Модель оцінки страхового випадку із використанням розподілу Твіді

Кінда Віталій

Дипломний керівник: к.т.н. старший викладач Терентьєв Олександр Миколайович

Національний технічний універститет України "Київський політехнічний інститут імені Ігоря Сікорського"
ІПСА

19 червня 2018 р.

Мета дослідження

- побудова моделей для оцінки страхових випадків із використанням розподілу Твіді в майнових страхових відносинах із використанням узагальнених лінійних моделей (GLM - Generalised Linear Model) та узагальнених оціночних рівнянь (GEE - Generalised Estimation Equation).

Об'єкт дослідження

- набір страхових даних по виплатах страхових вимог за рік для майнових страхових полісів.

Предмет дослідження

- Узагальнені лінійні моделі.
- Узагальнені оціночні рівняння.
- Розподіл Твіді.

Актуальність дослідження

- В роботі побудовано модель оцінки страхових випадків, що дозволяє страховим компаніям оцінювати величину можливих страхових вимог, що відбулися, але ще не відрапортовані. Це в свою чергу дозволяє оцінювати величину страхового тарифу в подальшому та виводити бізнес страхової в прибуток.

Постановка задачі

- Ознайомитись із теорією майнового страхування (non-life insurance).
- Провести дослідження в страховій сфері в рамках резервування фінансів на виплату страхових вимог.
- Побудувати необхідні моделі використовуючи GLM та GEE із використанням розподілу Твіді.
- Проаналізувати та порівняти одержані результати.

Вид даних

Accident year i	Development year j 1	2	$\cdots$	n-1	n
1	$Y_{1,1}$	$Y_{1,2}$	$\cdots$	$Y_{1,n-1}$	$Y_{1,n}$
2	$Y_{2,1}$	$Y_{2,2}$	$\cdots$	$Y_{2,n-1}$	$Y_{2,n}$
i	$\vdots$	$\vdots$	$Y_{i,n+1-i}$	$\vdots$	$\vdots$
n-1	$Y_{n-1,1}$	$Y_{n-1,2}$	$\cdots$		
n	$Y_{n,1}$	$Y_{n,2}$	$\cdots$		$Y_{n,n}$

Трикутний вигляд даних для нормалізованих додаткових виплат претензій $Y_{i,j}$

$$f_Y(y;\theta,\phi/w) = a(y,\phi/w)\exp\{\frac{w}{\phi}(y\theta - k(\theta))\}, y \in S \subseteq \mathbb{R},$$

де $a(\cdot;\cdot)$, $k(\cdot)$ - функції, $\theta \in \mathbb{R}$ - канонічний параметр, $\phi > 0$ - параметр дисперсії, $w > 0$ - ваги. $k(\cdot)$ двічі неперервно-диференційовна та із першою похідною оберненої функції.

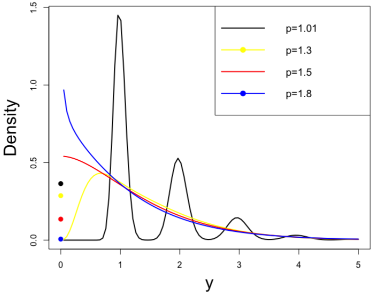
Розподіл	θ	$\frac{w}{\phi}$	θ	$\frac{W}{\Phi}$	$k(\theta)$
$Binomial(n,p)$	$ln(\frac{p}{1-p})$	n	$\mathbb{R}$	$\mathbb{N}$	$ln(\frac{1+\exp(\theta)}{2})$
$Poisson(\mu_N)$	$ln(\mu_N)$	$-$	$\mathbb{R}$	$-$	$\exp(\theta) - 1$
$Gamma(\alpha,\tau)$	$1-\tau$	α	$(-\infty,1)$	$\mathbb{R}$	$ln(\frac{1}{1-\theta})$
$N(\mu,\sigma^2)$	μ	σ^2	$\mathbb{R}$	$(0,\infty)$	$\frac{1}{2}\theta^2$

EDM параметризація деяких широко відомих розподілів

Особливий інтерес в експоненційних дисперсних моделях є клас розподілів з дисперсними функціями виду

$$V(\mu) = \mu^p,$$

для деяких p.



Щільність Твіді для різних значень p.

Математична частина

GLM

- Випадкові величини $Y_1,...Y_n$ є незалежними і розподіл Y_i залежить від описових змінних $\boldsymbol{x_i}$ через регресійні параметри $\boldsymbol{\beta}$.
- Змінна відгуку Y_i слідує з EDM із середнім μ_i, параметром дисперсії ϕ та вагами w_i

$$Y_i \sim ED(\mu_i, \frac{\phi}{w_i}).$$

- Розглядається лінійна комбінація описових змінних

$$\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} = \sum_{j=1}^{k} x_{ij}\beta_j,$$

де η_i називається лінійним предиктором.

- Існує строго монотонна і двічі неперервно-диференційовна функція зв'язку (link function) $g(\cdot)$ така, що

$$g(\eta_i) = \mu_i.$$

- Оцінкою максимальної правдоподібності $\widehat{\beta}$ буде рішення рівняння

$$\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{G}(\boldsymbol{y} - \boldsymbol{\mu}) = \mathbf{0},$$

  де $\boldsymbol{W} = diag\{w(\mu_1),...,w(\mu_n)\}$ та $\boldsymbol{G} = diag\{g'(\mu_1),...,g'(\mu_n)\}$ діагональні матриці розмірності $(n \times n)$.
- Щоб отримати оцінку максимальної правдоподібності використовуємо алгоритм ітераційно зважених найменших квадратів (IWLS - Iterative Weighted Least Squares).

- Припускаємо мультиплікативну структуру відгуку $\mu_{i,j} = \gamma_i \nu_j$.
- Функція зв'язку - $\eta_{i,j} = log(\mu_{i,j}) = log(\gamma_i) + log(\nu_j)$.
- Лінійний предиктор - $\eta_{i,j} = \boldsymbol{x}_{i,j} \boldsymbol{\beta}_{i,j}$, де $\boldsymbol{x}_{i,j} = (0, \delta_{2,i}, \cdots, \delta_{n,i}, \delta_{1,j}, \cdots, \delta_{n,j})$; $\delta_{j,i}$ - дельти Кронекера.
- Вектор невідомих параметрів має вигляд $\boldsymbol{\beta} = (0, log\gamma_2, \cdots, log\gamma_n, log\nu_1, \cdots, log\nu_n)$.
- Маючи оцінку невідомих регресійних параметрів, спрогнозовані платежі будуть визначатися наступним чином $\widehat{Y}_{i,j} = \widehat{\mu}_{i,j} = exp(\widehat{\eta}_{i,j}) = \widehat{\gamma}_i \widehat{\nu}_j$.

- Випадкові величини $\boldsymbol{Y_i},\ldots\boldsymbol{Y_K}$ є незалежними, однак, компоненти Y_i можуть бути корельованими.
- Розглядається лінійна комбінація описових змінних

$$\eta_{i,j} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta},$$

  де $\eta_{i,j}$ називається лінійним предиктором.
- Очікуване значення $\mu_{i,j}$ задовольняє тотожність

$$g(\mu_{i,j}) = \eta_{i,j},$$

  де $g(\cdot)$ строго монотонна і двічі неперервно-диференційовна функція зв'язку (link function). Лінійний предиктор разом із функцією зв'язку повністю визначає структуру середнього.

- Дисперсія відгуку $Y_{i,j}$ може бути виражена, як функція від середнього

$$var(Y_{i,j}) = \frac{\phi}{w_{i,j}} V(\mu_{i,j}),$$

  де $V(\cdot)$ відома функція дисперсії і $w_{i,j} > 0$ є попередньо відомі ваги, параметр дисперсії $\phi > 0$ може бути відомим або ні.

- Кореляція між компонентами Y_i представлена у вигляді кореляційної матриці $\boldsymbol{R_i} = \boldsymbol{R_i}(\nu)$ розміром $(n_i \times n_i)$.

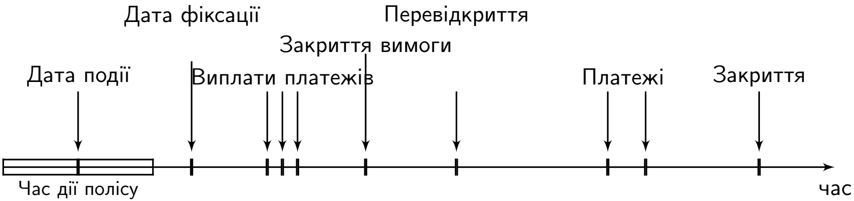
Набір даних

- Набір даних був взятий від шведського актуарного товариства (CAS - Casualty Actuarial Society).
- Мова програмування R.

```
calculate_Z = function(z_mu, z_y, n_obs){
  Z = matrix(ncol = 1, nrow = n_obs)
  for (i in 1: n_obs){
    Z[i,1] = log(z_mu[i]) + (z_y[i] - z_mu[i])/(z_mu[i])
  }
  return(Z)
}

calculate_Beta = function(b_X, b_W, b_Z){
  return( ginv(t(b_X) %*% b_W %*% b_X) %*% (t(b_X) %*% b_W %*% b_Z) )
}
```

Час життя страхової претензії в майновому страхуванні



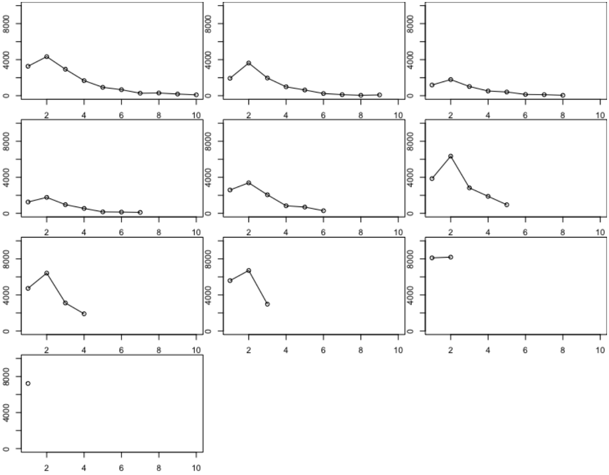
Набір даних

Було розглянуто набір даних шведської страхової компанії, що стосуються полісів автострахування протягом 1988-1997 років. Розмір датасету 13300 записів. Для дослідження обрано вибірку розміром 100 записів.
Відношення навчальна/перевірочна – 55/45 %.
Опис змінних:

- *GRCODE* - унікальний ідентифікатор компанії;
- *GRNAME* - назва страхової компанії;
- *AccidentYear* - рік настання страхового випадку;
- *DevelopmentYear* - рік виплати страхової претензії;
- *DevelopmentLag* - лаг виплати страхової претензії;
- *CumPaidLoss* - кумулятивні виплачені збитки та розподілені витрати на кінець року.

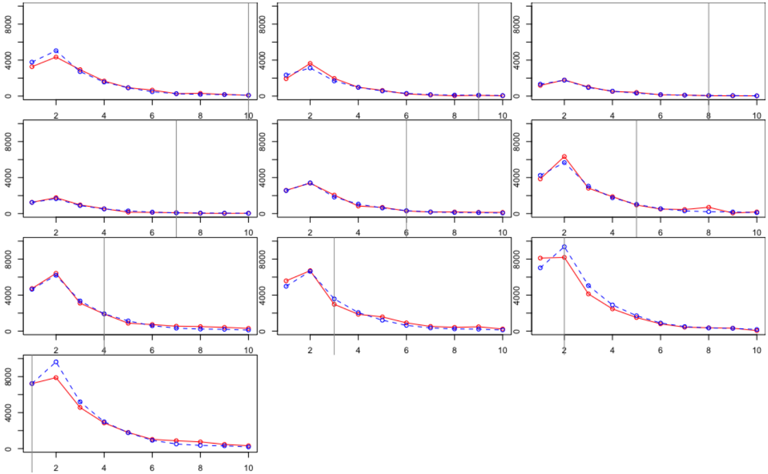
Набір даних

Графічний вигляд набору даних для страхової компанії

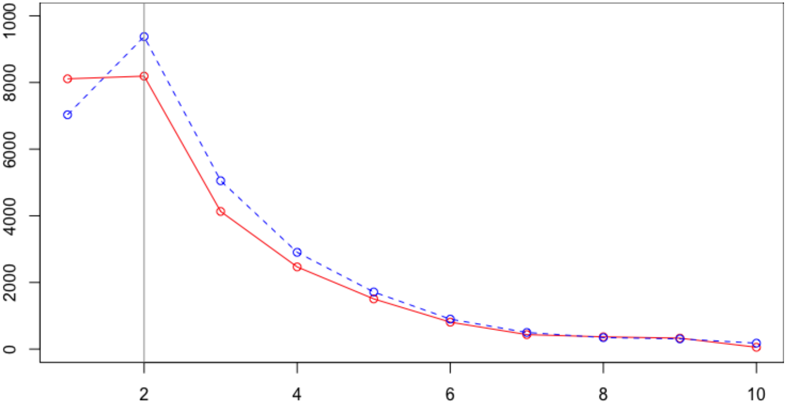


Вид верхньої трикутної матриці навчальної вибірки

Результати роботи
GLM

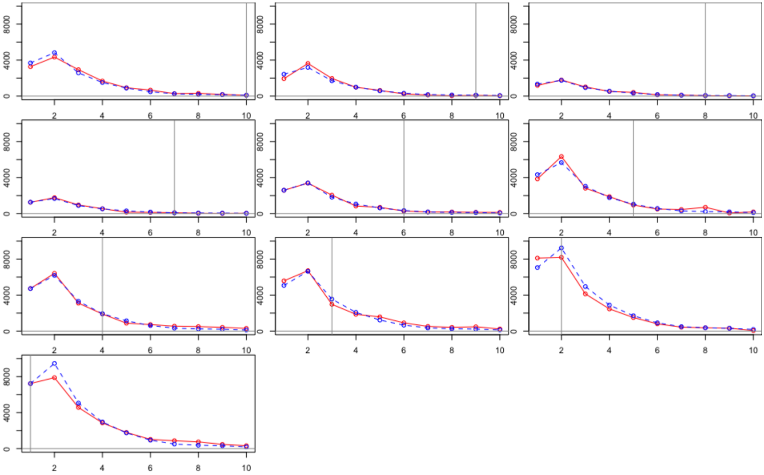


Загальний вигляд результатів роботи моделі Твіді GLM



Графік залежності лагу років від суми виплат. Червона крива - реальні значення; синя крива(пунктирна) - спрогнозовані; вертикальна лінія - це поріг між навчальною та тестовими вибірками.

Результати роботи
GEE



Загальний вигляд результатів роботи моделі Твіді GEE

Порівняння результатів

Model	Accident year j									
	1989	1990	1991	1992	1993	1994	1995	1996	1997	Total
Real	34	58	91	636	1917	3367	6009	10095	20540	42747
GLM	63	91	146	483	1346	2605	4847	11896	21864	43341
GEE	63	93	156	499	1395	2662	4938	11816	21558	43180

Реальні та спрогнозовані вартості страхових випадків із використанням моделей Твіді відносно року настання страхових випадків

Порівняння результатів

Model	MAPE(%)
GLM	31.8
GEE	32.0

Результати MAPE

Достатньо високий відсоток пояснюється тим, що запропоновані моделі доволі неякісно роблять прогноз для перших років, внаслідок цього суттєво збільшується значення критерію MAPE на 7 - 8 відсотків, коли на останньому році похибка становить менше одного відсотка.
Прогноз загальних необхідних резервів проводиться з похибкою близько 3 відсотків.

Висновки

Розглянуто такі методи моделювання оцінки величини страхових випадків: узагальнена лінійна модель, узагальнені оціночні рівняння.

Розроблено програмний продукт за допомогою мови програмування R для побудови моделей із використанням методу максимальної правдоподібності для оцінювання невідомих параметрів моделі.

Виконано порівняльний аналіз якості моделей на основі статистичного критерію MAPE.

Напрямок подальших досліджень

Вдосконалення розробленого методу побудови оцінки настання страхового випадку.

Розробка моделі оцінки страхового випадку для унікального клієнта із використанням розподілу Твіді.

Участь у конференціях:

XX міжнародна
науково-технічна конференція SAIT 2018:
"Модель оцінки страхового
випадку із використанням розподілу Твіді"
Кінда В.В.

Дякую за увагу.