scientific reports

OPEN

Machine learning and statistical models for analyzing multilevel patent data

Sunyun Qi[1], Yu Zhang[2], Hua Gu[1], Fei Zhu[1], Meiying Gao[1], Hongxiao Liang[3], Qifeng Zhang[1]✉ & Yanchao Gao[1]✉

A recent surge of patent applications among public hospitals in China has aroused significant research interest. A country's healthcare innovation capacity can be measured by its number of patents. This paper explores the link between the number of patents and ten independent variables. Multicollinearity was carefully detected and removed by using the variable selection method and LASSO regression, respectively. The Poisson model and the negative binomial model were proposed to analyze the patent data. Three goodness of fit tests, the Pearson test, the deviance test, and the DHARMa non-parametric dispersion test, were conducted to investigate if the model has a good fit. After discovering four clusters by conducting agglomerative hierarchical clustering, these two models were replaced by the negative binomial mixed model. The likelihood ratio test was used to determine which model is more appropriate and the results reveal that the negative binomial mixed model outperforms both the Poisson model and the negative binomial model. Three variables, number of health technicians per 10,000 people, financial expenditure on science and technology as well as number of patent applications per 10,000 health personnel, have a significantly positive relationship with the number of patents in Chinese tertiary public hospitals.

Over the past ten years, the number of patents application in China has been skyrocketing. China has surpassed all other countries worldwide in terms of patent application filings since 2011[1]. The number of applications for invention patents climbed from 526,412 in 2011 to 1,586,000 in 2021, with an average annual growth rate of 11.7%, according to the National Bureau of Statistics of China. Healthcare patents, one of the most important areas of the patent, provide important protections for intellectual property in the medical arena, which can further innovations that benefit everyone. Medical patents are defined broadly to include patents that relate to pharmaceuticals; methods of making and using them; medical treatment regimens; surgical procedures; medical devices; health care information technology for hospital and health care management systems; and combinations of them[2].

Medical patents holders are mainly tertiary public hospitals in China. In China, hospitals are organized in a three-tiered hierarchy, with primary hospitals providing general healthcare and preventive care to the population. Secondary hospitals provide complete health care to a region, accept referrals from primary hospitals, and are also responsible for teaching and research. Lastly, tertiary hospitals, often located in urban areas, are responsible for specialty care and act as medical centres for several regions[3–6]. This system is motivated by the expectation that tertiary hospitals can focus more on research and lead Chinese medical innovation to a higher level. As for the evaluation, patents are vital to measure hospitals' achievement. The number of applied patents can be related to many factors. For example, based on the scale of a region, the investment or the population size can be considered.

Following the acquisition of data on the number of patents, a linear regression model will be considered first to fit. However, we would only consider the linear regression process, the least absolute shrinkage and selection operator (LASSO) regression model, in the step of variable selection[7]. The number of patents is a counting value over a fixed period of time. In other words, we are counting the occurrences of the event that a patent application is submitted during a certain time interval. We are assuming that the event happens completely randomly and

[1]Zhejiang Provincial Center for Medical Science Technology and Education Development, Hangzhou 310000, Zhejiang, China. [2]Leuven Statistics Research Centre, Faculty of Science, KU Leuven (Katholieke Universiteit Leuven), 3001 Heverlee, Belgium. [3]Department of Public Utilities Management, Faculty of Humanities and Management, Zhejiang Chinese Medical University, Hangzhou 310000, Zhejiang, China. ✉email: kjzx@zjwjw.gov.cn; kjzxgyc@163.com

independently. Thus, the number of patents no longer follows a normal distribution, and hence a simple linear regression model will be dismissed under our assumption.

To overcome the problem of the non-normal distribution of dependent variables, generalized linear models like the Poisson regression model will be introduced[8]. A count variable is a variable that reflects the number of occurrences of an interested event in a fixed period of time[8]. Linear regression model is not appropriate for a count variable as a dependent variable and problems like biased standard errors will occur[9]. Poisson regression model offers an alternative analysis for count data. It can also be used for summarizing relative risk across strata of a covariate and for evaluating interactions between covariates[10]. The Poisson regression model is an example of generalized linear models (GLM). There are three components in a GLM: a random component, a systematic component and a link function. Random component is the probability distribution of the dependent variable, for example, Poisson distribution for response variable in the Poisson regression. Systematic component refers to the independent variables as a combination of linear predictors and link function specifies the link between random and systematic components[11–13]. Poisson regression with overdispersion can be replaced with the negative binomial model if the assumption must be met for the model to be valid[14]. Besides, this paper also proposed a mixed model, which is a combination of regression with clustering[15–17]. The summary of results is provided in the last part.

Materials and methods

Data description.

Patent number data from the effective patents for tertiary public hospitals in China was used as the dependent variable in this paper. The Baiten database (www.baiten.cn) was thoroughly searched for patent numbers in tertiary public hospitals. Since a time lag exists between the application time and publication time in patent authorization period[18], we select the year of application to conduct our count procedure. A total number of 165,262 patent was collected in 2243 tertiary public hospitals from year 2016 to year 2021. The independent variables, population, GDP, number of health technicians per 10,000 people, R &D, number of health personnel, financial expenditure on education, financial expenditure on science and technology, financial health care expenditure, health industry income per capita and number of patent applications per 10,000 health personnel, were obtained from the National Bureau of Statistics (NBS) website.

Summary statistics regarding variables in our data are shown in Table 1. Table 2 and Fig. 1show the number of patents from the year 2016 to the year 2021. The sharp increase in the number of patents in 2018 was due to the regulation announcement that tertiary public hospitals began assessing patents in 2018. The reason why only a half number of patents in 2021 compared with the year 2020 is that China carried out intellectual property quality improvement projects.

Abbreviations	Description	Mean	SD	Min.	Max.
NumPat	Number of patents	888.51	1446.50	0.00	9,951.00
Pop	Population	4526.72	2975.69	340.00	12,684.00
GDP	Gross domestic product	30,253.57	24,993.020	124,369.70	1173.00
NumHeaTec	Number of health technicians per 10,000 people	71.05	12.98	45.00	126.00
R &D	Research and development	99,422.66	144,492.69	190.00	700,017.00
NumHeaPer	Number of health personnel	40.37	25.34	2.92	102.79
FinEdu	Financial expenditure on education	1000.06	616.57	152.57	3,510.56
FinSci	Financial expenditure on science and technology	166.32	201.56	4.81	1168.79
FinHea	Financial health care expenditure	510.84	300.43	69.97	1772.99
HeaInc	Health industry income per capita	99,681.45	31,336.74	51,135.00	208,481.00
NumPatHeaPer	Number of patent applications per 10,000 health personnel	17.99	22.05	0.00	120.87
Pro	Province	categorical	categorical	categorical	categorical

Table 1. Variable description.

Year	Number of patents
2016	3325
2017	6613
2018	14,509
2019	39,131
2020	63,737
2021	37,947

Table 2. The number of patents from 2016 to 2021.

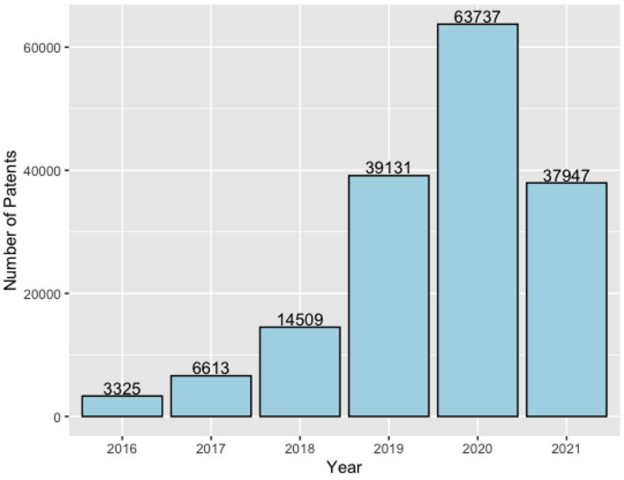

Figure 1. Bar plot of the number of patents every year between 2016 and 2021.

Variable selection.

Multicollinearity appears when independent variables (NumPat, Pop, GDP, NumHeaTec, R &D, NumHeaPer, FinEdu, FinSci, FinHea, HeaInc and NumPatHeaPer) in the regression model are highly correlated to each other. It generates high variance of the estimated coefficients and hence, the coefficient estimates corresponding to those interrelated explanatory variables will lead to wrong results. To detect multicollinearity, the correlation matrix is plotted. Variable selection and LASSO regression can be utilized to eliminate the problem of multicollinearity.

Variable selection via variance inflation factor. One of the two methods we applied here is variable selection method using variance inflation factor (VIF).

VIF is used to measure the multicollinearity among the our ten independent variables. When there is high correlation among the predictor variables, the standard errors of predictors coefficients will increase, then the variance of their coefficients will be inflated.

The VIF on NumPat, Pop, GDP, NumHeaTec, R &D, NumHeaPer, FinEdu, FinSci, FinHea, HeaInc, NumPatHeaPer is defined as

$$VIF_j = \frac{1}{1 - R_j^2} \tag{1}$$

where R_j is the coefficient determination for the regression of x_j on the remaining variables[19]. For example, to calculate R^2_{NumPat}, we regress *NumPat* against all other independent variables and then we can obtain R^2_{NumPat}, thus, VIF of *NumPat* can be derived. Normally, for VIF, a value of 10 and above indicates multicollinearity[20].

The variable selection procedure is done by removing the variable that has the highest VIF value. Then correlation matrix and VIF were calculated again to remove next variable that has the highest VIF value. The procedure was stopped when there is no VIF value larger than 10[21].

LASSO regression. Though variable selection via VIF can perform feature selection and make parsimonious models, with advancements in machine learning, Least Absolute Shrinkage and Selection Operator, known as LASSO, provides a good alternative as it gives much better output, requiring fewer tuning parameters and being automated to a large extend[22].

LASSO was proposed by Tibshirani (1996) for subset selection based in regression process. It puts a constraint on the sum of the absolute values of the model parameters. In other words, it apply a shrinking or regularization process where it penalizes the coefficients of the regression variables, shrinking some of them to zero.

LASSO solves the l_1-penalized regression problem of finding $\beta = \beta_j$ to minimize

$$\sum_{i=1}^{N}\left(y_i - \sum_j x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p}|\beta_j| \tag{2}$$

where x_{ij} are the standardized NumPat, Pop, GDP, NumHeaTec, R &D, NumHeaPer, FinEdu, FinSci, FinHea, HeaInc, NumPatHeaPer and y_i is the NumPat. The method of K-fold cross-validation is performed to tuning parameter values and we used ten-fold. This procedure begins with splitting the data into "folds." Then the prediction performance of each model is evaluated across the "left-out" fold using a sequence of tuning parameter settings. This procedure is done until every fold has been calculated as test data. Typically, the tuning parameter is set to the sequence value with the smallest cross-validation error[23].

Methods. The Poisson regression model was utilized to analyze the relationship between count data and the independent variables. In Poisson regression, the response variable is assumed to follow a Poisson distribution $NumPat \sim Poisson(\lambda_i)$, and hence the regression formula is defined as

$$\log(\lambda_i) = \beta_0 + \boldsymbol{\beta}'\mathbf{x} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{3}$$

or written as

$$\lambda_i = E(y_i \mid \mathbf{x}) = e^{\beta_0 + \boldsymbol{\beta}'\mathbf{x}} = e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}} = e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}} \tag{4}$$

where λ_i is the expected value or mean of the NumPat, $\mathbf{x}$ is a matrix of the NumPat, Pop, GDP, NumHeaTec, R &D, NumHeaPer, FinEdu, FinSci, FinHea, HeaInc, NumPatHeaPer and β_0, and $\boldsymbol{\beta}'$ is the set of regression coefficients to be estimated.

Different from linear regression models, the Poisson regression model uses Maximum Likelihood Estimation (MLE) method instead of OLS estimation. Since we have assumed that $NumPat \sim Poisson(\lambda_i)$, the probability density function (pdf) of y_i (NumPat) is

$$f(y_i \mid \boldsymbol{x}; \lambda_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}, y_i = 0, 1, 2, \ldots \tag{5}$$

then the likelihood function takes the form

$$L(\boldsymbol{\beta} \mid \boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{N} \frac{e^{y_i \boldsymbol{x}_i^T \boldsymbol{\beta}} e^{-e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}}{y_i!} \tag{6}$$

given N vectors $\boldsymbol{x}_i \in \mathbb{R}^{p+1}$ with $i = 1, \ldots, N$ along with a set of N values $y_1, \ldots, y_N \in \mathbb{N}$. Hence the log-likelihood is derived as

$$\ell(\boldsymbol{\beta} \mid \boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{N} (y_i \boldsymbol{x}_i^T \boldsymbol{\beta} - e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} - log(y_i!)) \sim \sum_{i=1}^{N} (y_i \boldsymbol{x}_i^T \boldsymbol{\beta} - e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}) \tag{7}$$

and finally some calculation methods would be taken to maximize the log-likelihood, selecting the best values of $\boldsymbol{\beta}$.

The interpretation of $\boldsymbol{\beta}'$ is that, in particular, given one unit change of the independent variable, the difference in the natural logarithm of expected counts is expected to change by the respective regression coefficient, with other independent variables in the model constant.

We should recognize that models can only approximate complete information or reality. Thus, we tried to find the model that minimize the loss of information. A useful criterion in model selection named Akaike's information criterion (AIC) is defined as

$$\mathrm{AIC} = -2[\,\mathrm{L}(\beta)] + 2\mathrm{k} \tag{8}$$

where $\mathrm{L}(\boldsymbol{\beta})$ is the log-likelihood function of the candidate model evaluated under $\boldsymbol{\beta}$ by using observations and k is the number of unknown parameters. Then, the resulting model is subjected to the log-likelihood value where $\mathrm{L}(\boldsymbol{\beta})$ is log-likelihood at convergence.

A basic assumption of the Poisson distribution is that, for the count data, the mean equals the variance. However, this assumption is not always satisfied. Real data often does not show this specific pattern and the variance tends to be larger than the mean for most of the data. This is known as overdispersion problem in statistics[24].

We have discussed that it is unrealistic to assume $Var(y_i) = E(y_i)$ for most cases, hence a more flexible relationship between the variance and the mean should be considered[25]. As a result, the negative binomial model is proposed.

In the negative binomial model, the number of patents y_i was assumed to follow the negative binomial distribution, that is, $y_i \sim NB\left(y_i \mid \lambda_i, \theta\right)$ with $\lambda_i \sim Gamma(\theta, \theta e^{-\boldsymbol{x}_i^T \boldsymbol{\beta}})$, whose pdf is

$$f(y_i \mid \theta) = \frac{\Gamma\left(y_i + \theta\right)}{\Gamma(\theta) y_i!} \cdot \left(\frac{\theta}{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} + \theta}\right)^{\theta} \cdot \left(\frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} + \theta}\right)^{y_i} \tag{9}$$

where λ_i is the mean and θ is the dispersion parameter that controls the amount of overdispersion. Thus, the variance and the mean of $y_i$can be derived as

$$E(y_i) = \mu = E[E(y_i \mid \lambda_i)] = E(\lambda_i) = e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} \tag{10}$$

$$Var(y_i) = E[Var(y_i \mid \lambda_i)] + Var(E[y_i \mid \lambda_i]) = E(\lambda_i) + Var(\lambda_i) = e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} + \frac{1}{\theta} e^{2\boldsymbol{x}_i^T \boldsymbol{\beta}} = \mu + \frac{1}{\theta}\mu^2 \tag{11}$$

As the above formula shows, the variance and the mean are not the same thing in the negative binomial regression assumption. This is not the case with the Poisson regression. Instead, the model assumes a quadratic relationship between the mean and the variance. The negative binomial model can be used for overdispersed count data and it can be considered as a generalization of Poisson regression since it has the same mean structure as Poisson regression and it has an extra parameter to model the overdispersion.

The MLE of parameters in the negative binomial model is straightforward[25]. Lawless has discussed about the efficiency and robustness properties of the inference procedure[26].

Hierarchical clustering, one of the most popular unsupervised learning methods, was conducted to find if there exists correlation in the data. In other words, within-cluster observations have high similarity in comparison to others. There are two approaches, agglomerative hierarchical clustering and divisive hierarchical clustering. Agglomerative hierarchical clustering is a bottom up approach that each observation starts from its own cluster, and then clusters are grouped as observations move up the hierarchy. Divisive hierarchical cluserting is a top down approach that observations start from one cluster, and then split as observations moving down the hierarchy. However, the complexity of divisive clustering is $\boldsymbol{O}(2^n)$, which makes it too slow for large data sets. Moreover, no provision can be made for a relocation of observations that may have been 'incorrectly' grouped at an early stage in divisive hierarchical cluserting[27]. Therefore, the agglomerative clustering was used to group data into clusters based on their similarity in this paper.

Euclidean distance matrix was used for hierarchical clustering and the linkage criterion determines the distance between sets of observations as a function of the pairwise distances between observations. Ward's method is based on the objective of minimizing the deterioration in the overall within sum of squares. The latter is the sum of squared deviations between the data in a cluster and the centroid:

$$WSS = \sum_{i \in C} (x_i - \bar{x}_C)^2 \tag{12}$$

with $\bar{x}_i$ is a data point and$\bar{x}_C$ as the centroid of cluster C. Since any merger of two existing clusters results in an overall WSS decline, Ward's method is intended to minimize this decline. In other words, it is utilized to minimize the difference between the new WSS in the merged cluster and the total WSS of the merged components.

Dendrogram is widely used in visualizing a clustering hierarchy and it is simple to interpret similarity and clustering. The horizontal axis of the dendrogram indicates the dissimilarity or distance between clusters, while the vertical axis represents the objects and clusters.

The Poisson mixed effects model can be an appropriate choice for clustered patent count data. However, it still suffers from the overdispersion problem. Therefore, the negative binomial mixed model was proposed based on the negative binomial model, to analyze the clustered count data, where the observations are no longer considered as independent with each other but are correlated on the counts. Observations are divided into several clusters. These clusters are also called "subjects" and observations in each cluster are seen as repeated measurements over a period of time, hence they have correlation, while observations from different clusters or subjects are independent. To be more specific, y_{ij} is the value of the patent count variable for i_{th} subject at j_{th} time point.

Compared with the negative binomial model, the only change in the negative binomial mixed model is that, in order to take the influence of within-cluster correlation into account, we add subject-specific random effects into the linear predictor. The mean of i_{th} subject μ_i is related to the host variables via the logarithm link function written in matrix notation:

$$\log(\mu_i) = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b} + \epsilon_i, \boldsymbol{b} \sim \boldsymbol{N}(0, \boldsymbol{\Psi}) \tag{13}$$

where $\boldsymbol{\beta}$ is the vector of fixed effects for the independent variables (NumPat, Pop, GDP, NumHeaTec, R &D, NumHeaPer, FinEdu, FinSci, FinHea, HeaInc, NumPatHeaPer)X_i, cluster$\boldsymbol{b}$ is the vector of random effects and ϵ_i is the random errors. The random effects are utilized to partition the multiple sources of variation, and thus to avoid biased inference on the effects of the NumPat, Pop, GDP, NumHeaTec, R &D, NumHeaPer, FinEdu, FinSci, FinHea, HeaInc and NumPatHeaPer. The vector of the random effects is usually assumed to follow the multivariate normal distribution and $\boldsymbol{\Psi}$ is a positive-definite variance-covariance matrix that determines the form and complexity of random effects b.

Results

Variable selection.

The correlation matrix plot was plotted to detect if multicollinearity exists. According to the heat map in Fig. 2, obviously there is severe collinearity among the predictor variables. Thus, before fitting models, some measures may need to be taken to avoid this problem, otherwise we might have irrelevant variables in our model, influencing the significance of coefficients.

A more formal way to test multicollinearity is the Variance Inflation Factor. The multicollinearity arises when VIF is larger than 10 and the detailed VIF results are shown in Table 3. A variable selection procedure was conducted to get rid of the high multicollinearity problem. It is done by removing the variable that has the highest VIF value, which is variable Population. Then the correlation matrix and VIF were calculated again to remove the next variable that has the highest VIF value. The procedure was stopped when there was no VIF value larger than 10. At the end of our procedure, the number of health technicians per 10,000 people, R &D, number of health personnel, financial expenditure on science and technology, health industry income per capita and number of patents applications per 10,000 people can be used with a further regression model, and the VIF values of these variables are shown in Table 4.

In LASSO regression, 5-fold cross-validation is performed to find an optimal lambda value, which equals 0.01. After using this optimal lambda value for LASSO regression, we have the coefficients of population, GDP, financial expenditure on education and Financial health care expenditure constrained to 0, leaving number of health technicians per 10,000 people, R &D, number of health personnel, financial expenditure on science and

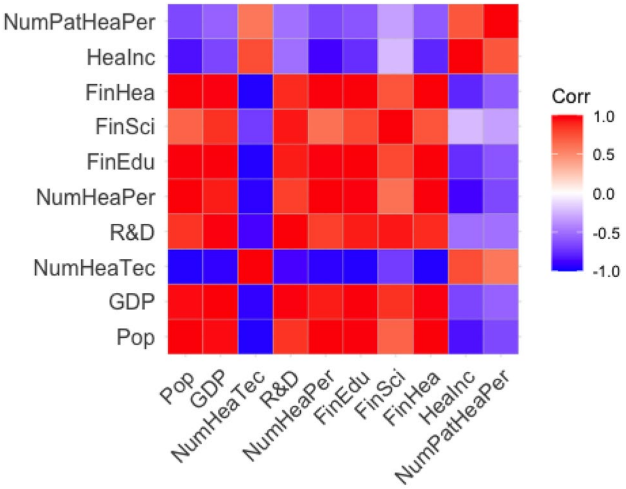

Figure 2. Correlation matrix heatmap for measuring of dispersion.

Variable	VIF	Detection
Pop	107.34	Collinearity
GDP	28.96	Collinearity
NumHeaTec	4.33	No collinearity
R &D	13.58	Collinearity
NumHeaPer	94.16	Collinearity
FinEdu	43.29	Collinearity
FinSci	12.85	Collinearity
FinHea	32.37	Collinearity
HeaInc	3.76	No collinearity
NumPatHeaPer	2.63	No collinearity

Table 3. VIF with all variables.

Variable	VIF
NumHeaTec	1.78
R &D	5.58
NumHeaPer	2.55
FinSci	5.98
HeaInc	3.40
NumPatHeaPer	1.98

Table 4. VIF with selected variables.

technology, health industry income per capita and number of patents applications per 10,000 health personnel. It is the same result when we use the variable selection method via the VIF.

Poisson regression model and negative binomial model. After backward selection based on the AIC used, the number of health personnel was dropped and we used left variables to generate our Poisson regression model, negative binomial model and negative binomial mixed model. Goodness of fit tests are performed to check if the model is correct given the data. First, the Pearson and deviance goodness of fit tests were utilized to test the goodness of fit of the Poisson regression model. The p-values of both these two tests are close to 0. Thus, we don't have enough evidence to show our model is good. Moreover, Kolmogorov-Smirnov test were used to test the goodness of fitness. The p-value of KS test is close to 0 and the residuals of the Poisson model didn't follow a uniform distribution.

DHARMa non-parametric dispersion has a very nice test for dispersion and the result showed that the observed value was much larger than what we could expect under the model. Thus, the Poisson regression model suffers from overdispersion.

In order to deal with the overdispersion problem, a negative binomial model was used. As we can see from the Fig. 3, the *p*-value of the Kolmogorov-Smirnov test showed the residuals of the negative binomial model follow the uniform distribution and the DHARMa nonparametric dispersion test showed the overdispersion problem no longer exists.

Hierarchical clustering. To detect if there are correlations or clusters among the 31 provinces in China, the data of the year 2021 was selected to conduct the hierarchical clustering method. After plotting the dendrogram in Fig. 4, four clusters can be summarized from the dendrogram plot. After doing agglomerative hierarchical clustering among the 31 provinces, and the detailed provinces in the four clusters are shown in Table 5.

Negative binomial mixed model. Taking the clustering and the overdispersion problem into account, a negative binomial mixed model was conducted in this paper. The results of the coefficient estimator and *p*-values of Poisson model, negative binomial model (abbreviated as NBM) and negative binomial mixed model's fix effect (abbreviated as NBMM FE) are shown in Table 6.

A likelihood ratio test was used to determine if the negative binomial model is more appropriate statistically than the Poisson model, and the results <0.001 suggest that the negative binomial model is a better fit. The same conclusion can be drawn from the Pearson goodness of fit test (*p*-value = 0.98), deviance goodness of fit test (*p*-value = 0.05) and the DHARMa non-parametric dispersion test (*p*-value = 0.32) in the negative binomial model. The likelihood ratio test between the negative binomial model and the negative binomial mixed model suggests that the multilevel model provides a better fit. The variance of the random effects equals 0.2908, which

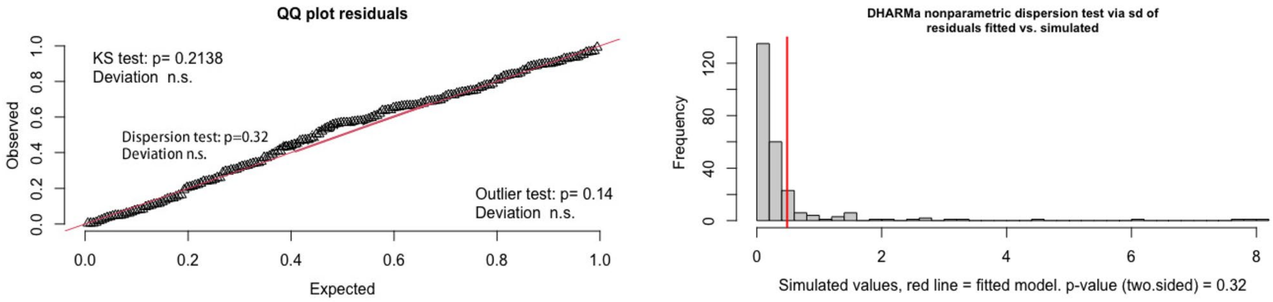

Figure 3. Goodness of fit for negative binomial model. Left: the Kolmogorov-Smirnov test indicates that the residuals of the negative binomial model follow a uniform distribution; Right:the DHARMa nonparametric dispersion test shows that the overdispersion problem no longer exists in the negative binomial model.

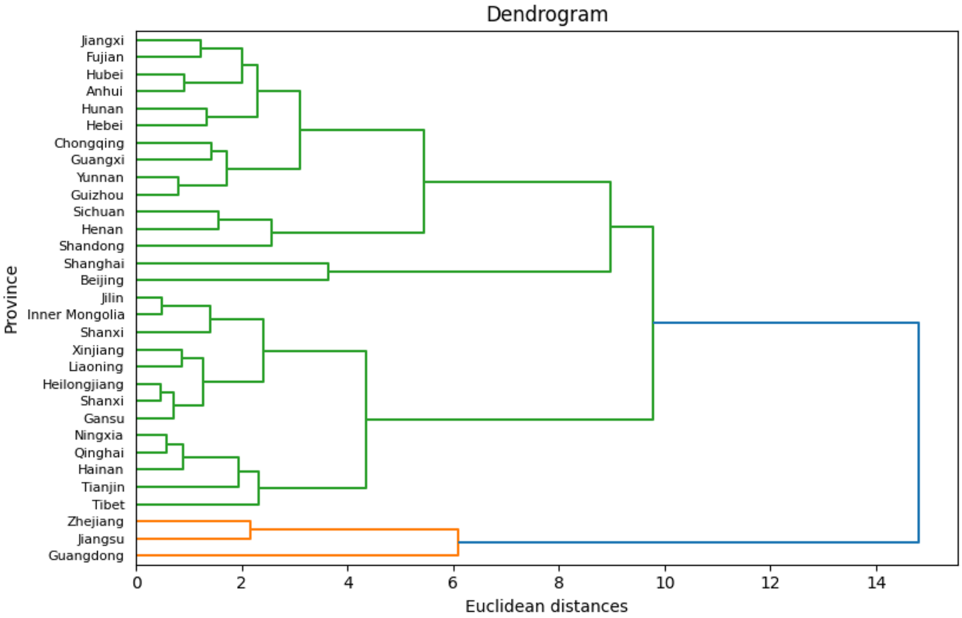

Figure 4. Dendrogram of a hierarchical clustering using agglomerative clustering.

Cluster	Province
Cluster 1	Jiangsu, Zhejiang, Guangdong
Cluster 2	Tianjin, Shanxi, Neimenggu, Liaoning, Jilin, Heilongjiang, Hainan, Xizang, Shanxi, Gansu, Qinghai, Ningxia, Xinjiang
Cluster 3	Beijing, Shanghai
Cluster 4	Hebei, Anhui, Fujian, Jiangxi, Shandong, Henan, Hubei, Hunan, Guangxi, Chongqing, Sichuan, Guizhou, Yunnan

Table 5. Clustering results.

Variable	Poisson		NBM		NBMM FE	
	ES (95% CI)	*p*-value	ES (95% CI)	*p*-value	ES (95% CI)	*p*-value
NumHeaTec	0.20 (0.20,0.21)	< 0.001	0.35 (0.16,0.55)	< 0.001	0.72 (0.54,0.92)	< 0.001
R &D	0.06 (0.05,0.06)	< 0.001	0.39 (0.06,0.72)	< 0.01	0.34 (− 0.08,0.76)	0.11
FinSci	0.34 (0.33,0.35)	< 0.001	0.20 (− 0.14,0.56)	0.2	0.36 (0.03,0.70)	< 0.05
HeaInc	− 0.32 (− 0.33, − 0.31)	< 0.001	− 0.41 (− 0.64,-0.17)	< 0.001	− 0.38 (− 0.57,-0.18)	< 0.001
NumPatHeaPer	0.64 (0.64,0.65)	< 0.001	1.31 (1.09,1.55)	< 0.001	1.12 (0.92,1.32)	< 0.001

Table 6. Model Estimated Cficients.

is not too close to 0 and thus, we cannot assume all provinces are independent of each other and the clustering needs to be considered in this data.

The AIC of three models, Poisson model, negative binomial model and negative binomial mixed model, are 71270, 2588.1 and 2533.6, respectively. It is also proved that negative binomial mixed model outperformed than these two models in our data.

Discussion

Regarding to the count variable, number of patents, Poisson regression model is a commonly used analysis. The Poisson regression is performed based on the assumption that the mean and the variance of the dependent variable are the same. Overdispersion appears when there is more variability around the variance than the mean[28]. Therefore, the negative binomial model is better than the Poisson model when the data shows evidence of overdisperson[29,30]. After conducting the agglomerative hierarchical clustering, four clusters among the 31 provinces were demonstrated in Table 5 and Fig. 4. Therefore, we have solid evidence to take the clustering into account and use multilevel model in this data. The R &Ds of Jiangsu, Zhejiang and Guangdong provinces are almost 10 times larger than other provinces which is the primary reason that these three provinces are grouped. It means that Jiangsu, Zhejiang and Guangdong have a relatively high economy and undertake to innovate and introduce new products and services. Beijing and Shanghai are the two major and most well-known provinces in China. They have the similar economic status, policies and wealth structures. According to our data, all of variables are similar between Beijing and Shanghai and it is reasonable to group them together. Another two clusters can be summarized by saying that they are separated by region—most of the provinces in cluster 2 are from northern China and most of the provinces in cluster 4 are from southern China.

Considering both the clustering and the overdispersion problem, a negative binomial mixed model was proposed. The comparison among estimated coefficients with confidence intervals and p-values of the Poisson model, negative binomial model and negative binomial mixed model's fix effect are shown in Table 6. From the results of negative binomial model, three variables, number of health technicians per 10,000 people, financial expenditure on science and technology as well as number of patents applications per 10,000 health personnel, have significantly positive relationship with the number of patents in Chinese tertiary public hospitals. To be specific, holding all other variables constant in the negative binomial mixed model , by increasing the number of patents applications per 10,000 health personnel by 1 unit, the number of patents will increase more than two times . The variable, health industry income per capita, has a significantly negative relationship with the patent number and when it increases by 1 unit, the number of patents decreases by 32%. Moreover, the likelihood ratio test and AIC are used to compare two models based on the ratio of the likelihoods[31,32]. The results of likelihood ratio test revealed that negative binomial mixed model outperformed both the Poisson model and the negative binomial model in the data. The variance of the random effects in the negative binomial model is not close to 0 and it also proved that there were associations among the provinces.

The reason why R &D is not significant in the negative binomial mixed model is that we add the province as a random effect in this model. This result is consistent with the finding that R &D is very imprecisely estimated when comparing France and the USA[33]. It may be because the R &D is a significant variable when we group the provinces, but after we clustered, each group had close R &D and thus, the p-value of R &D is not significant after we grouped.

We suggest that each province should adjust the proportion of financial investment and health technicians in different regions according to the local conditions. For example, relatively wealthy provinces can devote more resources to increase the number of health technicians and the protection and supervision of the patent exchange market, while less wealthy provinces should invest more finance in medical science and technology to increase their research and development ability.

The literature on comparison in terms of the traditional variable selection method and the LASSO regression remains scarce. In our study, we compared these methods in the selection of the feature. Moreover, to the best of our knowledge, this paper is the first work to use hierarchical clustering before conducting the multilevel model. One potential limitation of our negative binomial mixed model is that it is not designed to explicitly detect the association between the number of patents and one specific independent variable. The variance function versus the observed variance in our negative binomial mixed model needs to be further investigated and if the variance is wrong, the random effects need to be fixed since it models the variance and covariance structure.

Data availability

The datasets generated during and analysed during the current study are available from the corresponding author on reasonable request.

Received: 18 January 2023; Accepted: 29 June 2023
Published online: 07 August 2023

References

1. Chen, Z. & Zhang, J. Types of patents and driving forces behind the patent growth in China. *Econ. Modell.* **80**, 294–302 (2019).
2. Mayfield, D. L. Medical patents and how new instruments or medications might be patented. *Missouri Med.* **113**(6), 456 (2016).
3. Feng, X. L. *et al.* Extending access to essential services against constraints: The three-tier health service delivery system in rural china (1949–1980). *Int. J. Equity Heal.* **16**, 1–18 (2017).
4. Tu, J., Wang, C. & Wu, S. The internet hospital: An emerging innovation in china. *Lancet Glob. Health* **3**, e445–e446 (2015).
5. YE, T. *et al.* A reflection on the continuity of the three-tier healthcare network in rural china. *Chin. J. Hosp. Adm.* 184–187 (2011).
6. Yip, W.C.-M., Hsiao, W., Meng, Q., Chen, W. & Sun, X. Realignment of incentives for health-care providers in china. *The Lancet* **375**, 1120–1130 (2010).
7. Tibshirani, R. Regression shrinkage and selection via the lasso: A retrospective. *J. R. Stat. Soc. Ser. B* **73**, 273–282 (2011).
8. Coxe, S., West, S. G. & Aiken, L. S. The analysis of count data: A gentle introduction to poisson regression and its alternatives. *J. Pers. Assess.* **91**, 121–136 (2009).
9. Gardner, W., Mulvey, E. P. & Shaw, E. C. Regression analyses of counts and rates: Poisson, overdispersed poisson, and negative binomial models. *Psychol. Bull.* **118**, 392 (1995).
10. Frome, E. L. & Checkoway, H. Use of poisson regression models in estimating incidence rates and ratios. *Am. J. Epidemiol.* **121**, 309–323 (1985).
11. Dobson, A. J. & Barnett, A. G. *An Introduction to Generalized Linear Models* (Chapman and Hall/CRC, 2018).
12. Faraway, J. J. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models* (Chapman and Hall/CRC, 2016).
13. Jiang, J. & Nguyen, T. *Linear and Generalized Linear Mixed Models and Their Applications* Vol. 1 (Springer, 2017).
14. Hinde, J. & Demétrio, C. G. Overdispersion: Models and estimation. *Comput. Stat. Data Anal.* **27**, 151–170 (1998).
15. Yirga, A. A., Melesse, S. F., Mwambi, H. G. & Ayele, D. G. Negative binomial mixed models for analyzing longitudinal cd4 count data. *Sci. Rep.* **10**, 1–15 (2020).
16. Zhang, X. *et al.* Negative binomial mixed models for analyzing longitudinal microbiome data. *Front. Microbiol.* **9**, 1683 (2018).
17. Zhang, X. *et al.* Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinform.* **18**, 1–10 (2017).
18. Dang, J. & Motohashi, K. Patent statistics: A good indicator for innovation in china? patent subsidy program impacts on patent quality. *China Econ. Rev.* **35**, 137–155 (2015).
19. Chan, J.Y.-L. *et al.* Mitigating the multicollinearity problem and its machine learning approach: A review. *Mathematics* **10**, 1283 (2022).
20. Daoud, J. I. Multicollinearity and regression analysis. *J. Phys: Conf. Ser.* **949**, 012009 (2017).
21. Prahutama, A., Ispriyanti, D. & Warsito, B. Modelling generalized poisson regression in the number of dengue hemorrhagic fever (DHF) in east nusa tenggara. In *E3S Web of Conferences*, vol. 202, 12017 (EDP Sciences, 2020).
22. Muthukrishnan, R. & Rohini, R. Lasso: A feature selection technique in predictive modeling for machine learning. In *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, 18–20 (2016).
23. Roberts, S. & Nowak, G. Stabilizing the lasso against cross-validation variability. *Comput. Stat. Data Anal.* **70**, 198–211 (2014).
24. Moksony, F. & Hegedűs, R. The use of Poisson regression in the sociological study of suicide. *Corvinus J. Sociol. Soc. Policy* **5**, 97 (2014).
25. Greene, W. Functional forms for the negative binomial model for count data. *Econ. Lett.* **99**, 585–590 (2008).
26. Lawless, J. F. Negative binomial and mixed Poisson regression. *Can. J. Stat.* **15**, 209–225 (1987).
27. Day, W. H. & Edelsbrunner, H. Efficient algorithms for agglomerative hierarchical clustering methods. *J. Classif.* **1**, 7–24 (1984).
28. Nelder, J. A. & Wedderburn, R. W. Generalized linear models. *J. R. Stat. Soc. Ser. A (General)* **135**, 370–384 (1972).
29. Hilbe, J. M. *Negative Binomial Regression* (Cambridge University Press, 2011).
30. Paternoster, R. & Brame, R. Multiple routes to delinquency? A test of developmental and general theories of crime. *Criminology* **35**, 49–84 (1997).
31. Lewis, F., Butler, A. & Gilbert, L. A unified approach to model selection using the likelihood ratio test. *Methods Ecol. Evol.* **2**,155–162 (2011).
32. Vuong, Q. H. Likelihood ratio tests for model selection and non-nested hypotheses. *Econom. J. Econ. Soc.* **57**, 307–333 (1989).
33. Hall, B. H. & Mairesse, J. Exploring the relationship between r &d and productivity in French manufacturing firms. *J. Econ.* **65**, 263–293 (1995).

Author contributions

Conceptualization and methodology, S.Q. and Y.Z; Formal analysis, S.Q. and Y.Z; Data collection, H.L.; Supervision, Y.G. and H.G.; Writing-original draft preparation, S.Q.; Writing-review and editing, F.Z., Q.Z. and M.G.. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by Zhejiang Province Pointer Leader Plan Major Social Welfare Project grant number 2022C03124 and Zhejiang Province Soft Science Key Project grant number 2020C25002.

Competing interersts

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Q.Z. or Y.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.