Rare events bias of logistic regression

Rok Blagus[1], Lara Lusa[1], Jelle J. Goeman[2]

[1]Institute for Biostatistics and Medical Informatics
University of Ljubljana

[2]Department of Medical Statistics and Bioinformatics
Leiden University Medical Center

rok.blagus@mf.uni-lj.si

Bled, 2016

Outline

- Rare events bias explored (simulations) and explained
- Solutions: (multiple) undersampling
- Real data example (EM data)
- Conclusions

Class-prediction with logistic regression

- We are interested in predicting, for new data (random design), if the event ($Y_i = \{0, 1\}$) will occur, given the characteristics of the subjects ($\mathbf{X}_i$).

Class-prediction with logistic regression

- We are interested in predicting, for new data (random design), if the event ($Y_i = \{0, 1\}$) will occur, given the characteristics of the subjects ($\mathbf{X}_i$).
- We are in the setting where the proportion of events, $Y_i = 1$, is small $\Rightarrow$ rare events.

Class-prediction with logistic regression

- We are interested in predicting, for new data (random design), if the event ($Y_i = \{0, 1\}$) will occur, given the characteristics of the subjects ($\mathbf{X}_i$).
- We are in the setting where the proportion of events, $Y_i = 1$, is small $\Rightarrow$ rare events.
- (penalized) logistic regression will be used to obtain

$$\hat{\pi}_i = P(\hat{Y}_i = 1|\mathbf{X}_i) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2+\ldots+\hat{\beta}_p X_p}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2+\ldots+\hat{\beta}_p X_p}}$$

- The $\hat{\pi}_i$ will be the basis for prediction:

Class-prediction with logistic regression

- We are interested in predicting, for new data (random design), if the event ($Y_i = \{0, 1\}$) will occur, given the characteristics of the subjects ($\mathbf{X}_i$).
- We are in the setting where the proportion of events, $Y_i = 1$, is small $\Rightarrow$ rare events.
- (penalized) logistic regression will be used to obtain

$$\hat{\pi}_i = P(\hat{Y}_i = 1|\mathbf{X}_i) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p}}$$

- The $\hat{\pi}_i$ will be the basis for prediction:
  - An event is predicted if $\hat{\pi}_i > \tau$.

Class-prediction with logistic regression

- We are interested in predicting, for new data (random design), if the event ($Y_i = \{0, 1\}$) will occur, given the characteristics of the subjects ($\mathbf{X}_i$).
- We are in the setting where the proportion of events, $Y_i = 1$, is small $\Rightarrow$ rare events.
- (penalized) logistic regression will be used to obtain

$$\hat{\pi}_i = P(\hat{Y}_i = 1|\mathbf{X}_i) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p}}$$

- The $\hat{\pi}_i$ will be the basis for prediction:
  - An event is predicted if $\hat{\pi}_i > \tau$.
  - When $\hat{\pi}_i = \tau$ the class is randomly assigned.

Class-prediction with logistic regression

- We are interested in predicting, for new data (random design), if the event ($Y_i = \{0, 1\}$) will occur, given the characteristics of the subjects ($\mathbf{X}_i$).
- We are in the setting where the proportion of events, $Y_i = 1$, is small $\Rightarrow$ rare events.
- (penalized) logistic regression will be used to obtain

$$\hat{\pi}_i = P(\hat{Y}_i = 1|\mathbf{X}_i) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2+\ldots+\hat{\beta}_p X_p}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2+\ldots+\hat{\beta}_p X_p}}$$

- The $\hat{\pi}_i$ will be the basis for prediction:
  - An event is predicted if $\hat{\pi}_i > \tau$.
  - When $\hat{\pi}_i = \tau$ the class is randomly assigned.
  - We will use $\tau = \bar{y}$.

$$\sum_{i=1}^{n}(y_i - \hat{\pi}_i) = 0,$$

$$\bar{y} = \bar{\hat{\pi}}.$$

Rare events bias explored, $Y|X$, $p = 1$, different effect size

- The response for $n = 100$ training data was simulated from,

$$Y \sim \text{Bernoulli}\left(\frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}\right),$$

$$X \sim N(0, 1),$$

β_1:	0	0.1	0.2	0.5	1	2
β_0:	-2.20	-2.20	-2.25	-2.30	-2.60	-3.40

- Proportion of events: 0.10
- The model's performance was evaluated on 2000 independent test data simulated from the same model

Results, $Y|X$, $p = 1$, ML

	β_0	β_1	PA_0^T	PA_1^T	PA_0	PA_1	$\hat{\beta}_0$	$\hat{\beta}_1$
ML	-2.20	0.00	0.50	0.50	0.55	0.45	-2.31	0.01
	-2.20	0.10	0.52	0.52	0.55	0.46	-2.30	0.10
	-2.25	0.20	0.54	0.54	0.56	0.48	-2.37	0.21
	-2.30	0.50	0.60	0.60	0.60	0.57	-2.41	0.53
	-2.60	1.00	0.68	0.69	0.68	0.68	-2.76	1.09
	-3.40	2.00	0.79	0.81	0.79	0.80	-3.83	2.30

We can see that for small β_1,

$$|PA_0 - PA_1| > |PA_0^T - PA_1^T|$$

$\Rightarrow$ rare events bias

Results, $Y|X$, $p = 1$, ML

	β_0	β_1	PA_0^T	PA_1^T	PA_0	PA_1	$\hat{\beta}_0$	$\hat{\beta}_1$
ML	-2.20	0.00	0.50	0.50	0.55	0.45	-2.31	0.01
	-2.20	0.10	0.52	0.52	0.55	0.46	-2.30	0.10
	-2.25	0.20	0.54	0.54	0.56	0.48	-2.37	0.21
	-2.30	0.50	0.60	0.60	0.60	0.57	-2.41	0.53
	-2.60	1.00	0.68	0.69	0.68	0.68	-2.76	1.09
	-3.40	2.00	0.79	0.81	0.79	0.80	-3.83	2.30

We can see that for small β_1,

$$|\text{PA}_0 - \text{PA}_1| > |\text{PA}_0^T - \text{PA}_1^T|$$

$\Rightarrow$ rare events bias$\uparrow$:$\beta \downarrow$, $\bar{y} \downarrow$, $n/p \downarrow$.

Rare events bias $\neq$ small sample bias

- Firth's bias correction: Firth suggested to introduce bias in the score function in order to remove the small sample bias in regression coefficients

	β_0	β_1	PA_0^T	PA_1^T	PA_0	PA_1	$\hat{\beta}_0$	$\hat{\beta}_1$
Firth	-2.20	0.00	0.50	0.50	0.33	0.67	-2.20	0.00
	-2.20	0.10	0.52	0.52	0.34	0.66	-2.21	0.09
	-2.25	0.20	0.54	0.54	0.36	0.67	-2.24	0.20
	-2.30	0.50	0.60	0.60	0.47	0.68	-2.30	0.50
	-2.60	1.00	0.68	0.69	0.62	0.72	-2.60	0.99
	-3.40	2.00	0.79	0.81	0.76	0.83	-3.40	2.00

Rare events bias $\neq$ small sample bias

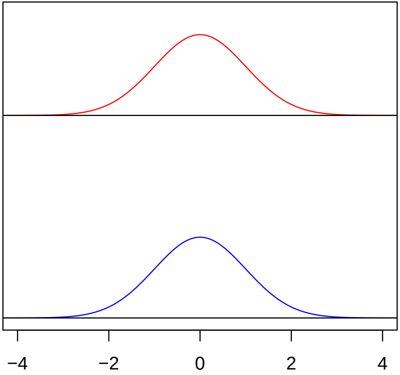
- Firth's bias correction: Firth suggested to introduce bias in the score function in order to remove the small sample bias in regression coefficients

	β_0	β_1	PA_0^T	PA_1^T	PA_0	PA_1	$\hat{\beta}_0$	$\hat{\beta}_1$
Firth	-2.20	0.00	0.50	0.50	0.33	0.67	-2.20	0.00
	-2.20	0.10	0.52	0.52	0.34	0.66	-2.21	0.09
	-2.25	0.20	0.54	0.54	0.36	0.67	-2.24	0.20
	-2.30	0.50	0.60	0.60	0.47	0.68	-2.30	0.50
	-2.60	1.00	0.68	0.69	0.62	0.72	-2.60	0.99
	-3.40	2.00	0.79	0.81	0.76	0.83	-3.40	2.00

$$|PA_0 - PA_1| > |PA_0^T - PA_1^T| \Rightarrow \text{rare events bias}$$

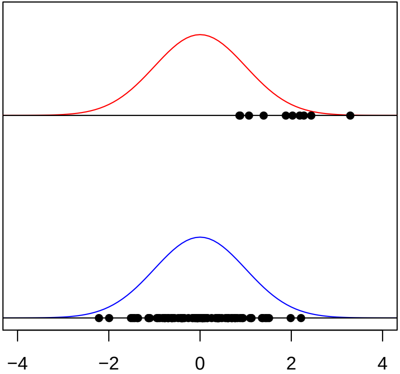
Explanation of the rare events bias of ML

- $\pi = 0.1$
- We explain this for $\beta_1 = 0$ and $X \sim N(0,1)$:
  - $(X|Y = 0) \sim N(0,1)$
  - $(X|Y = 1) \sim N(0,1)$



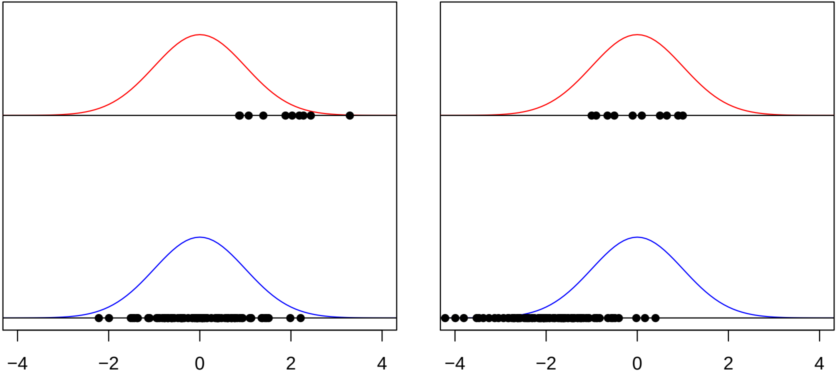
Explanation of the rare events bias of ML

- $\pi = 0.1$
- We explain this for $\beta_1 = 0$ and $X \sim N(0, 1)$:
  - $(X|Y = 0) \sim N(0, 1)$
  - $(X|Y = 1) \sim N(0, 1)$



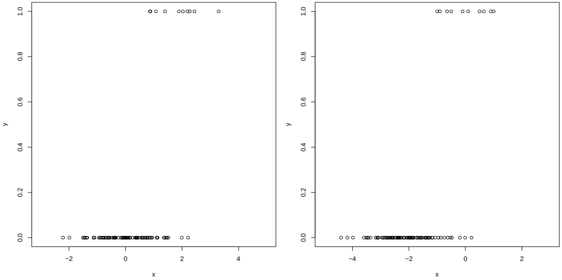
Explanation of the rare events bias of ML

- $\pi = 0.1$
- We explain this for $\beta_1 = 0$ and $X \sim N(0,1)$:
  - $(X|Y=0) \sim N(0,1)$
  - $(X|Y=1) \sim N(0,1)$



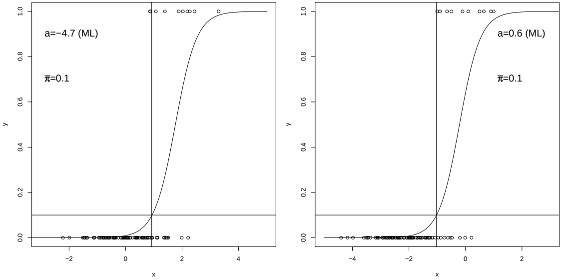
Explanation of the rare events bias of ML

$\hat{\beta}_1$ will never be (exactly) zero. $\hat{\beta}_1 \neq 0$ occurred only due to random variation of the events (left) and non-events (right)



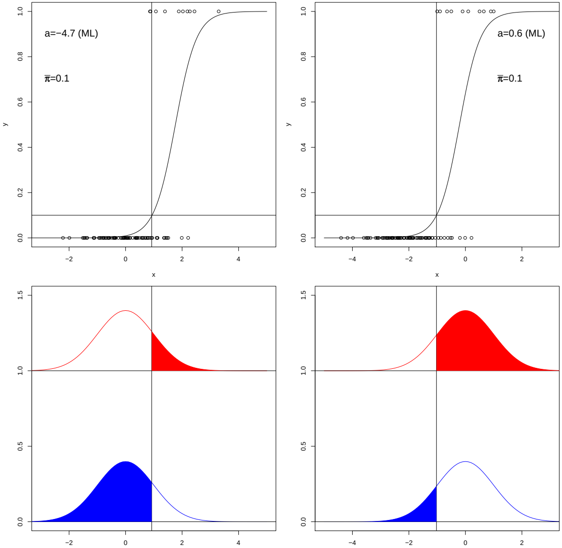
Explanation of the rare events bias of ML

$\hat{\beta}_1$ will never be (exactly) zero. $\hat{\beta}_1 \neq 0$ occurred only due to random variation of the events (left) and non-events (right)



Explanation of the rare events bias of ML

$\hat{\beta}_1$ will never be (exactly) zero. $\hat{\beta}_1 \neq 0$ occurred only due to random variation of the events (left) and non-events (right)



Explanation of the rare events bias of ML

- The situation where $\hat{\beta}_1 \neq 0$ occurred only due to random variation of the events (left) is more likely to appear due to larger sampling variability: systematic overfitting $\rightarrow$ rare events bias
- There is overfitting also when $\bar{y} = 0.5$, however there is no rare events bias as there is no systematic overfitting
- rare events bias $\uparrow$: $\bar{y} \downarrow$, $n/p \downarrow$, $\beta \downarrow$.

Results, $Y|X$, $p = 1$, $\beta_1 = 0$, λ optimized with CV

ML: $\text{PA}_0 = 0.55$, $\text{PA}_1 = 0.45$

	β_0	β_1	PA_0	PA_1	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_1 = 0$
L1	-2.20	0.00	0.52	0.48	-2.28	0.00	0.77

Simulation results considering only simulations with $\hat{\beta}_1 \neq 0$:

	PA_0	PA_1	$\hat{\beta}_0$	$\hat{\beta}_1$
L1	0.56	0.44	-2.31	-0.05

Solution: (multiple) undersampling (US)

- multiple under-sampling (MUS):
  - 100 random samples of the non-events were taken so that the number of events and non-events was equal in each sample
  - the model was estimated on each selected subset
  - the class was determined by majority voting

Solution: (multiple) undersampling (US)

- multiple under-sampling (MUS):
  - 100 random samples of the non-events were taken so that the number of events and non-events was equal in each sample
  - the model was estimated on each selected subset
  - the class was determined by majority voting
- classification threshold was set to 0.5

Results, $Y|X$, $p = 1$, multiple under-sampling

	β_0	β_1	PA_0^T	PA_1^T	PA_0	PA_1	$\hat{\beta}_0$	$\hat{\beta}_1$
ML	-2.20	0.00	0.50	0.50	0.50	0.50	0.00	0.00
	-2.20	0.10	0.52	0.52	0.50	0.50	-0.00	0.11
	-2.25	0.20	0.54	0.54	0.52	0.52	-0.02	0.21
	-2.30	0.50	0.60	0.60	0.59	0.59	-0.10	0.53
	-2.60	1.00	0.68	0.69	0.67	0.68	-0.38	1.02
	-3.40	2.00	0.79	0.81	0.79	0.81	-1.21	1.98

$$|PA_0 - PA_1| = |PA_0^T - PA_1^T| \Rightarrow \text{no rare events bias}$$

EM data

- Data from a study on erythema migrans (early Lyme borreliosis).

EM data

- Complete response was defined as a return to pre-Lyme borreliosis health status.
- The goal was to predict the incomplete response 6 months after treatment
- 133 (n_0) patients with complete response 6 months after treatment.
- 15 (n_1) patients with incomplete response 6 months after treatment.
- The outcome was predicted using 87 covariates (p, 80 numeric and 7 binary).

EM data

	PA_0	PA_1
L1	0.74	0.53
MUS.L1	0.67	0.67

Conclusions

- rare events bias: $|PA_0 - PA_1| > |PA_0^T - PA_1^T|$.

Conclusions

- rare events bias: $|PA_0 - PA_1| > |PA_0^T - PA_1^T|$.
- small sample bias $\neq$ rare events bias (Firth's bias correction)

Conclusions

- rare events bias: $|PA_0 - PA_1| > |PA_0^T - PA_1^T|$.
- small sample bias $\neq$ rare events bias (Firth's bias correction)
- rare events bias is caused by larger sampling variability of events $\Rightarrow$ systematic deviations in favor of the non-events (systematic overfitting)

Conclusions

- rare events bias: $|PA_0 - PA_1| > |PA_0^T - PA_1^T|$.
- small sample bias $\neq$ rare events bias (Firth's bias correction)
- rare events bias is caused by larger sampling variability of events $\Rightarrow$ systematic deviations in favor of the non-events (systematic overfitting)
- rare events bias disappears when the effect is large enough

Conclusions

- rare events bias: $|PA_0 - PA_1| > |PA_0^T - PA_1^T|$.
- small sample bias $\neq$ rare events bias (Firth's bias correction)
- rare events bias is caused by larger sampling variability of events $\Rightarrow$ systematic deviations in favor of the non-events (systematic overfitting)
- rare events bias disappears when the effect is large enough
- in high-dimensional setting the rare events bias is large due to many null variables (in the $p > n$ setting it is easier to overfit the data hence also larger systematic overfitting)

Conclusions

- rare events bias: $|PA_0 - PA_1| > |PA_0^T - PA_1^T|$.
- small sample bias $\neq$ rare events bias (Firth's bias correction)
- rare events bias is caused by larger sampling variability of events $\Rightarrow$ systematic deviations in favor of the non-events (systematic overfitting)
- rare events bias disappears when the effect is large enough
- in high-dimensional setting the rare events bias is large due to many null variables (in the $p > n$ setting it is easier to overfit the data hence also larger systematic overfitting)
- (multiple) undersampling removes the rare events bias

Conclusions

- rare events bias: $|PA_0 - PA_1| > |PA_0^T - PA_1^T|$.
- small sample bias $\neq$ rare events bias (Firth's bias correction)
- rare events bias is caused by larger sampling variability of events $\Rightarrow$ systematic deviations in favor of the non-events (systematic overfitting)
- rare events bias disappears when the effect is large enough
- in high-dimensional setting the rare events bias is large due to many null variables (in the $p > n$ setting it is easier to overfit the data hence also larger systematic overfitting)
- (multiple) undersampling removes the rare events bias
- what about categorical covariate(s)?
- what about the bias of the estimated probabilities?