Simulations description

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 glm. 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. The exact results are presented in the table below

no.	JE1000	JEUP1000	GLM1000	JE5000	JEUP5000	GLM5000
1	0.4132	0.3739	0.3390	0.0079	0.0084	0.4648
2	0.2408	0.2400	0.3136	0.0065	0.0055	1.1799
3	0.0944	0.0959	0.3223	0.0156	0.0168	0.0307
4	0.6448	0.4262	0.0997	0.0193	0.0258	0.6300
5	0.0578	0.0523	0.2284	0.0114	0.0051	0.5453
6	0.0682	0.0856	0.2633	0.0054	0.0053	0.2713
7	0.1491	0.3416	0.1041	0.0044	0.0043	0.2102
8	0.2437	0.2321	0.1435	0.0070	0.0074	0.3235
9	0.0600	0.0716	0.1548	0.0048	0.0050	0.5858
10	0.0822	0.0722	0.3007	0.0110	0.0090	1.4192
11	0.1228	0.5214	0.2833	0.0015	0.0015	0.0750
12	0.6709	0.6343	0.1508	0.0084	0.0065	0.4596
13	0.1047	0.1425	0.0400	0.0273	0.0266	0.2776
14	0.0710	0.0900	0.1076	0.0056	1.2145	0.0419
15	0.2827	0.2742	0.2097	0.0278	0.0440	0.8097
16	0.1242	0.1300	0.3053	0.0010	0.0012	0.1417
17	0.1042	0.1254	0.0949	0.0046	0.0044	1.1150
18	0.1651	0.1205	0.1088	0.0103	0.0095	0.5859
19	0.0668	0.0568	0.0606	0.0021	0.0027	0.3022
20	0.0294	0.0308	0.1232	0.0186	0.0228	0.8205
21	0.0947	0.0864	0.7771	0.0140	0.0155	0.2564
22	0.1119	0.0836	0.3346	0.0128	0.0134	0.5104
23	0.1021	0.1011	0.1720	0.0021	0.0387	0.2155
24	0.1469	0.1384	0.1274	0.0045	0.0033	0.5891
25	0.0675	0.0602	0.2215	0.0118	0.5053	0.2359
Mean	0.1728	0.1835	0.2155	0.0098	0.0801	0.4839

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.