Simulations description

Joint estimation

Estimator

Method used in simulations is based on joint estimation described in article 'Inference for Regression with Variables Generated by AI or Machine Learning'. Instead of using linear model as authors did

$$Y_i = \gamma^T \theta_i + \alpha^T q_i + \varepsilon_i$$

I used generalised linear model with Poisson distribution

$$Y_i \sim \mathcal{P}(\mu_i), \qquad \log \mu_i = \gamma^T \theta_i + \alpha^T q_i.$$

In model I assumed that we have 10 classes so θ_i's are 10-dimensional vectors with one 1 and nine 0's and that q_i (data without measurement error) are one-dimensional. We also need error matrix Ω that on i-th row and j-th column has estimated probability of model to classify true value j of θ as i.

Having this assumed we can compute likelihood by adding over all possible true values of θ_i

$$l(Y_i, \hat{\theta}_i = d; \gamma, \alpha, \omega) = \sum_{k=1}^{10} \omega_{dk} \cdot \frac{e^{-\mu_{ik}} \mu_{ik}^{Y_i}}{Y_i!}$$

with μ_{ik} being k-th element of γ plus αq_i. and $\omega_{dk} = P(\hat{\theta}_i = d, \theta_i = k)$, that is probability of model classifying value of θ_i as d when true value is k.

The total log-likelihood is

$$L = \sum_{i=1}^{n} \log l(Y_i, \hat{\theta}_i = d; \gamma, \alpha, \omega). \tag{*}$$

After search for values $\hat{\gamma}, \hat{\alpha}$ that maximizes L.

Simulation

I made two simulation: repeated 25 simulation of 1000 observations and repeated 25 simulation of 5000 observations.

In smaller data set I drawn gamma from $U(1, 2)$, α from $U(-0.5, 0.5)$, error matrix as scaled matrix with accuracy of 70% and errors following Poisson distribution, θ_i were drawn with given probability p, responses of AI model were drawn using θ_i's and error matrix and Y_i were drawn from Poisson distribution.

Similar case was for bigger data set but this time probability for drawing θ_i was also randomized and so was accuracy of predictions in error matrix (from $U(60, 90)$)

Then I applied R function *optimx* to maximize likelihoods given by (*) with added square penalty function to prevent overfitting. This method was used twice – firstly in case when probabilities of occurring true classes were known and secondly when they were estimated from joint estimation.

Results of implemented methods were compared with the results of R function *glm* that ignores potential measurement errors. For each simulation was computed sum of squares of prediction errors that is for k-th repetition of simulation error SSE_k was given by formula

$$SSE_k = (\hat{\gamma}_k - \gamma_k)^T(\hat{\gamma}_k - \gamma_k) + (\hat{\alpha}_k - \alpha_k)^2.$$

Results

In both bigger and smaller number of observations joint estimation was better option. In smaller set it achieved mean error 0.1728 with known probabilities and 0.1835 with unknown probabilities compared to standard GLM with error 0.2155. In set of 5000 observations difference was significantly bigger – joint estimation with known probabilities had mean error 0.0098, joint estimation with predicted probabilities has mean error 0.0801 when *glm* had mean error 0.4839. More statistics is presented in the table below

Table 1: Results for joint estimation

Statistic	JE1000	JEUP1000	GLM1000	JE5000	JEUP5000	GLM5000
Mean	0.1728	0.1835	0.2155	0.0098	0.0801	0.4839
Median	0.1083	0.1230	0.1908	0.0082	0.0087	0.4622
St. Dev.	0.1656	0.1567	0.1467	0.0072	0.2511	0.3537
Min	0.0294	0.0308	0.0400	0.0010	0.0012	0.0307
Max	0.6709	0.6343	0.7771	0.0278	1.2145	1.4192

where JE stands for joint estimation, JEUP means joint estimation with unknown probabilities and GLM are results for standard R *glm* function. Number in name describe number of observations used in the simulation.

Bias corrected estimators

Estimator

Here in the same generalised linear model I used formulas from chapter 5.1 from the article 'Inference for Regression with Variables Generated by AI or Machine Learning':

$$\hat{\psi}^{bca} = \left(I + \frac{\hat{\kappa}}{\sqrt{n}} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\xi}_i \hat{\xi}_i^T \right)^{-1} \begin{bmatrix} \hat{\Omega} & 0 \\ 0 & 0 \end{bmatrix} \right) \hat{\psi},$$

$$\hat{\psi}^{bcm} = \left(I - \frac{\hat{\kappa}}{\sqrt{n}} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\xi}_i \hat{\xi}_i^T \right)^{-1} \begin{bmatrix} \hat{\Omega} & 0 \\ 0 & 0 \end{bmatrix} \right)^{-1} \hat{\psi}$$

Because in GLM there is not closed formula for least square estimators I used results of R function *glm* as $\hat{\psi}$. Parameter $\frac{\kappa}{\sqrt{n}}$ was estimated by the total false-positive ratio

$$\frac{\hat{\kappa}}{\sqrt{n}} = FP = \sum_{k=2}^{10} FP_k = \sum_{k=2}^{10} \pi_k (1 - p_k)$$

The $\hat{\Omega}$ in simulation was a 10×10 matrix with zeros on the upper left corner (representing intercept) and

$$\frac{P(\hat{\theta} = i)\delta_{ij} - P(\hat{\theta} = i, \theta = j)}{FP}$$

on the $i-$th row and $j-$th column.

Simulation

There were 3 simulations, each of them repeated 50 times. The first one used 1000 observations, the second one used 10000 and the third one 100000. For each of them parameter γ was drawn from $U(1, 2)$, α was drawn from $U(-0.5, 0.5)$, true probabilities were fixed

$$p = (0.25, 0.2, 0.15, 0.05, 0.05, 0.1, 0.02, 0.03, 0.05, 0.1),$$

default AI model accuracy was set to be 70% with misclassification errors drawn from the Poisson distribution, and values of q_is were drawn from Bernoulli distribution with probability 0.7.

There were considered three estimators – bias corrected additive estimator $\hat{\psi}^{bca}$, bias corrected multiplicative estimator $\hat{\psi}^{bcm}$ and standard GLM estimator taken from R function *glm*. Results were compared with true values for each observation and for each of them SSE_k was computed by formula

$$SSE_k = (\hat{\gamma}_k - \gamma_k)^T (\hat{\gamma}_k - \gamma_k) + (\hat{\alpha}_k - \alpha_k)^2.$$

Results

In all three simulations bca estimator has lower mean error and standard deviation than standard GLM estimator. The difference was the bigger, the more observations were used.

Bcm estimator came out to be much more unstable – even though its median error was smaller than bca in case of 10000 and 100000 observations, its standard deviation was much bigger and its maximum errors were enormous. The results are shown in the table below.

Table 2: Results for bias corrected estimators

Statistic	bca1000	bcm1000	GLM1000
Mean	0.1447	1.1788	0.1887
Median	0.1292	0.2851	0.1633
St. Dev.	0.0972	3.2714	0.1115
Min	0.0176	0.0251	0.0282
Max	0.4612	21.9636	0.5002
Statistic	bca10000	bcm10000	GLM10000
Mean	0.0672	0.1510	0.1670
Median	0.0603	0.0553	0.1548
St. Dev.	0.0400	0.4175	0.0723
Min	0.0144	0.0067	0.0607
Max	0.1959	2.9502	0.3872
Statistic	bca100000	bcm100000	GLM100000
Mean	0.0504	0.3771	0.1412
Median	0.0373	0.0345	0.1195
St. Dev.	0.0448	1.8176	0.0894
Min	0.0058	0.0047	0.0225
Max	0.2423	12.8195	0.4717