Properties of uplift estimators for small samples

Krzysztof Rudaś [1,2]

Szymon Jaroszewicz [2]

[1] Warsaw University of Technology

[2] Institute of Computer Science

Polish Academy of Sciences

XLV Conference on Mathematical Statistics

Będlewo, Poland

What is uplift modeling?

- We study efficiency of new medicine reducing blood pressure,
- We can divide patients into three groups
  1. having lower blood pressure **because** of medicine (**positive impact**)
  2. having the same blood pressure **anyway** (**unnecessary costs**)
  3. having higher blood pressure **because** of medicine (**negative impact**)
- Problem: Select targets for receiving a medicine.

Usually used approaches

Widely used machine learning algorithms are not suitable for this problem.

- We have answer for a question how high is blood pressure of patient after obtaining a medicine.
- We want patient who will have lower blood pressure because of medicine.

Solution: Uplift modeling

We want to predict:

Uplift models predict change in behaviour resulting from the action

$$\mathrm{E}^{Target}(Y \mid X_1, \dots, X_m) - \mathrm{E}^{Not\ target}(Y \mid X_1, \dots, X_m)$$

Problem: we never have this two pieces of information at the same time.

Solution:

- two training sets:
  1. the treatment group
     on which the action was taken
  2. the control group
     on which no action was taken

- Build a model which predicts the difference between responses in the treatment and control groups

Randomization

- We will assign observations to treatment/control group randomly
- Random assignment allows for causal interpretation

Linear uplift regression: assumptions

Assume linear responses in treatment and control

$$y^C = X^C \beta^C + \varepsilon^C$$
$$y^T = X^T \beta^T + \varepsilon^T = X^T \beta^C + X^T \beta^U + \varepsilon^T$$

- $\mathrm{E}\, \varepsilon_i^T = 0$ and $\mathrm{Var}\, \varepsilon_i^T = \sigma^2$
- $\mathrm{E}\, \varepsilon_i^C = 0$ and $\mathrm{Var}\, \varepsilon_i^C = \sigma^2$
- equal sizes of treatment and control groups $n^T = n^C = \frac{n}{2}$

Goal

Estimate β^U.

Double regression

Build linear regression models in treatment and control separately. We get two estimators

$$\hat{\beta}^T = (X^{T'} X^T)^{-1} X^T y^T$$
$$\hat{\beta}^C = (X^{C'} X^C)^{-1} X^C y^C$$

The double regression estimator

The double regression estimator is defined as

$$\hat{\beta}_d^U = \hat{\beta}^T - \hat{\beta}^C$$

Uplift regression

Do we have a single model approach to regression?

Uplift regression estimator

Define

$$\tilde{y}_i = \begin{cases} 2y_i^T & \text{if } g_i = T, \\ -2y_i^C & \text{if } g_i = C. \end{cases}$$

The uplift regression estimator is:

$$\hat{\beta}_u^U = (X'X)^{-1}X'\tilde{y}$$

Comparing models

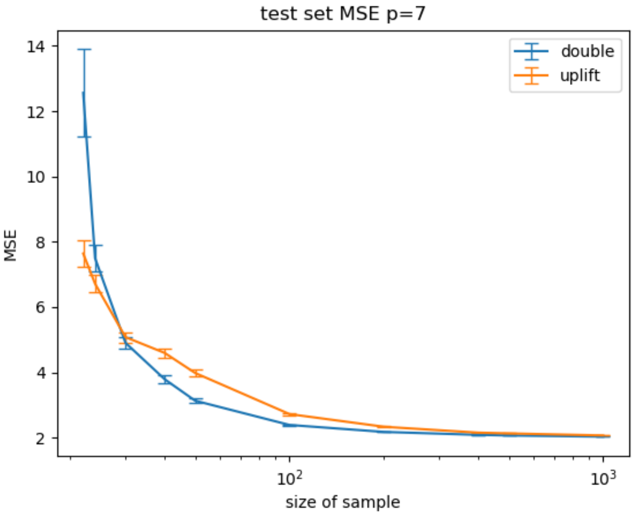
We will look at MSE of uplift estimator:

$$MSE\left(\hat{\beta}^U\right) = \mathrm{E}\,||\left(y^T_{test} - y^C_{test}\right) - X_{test}\hat{\beta}^U||^2$$

- MSE depends on bias and trace of variance of $\hat{\beta}^U$

MSE comparison

Angle between β^T and β^C equal to $\frac{\pi}{10}$.



Properties of double regression estimator

Theorem

Let $\hat{\beta}_d^U$ be the double regression estimator, then

1. $\hat{\beta}_d^U$ is unbiased,
2. If $\mathrm{E}\, X_i = 0$ and $\mathrm{Var}\, X_i = \Sigma$,

$$\sqrt{n}\left(\hat{\beta}_d^U - \beta^U\right) \xrightarrow{d} N\left(0, 4\sigma^2\Sigma^{-1}\right).$$

Uplift regression

Theorem

Let $\hat{\beta}^U_u$ be the uplift regression estimator

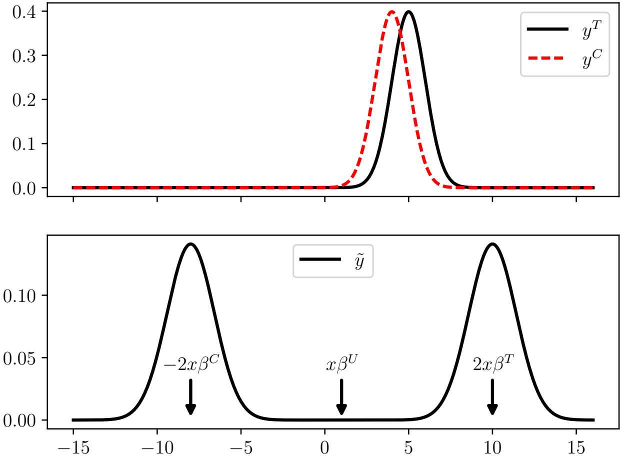
1. $\hat{\beta}^U_u$ is unbiased,
2. If $\mathrm{E}\, X_i = 0$ and $\mathrm{Var}\, X_i = \Sigma$ the

$$\sqrt{n}\left(\hat{\beta}^U_u - \beta^U\right) \xrightarrow{d} N\left(0, 4\sigma^2\Sigma^{-1} + bb' + \Sigma^{-1}\,\mathrm{Tr}(bb'\Sigma)\right)$$

where $b = \beta^T + \beta^C$.

- There is an additional > 0 term depending on $\beta^T + \beta^C$
- If $\beta^T + \beta^C = 0$, then asymptotic distribution is the same as in double.

Uplift regression – general case – intuition



Corrected uplift regression

- Can we get a single uplift regression model without this problem?
- If we subtract some β^* from β^T, β^C uplift does not change

$$(\beta^T - \beta^*) - (\beta^C - \beta^*) = \beta^T - \beta^C = \beta^U$$

- If we pick $\beta^* = \frac{\beta^T + \beta^C}{2}$ we additionally get

$$b = (\beta^T - \beta^*) + (\beta^C - \beta^*) = 0$$

- How can we modify the original problem? We don't even know true β^T and β^C needed for β^*

Corrected uplift regression

1. Estimate β^*:
$$\hat{\beta}^* = (X'X)^{-1}X'y$$
2. Correct the original y
$$y^{corr} = y - X\hat{\beta}^*$$
3. Build uplift regression on corrected data
$$\hat{\beta}^U_{corr} = (X'X)^{-1}X'\widetilde{y^{corr}}$$

Properties of the corrected uplift regression

Theorem

Let $\hat{\beta}^U_{corr}$ be the corrected uplift regression estimator. Then

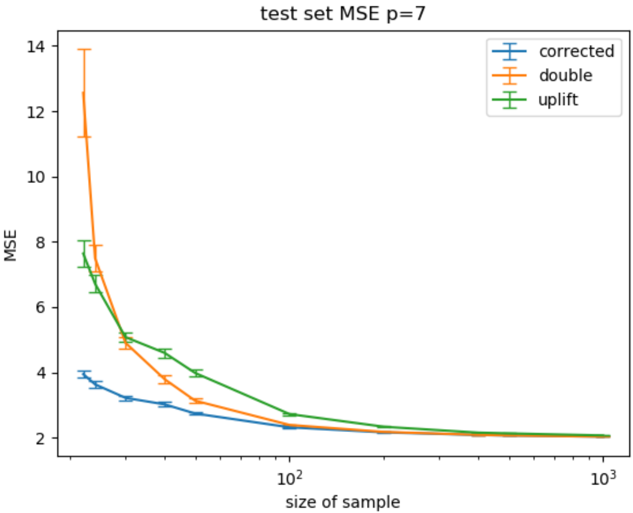
1. $\hat{\beta}^U_{corr}$ is asymptotically unbiased
2. If $\mathrm{E}\, X_i = 0$ and $\mathrm{Var}\, X_i = \Sigma$,

$$\sqrt{n}\left(\beta^{\hat{U}}_{corr} - \beta^U\right) \xrightarrow{d} N\left(0, 4\sigma^2\Sigma^{-1}\right)$$

- Asymptotic behaviour identical to double regression (correction works)
- Experiments show it is also better for small n.

Experiments

Angle between β^T and β^C equal to $\frac{\pi}{10}$.



MSE for small samples

- The biggest differences between MSE of estimators are observed for small samples,
- For small samples MSE of lower regression is higher than double.
- Corrected model works very well for $\beta^T = \beta^C$.

Double and uplift MSE

Theorem

Let $\hat{\beta}_d^U$ be the double regression estimator, then

1. If $X_i \sim N_p(0, I)$,

$$\operatorname{Tr} \operatorname{Var} \hat{\beta}_d^U = \frac{2p}{n/2 - p - 1}\sigma^2$$

Theorem

Let $\hat{\beta}_u^U$ be the uplift regression estimator, then

1. If $X_i \sim N_p(0, I)$,

$$\operatorname{Tr} \operatorname{Var} \hat{\beta}_u^U = \frac{4p}{n - p - 1}\sigma^2 + (4c(n,p) - 1)\|\beta^T + \beta^C\|^2,$$

where $(4c(n,p) - 1) > 0$

Comparison of MSE

If $\beta^T + \beta^C = 0$ then:

$$\operatorname{Tr}\operatorname{Var}\hat{\beta}_u^U = \frac{4p}{n-p-1}\sigma^2 \leqslant \frac{2p}{n/2-p-1}\sigma^2 = \operatorname{Tr}\operatorname{Var}\hat{\beta}_d^U$$

Special case $\beta^T = \beta^C$:

$$\operatorname{Tr}\operatorname{Var}\hat{\beta}_u^U = \frac{4p}{n-p-1}\sigma^2 + 16c\|\beta\|^2 - 4\|\beta\|^2.$$

Corrected MSE

Theorem

Let $\hat{\beta}^U_{corr}$ be the double regression estimator, then

1. If $X_i \sim N_p(0, I)$, $\mathrm{E}\,\hat{\beta}^U_{corr} = \frac{n(n-p)}{n^2+n-2}\beta^U$
2. If $X_i \sim N_p(0, I)$ and $\beta^T = \beta^C$,

$$\mathrm{Tr}\,\mathrm{Var}\,\hat{\beta}^U_{corr} = \frac{4p}{n-p-1}\frac{n^2-np}{n^2+n-2}\sigma^2$$

- If $\beta^T = \beta^C$, then $\hat{\beta}^U_{corr}$ is unbiased.
- $\mathrm{Tr}\,\mathrm{Var}\,\hat{\beta}^U_{corr} = \frac{4p}{n-p-1}\frac{n^2-np}{n^2+n-2}\sigma^2 \leqslant \frac{4p}{n-p-1}\sigma^2 \leqslant \mathrm{Tr}\,\mathrm{Var}\,\hat{\beta}^U_u$
- $\mathrm{Tr}\,\mathrm{Var}\,\hat{\beta}^U_{corr} = \frac{4p}{n-p-1}\frac{n^2-np}{n^2+n-2}\sigma^2 \leqslant \frac{2p}{\frac{n}{2}-p-1}\sigma^2 = \mathrm{Tr}\,\mathrm{Var}\,\hat{\beta}^U_d$

Test procedure

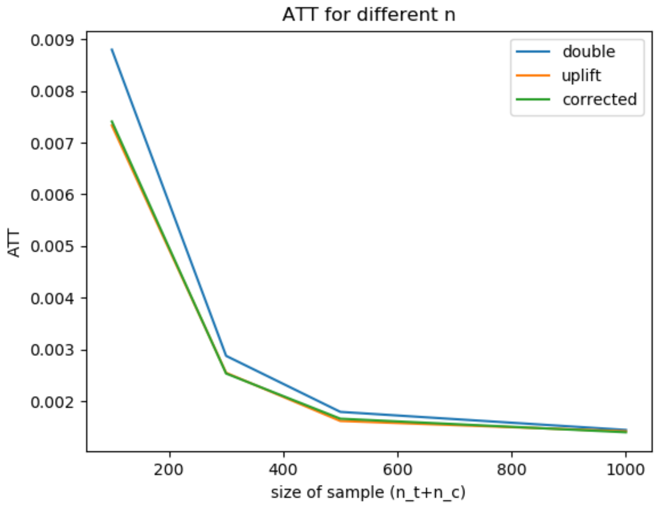
- Problem: we have only half of the information on each observation
  - We cannot to compare: $X_i\hat{\beta^U}$ with $y_i^T - y_i^C$
- Actually we compare sample ATT:

$$\frac{1}{n^T}\sum_{i=1}^{n^T} X_i\hat{\beta^U} \text{ with } \frac{1}{n^T}\sum_{i=1}^{n^T} y_i^T - \frac{1}{n^C}\sum_{i=1}^{n^C} y_i^C$$

Experiment on real data - Hillstrom

- Hillstrom experiment - results of an e-mail campaign for an Internet based retailer,
- Target: - find customers who spent much more money because of campaign.

Results


ATT for different n
0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
ATT
200
400
600
800
1000
size of sample (n_t+n_c)
double
uplift
corrected

Conclusions

- Uplift model works better than double for small samples
- But improvements are possible
- Corrected model works better than uplift and double for $\beta^T = \beta^C$ (realistic case)

Bibliography

[1] Gérard Letac and Hélène Massam. Quadratic and Inverse Regressions for Wishart Distributions, The Annals of Statistics , Vol. 26, No. 2, 573–595, page 582, 1998.

[2] K. Rudaś, S. Jaroszewicz. Linear regression for uplift modeling. *Data Mining and Knowledge Discovery*, 32(5), pages 1275-1305, Sep, 2018.

[3] K. Rudaś, S. Jaroszewicz. Shrinkage Estimators for Uplift Regression. In *Proc. of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD'19)*, Würzburg, Germany, Sep, 2019.

Thank you for attention!