ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2021-81523

Investigation of Factors Affecting the Performance of Poisson Regression Model: A Simulation Study

Poisson Regresyon Modelinin Performansını Etkileyen Faktörlerin Araştırılması: Bir Benzetim Çalışması

Didem DERİCİ YILDIRIM[a], Gülhan TEMEL[a], İrem ERSÖZ KAYA[b]

[a]Department of Biostatistics and Medical Informatics, Mersin University Faculty of Medicine, Mersin, TURKEY
[b]Department of Computer Engineering, Tarsus University Faculty of Engineering, Mersin, TURKEY

ABSTRACT Objective: The aim of this study is to compare the performance of the regression correlation coefficient (RCC) and its estimators, which is the measure of the power of the regression model, in terms of bias and root mean square error (RMSE), considering the multicollinearity for different sample sizes, regression intercept (α) parameters and missing observation rates. **Material and Methods:** Data were produced with MATLAB for the combination of different sample sizes (n: 50, 100, 200, 500), correlations between independent variables (r: 0.0, 0.70), α parameters (α: 0.0, 0.80) and missing observation ratios (m: 0%, 5%, 10%, 20%) for estimators of RCC. Then, randomly subtracted data were analyzed with Poisson regression model and this process was repeated 1,000 times. **Results:** Simulation results showed that leave-one-out-cross validation estimator ($\hat{R}_{crs}$) had the smallest bias and RMSE compared with the other estimators for n=50. All estimators provided similar bias and RMSE values depending on the increase in sample size. It was observed that $\hat{R}_{crs}$ was more robust than the others against to missing ratio for small sample size. In contrary to missing ratio, the most adversely affected estimator by multicollinearity was $\hat{R}_{crs}$. **Conclusion:** The usage of Poisson regression model in clinical trials is wide spreading. Therefore, it is important to evaluate the success of the model in the best way. The study was the first study which investigated the effect of missing ratio on the estimators of RCC for Poisson regression model. It was detected that $\hat{R}_{crs}$ was the most successful estimator in the case of missing observations.

Keywords: Regression correlation coefficient;
Poisson regression; missing ratio; re-sampling

ÖZET Amaç: Bu çalışmanın amacı, Poisson regresyon modelinin gücünü gösteren regresyon korelasyon katsayısı (RKK) ve tahmin edicilerinin, farklı örnek genişlikleri, regresyon sabiti (α) parametreleri ve eksik gözlem oranları için çoklu bağlantı dikkate alınarak performanslarının yanlılık ve kök ortalama kare hatası [root mean square error (RMSE)] cinsinden karşılaştırılmasıdır. **Gereç ve Yöntemler:** Bu çalışmada, RKK'nın tahmin edicileri farklı örnek genişliği (n: 50, 100, 200, 500), bağımsız değişkenler arası korelasyon (r: 0,0, 0,70), α parametresi (α: 0,0, 0,80) ve eksik gözlem oranı (m: %0, %5, %10, %20) olmak üzere oluşturulan her bir kombinasyon için MATLAB programında üretilen ve tamamen rastgele olarak eksiltilen veriler Poisson regresyon modeli ile analiz edilmiş ve bu işlem 1.000 kere tekrarlanmıştır. **Bulgular:** Yapılan simülasyon sonuçlarına göre n=50 için çapraz geçerlilik tahmin edicisi ($\hat{R}_{crs}$) diğer tahmin edicilere göre en düşük yanlılık ve RMSE değerine sahiptir. Örnek genişliği arttıkça tüm tahmin edicilerin yanlılık ve RMSE değerleri tüm kombinasyonlar için birbirine yaklaşmaktadır. Küçük örnek genişliğinde eksik gözleme en dayanıklı tahmin edici $\hat{R}_{crs}$dir. Bu durumun aksine çoklu bağlantıdan en olumsuz etkilenen tahmin edici de $\hat{R}_{crs}$dir. **Sonuç:** Poisson regresyon modelinin kullanımı klinik araştırmalarda gün geçtikçe artmaktadır. Model başarısını en doğru şekilde değerlendirmek de önemli hâle gelmektedir. Yapılan çalışma model performansını gösteren tahmin edicilerin başarısına eksik gözlemin etkisini Poisson regresyon modeli için inceleyen ilk çalışma olup, eksik gözlem durumunda en başarılı tahmin edicinin $\hat{R}_{crs}$ olduğu tespit edilmiştir.

Anahtar kelimeler: Regresyon korelasyon katsayısı;
Poisson regresyon; eksik gözlem;
yeniden örnekleme

Usage of correct type of regression analysis according to the type of outcome variable is too important to model the relationship between variables correctly. Most of the outcome variables in medicine are count data type that are not normally distributed. But they are analyzed as continuous or binary unfortunately. Es-

Correspondence: Didem DERİCİ YILDIRIM
Department of Biostatistics and Medical Informatics, Mersin University Faculty of Medicine, Mersin, TURKEY/TÜRKİYE
E-mail: didemderici@hotmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 21 Jan 2021 ***Received in revised form:*** 08 Apr 2021 ***Accepted:*** 18 May 2021 ***Available online:*** 25 May 2021

pecially in regression analysis, this will cause loss of information and false results when modeling with linear or binary logistic regression analysis.[1] Therefore, Poisson regression models are often used for modeling the count data.[2]

In regression analysis, researchers are interested in not only parameter estimation, but also the ratio of explanation outcome by independent variables called as coefficient of determination (R^2). Because R^2 was an indicator of the power of the regression model. Firstly, R^2 based on residuals was suggested for Poisson regression model.[3] But suggested R^2 was extremely inflated when sample size was small in comparison to number of independent variables. Therefore, an adjusted R^2 was suggested to figure out this problem by Mittlobock and Waldhor.[4] In addition, similar coefficient called as entropy correlation coefficient was suggested by Eshima and Tabata.[5]

The most recent regression coefficient introduced for the Poisson regression model in the literature was the regression correlation coefficient (RCC) between the dependent variable Y and the conditional expected value of Y given X (E(Y|X)) that was suggested by Takahashi and Kurasowa.[6] They used three other estimators of RCC that were sample correlation, Jack-knife and leave-one-out cross validation estimators which were suggested by Zheng and Agresti.[7] Then, Kaengtong and Domthong proposed an estimator of RCC robust to the correlation between independent variables in accordance with RCC proposed by Takahashi and Kurasowa in 2016.[7,8]

In the light of these information, the aim of this study was to compare the estimators of RCC for Poisson regression model mentioned above paragraph in terms of bias and Root Mean Square of Error (RMSE) for different sample sizes, α parameters and missing ratios considering the multicollinearity.

MATERIAL AND METHODS

POISSON REGRESSION ANALYSIS

Poisson regression is used to model count data types. Poisson regression is a method used for positively skewed data with equal mean and variance under the Poisson distribution assumption. The most important assumption of this method is the variance of variable is not higher than mean of its.[9] The linear function of the estimators of the Poisson regression model is provided by logarithmic transformation. The Poisson regression model is given in Equation 1. Maximum likelihood estimation method is used in parameter estimation.

$$log_e(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots . + \beta_n x_n \qquad (1)$$

REGRESSION CORRELATION COEFFICIENT ($\hat{\boldsymbol{R}}$)

RCC for Poisson model is calculated to evaluate the performance of model. It is the relationship between value of *Y* and the conditional expected value of Y given X *(E(Y|X))*. If this value is close to 1, this means that the model is good, otherwise if this value is close to 0, model is not good. It is given in Equation 2.

$$RCC(Y,X)=cor(Y, E(Y|X)) \qquad (2)$$

In our study, the explicit form of RCC for Poisson regression model ($\hat{R}$) was used where *μ (px1)* was the mean vector of *X, Σ (pxp)* was the covariance matrix of *X, α (1x1)* was the intercept of the model and *β (px1)* was the regression coefficients vector. The estimator $\hat{R}$ was calculated as in Equation 3.[6]

$$\hat{R} = \left(\frac{\exp\,(\hat{\alpha}+\mu^T\hat{\beta}+\frac{1}{2}\hat{\beta}^T\Sigma\hat{\beta})(\exp\,(\hat{\beta}^T\Sigma\hat{\beta})-1}{1+\exp\,(\hat{\alpha}+\mu^T\hat{\beta}+\frac{1}{2}\hat{\beta}^T\Sigma\hat{\beta})(\exp\,(\hat{\beta}^T\Sigma\hat{\beta})-1} \right)^{1/2} \qquad (3)$$

MODIFIED REGRESSION CORRELATION COEFFICIENT($\hat{\boldsymbol{R}}_{\boldsymbol{M}}$)

Adjusted $\hat{R}$ is introduced as modified correlation coefficient ($\hat{R}_M$) for multicollinearity in Poisson model.[8] The formula of $\hat{R}_M$ is given in Equation 4.

$$\hat{R}_M=\frac{(n-p-1)}{(n-1)}\left(\frac{\exp\,(\hat{\alpha}+\mu^T\hat{\beta}+\frac{1}{2}\hat{\beta}^T\sum\hat{\beta})(\exp\,(\hat{\beta}^T\sum\hat{\beta})-1}{1+\exp\,(\hat{\alpha}+\mu^T\hat{\beta}+\frac{1}{2}\hat{\beta}^T\sum\hat{\beta})(\exp\,(\hat{\beta}^T\sum\hat{\beta})-1}\right)^{1/2} \quad (4)$$

Traditional methods obtain statistics using theoretical sampling distributions. In re-sampling methods, unlike theoretical distributions, it has a revolutionary methodology. Statistical inferences are made by re-sampling. The re-sampling methods used as estimators of RCC in our study are summarized below.[7,10] In order to investigate the success of these methods against traditional method, sample correlation estimator is also considered.

SAMPLE CORRELATION ESTIMATOR ($\hat{\boldsymbol{R}}_{\text{cor}}$)

The simplest correlation coefficient obtained from Poisson regression model is given in Equation 5.

$$\hat{R}_{cor}=Cor\left(Y,\hat{Y}\right) \quad (5)$$

JACK-KNIFE ESTIMATOR ($\hat{\boldsymbol{R}}_{jck}$)

This estimator is a correlation without ith element of the sample and calculated as Equation 6.[6,8]

$$\hat{R}_{jck}=NCor\left(Y,\hat{Y}\right)-\frac{(N-1)}{N}\sum_{i=1}^{N}Cor(Y,\hat{Y})^{(-i)} \quad (6)$$

LEAVE-ONE-OUT-CROSS VALIDATION ($\hat{\boldsymbol{R}}_{\boldsymbol{crs}}$)

This coefficient is calculated of the vector of Y_i obtained from maximum estimates of $\hat{\beta}$ and $\hat{\alpha}$ by deleting ith element of sample. The formula is given in Equation 7.[8]

$$\hat{Y}_{crs}\ is\ \hat{Y}^{(-1)},\hat{Y}^{(-2)},\dots\dots,\hat{Y}^{(-N)} \qquad \hat{R}_{crs}=Cor(Y,\hat{Y}_{crs}) \quad (7)$$

SIMULATION STUDY

The simulation study was performed in four steps given as below. In this study, simulations were carried out by increasing the correlation coefficient between variables, from no correlation (r=0) and increasing in units of 0.10. Regression intercepts were determined by the reference articles.[6,8] Each combination for correlation coefficients, missing ratio and sample sizes were tested and only breakpoints were reported in this study.

1. Data were generated with Poisson distribution for two independent variables (p=2). The correlations of independent variables (r) were 0.0 and 0.70. The sample sizes (n) were 50, 100, 200 and 500. The model parameters α were set as 0.0 and 0.8 while β parameters were set as 0.30. (β_1 and β_2 parameters were the same for each scenario).
2. Poisson regression was done with the model given in Equation 8.

$$\log\left(E(Y|X)\right)=\alpha+\beta_1X_1+\beta_2X_2\ ,\quad Y|X\sim Poisson(\theta),\ X_1,X_2\sim N(0,1) \quad (8)$$

3. Missing data sets were created by deleting data completely randomly from the complete data at different rates (m= 0%, 5%, 10%, 20%).
4. Bias and RMSE values of RCC were obtained for all scenarios. The simulation applications were executed with MATLAB R2017b and the number of simulation was determined as 1,000.

RESULTS

Firstly, we compared the coefficients with small sample size (n=50) for all estimators according to missing ratios and correlation between independent variables in terms of bias and RMSE that was given in Table 1. For α=0.0, and β$_1$,β$_2$=0.30, if the missing ratio increased, bias and RMSE values of all estimators except $\hat{R}_{\text{crs}}$ increased. $\hat{R}_{\text{crs}}$ had the smallest bias and RMSE compared with the other estimators. When α increased to 0.80, a little decrease had been observed in bias and RMSE values in the situation of both correlation and no correlation.

TABLE 1: Comparison of the regression correlation coefficient estimators according to correlation between independent variables and missing ratios for n=50.

		n=50							
		α=0				α=0.80			
		Bias		RMSE		Bias		RMSE	
	m	r=0	r=0.70	r=0	r=0.70	r=0	r=0.70	r=0	r=0.70
$\hat{R}$	0	0.2629	0.2630	0.0736	0.0760	0.2572	0.2275	0.0712	0.0600
	5	0.2663	0.2659	0.0765	0.0791	0.2593	0.2279	0.0740	0.0615
	10	0.2700	0.2723	0.0789	0.0837	0.2618	0.2294	0.0759	0.0642
	20	0.2832	0.2823	0.0907	0.0956	0.2643	0.2310	0.0821	0.0678
$\hat{R}_M$	0	0.2455	0.2405	0.0645	0.0647	0.2341	0.1996	0.0599	0.0482
	5	0.2484	0.2436	0.0671	0.0677	0.2362	0.2000	0.0626	0.0498
	10	0.2522	0.2502	0.0693	0.0722	0.2387	0.2012	0.0644	0.0523
	20	0.2643	0.2592	0.0800	0.0831	0.2407	0.2031	0.0703	0.0559
$\hat{R}_{cor}$	0	0.2632	0.2624	0.0701	0.0696	0.2609	0.2287	0.0688	0.0541
	5	0.2681	0.2660	0.0728	0.0717	0.2646	0.2309	0.0709	0.0555
	10	0.2729	0.2708	0.0756	0.0744	0.2681	0.2312	0.0729	0.0562
	20	0.2877	0.2802	0.0844	0.0808	0.2768	0.2415	0.0789	0.0620
$\hat{R}_{jck}$	0	0.2717	0.2768	0.0752	0.0776	0.2715	0.2420	0.0748	0.0601
	5	0.2786	0.2810	0.0793	0.0805	0.2750	0.2474	0.0766	0.0633
	10	0.2839	0.2868	0.0828	0.0840	0.2806	0.2485	0.0802	0.0642
	20	0.3069	0.3052	0.0982	0.0970	0.2962	0.2680	0.0905	0.0761
$\hat{R}_{crs}$	0	0.0991	0.1426	0.0202	0.0232	0.1613	0.1579	0.0281	0.0261
	5	0.0726	0.1180	0.0191	0.0204	0.1447	0.1424	0.0240	0.0221
	10	0.0524	0.1035	0.0222	0.0178	0.1267	0.1343	0.0213	0.0202
	20	-0.0068	0.0429	0.0294	0.0195	0.0770	0.0961	0.0160	0.0149

*n: Total sample size; r: Correlation coefficient between independent variables; m: Missing ratio; α: Intercept; $\hat{R}$: Regression correlation coefficient; $\hat{R}_M$: Modified regression correlation coefficient; $\hat{R}_{cor}$: Sample correlation estimator; $\hat{R}_{jck}$: Jack-knife estimator; $\hat{R}_{crs}$: Leave-one-out cross validation estimator; RMSE: Root mean square error.

Table 2 showed that the results were so similar with n=50 for n=100. Only the bias and RMSE values of $\hat{R}_{crs}$ estimations increased with the rise in sample size. For α=0,80, bias values increased when there is no correlation. But with the correlation, there was a decrease in bias values for all missing ratios.

TABLE 2: Comparison of the regression correlation coefficient estimators according to correlation between independent variables and missing ratios for n=100.

		n=100							
		α=0				**α=0.80**			
		Bias		RMSE		Bias		RMSE	
	m	r=0	r=0.70	r=0	r=0.70	r=0	r=0.70	r=0	r=0.70
$\hat{R}$	0	0.2490	0.2516	0.0638	0.0672	0.2493	0.2209	0.0640	0.0533
	5	0.2511	0.2528	0.0654	0.0681	0.2520	0.2208	0.0664	0.0542
	10	0.2540	0.2579	0.0677	0.0712	0.2520	0.2226	0.0669	0.0557
	20	0.2571	0.2601	0.0707	0.0754	0.2522	0.2210	0.0700	0.0579
$\hat{R}_M$	0	0.2403	0.2406	0.0595	0.0619	0.2408	0.2070	0.0598	0.0474
	5	0.2424	0.2418	0.0611	0.0627	0.2405	0.2069	0.0607	0.0483
	10	0.2454	0.2470	0.0633	0.0657	0.2405	0.2087	0.0613	0.0497
	20	0.2482	0.2491	0.0662	0.0698	0.2406	0.2070	0.0643	0.0521
$\hat{R}_{cor}$	0	0.2524	0.2540	0.0640	0.0648	0.2529	0.2207	0.0643	0.0496
	5	0.2542	0.2555	0.0651	0.0656	0.2541	0.2214	0.0649	0.0501
	10	0.2564	0.2580	0.0663	0.0670	0.2561	0.2222	0.0659	0.0507
	20	0.2634	0.2614	0.0701	0.0692	0.2585	0.2240	0.0676	0.0521
$\hat{R}_{jck}$	0	0.2572	0.2626	0.0666	0.0693	0.2576	0.2284	0.0669	0.0529
	5	0.2597	0.2647	0.0682	0.0705	0.2597	0.2298	0.0677	0.0537
	10	0.2630	0.2681	0.0700	0.0724	0.2623	0.2317	0.0691	0.0548
	20	0.2733	0.2750	0.0759	0.0768	0.2687	0.2377	0.0730	0.0582
$\hat{R}_{crs}$	0	0.1750	0.1945	0.0323	0.0385	0.1746	0.1857	0.0323	0.0351
	5	0.1631	0.1867	0.0301	0.0356	0.1995	0.1810	0.0402	0.0335
	10	0.1475	0.1758	0.0264	0.0322	0.1917	0.1738	0.0373	0.0312
	20	0.1110	0.1447	0.0209	0.0241	0.1647	0.1571	0.0288	0.0259

*n: Total sample size; r: Correlation coefficient between independent variables; m: Missing ratio; α: Intercept; $\hat{R}$: Regression correlation coefficient; $\hat{R}_M$: Modified regression correlation coefficient; $\hat{R}_{cor}$: Sample correlation estimator; $\hat{R}_{jck}$=Jack-knife estimator; $\hat{R}_{crs}$: Leave-one-out cross validation estimator, RMSE: Root mean square error.

Bias and RMSE values were systematically increased dependent to the increase in missing ratio for sample size 200 and the results were given in Table 3. These values of $\hat{R}_{crs}$ estimator increased dependent to the sample size. But they were still decreases although missing ratio increases for $\hat{R}_{crs}$.

TABLE 3: Comparison of the regression correlation coefficient estimators according to correlation between independent variables and missing ratios for n=200.

		n=200							
		α=0				**α=0.80**			
		Bias		RMSE		Bias		RMSE	
	m	r=0	r=0.70	r=0	r=0.70	r=0	r=0.70	r=0	r=0.70
$\hat{R}$	0	0.2478	0.2521	0.0623	0.0654	0.2460	0.2177	0.0618	0.0499
	5	0.2481	0.2495	0.0626	0.0645	0.2497	0.2183	0.0638	0.0507
	10	0.2476	0.2489	0.0627	0.0647	0.2499	0.2163	0.0641	0.0502
	20	0.2498	0.2530	0.0644	0.0680	0.2475	0.2193	0.0642	0.0532
$\hat{R}_M$	0	0.2436	0.2467	0.0602	0.0627	0.2403	0.2108	0.0590	0.0469
	5	0.2439	0.2441	0.0605	0.0618	0.2439	0.2114	0.0610	0.0477
	10	0.2433	0.2434	0.0605	0.0620	0.2442	0.2093	0.0613	0.0473
	20	0.2455	0.2476	0.0622	0.0653	0.2418	0.2124	0.0614	0.0502
$\hat{R}_{cor}$	0	0.2475	0.2512	0.0615	0.0632	0.2485	0.2162	0.0619	0.0472
	5	0.2491	0.2516	0.0622	0.0634	0.2496	0.2175	0.0624	0.0479
	10	0.2493	0.2514	0.0624	0.0634	0.2504	0.2167	0.0628	0.0476
	20	0.2531	0.2533	0.0644	0.0644	0.2522	0.2193	0.0639	0.0490
$\hat{R}_{jck}$	0	0.2500	0.2555	0.0627	0.0654	0.2512	0.2201	0.0632	0.0488
	5	0.2521	0.2572	0.0638	0.0663	0.2528	0.2220	0.0640	0.0498
	10	0.2529	0.2577	0.0643	0.0666	0.2539	0.2224	0.0646	0.0500
	20	0.2577	0.2610	0.0668	0.0684	0.2577	0.2265	0.0666	0.0520
$\hat{R}_{crs}$	0	0.2092	0.2222	0.0442	0.0495	0.2259	0.1992	0.0511	0.0400
	5	0.2043	0.2171	0.0423	0.0473	0.2226	0.1974	0.0497	0.0394
	10	0.1972	0.2110	0.0396	0.0448	0.2184	0.1935	0.0478	0.0379
	20	0.1777	0.1963	0.0332	0.0391	0.2066	0.1863	0.0429	0.0353

*n: total sample size; r: Correlation coefficient between independent variables; m: Missing ratio; α: intercept; $\hat{R}$: Regression correlation coefficient; $\hat{R}_M$: Modified regression correlation coefficient; $\hat{R}_{cor}$: Sample correlation estimator; $\hat{R}_{jck}$: Jack-knife estimator; $\hat{R}_{crs}$: Leave-one-out cross validation estimator; RMSE: Root mean square error.

$\hat{R}, \widehat{R}_M$, $\hat{R}_{cor}, \widehat{R}_{jck}$ and $\hat{R}_{crs}$ had similar results for the same conditions when the sample size was 500. The results were given in Table 4. $\hat{R}_{crs}$ values were quite close to others with the increasing of sample size.

TABLE 4: Comparison of the regression correlation coefficient estimators according to correlation between independent variables and missing ratios for n=500.

		n=500							
		α=0				**α=0.80**			
		Bias		RMSE		Bias		RMSE	
	m	r=0	r=0.70	r=0	r=0.70	r=0	r=0.70	r=0	r=0.70
$\hat{R}$	0	0.2449	0.2507	0.0603	0.0636	0.2458	0.2451	0.0609	0.0606
	5	0.2442	0.2496	0.0600	0.0631	0.2448	0.2133	0.0606	0.0469
	10	0.2461	0.2498	0.0611	0.0636	0.2462	0.2152	0.0613	0.0477
	20	0.2477	0.2509	0.0621	0.0645	0.2486	0.2139	0.0630	0.0478
$\hat{R}_M$	0	0.2432	0.2485	0.0595	0.0625	0.2435	0.2428	0.0598	0.0595
	5	0.2425	0.2474	0.0592	0.0621	0.2425	0.2105	0.0594	0.0457
	10	0.2444	0.2476	0.0602	0.0625	0.2439	0.2124	0.0602	0.0465
	20	0.2460	0.2487	0.0613	0.0634	0.2464	0.2111	0.0618	0.0466
$\hat{R}_{cor}$	0	0.2453	0.2494	0.0602	0.0622	0.2465	0.2465	0.0608	0.0608
	5	0.2457	0.2498	0.0604	0.0624	0.2465	0.2137	0.0608	0.0459
	10	0.2466	0.2495	0.0609	0.0623	0.2473	0.2145	0.0612	0.0463
	20	0.2472	0.2505	0.0612	0.0628	0.2483	0.2149	0.0617	0.0465
$\hat{R}_{jck}$	0	0.2463	0.2512	0.0607	0.0631	0.2476	0.2477	0.0613	0.0614
	5	0.2469	0.2518	0.0610	0.0634	0.2479	0.2159	0.0615	0.0468
	10	0.2481	0.2520	0.0616	0.0635	0.2488	0.2169	0.0619	0.0473
	20	0.2491	0.2539	0.0622	0.0645	0.2505	0.2184	0.0628	0.0480
$\hat{R}_{crs}$	0	0.2303	0.2377	0.0531	0.0565	0.2375	0.2373	0.0564	0.0564
	5	0.2280	0.2360	0.0521	0.0557	0.2360	0.2059	0.0557	0.0426
	10	0.2262	0.2336	0.0513	0.0546	0.2347	0.2053	0.0551	0.0423
	20	0.2179	0.2278	0.0477	0.0520	0.2304	0.2018	0.0531	0.0410

*n: Total sample size; r: Correlation coefficient between independent variables; m: Missing ratio; α: Intercept; $\hat{R}$: Regression correlation coefficient; $\hat{R}_M$: Modified regression correlation coefficient; $\hat{R}_{cor}$: Sample correlation estimator; $\hat{R}_{jck}$: Jack-knife estimator; $\hat{R}_{crs}$: Leave-one-out cross validation estimator; RMSE: Root mean square error.

Regardless of the sample size, the ranking of estimators were detected in terms of bias as $\hat{R}_{crs}$<$\hat{R}_M$<$\hat{R}$<$\hat{R}_{cor}$<$\hat{R}_{jck}$ in Figure 1.

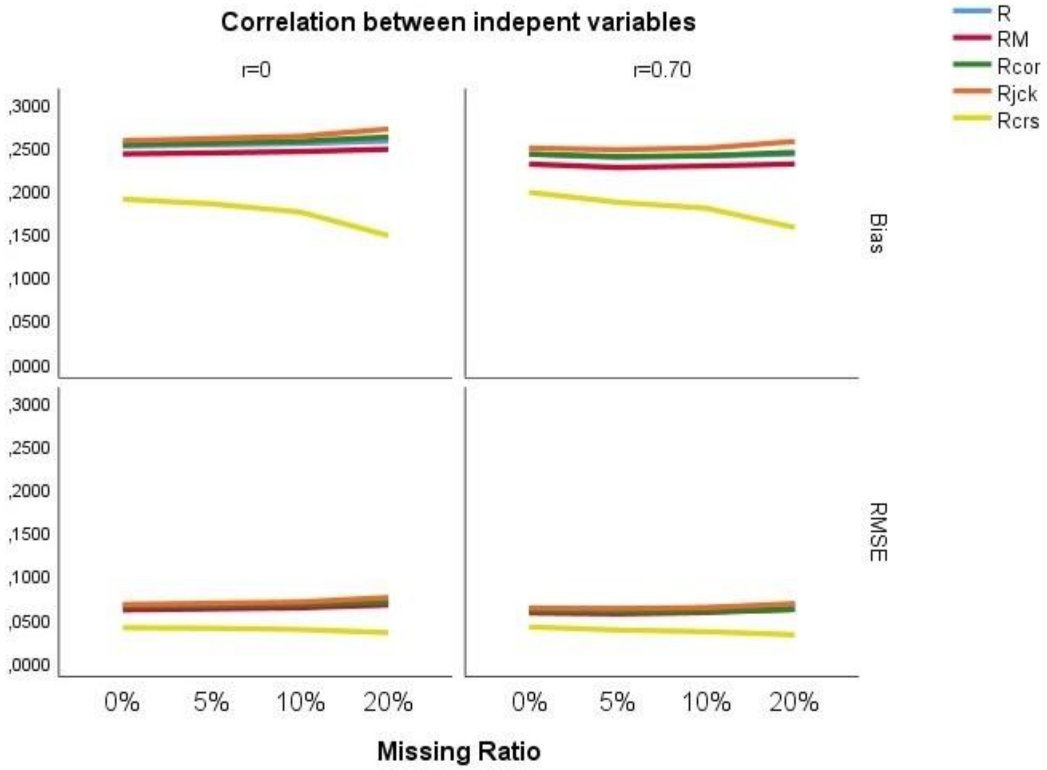

FIGURE 1: Comparison of estimators regardless of sample size and parameters of model.

Except $\hat{R}_{crs}$, the other estimators had similar results in the condition of correlation or not. $\hat{R}_{crs}$ had the smallest bias and RMSE values for all simulations. But bias of $\hat{R}_{crs}$ with correlation was higher than without correlation. Missing structure of the data increased bias of the estimators except $\hat{R}_{crs}$ and especially for 20% missing ratio, there was an increase in bias of the estimators. It was observed that $\hat{R}_{crs}$ is more robust than the others against to missing ratio for small sample size, especially.

DISCUSSION

The power of the regression model was widely evaluated by R-squared (R^2). The inclusion of more variables to the model increased R^2 but this was not generally meant that the power of the model increased. Because mean square error of the model increased at the same time. Therefore, new estimators of R^2 were suggested for Poisson regression model to prevent the increase in bias.[8]

There were many measures of model performance in regression analysis such as AIC, R^2, adjusted R^2 but they had some limitations. Therefore, a new and strong estimator of RCC was suggested that was the correlation between Y and conditional expectation E(Y|X) by Zheng and Agresti. In their study, RCC and jack-knife, sample correlation estimator, leave one out cross validation estimators of RCC were compared in terms of bias and RMSE in general linear model.[7] In a simulation study for Poisson regression model by Takahashi and Kurosawa, regardless of sample size, there was a similar ranking ($\hat{R}$ <$\hat{R}_{cor}$ <$\hat{R}_{jck}$) with our study in terms of RMSE.[6] Moreover, in our study we investigated the missing effect on the estimators of RCC and $\hat{R}_{crs}$ the most robust estimator against missing ratio($\hat{R}_{crs}$<$\hat{R}$ <$\hat{R}_{cor}$ <$\hat{R}_{jck}$).

The estimators of RCC were compared with multicollinearity between independent variables by Kaengthong and Domthong. Moreover, they were introduced a modified estimator of RCC ($\hat{R}_M$) in case of multicollinearity. The proposed RCC and modified RCC were defined as $\hat{R}$ ve $\hat{R}_M$ in our study. They had found that the bias and RMSE values of $\hat{R}$ were higher than $\hat{R}_M$'s when independent variable number was 2, α=0.80 and r=0.70 which was the similar scenario with us.[8] The same situation was observed in our study for the case where the missing ratio was 0. The bias and RMSE values for $\hat{R}_M$ were lower in the correlated structure. Turkan and Ozel suggested a new modified Jacknifed poisson ridge regression estimator (MJPR) for different sample sizes in case of high multicollinearity to evaluate the performance of the model.[11] Kibria-Lukman (K-L) estimator for poisson regression model was proposed by Lukman et al. for different sample sizes, intercepts and number of independent variables in case of high multicollinearity.[12] In another study conducted by Kristofer and Shukur, they compared the traditional maximum likelihood estimator and poisson ridge estimator in case of multicollinearity.[13] There were many studies about poisson regression model to tackle multicollinearity. In our study, missing effect on estimators was especially investigated in addition to multicollinearity problem.

According to our results, it was observed that $\hat{R}$, $\hat{R}_M$, $\hat{R}_{cor}$ and $\hat{R}_{jck}$ were affected from missing ratio negatively. In this model, the simulation study was performed without changing β parameters. Regardless of sample size and α parameters, only missing ratios were evaluated in terms of bias and RMSE in Figure 1. $\hat{R}_{crs}$ was the strongest estimator against missing ratio in comparison to other estimators. But it was affected from multicollinearity more than the others. The differences between the other estimators except $\hat{R}_{crs}$ were slightly more obvious for bias values. There was an overall increase in bias and RMSE values with missing in data.

The study conducted by Temel et al. showed that re-sampling methods had better results for small sample sizes than large sample sizes. The better classification in small samples did not mean that these methods were less successful in large samples. The accuracy of classification for all methods decreased slightly because the variation increased dependent to the increase in sample size.[14] Especially in our study, very low bias and RMSE values were obtained for $\hat{R}_{crs}$ at small sample sizes.

The strength aspect of our study was considering the missing values in data. In previous studies, sample size, multicollinearity and different α and β parameters were investigated but this was the first study that evaluated the missing effect for RCC estimators of Poisson regression model. The limitation of our study was the fact that different number of independent variables and β parameters were not considered for simulation plan.

CONCLUSION

Poisson regression models have been frequently used in clinical studies. Among the coefficients proposed for evaluating the predictive power of the model, $\hat{R}_{crs}$ was the most successful in case of missing observations. However, if there was a relationship between independent variables, the success of $\hat{R}_{crs}$ decreased.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Didem Derici Yıldırım, Gülhan Temel;* ***Design:*** *Didem Derici Yıldırım, Gülhan Temel;* ***Control/Supervision:*** *Didem Derici Yıldırım, Gülhan Temel, İrem Ersöz Kaya;* ***Data Collection and/or Processing:*** *Didem Derici Yıldırım, Gülhan Temel, İrem Ersöz Kaya;* ***Analysis and/or Interpretation:*** *Didem Derici Yıldırım, Gülhan Temel;* ***Literature Review:*** *Didem Derici Yıldırım, Gülhan Temel;* ***Writing the Article:*** *Didem Derici Yıldırım, Gülhan Temel;* ***Critical Review:*** *Didem Derici Yıldırım, Gülhan Temel, İrem Ersöz Kaya.*

REFERENCES

1. Byers AL, Allore H, Gill TM, Peduzzi PN. Application of negative binomial modeling for discrete outcomes: a case study in aging research. J Clin Epidemiol. 2003;56(6):559-64. [Crossref] [PubMed]
2. Kusuma RD, Purwono Y. Zero-Inflated Poisson Regression Analysis on Frequency of Health Insurance Claim PT. XYZ. Advances in Economics, Business and Management Research. 2018;72:321-5. [Crossref]
3. Cameron AC, Windmeijer F. R-squared measures for count data regression models with applications to health care utilization. Journal of Business and Economic Statistics. 1996;14(2):209-20. [Crossref]
4. Mittlbock M, Waldhor T. Adjustments for R2-measures for poisson regression model. Comput. Statist. Data Anal. 2000;34(4):461-72. [Crossref]
5. Eshima N, Tabata M. Entropy correlation coefficient for measuring predictive power of generalized linear model. Statist. Probab. Let. 2007;77(6):588-93. [Crossref]
6. Takahashi A, Kurosawa T. Regression correlation coefficient for a poisson regression model. Comput. Statist. Data Anal. 2016;98:71-8. [Crossref]
7. Zheng B, Agresti A. Summarizing the predictive power of a generalized linear model. Stat Med. 2000;19(13):1771-81. [Crossref] [PubMed]
8. Kaengthong N, Domthong U. Modified regression correlation coefficient for poisson regression model. Journal of Physics Conference Series. 2017;890(1):012155. [Crossref]
9. Elhai JD, Calhoun PS, Ford JD. Statistical procedures for analyzing mental health services data. Psychiatry Res. 2008;160(2):129-36. [Crossref] [PubMed]
10. Yu CH. Resampling methods: Concepts, applications and justification. Practical Assessment, Research and Evaluation. 2002;8(19):1-16. [Link]
11. Turkan S, Ozel G. A new modified jackknifed estimator for the poisson regression model. Journal of Applied Statistics. 2015;43(10):1892-1905. [Crossref]
12. Lukman AF, Adewuyi E, Mansson K, Kibria BMG. A new estimator for the multicollinear poisson regression model: Simulation and application. Scientific Reports. 2021;11:3732-43. [Crossref]
13. Mansson K, Shukur G. A Poisson ridge regression estimator. Economic Modelling. 2011;28(4):1475-81. [Crossref]
14. Orekici Temel G, Erdogan S, Ankaralı H. Sınıflama modelinin performansını değerlendirmede yeniden örnekleme yöntemlerinin kullanımı [Usage of resampling methods for evaluating the performance of classification model]. Bilişim Teknolojileri Dergisi. 2012;5(3):1-8.