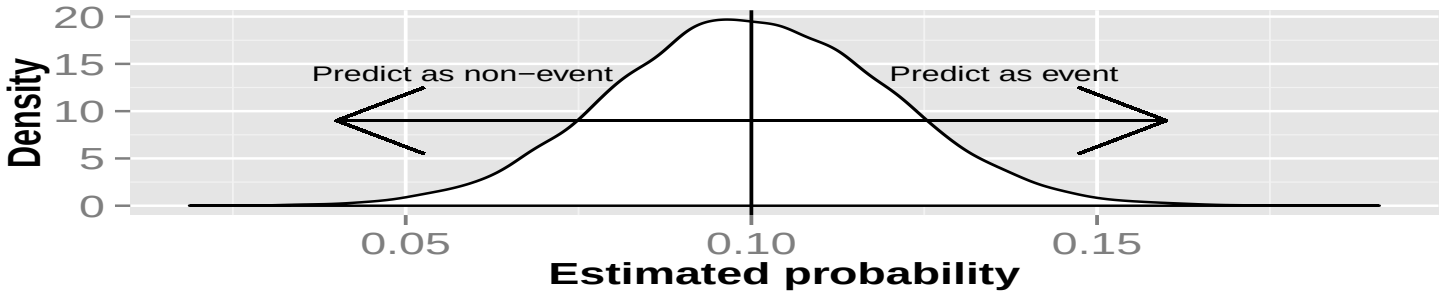
Impact of rare events on predicted probabilities from logistic regression

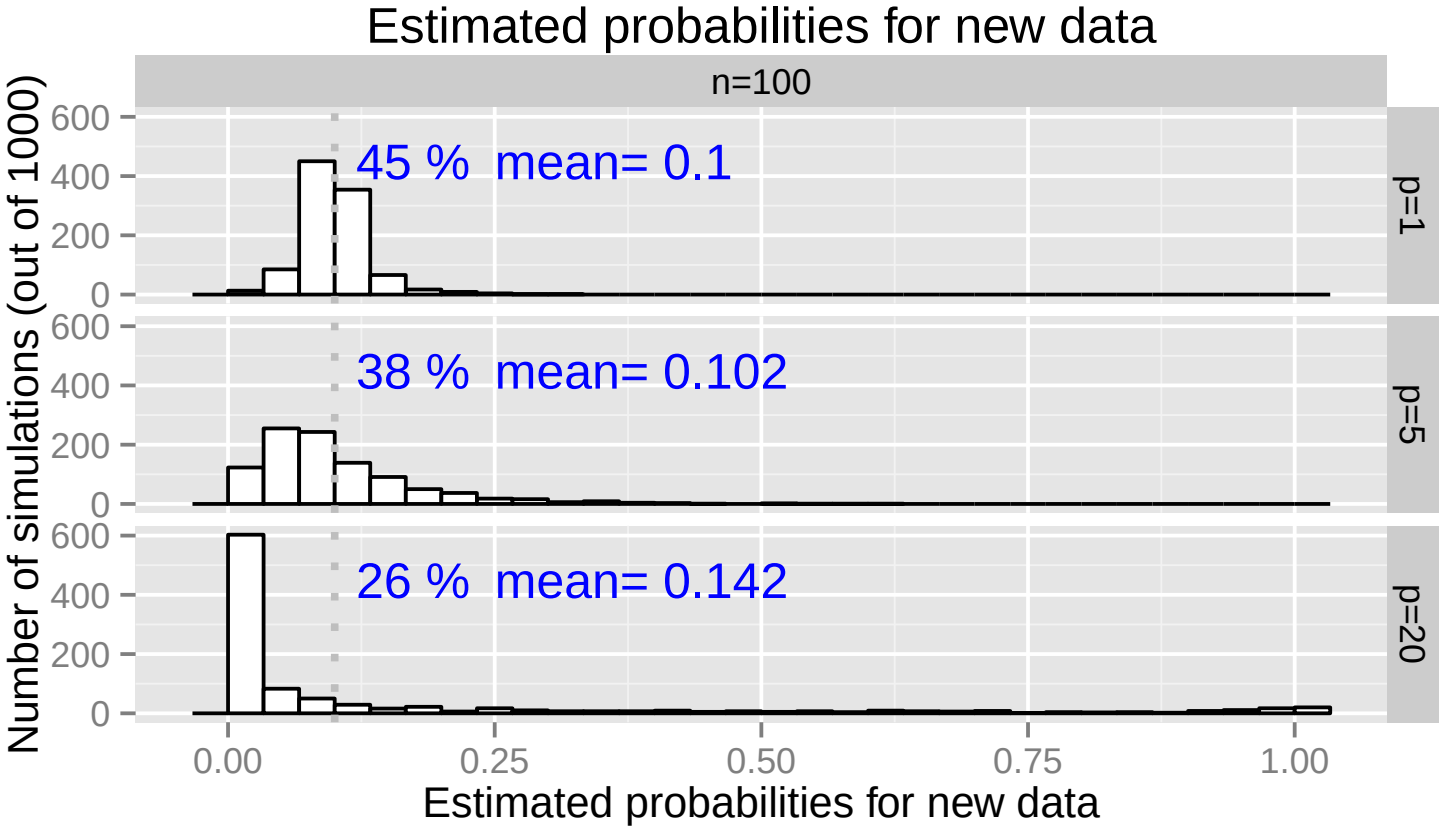
Lara Lusa and Rok Blagus

Institute for Biostatistics and Medical Informatics, University of Ljubljana

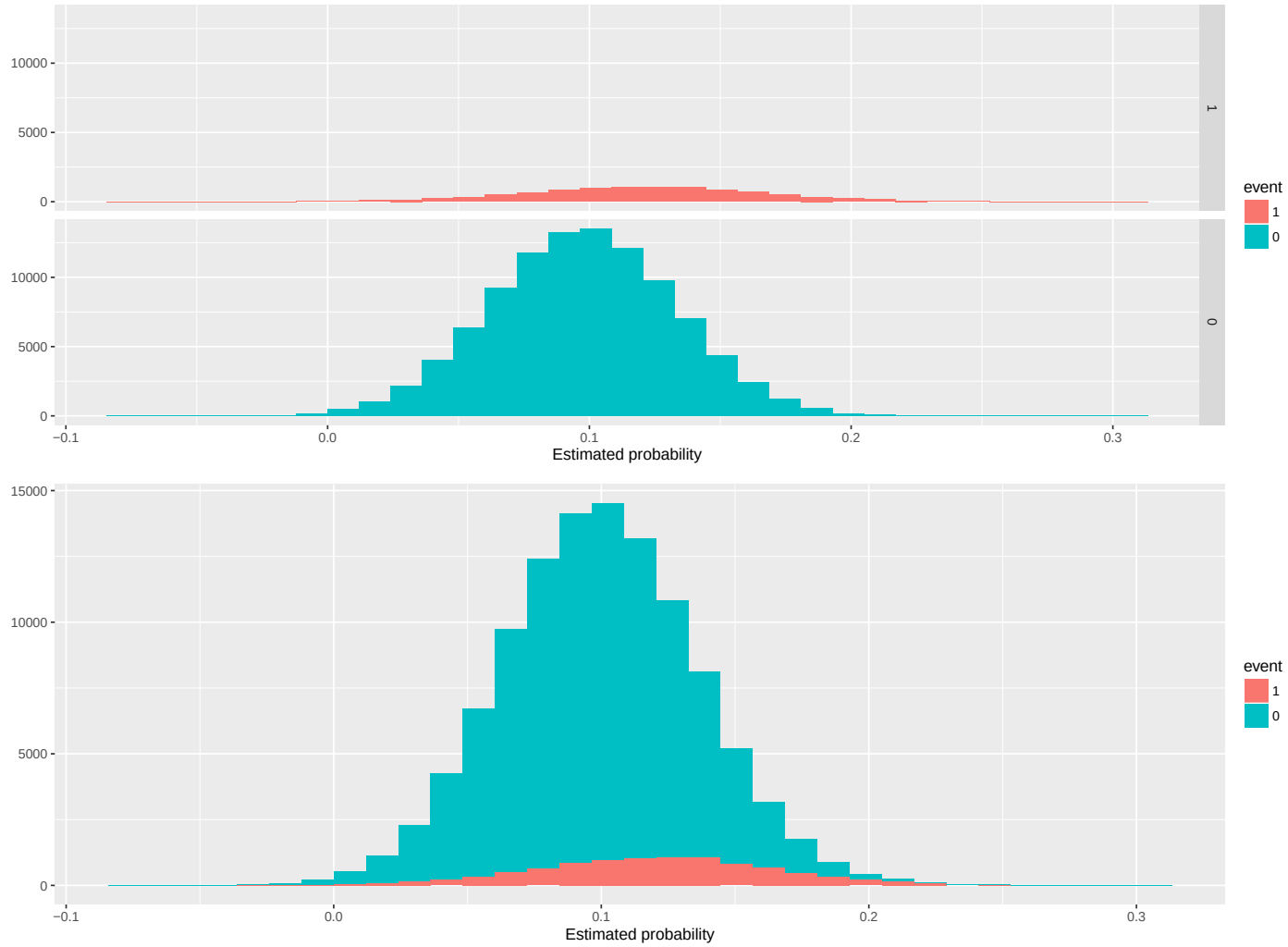
September 2016

From AS2015: Prediction: Can the estimated probabilities be used to predict events for new data?



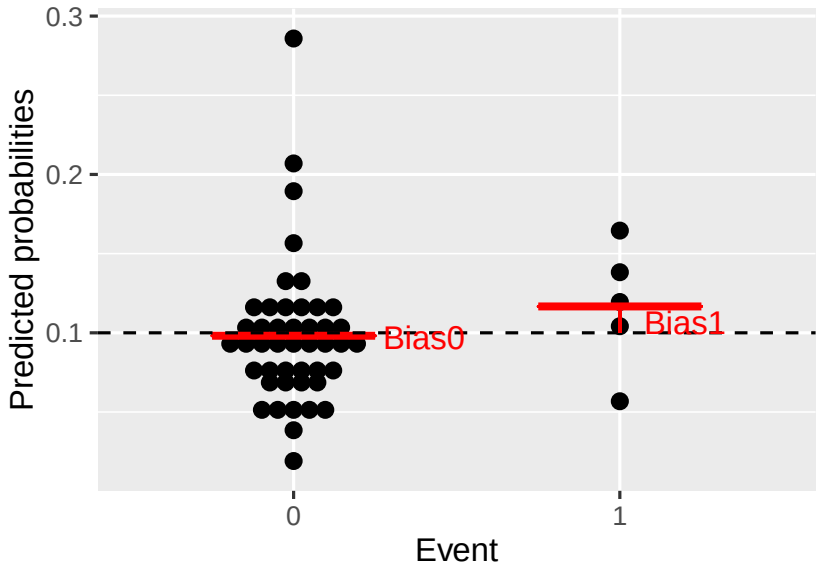


Can we explain the asymmetric distribution of the probabilities (at least in the null case)?



Positive asymmetry: Median $<$ mean $(= \pi) \rightarrow$ the majority of the samples are predicted as non-events when the marginal event probability is used as the classification threshold

Can we explain how rare events influence the predicted probabilities?

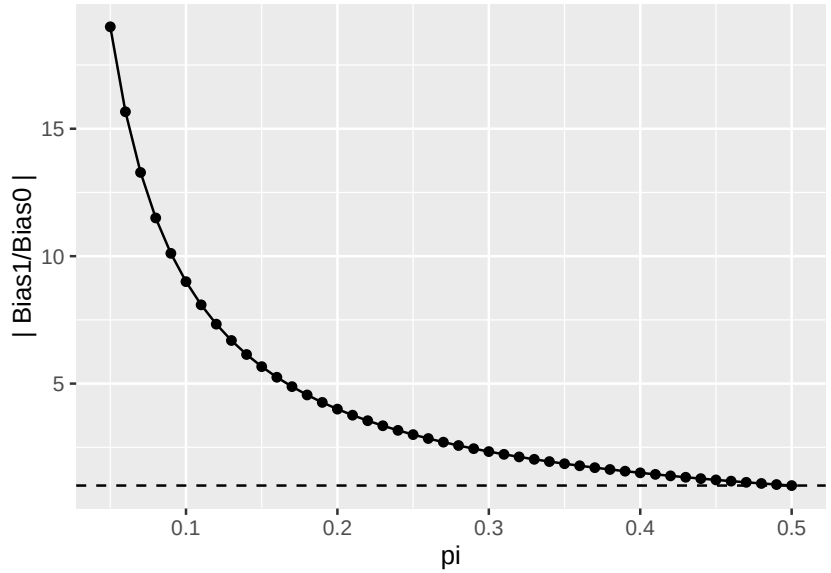


overall	events	non-events
$E(\hat{\pi}) = \pi$	$E(\hat{\pi}_1) = E(\pi \mid Y = 1)$	$E(\hat{\pi}_0) = E(\pi \mid Y = 0)$
$Bias(\hat{\pi}) = E(\hat{\pi} - \pi) = 0$	$Bias_1(\hat{\pi}) = E(\hat{\pi}_1 - \pi)$	$Bias_0(\hat{\pi}) = E(\hat{\pi}_0 - \pi)$

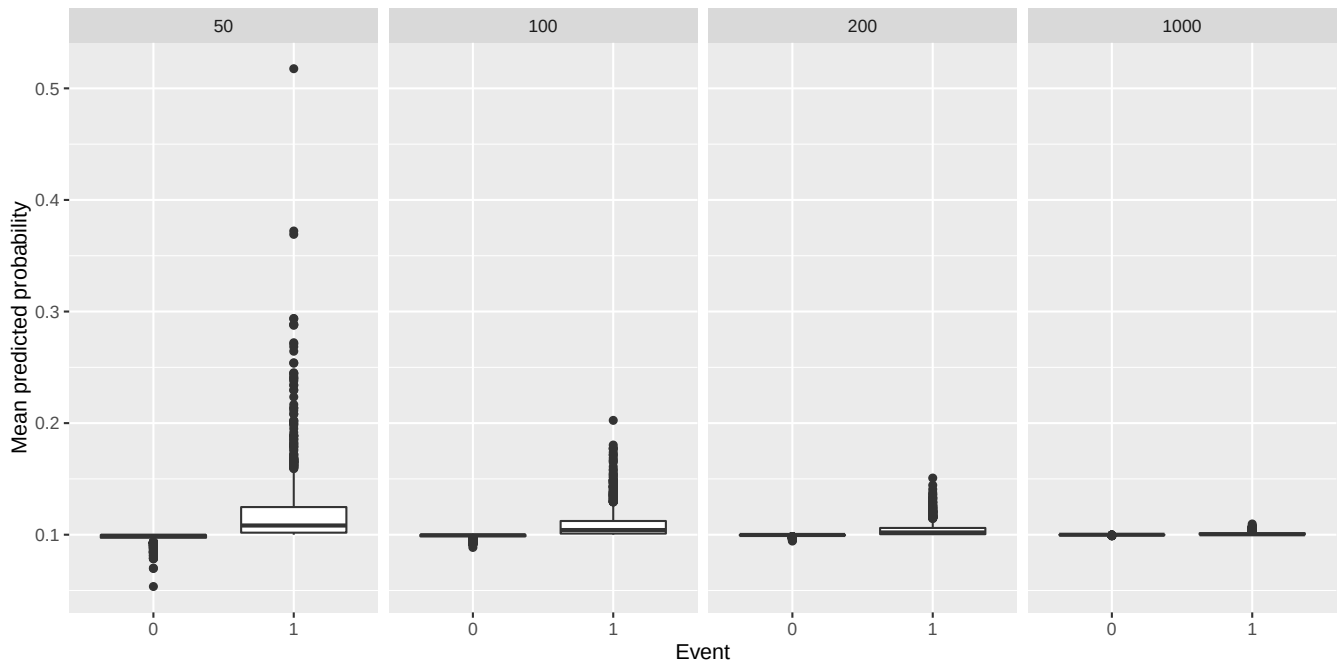
$$\frac{Bias_1(\hat{\pi})}{Bias_0(\hat{\pi})} = \frac{\pi - 1}{\pi}$$

Can we explain how rare events influence the predicted probabilities?

Varying π



Varying n



With categorical covariates the proportion of samples predicted as events can be derived (under the null)

One binary covariate

Example ($n = 100, \pi = 0.10$)

	$x = 0$ ($n = 50$)	$x = 1$ ($n = 50$)
Observed proportion of events	$4/50$	$6/50$
Estimated probability of events	$4/50$ (k_0/n_0)	$6/50$ (k_1/n_1)
Estimated class (event is predicted if $\hat{\pi}_i > k/n = 5/50$)	0	1

Exactly 50% will be classified as events (no bias)

Also for this example: $\frac{Bias_1}{Bias_0} = -9$

$P(\hat{Y} = 1|Y = 1) = \frac{4/50 \cdot 4 + 6/50 \cdot 6}{10} = 0.104$

$P(\hat{Y} = 1|Y = 0) = \frac{4/50 \cdot 46 + 6/50 \cdot 44}{90} = 0.0995$

With categorical covariates the proportion of samples predicted as events can be derived (under the null)

Example ($n = 100, \pi = 0.10$)

	$x_0 = 0, x_2 = 0$ ($n = 25$)	$x_0 = 1, x_2 = 0$ ($n = 25$)	$x_0 = 0, x_2 = 1$ ($n = 25$)	$x_0 = 1, x_2 = 1$ ($n = 25$)
Observed	$4/25$ (k_{00}/n_{00})	$3/25$	$2/25$	$1/25$
Estimated $\hat{\pi}$	$4/25$=0.16 ($\hat{\pi}_{00}$)	$3/25$=0.12	$2/25$=0.08	$1/25$=0.04
Estimated class $\hat{\pi}_i > k/n = 10/100$	1	1	0	0

With categorical covariates the proportion of samples predicted as events can be derived (under the null)

Example ($n = 100, \pi = 0.10$)

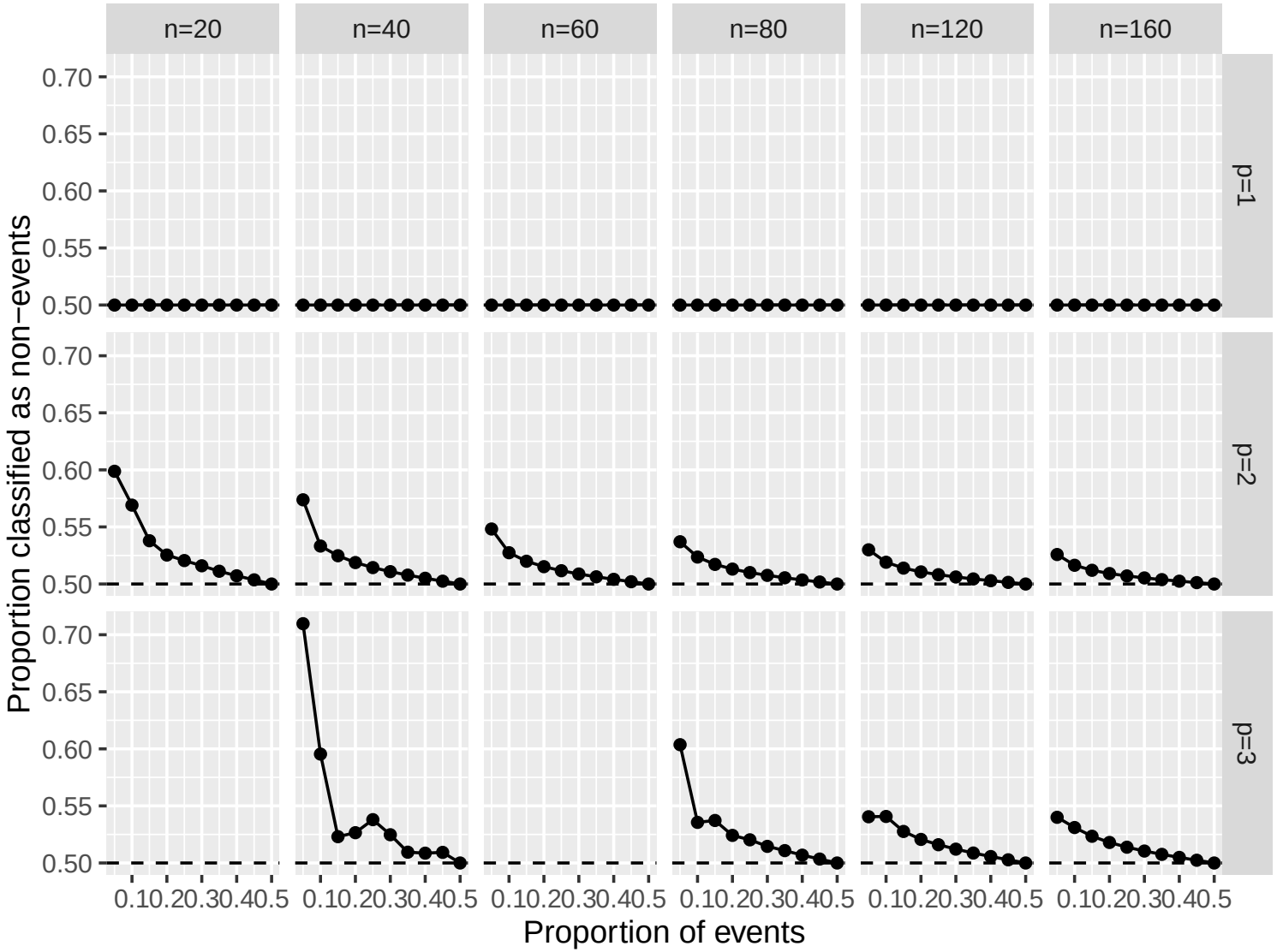
	$x_0 = 0, x_2 = 0$ ($n = 25$)	$x_0 = 1, x_2 = 0$ ($n = 25$)	$x_0 = 0, x_2 = 1$ ($n = 25$)	$x_0 = 1, x_2 = 1$ ($n = 25$)
Observed	4/25 (k_{00}/n_{00})	3/25	2/25	1/25
Estimated $\hat{\pi}$	4/25=0.16 ($\hat{\pi}_{00}$)	3/25=0.12	2/25=0.08	1/25=0.04
Estimated class $\hat{\pi}_i > k/n = 10/100$	1	1	0	0

$$P(\hat{Y} = 1 | x_1 = i, x_2 = j) = P(\hat{\pi}_{ij} > K/n) + 1/2P(\hat{\pi}_{ij} = K/n)$$

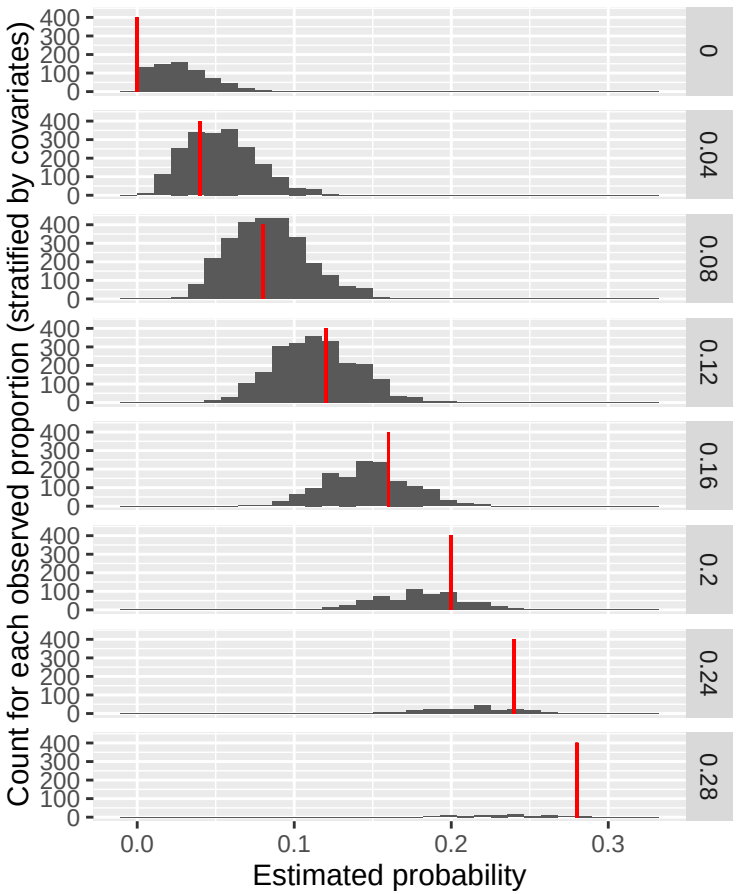
$$= P(K_{ij} > K/n) + 1/2P(K_{ij} = K/n) = f(\pi, n, n_{ij}(p))$$

- $\hat{\pi}_{ij} = k_{ij}/n_{ij}$
- $K \sim Bin(n, \pi)$, $K_{ij} \sim Bin(n_{ij}, \pi)$, $K_{-ij} = K - K_{ij} \sim Bin(n - n_{ij}, \pi)$
- $K_{ij}/n_{ij} = K/n = \frac{K_{00}+K_{01}+K_{10}+K_{11}}{n} \sim K_{ij} = \frac{n_{ij}}{n} K_{-ij}$
- $P(\hat{\pi}_{ij} = K/n) = P(K_{ij}/n_{ij} = K/n) = P(K_{ij} = n_{ij}/nK_{-ij}) = \sum_{k_{-ij}=0}^{n-n_{ij}} P(K_{ij} = n_{ij}/nK_{-ij} | K_{-ij} = k_{-ij}) P(K_{-ij} = k_{-ij})$
- $P(\hat{\pi}_{ij} > K/n) = P(K_{ij} > n_{ij}/nK) = P(K_{ij} > n_{ij}/nK_{-ij}) = \sum_{k_{-ij}=0}^{n-n_{ij}} P(K_{ij} > n_{ij}/nk_{-ij} | K_{-ij} = k_{-ij}) P(K_{-ij} = k_{-ij})$

Classification results (null case)



Does the problem persist also in models with main effects only?



- The problem is even more severe (more bias towards non-events)
- The estimated probabilities in each cell are not equal to the observed proportion
- Compared to the saturated model: there are less ties, $Bias_1$ is smaller (less overfitting)
- Predicted events $\rightarrow$ non-events is more common than the opposite
- On average the estimated probabilities are correct for each combination of the covariates

Conclusions

- The overfitting and the rare events problems are closely related
- Categorical covariates (and saturated models) allow the derivation of theoretical results, which do not seem obtainable with continuous covariates.
- All the examples used few variables. Increasing the number of variables all the problems become more severe.
- The presented theoretical results can be helpful for understanding and isolating the sources of the rare event bias
- We hope to be able to use these results to propose approaches that can diminish the rare event bias