Penalized logistic regression with rare events: preliminary results

Lara Lusa, Rok Blagus, Angelika Geroldinger and Georg Heinze

Institute for Biostatistics and Medical Informatics, University of Ljubljana
CeMSIIS, Medical University of Vienna

29th of September 2015

Background

- We are interested in estimating the probability that a rare event will occur, given the characteristics of a subject: $\pi_i = P(Y_i = 1|X_i)$.

Background

- We are interested in estimating the probability that a rare event will occur, given the characteristics of a subject: $\pi_i = P(Y_i = 1|X_i)$.
- The π_i will be the basis for prediction.

Background

- We are interested in estimating the probability that a rare event will occur, given the characteristics of a subject: $\pi_i = P(Y_i = 1|X_i)$.
- The π_i will be the basis for prediction.
- Logistic regression can be used if the number of considered covariates (p) is reasonably small compared to the number of subjects (n).
    - We assume: $\log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$
    - Maximum likelihood method is used to obtain the estimates for the intercept ($\hat{\beta}_0$) and the regression coefficients ($\hat{\beta}$).
- $\hat{\pi}_i = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p}}{1+e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p}}$ can be used to estimate a class membership for each of the samples

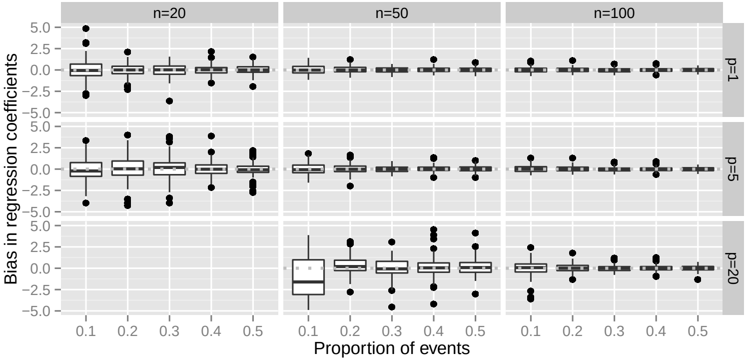
Background

- We are interested in estimating the probability that a rare event will occur, given the characteristics of a subject: $\pi_i = P(Y_i = 1|X_i)$.
- The π_i will be the basis for prediction.
- Logistic regression can be used if the number of considered covariates (p) is reasonably small compared to the number of subjects (n).
  - We assume: $\log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$
  - Maximum likelihood method is used to obtain the estimates for the intercept ($\hat{\beta}_0$) and the regression coefficients ($\hat{\beta}$).
- $\hat{\pi}_i = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p}}{1+e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p}}$ can be used to estimate a class membership for each of the samples
- An event is predicted if $\hat{\pi}_i > \pi$ (marginal event proportion).

Background

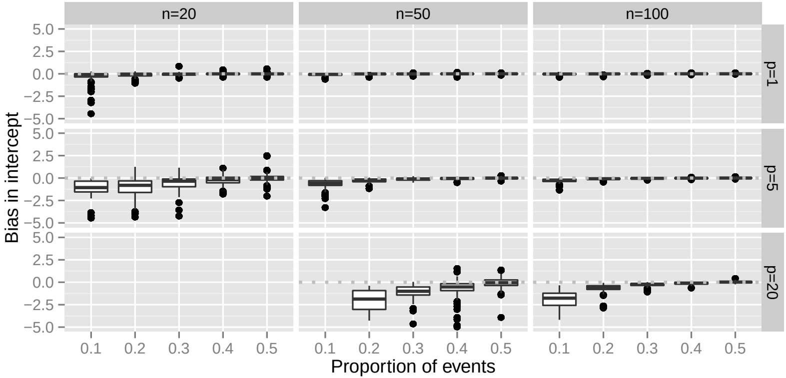
- We are interested in estimating the probability that a rare event will occur, given the characteristics of a subject: $\pi_i = P(Y_i = 1|X_i)$.
- The π_i will be the basis for prediction.
- Logistic regression can be used if the number of considered covariates (p) is reasonably small compared to the number of subjects (n).
  - We assume: $\log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$
  - Maximum likelihood method is used to obtain the estimates for the intercept ($\hat{\beta}_0$) and the regression coefficients ($\hat{\beta}$).
- $\hat{\pi}_i = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p}}{1+e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p}}$ can be used to estimate a class membership for each of the samples
- An event is predicted if $\hat{\pi}_i > \pi$ (marginal event proportion).
- Simple simulations under the null ($\beta = 0$) will be used to explore the properties of some models.

Null case simulation results



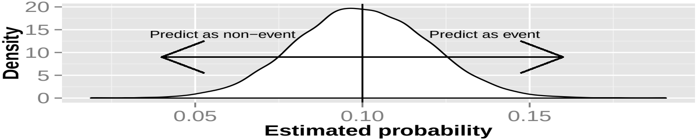
$X \sim N(0,1)$ i.i.d, Y independent from X, $\beta = 0$

Null case simulation results

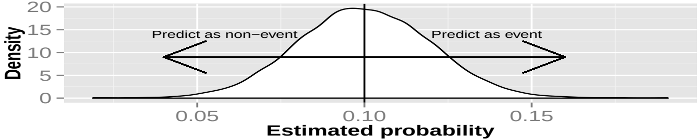


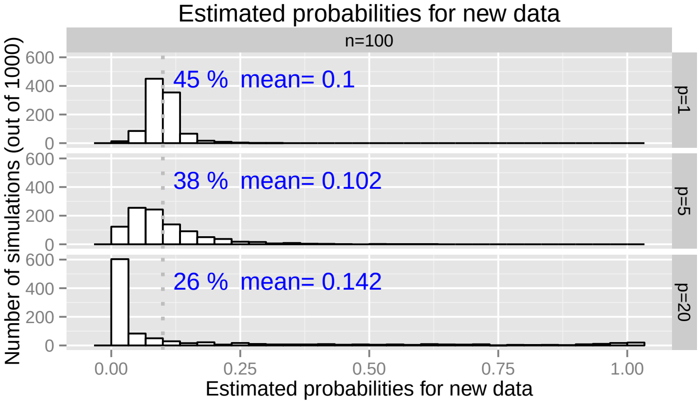
$X \sim N(0,1)$ i.i.d, Y independent from X, $\beta_0 = logit\pi$

Prediction: Can the estimated probabilities be used to predict events for new data?



Prediction: Can the estimated probabilities be used to predict events for new data?





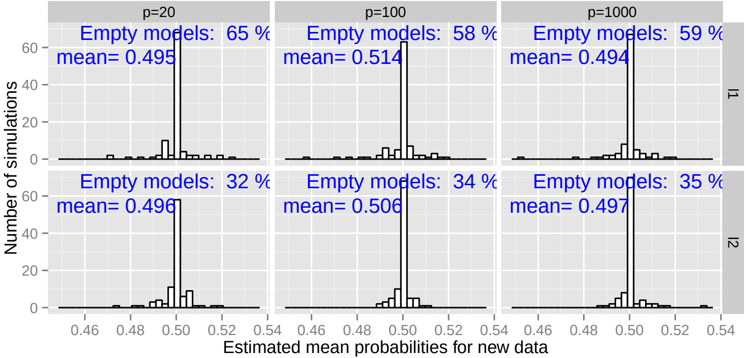
Logistic regression and rare events

- Bad properties when dealing with rare events: biased and imprecise estimates
- Unclear how to use the results from logistic regression for prediction purposes (which threshold?)
- Cannot be used for high-dimensional data ($p > n$)

Logistic regression and rare events

- Bad properties when dealing with rare events: biased and imprecise estimates
- Unclear how to use the results from logistic regression for prediction purposes (which threshold?)
- Cannot be used for high-dimensional data ($p > n$)
- Penalized logistic regression (PLR) with lasso (l1) or ridge penalty (l2) can be used with high-dimensional data and might solve some of the problems observed for logistic regression
- Estimation of PLR in R: `glmnet` or `penalized` package

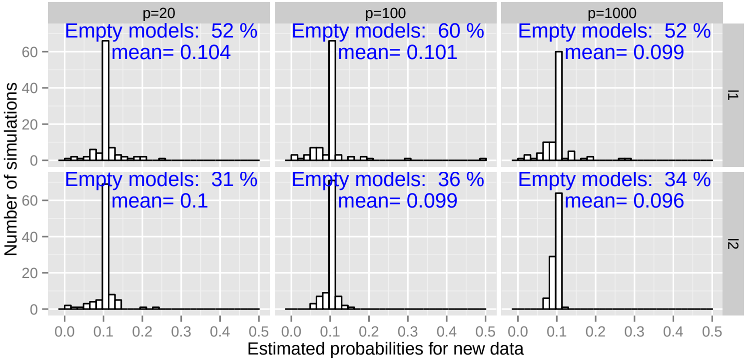
Prediction with penalized models with balanced data ($\pi = 0.50$)



Empty models: all regression coefficients set to zero (l1: exactly, l2: approximately as $\hat{\lambda} \to \infty$).

$X \sim N(0,1)$ i.i.d, Y independent from X, $n_{train} = 100$, $\pi = 0.50$

Prediction with penalized models with rare events ($\pi = 0.10$)

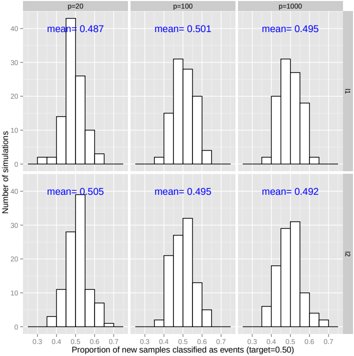


Empty models: all regression coefficients set to zero (l1: exactly, l2: approximately as $\hat{\lambda} \to \infty$).

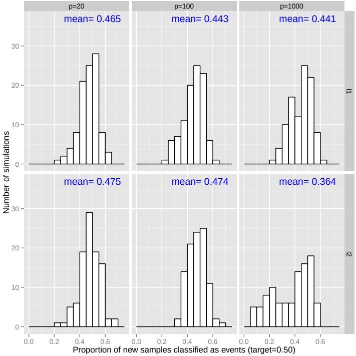
$X \sim N(0, 1)$ i.i.d, Y independent from X, $n_{train} = 100$, $\pi = 0.10$

Proportion of samples classified as events (target=0.50)

Balanced ($\pi = 0.50$)



Rare events ($\pi = 0.10$)



Weighted penalized models

- The likelihood is weighted
  $L(\beta|X) = \prod \pi_i^{y_i w_1} (1 - \pi_i)^{1 - y_i w_0}$

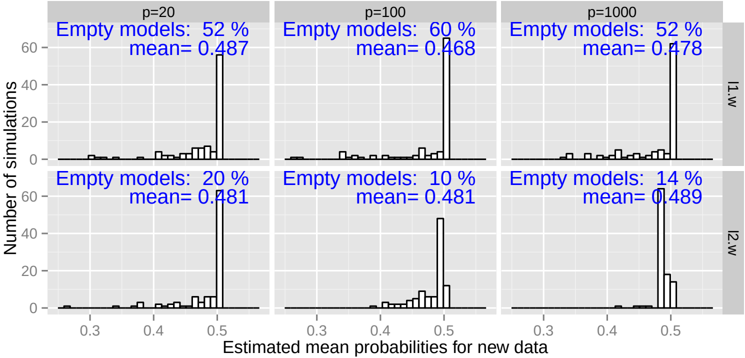
Weighted penalized models

- The likelihood is weighted
  $L(\beta|X) = \prod \pi_i^{y_i w_1}(1-\pi_i)^{1-y_i w_0}$
- $\sum_{i=1}^{n} y_1 w_1 = \sum_{i=1}^{n} (1-y_i) w_0$ gives the same weights to events and non-events.

Weighted penalized models

- The likelihood is weighted
  $L(\beta|X) = \prod \pi_i^{y_i w_1}(1-\pi_i)^{1-y_i w_0}$
- $\sum_{i=1}^{n} y_1 w_1 = \sum_{i=1}^{n} (1-y_i) w_0$ gives the same weights to events and non-events.
- These type of models can be fitted using standard software.

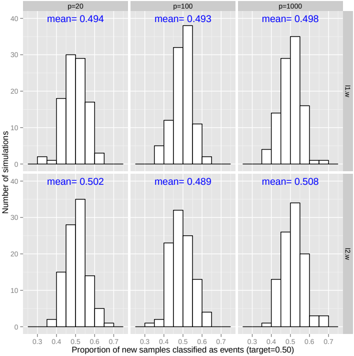
Prediction with weighted penalized models with rare events



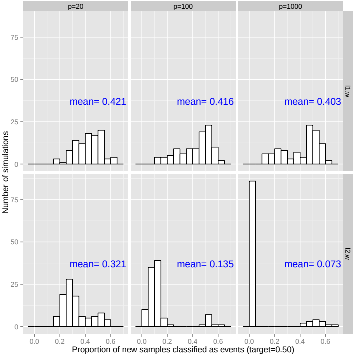
$X \sim N(0,1)$ i.i.d, Y independent from X, $n_{train} = 100$, $\pi = 0.10$

Weighted PLR: Proportion of samples classified as events (target=0.50)

Balanced ($\pi = 0.50$)



Rare events ($\pi = 0.10$)



Conclusions

- Are we really interested in binary event prediction?

Conclusions

- Are we really interested in binary event prediction?
- Under the null l1 identifies the empty model in about 60% of the cases, l2 in about 35% of the cases.

Conclusions

- Are we really interested in binary event prediction?
- Under the null l1 identifies the empty model in about 60% of the cases, l2 in about 35% of the cases.
- l1 performs better than l2 also in the alternative case and in the analyses of real high-dimensional data (like gene-expression microarrays).
- The classification based on l1 and l2 is biased towards the majority class (non-events). l1 is less biased than l2 when the number of variables is large.

Conclusions

- Are we really interested in binary event prediction?
- Under the null l1 identifies the empty model in about 60% of the cases, l2 in about 35% of the cases.
- l1 performs better than l2 also in the alternative case and in the analyses of real high-dimensional data (like gene-expression microarrays).
- The classification based on l1 and l2 is biased towards the majority class (non-events). l1 is less biased than l2 when the number of variables is large.
- Weighted PLR does not seem to increase the accuracy in the prediction of the probability of events

Conclusions

- Are we really interested in binary event prediction?
- Under the null l1 identifies the empty model in about 60% of the cases, l2 in about 35% of the cases.
- l1 performs better than l2 also in the alternative case and in the analyses of real high-dimensional data (like gene-expression microarrays).
- The classification based on l1 and l2 is biased towards the majority class (non-events). l1 is less biased than l2 when the number of variables is large.
- Weighted PLR does not seem to increase the accuracy in the prediction of the probability of events and it increases the bias towards non-event classification, especially for l2.