics 2014;70:954–61.
[28] Mc Cullagh P, Nelder JA. Generalized linear models. New York: Chapman and Hall/CRC Press; 1989.
[29] Wedderburn RWM. Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. Biometrika 1974;61(3):439–47.
[30] Winkelmann R. Econometric analysis of count data. Springer; 2008.
[31] Lindén A, Mäntyniemi S. Sing the negative binomial distribution to model overdispersion in ecological count data. U Ecol 2011;92(7):1414–21.
[32] Bonnini S, Borghesi M. Relationship between mental health and socio-economic, demographic and environmental factors in the COVID-19 lockdown period-A multivariate regression analysis. Mathematics 2022;10(18):3237.
[33] Bonnini S, Corain L, Marozzi M, Salmaso L. Nonparametric hypothesis testing. In: Rank and permutation methods with applications in R. Chichester: Wiley; 2014.
[34] Pesarin F, Salmaso L. Permutation tests for complex data: theory, applications and software. Chichester: Wiley; 2010.
[35] Mosconi F, D’Ingiullo D. Institutional quality and innovation: evidence from Emilia–Romagna. Econ Innov New Technol 2023;32(2):165–97.
[36] Giacalone M, Agata Z, Cozzucoli PC, Alibrandi A. Bonferroni-Holm and permutation tests to compare health data: methodological and applicative issues. BMC Med Res Methodol 2018;18(1):1–9.
[37] Westfall PH, Young SS. Resampling-based multiple testing: examples and methods for p-value adjustment. New York, NY, USA: Wiley-Interscience; 1992.
[38] Westfall PH, Young SS. On adjusting P-values for multiplicity. Biometrics 1992;49:941–5.

Stefano Bonnini is Associate Professor in "Statistics". He received the Ph.D. in "Statistics applied to Economic and Social Sciences" from the University of Padua. His main research interests concern: Permutation tests, Nonparametric inference, Statistics applied to Economics and Business, Multivariate Analysis and Biostatistics. He is a member of the Italian Statistical Society. He wrote more than 100 papers published in international scientific journals, some books about methodological statistics, published by prestigious international publishers and took part to numerous scientific international conferences.



Michela Borghesi is currently a research fellow in Statistics at the University of Ferrara, Department of Economics and Management. She received the Ph.D. in "Economics and Management of Innovation and Sustainability" from the University of Ferrara. Her main research interests concern: nonparametric statistics, multivariate analysis, complex test of hypotheses. She also deals with statistics applied to economics and business, social sciences, health sciences, engineering and bio-sciences. She is a member of the Italian Statistical Society and a member of the Scientific Committee of the Center for Modelling Computing and Statistics (CMCS) of the University of Ferrara. She is author of some papers published in international scientific journals and took part as a speaker to some scientific conferences.

Massimiliano Giacalone is currently qualified for the functions of Associate Professor in "Statistics" and in "Economic Statistics". Actually, he is a researcher and teaching staff member of the Department of Economics, University of Campania "Luigi Vanvitelli". He received the Ph.D. in "Computational Statistics and Applications" from the University of Naples "Federico II". He was "visiting professor" at University Politecnica de Catalunja, University of Bucarest and Albanian University. His research area encompasses the following topics: Normp nonlinear regression, Non-Parametric Combination tests; Clustering time series analysis, Skewed Generalized Error Distribution, Forecasting and Applications of Statistics in Economics. He is member of "Italian Statistical Society", "International Association for Statistical Computing" and of the "Royal Statistical Society". He has been Adjunct Professor of "Statistics", "Probability", "Economic Statistics", at various Italian Universities. He is author of over one hundred scientific papers concerning Methodological and Applied Statistics.