Balance and Calibration of Probabilistic Scores
From GLM to Machine Learning

Arthur Charpentier
(with Agathe Fernandes Machado, Emmanuel Flachaire,
Ewen Gallic, François Hu)

Agenda and motivation

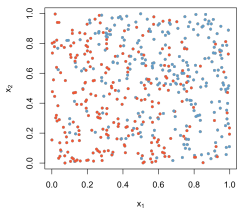
1. The logistic regression and classical probabilistic classifiers
   - Logistic regression (and Generalized Linear Models)
   - Epistemology of probabilistic interpretation with single case events
2. Balance and calibration
   - Calibration curves and measures
   - Using calibration curbes for recalibration
3. Application on real data (motor insurance)
   - Real data: accuracy, calibration
   - Synthetic data: accuracy, calibration and distance to grand truth

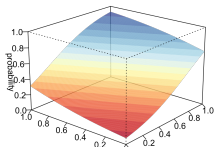
Based on [Fernandes Machado et al., 2024a, Fernandes Machado et al., 2024b, Fernandes Machado et al., 2024c]

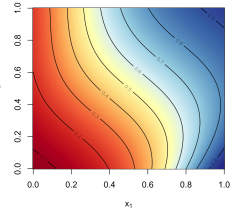
Scores from classical Models: a simple example

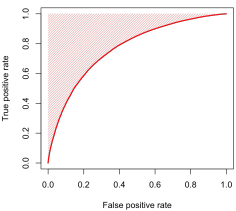
plain linear — nonlinear component

$$(y_i, x_{1,i}, x_{2,i}), \text{ where } \mu(x_1, x_2) = \frac{\exp[\, x_1 + x_2 + \psi(x_1, x_2)\,]}{1 + \exp[\, x_1 + x_2 + \psi(x_1, x_2)\,]},$$









scatterplot $(y_i, x_{1,i}, x_{2,i})$, where $\mathbb{P}[Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}] = \mu(\boldsymbol{x}) = 1 - \mathbb{P}[Y = 0 \mid \boldsymbol{X} = \boldsymbol{x}]$

$$\text{AUC} = \int_{\mathbb{R}} \text{ROC}(t)\,\mathrm{d}t, \text{ where } \begin{cases} \overline{F}_0(p) = \mathbb{P}[s(\boldsymbol{X}) > p \mid Y = 0] = \text{FPR} \\ \overline{F}_1(p) = \mathbb{P}[s(\boldsymbol{X}) > p \mid Y = 1] = \text{TPR} \end{cases}$$

$$\text{ROC} : t \mapsto \overline{F}_1 \circ \overline{F}_0^{-1}(t)$$

Scores from classical Models: logistic regression

$\{(y_i, \boldsymbol{x}_i)\}$ sample, realizations of i.i.d. vectors $(Y_i, \boldsymbol{X}_i)$ where $(Y \mid \boldsymbol{X} = \boldsymbol{x}) \sim \mathcal{B}(p(\boldsymbol{x}))$,

$$\text{where } \mathbb{P}[Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}] = p(\boldsymbol{x}) = \frac{e^{\boldsymbol{x}^\top \beta}}{1 + e^{\boldsymbol{x}^\top \beta}} = s_{\text{logistic}}(\boldsymbol{x})$$

Inference: **maximum likelihood**,

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \boldsymbol{x}_i^\top \beta \text{ or } p_i = s_\beta(\boldsymbol{x}_i) = \text{logit}^{-1}(\boldsymbol{x}_i^\top \beta) = \frac{\exp[\boldsymbol{x}_i^\top \beta]}{1 + \exp[\boldsymbol{x}_i^\top \beta]}$$

$$\log \mathcal{L}(\beta) = \sum_{i=1}^{n} \underbrace{y_i \log(p_i(\beta)) + (1 - y_i) \log(1 - p_i(\beta))}_{-\ell(y_i, p_i)}$$

Scores from classical Models: logistic regression

First order conditions yield

$$\frac{\partial \log \mathcal{L}(\beta)}{\partial \beta_k}\bigg|_{\beta=\widehat{\beta}} = \sum_{i=1}^{n} \frac{y_i}{p_i(\beta)} \frac{\partial p_i(\beta)}{\partial \beta_k} - \frac{1-y_i}{p_i(\beta)} \frac{\partial p_i(\beta)}{\partial \beta_k} = 0$$

or because of the analytical expression of $p_i(\beta)$,

$$\frac{\partial p_i(\beta)}{\partial \beta_k} = p_i(\beta)[1-p_i(\beta)]x_{k,i}$$

we obtain

$$\boldsymbol{X}^\top(\boldsymbol{y}-\widehat{\boldsymbol{p}}) = \boldsymbol{0}$$

$$\frac{\partial \log \mathcal{L}(\beta)}{\partial \beta_k}\bigg|_{\beta=\widehat{\beta}} = \sum_{i=1}^{n} x_{k,i}[y_i - p_i(\widehat{\beta})] = 0 \ , \ \forall k.$$

Scores from classical Models: logistic regression

Side note :

$$\left.\frac{\partial \log \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_0}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} = 0 \iff \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} p_i(\widehat{\boldsymbol{\beta}})$$

This corresponds to **balance in-sample**

(more generally, MLE in GLMs under the canonical link ensures balance in-sample)

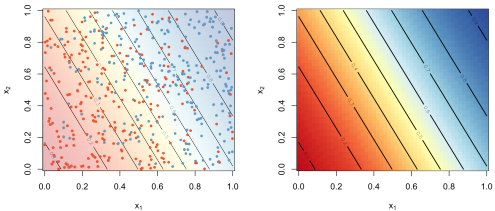
Out of sample, set $D = \sum_{i=1}^{m} (Y_i - s_{\widehat{\boldsymbol{\beta}}_n}(\boldsymbol{X}_i))$, using Taylor's expansion,

$$\mathbb{E}[D \mid \widehat{\boldsymbol{\beta}}_n] \approx -\sum_{i=1}^{m} s_{\widehat{\boldsymbol{\beta}}_n}(\boldsymbol{X}_i)(1 - s_{\widehat{\boldsymbol{\beta}}_n}(\boldsymbol{X}_i))\boldsymbol{X}_i^\top(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}).$$

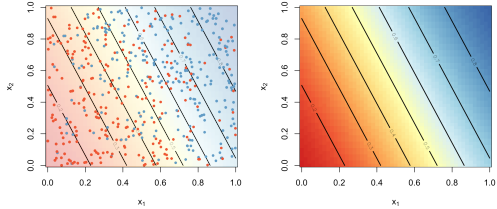
$$\mathrm{Var}(D \mid \widehat{\boldsymbol{\beta}}_n) = \sum_{i=1}^{m} s_{\widehat{\boldsymbol{\beta}}_n}(\boldsymbol{X}_i)(1 - s_{\widehat{\boldsymbol{\beta}}_n}(\boldsymbol{X}_i)).$$

Scores from classical Models: logistic regression

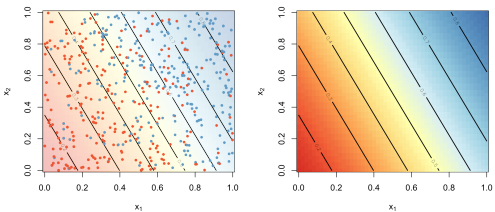
- (plain) Logistic



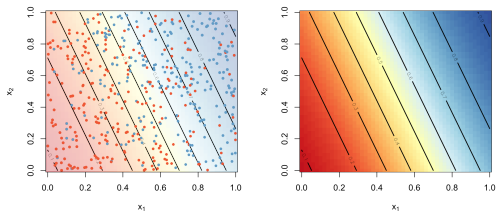
- Logistic with **lasso** (ℓ_1 penalty)



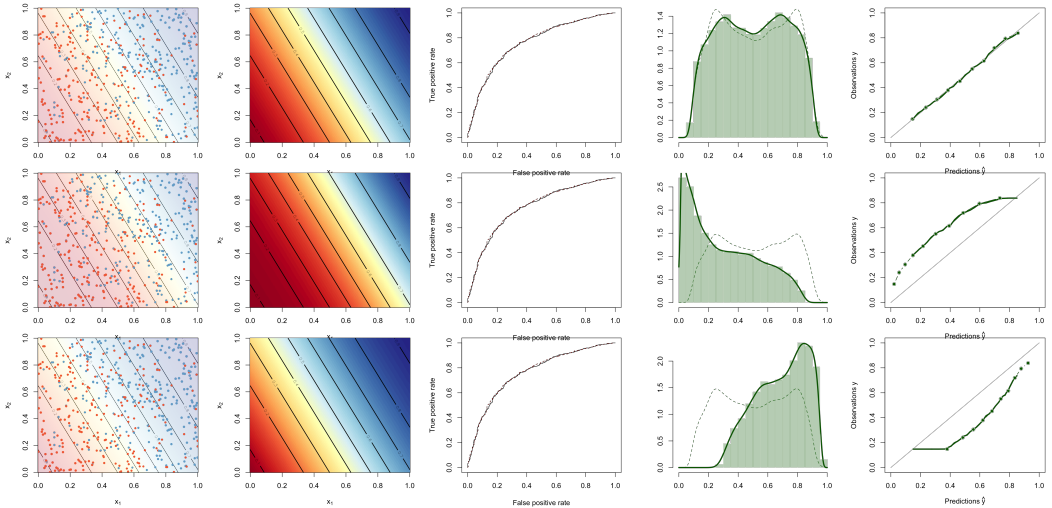
- Logistic with **Ridge** (ℓ_2 penalty)



- **Support Vector Machine** (SVM) plain vanilla

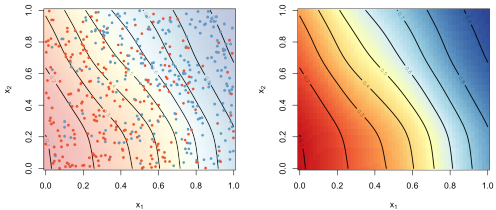


Scores from classical Models: logistic, p, p^2, $\sqrt{p}$

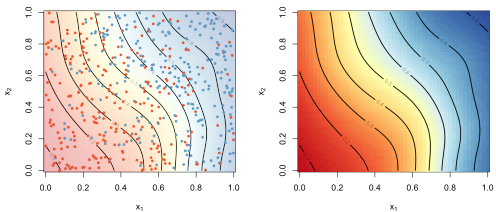


Scores from classical Models: GAM

- Logistic **GAM** with additive splines



- Logistic **GAM** with bivariate splines



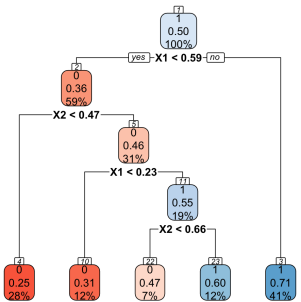
GAM (Generalized Addive Models)
$\mathbb{E}[Y \mid X_1 = x_1, X_2 = x_2]$ is here

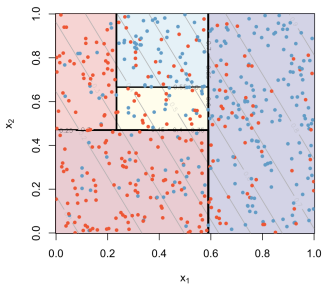
$$\frac{\exp(\beta_0 + \psi_1(x_1) + \psi_2(x_2))}{1 + \exp(\beta_0 + \psi_1(x_1) + \psi_2(x_2))}.$$

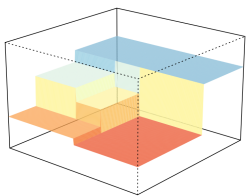
where

$$\begin{aligned}\psi_j(x) &= a_{j1} + a_{j2}x + a_{j3}x^2 + a_{j4}x^3 \\ &\quad + a_{j5}\,(x - s_{1j})_+^3 + a_{j6}\,(x - s_{2j})_+^3\end{aligned}$$

Scores from classical Models: Trees

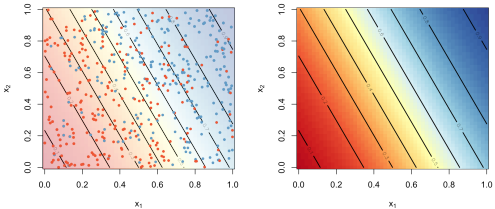






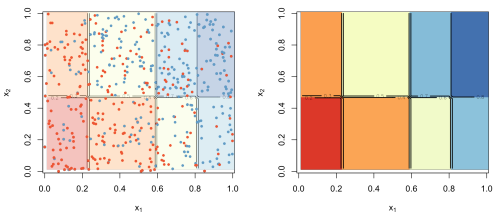
Scores from classical Models: Trees

- Linear **discriminant analysis**

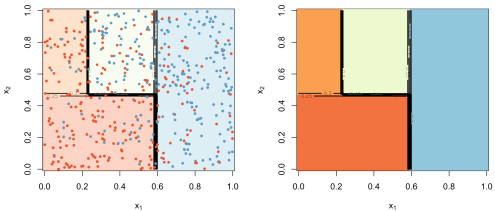


- Logistic with **categorical variables**

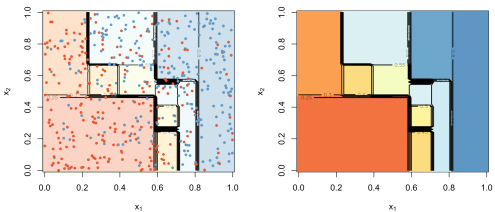
  (cut, $x_{j,k} = \mathbf{1}(x_j \in [a_k, a_{k+1}))$)



- **Classification Tree (1)**



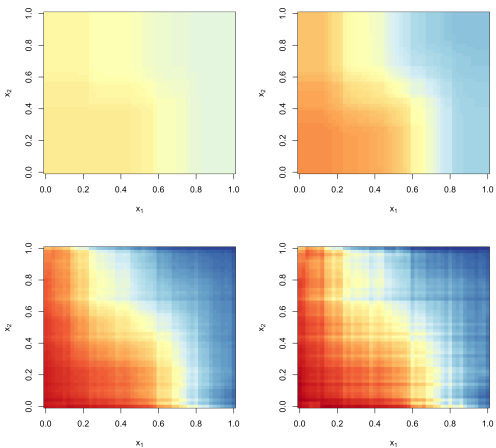
- **Classification Tree (2)**

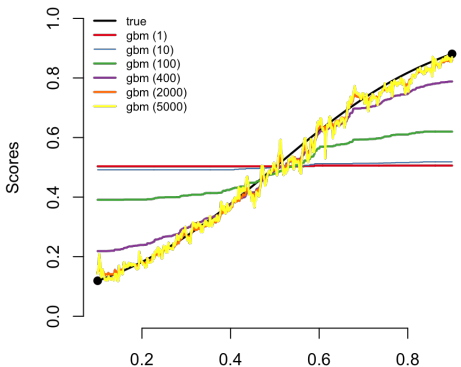


Scores from classical Models: Boosting

- **Ensemble learning**

  Boosting (sequential trees)

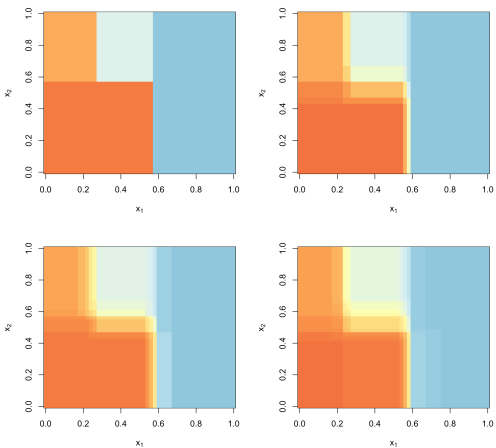


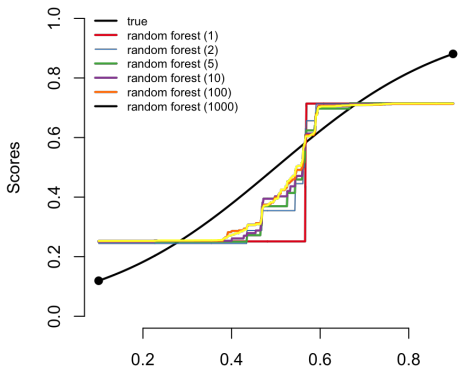


Scores from classical Models: Random Forest

- **Ensemble learning**

  Bagging (bootstrap+trees)

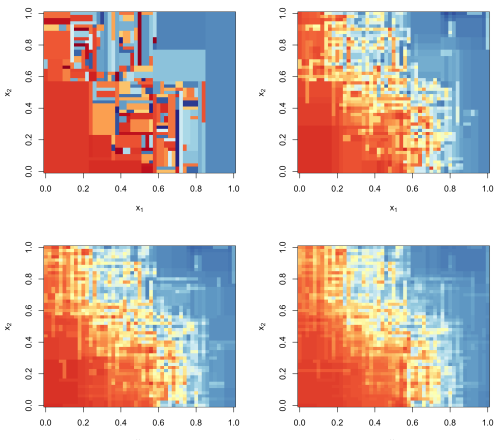


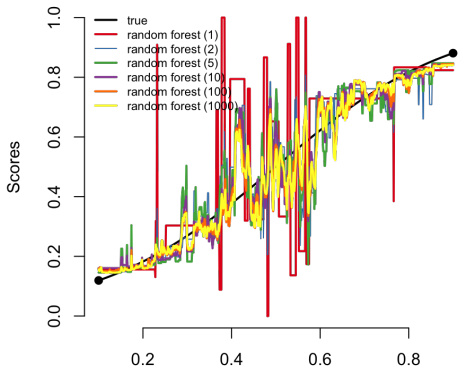


Scores from classical Models: Random Forest

- **Ensemble learning**

  Bagging (bootstrap+trees)

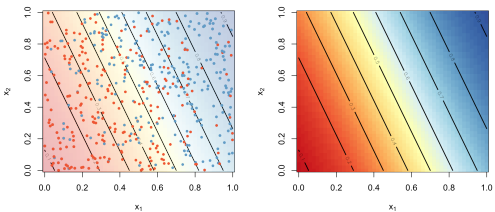




Scores from classical Models: Random Forest

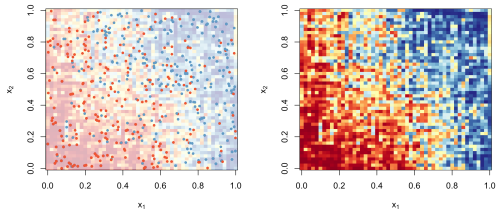
- **Support Vector Machine**

  (SVM) plain vanilla

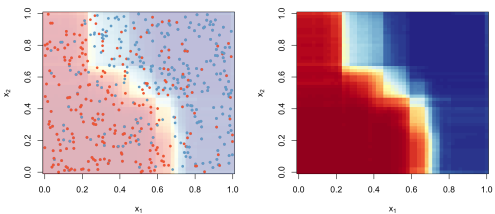


- **Classification Random Forest**

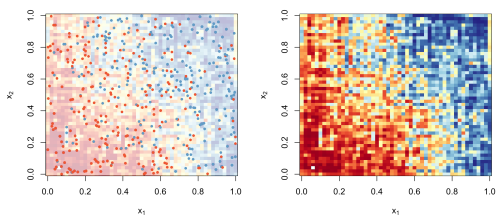
  with maximum nodes option



- **Classification Random Forest**

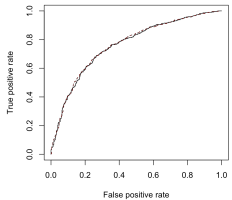


- **Regression Random Forest**

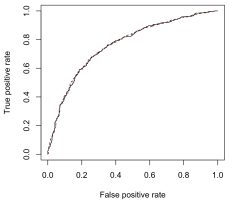


Accuracy of scores, and predictions

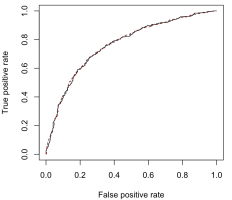
logistic 0.76197



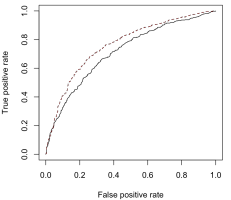
GAM 0.76290



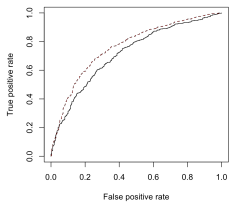
GAM 0.76294



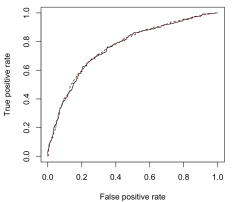
random forest 0.7148



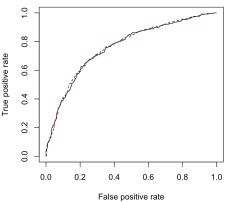
random forest 0.71684



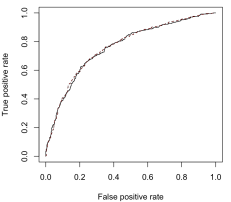
boosting 0.76151



boosting 0.76151

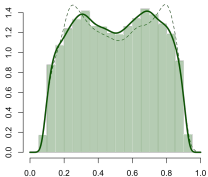


boosting 0.76151

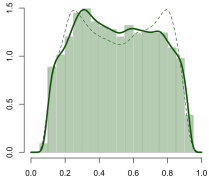


Distributions of scores

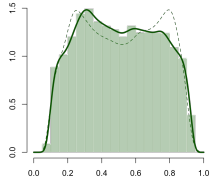
logistic



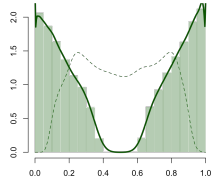
GAM



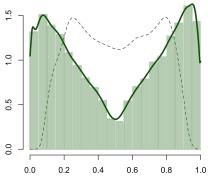
GAM



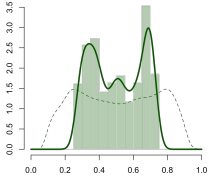
random forest



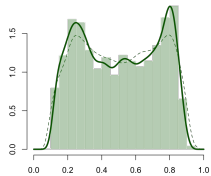
random forest



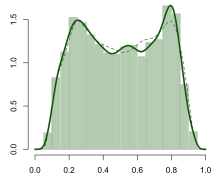
boosting



boosting



boosting



Scores?

"*The individual characteristics are an essential part of any model for individual risk assessment. Their statistical summary is called a score*," [Gourieroux and Jasiak, 2015]

In the context of a logistic regression, the "*canonical score*" is $s : \mathbb{R}^k \to [0, 1]$,

$$s(\boldsymbol{x}) := \text{logit}(\boldsymbol{x}^\top \boldsymbol{\beta}) = \frac{\exp[\boldsymbol{x}^\top \boldsymbol{\beta}]}{1 + \exp[\boldsymbol{x}^\top \boldsymbol{\beta}]}.$$

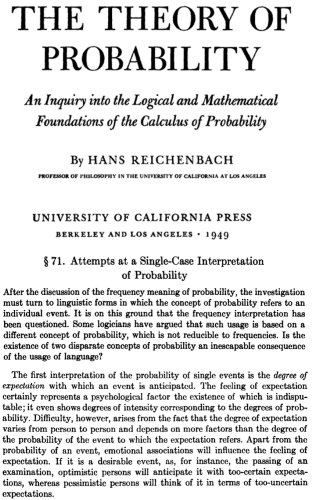
Regression function

The **regression function** is $\mu(\boldsymbol{x}) := \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$.

Scores?

"*If we are asked to find the probability holding for an individual future event, we must first incorporate the case in a suitable reference class*," [Reichenbach, 1971]

"*When we speak of the 'probability of death', the exact meaning of this expression can be defined in the following way only.* **We must not think of an individual, but of a certain class as a whole**, *e.g., 'all insured men forty-one years old living in a given country and not engaged in certain dangerous occupations'. A probability of death is attached to the class of men or to another class that can be defined in a similar way.* **The phrase 'probability of death', when it refers to a single person, has no meaning for us at all**," [von Mises, 1928, von Mises, 1939]



THE THEORY OF PROBABILITY

An Inquiry into the Logical and Mathematical Foundations of the Calculus of Probability

By HANS REICHENBACH

PROFESSOR OF PHILOSOPHY IN THE UNIVERSITY OF CALIFORNIA AT LOS ANGELES

UNIVERSITY OF CALIFORNIA PRESS

BERKELEY AND LOS ANGELES · 1949

§ 71. Attempts at a Single-Case Interpretation of Probability

After the discussion of the frequency meaning of probability, the investigation must turn to linguistic forms in which the concept of probability refers to an individual event. It is on this ground that the frequency interpretation has been questioned. Some logicians have argued that such usage is based on a different concept of probability, which is not reducible to frequencies. Is the existence of two disparate concepts of probability an inescapable consequence of the usage of language?

The first interpretation of the probability of single events is the *degree of expectation* with which an event is anticipated. The feeling of expectation certainly represents a psychological factor the existence of which is indisputable; it even shows degrees of intensity corresponding to the degrees of probability. Difficulty, however, arises from the fact that the degree of expectation varies from person to person and depends on more factors than the degree of the probability of the event to which the expectation refers. Apart from the probability of an event, emotional associations will influence the feeling of expectation. If it is a desirable event, as, for instance, the passing of an examination, optimistic persons will anticipate it with too-certain expectations, whereas pessimistic persons will think of it in terms of too-uncertain expectations.

Scores?

As explained in [Van Calster et al., 2019], "*among patients with an estimated risk of 20%, we expect 20 in 100 to have or to develop the event*,"

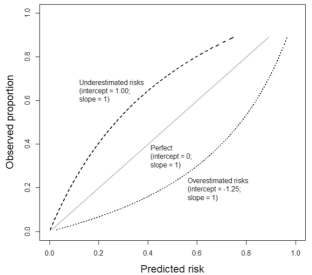
- If 40 out of 100 in this group are found to have the disease, the risk is underestimated
- If we observe that in this group, 10 out of 100 have the disease, we have overestimated the risk.

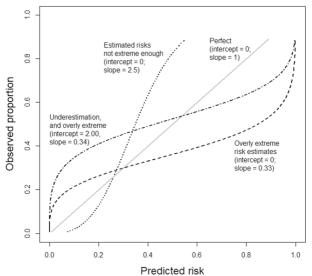
The prediction $\widehat{m}(\boldsymbol{X})$ of Y is a well-calibrated prediction if

20 out of 100 (proportion $y = 1$)

$$\mathbb{E}\big[\, Y \,\big|\, \widehat{Y} = \widehat{y} \,\big] = \widehat{y}, \ \forall \widehat{y}$$

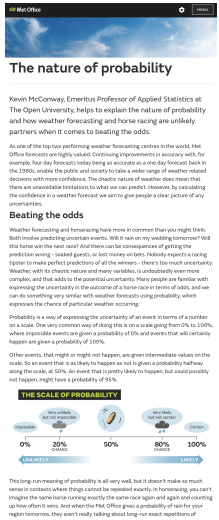
estimate risk $\widehat{y} = 20\%$





Scores?

"*Suppose the Met Office says that the probability of rain tomorrow in your region is 80%. They aren't saying that it will rain in 80% of the land area of your region, and not rain in the other 20%. Nor are they saying it will rain for 80% of the time. What they are saying is there is an 80% chance of rain occurring at any one place in the region, such as in your garden.* [...] *A forecast of 80% chance of rain in your region should broadly mean that, on about 80% of days when the weather conditions are like tomorrow's, you will experience rain where you are. [...] If it doesn't rain in your garden tomorrow, then the 80% forecast wasn't wrong, because it didn't say rain was certain. But* **if you look at a long run of days, on which the Met Office said the probability of rain was 80%, you'd expect it to have rained on about 80% of them.**" [McConway, 2021]


Met Office

MENU

The nature of probability

Kevin McConway, Emeritus Professor of Applied Statistics at The Open University, helps to explain the nature of probability and how weather forecasting and horse racing are unlikely partners when it comes to beating the odds.

As one of the top two performing weather forecasting centres in the world, Met Office forecasts are highly valued. Continuing improvements in accuracy with, for example, four-day forecasts today being as accurate as a one-day forecast back in the 1980s, enable the public and society to take a wider range of weather related decisions with more confidence. The chaotic nature of weather does mean that there are unavoidable limitations to what we can predict. However, by calculating the confidence in a weather forecast we aim to give people a clear picture of any uncertainties.

Beating the odds

Weather forecasting and horseracing have more in common than you might think. Both involve predicting uncertain events. Will it rain on my wedding tomorrow? Will this horse win the next race? And there can be consequences of getting the prediction wrong - soaked guests, or lost money on bets. Nobody expects a racing tipster to make perfect predictions of all the winners – there's too much uncertainty. Weather, with its chaotic nature and many variables, is undoubtedly even more complex, and that adds to the potential uncertainty. Many people are familiar with expressing the uncertainty in the outcome of a horse race in terms of odds, and we can do something very similar with weather forecasts using probability, which expresses the chance of particular weather occurring.

Probability is a way of expressing the uncertainty of an event: in terms of a number on a scale. One very common way of doing this is on a scale going from 0% to 100%, where impossible events are given a probability of 0% and events that will certainly happen are given a probability of 100%.

Other events, that might or might not happen, are given intermediate values on the scale. So an event that is as likely to happen as not is given a probability halfway along the scale, at 50%. An event that is pretty likely to happen, but could possibly not happen, might have a probability of 95%.

This long-run meaning of probability is all very well, but it doesn't make so much sense in contexts where things cannot be repeated exactly. In horseracing, you can't imagine the same horse running exactly the same race again and again and counting up how often it wins. And when the Met Office gives a probability of rain for your region tomorrow, they aren't really talking about long-run exact repetitions of tomorrow. Tomorrow's only going to happen once.

Scores: Balance

premium collected

losses paid

(Theoretical) Global balance, $\mathbb{E}\big[\widehat{s}(\boldsymbol{X})\big] = \mathbb{E}\big[Y\big]$

$$\mathbb{E}[Y] = \mathbb{E}[\widehat{s}(\boldsymbol{X})] \iff \mathbb{E}[Y - \widehat{s}(\boldsymbol{X})] = \mathbb{E}[\mu(\boldsymbol{x}) - \widehat{s}(\boldsymbol{X})] = 0.$$

Economically, if $\widehat{s}(\boldsymbol{x})$ is the price, the portfolio is self-financing (for random losses Y).

Empirical global balance,

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{s}(\boldsymbol{x}_i).$$

If $\widehat{s}$ is obtained with a logistic regression*, this is valid, on the training dataset.

* estimated using maximum likelihood techniques

Scores: Balance

Well-calibration (or "marginal balance", w.r.t. $\widehat{s}(\boldsymbol{x})$)

$$\mathbb{E}\big[Y - \widehat{s}(\boldsymbol{X}) \mid \widehat{s}(\boldsymbol{X})\big] = \mathbb{E}\big[\mu(\boldsymbol{x}) - \widehat{s}(\boldsymbol{X}) \mid \widehat{s}(\boldsymbol{X})\big] = 0.$$

Economically, price-based subgroups $\widehat{s}(\boldsymbol{x})$ are self-financing (for random losses Y).

Calibration: Curve g, or "calibration curve"

[Schervish, 1989] defined well-calibrated as

$$\mathbb{E}[Y \mid \widehat{s}(\boldsymbol{X}) = p] = p, \quad \forall p \in [0,1].$$

Thus, based on that previous expression, consider the calibration curve, named "reliability diagrams" in [Sanders, 1963, Wilks, 1990]

Calibration curve

The **calibration curve** is defined as

$$g : \begin{cases} [0,1] \to [0,1] \\ p \mapsto g(p) := \mathbb{E}[Y \mid \widehat{s}(\boldsymbol{X}) = p] \end{cases}$$

The g function for a well-calibrated model $\widehat{s}$ is the identity function $g(p) = p$.

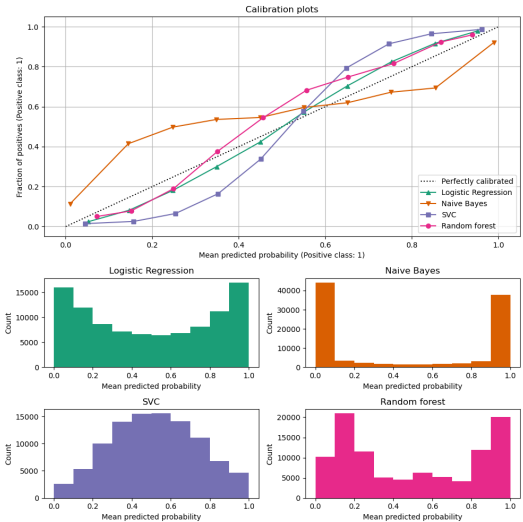
Calibration: Curve g, or "calibration curve", Inference

[Wilks, 1990], [Pakdaman Naeini et al., 2015] and [Kumar et al., 2019] considered quantile-based bins : $\overline{g}$ is the continuous piecewise linear function, interpolating linearly between the points

$$\{(\overline{s}_k, \overline{y}_k)\} \text{ where } k = 1, \cdots, 10,$$

$$\overline{s}_k = \frac{10}{n} \sum_{i \in I_k} \widehat{s}(\boldsymbol{x}_i) \text{ and } \overline{y}_k = \frac{10}{n} \sum_{i \in I_k} y_i,$$

$$I_k = \left\{ i : \left\lceil \frac{k-1}{10} \cdot n \right\rceil \leq \text{rank}(\widehat{s}(\boldsymbol{x}_i)) \leq \left\lfloor \frac{k}{10} \cdot n \right\rfloor \right\}$$



Calibration: Curve g, or "calibration curve", Inference

Given sample $\{(\boldsymbol{x}_i, y_i)\}$ and score $\widehat{s}$, consider a **local regression** of y's against $\widehat{s}(\boldsymbol{x})$'s, as in [Loader, 2006], see [Austin and Steyerberg, 2019, Denuit et al., 2021]. E.g.

$$\widehat{g}(p) := \frac{\sum_{i=1}^{n} K_h(p - \widehat{s}(\boldsymbol{x}_i)) \cdot y_i}{\sum_{i=1}^{n} K_h(p - \widehat{s}(\boldsymbol{x}_i))}, \ \forall p \in [0,1],$$

based on [Nadaraya, 1964, Watson, 1964], for some kernel K and some bandwidth h.

Calibration: Curve g, or "calibration curve", Inference

Since g should be increasing, quite naturally, we could consider an **isotonic regression** of y's against $\widehat{s}(\boldsymbol{x})$'s, as in [Kruskal, 1964], see [Niculescu-Mizil and Caruana, 2005, Wüthrich and Ziegel, 2024], $\widetilde{g}$ is the continuous piecewise linear function, interpolating linearly between the points $(\widehat{s}(\boldsymbol{x}_i), \hat{y}_i)$, where $\widehat{s}(\boldsymbol{x}_i)$'s are sorted,

$$\widetilde{g}(p) := \begin{cases} \hat{y}_1 & \text{if } p \leq \widehat{s}(\boldsymbol{x}_1) \\ \hat{y}_i + \dfrac{p - \widehat{s}(\boldsymbol{x}_i)}{\widehat{s}(\boldsymbol{x}_{i+1}) - \widehat{s}(\boldsymbol{x}_i)}(\hat{y}_{i+1} - \hat{y}_i) & \text{if } \widehat{s}(\boldsymbol{x}_i) \leq x \leq \widehat{s}(\boldsymbol{x}_{i+1}) \\ \hat{y}_n & \text{if } x \geq \widehat{s}(\boldsymbol{x}_n) \end{cases}$$

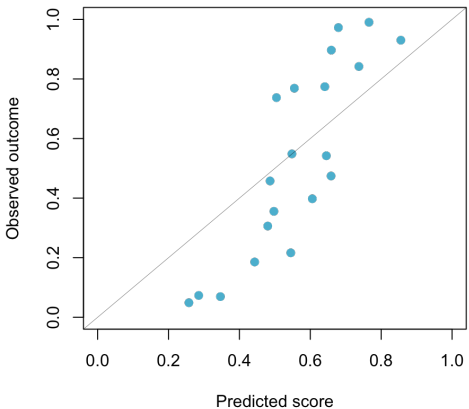
where

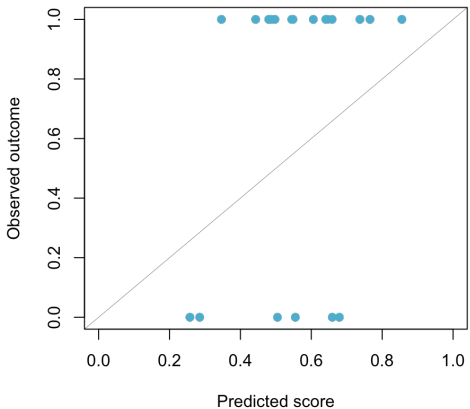
$$\min_{\hat{y}_1,\cdots,\hat{y}_n} \sum_{i=1}^{n} \left(\hat{y}_i - y_i\right)^2 \text{ subject to } \hat{y}_i \leq \hat{y}_j \text{ for all } (i,j) \in E,$$

$E = \{(i,j) : \widehat{s}(\boldsymbol{x}_i) \leq \widehat{s}(\boldsymbol{x}_j)\}$ specifies the partial ordering of the observed inputs $\widehat{s}(\boldsymbol{x}_i)$.

Calibration: Curve g

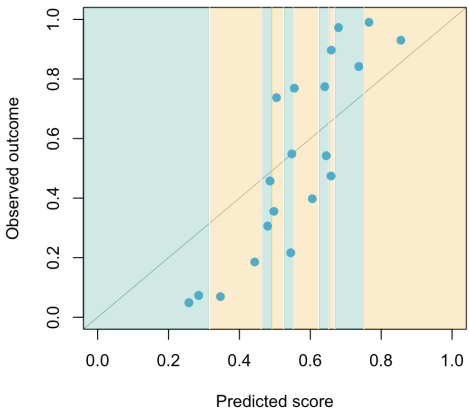
Scatterplot $(\widehat{s}(\boldsymbol{x}_i), y_i)$, $y_i \in \mathbb{R}$ (regression) and $y_i \in \{0, 1\}$ (classification)

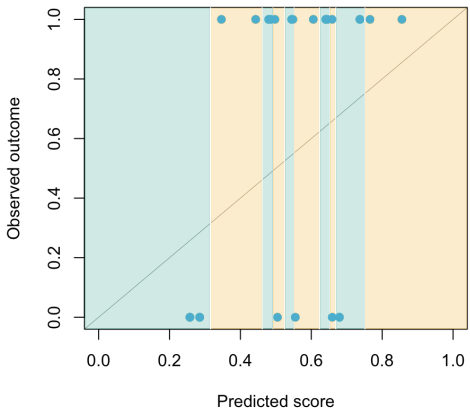




Calibration: Curve g

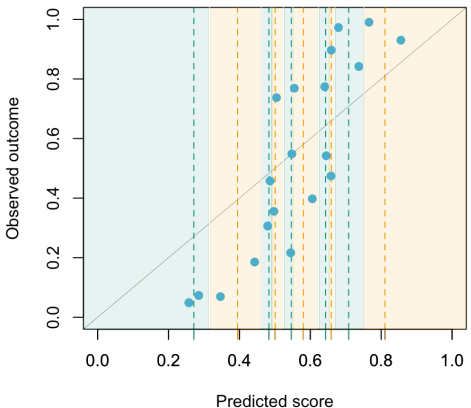
- **quantile-based bins**

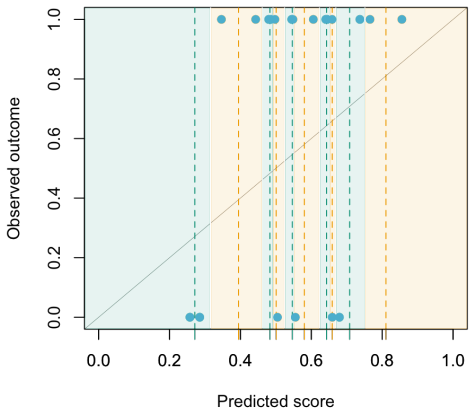




Calibration: Curve g

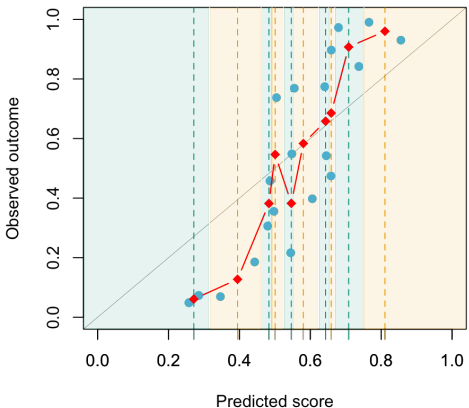
- **quantile-based bins**, compute $\overline{s}_k$ (here $k \in \{1, 2, \cdots, 10\}$

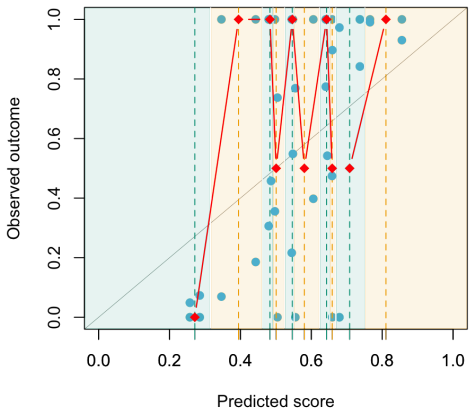




Calibration: Curve g

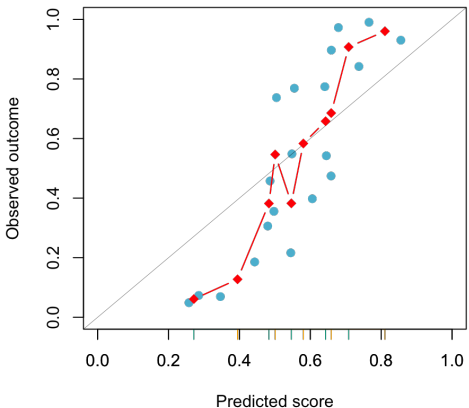
- **quantile-based bins**, compute $\overline{y}_k$ (here $k \in \{1, 2, \cdots, 10\}$

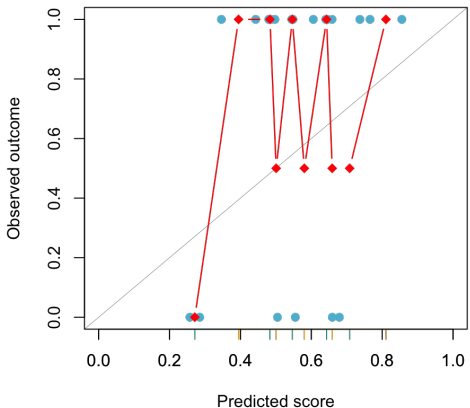




Calibration: Curve g

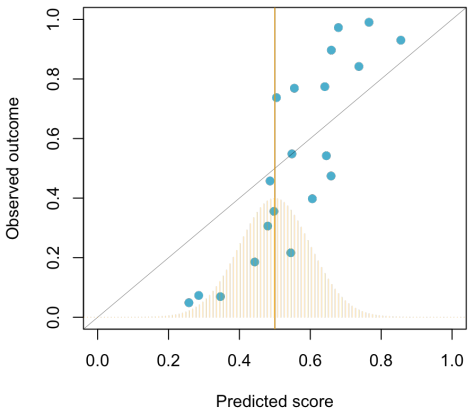
- **quantile-based bins**, interpolate from $(\overline{s}_k, \overline{y}_k)$ (here $k \in \{1, 2, \cdots, 10\}$

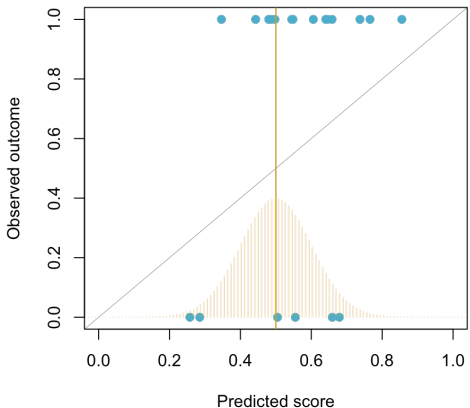




Calibration: Curve g

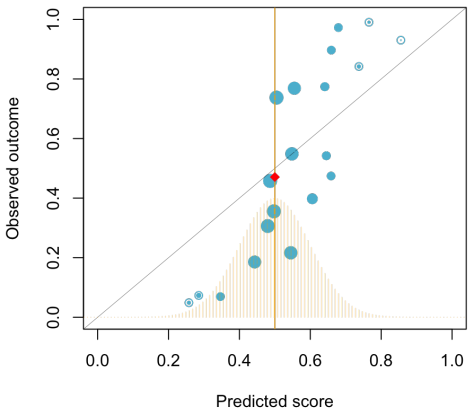
- **local average or regression**, here $p = 1/2$

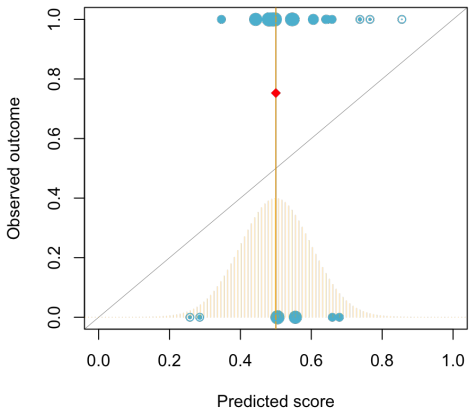




Calibration: Curve g

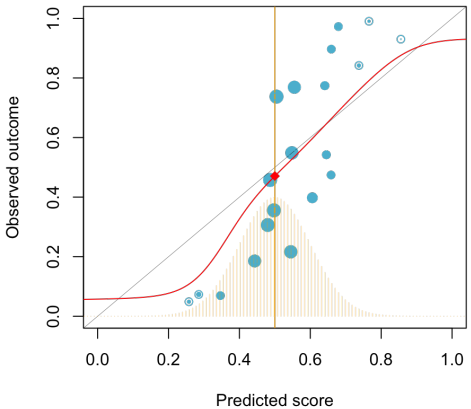
- **local average or regression**, compute $w_i \propto K_h(p - \widehat{s}(\boldsymbol{x}_i))$ and weighted average

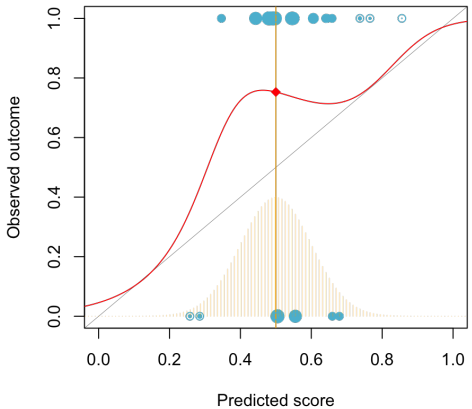




Calibration: Curve g

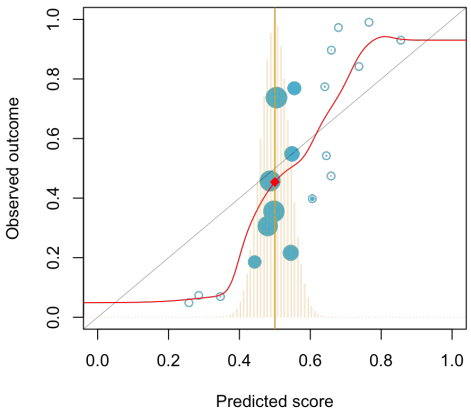
- **local average or regression**, then change p

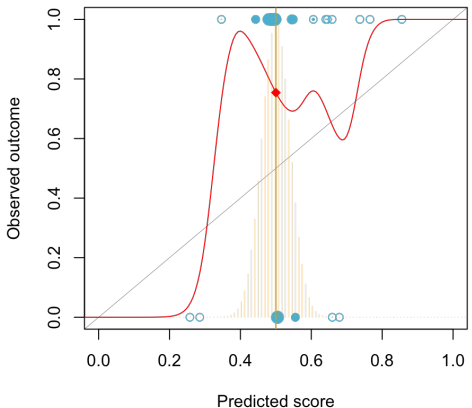




Calibration: Curve g

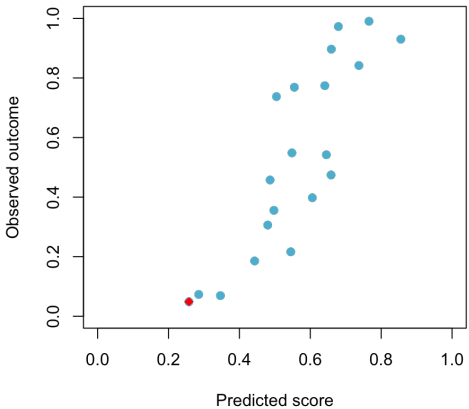
- **local average or regression**, find an appropriate bandwidth h

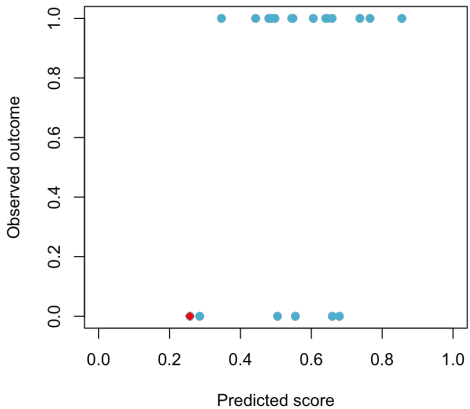




Calibration: Curve g

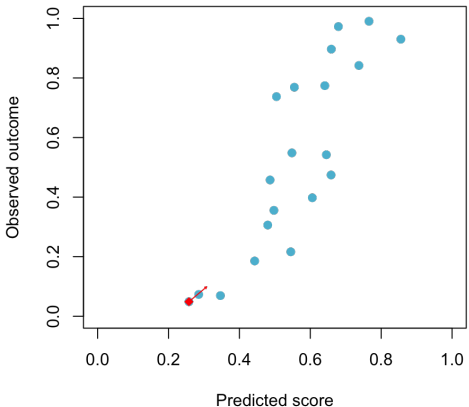
- **isotonic regression**, based on a sequential algorithm

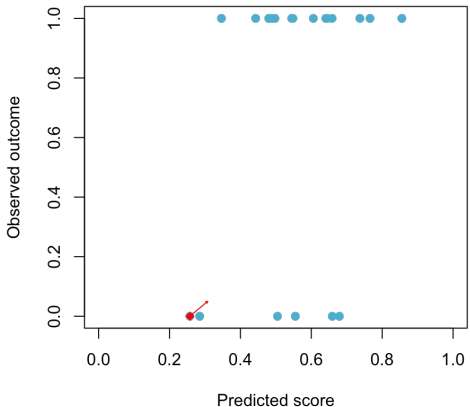




Calibration: Curve g

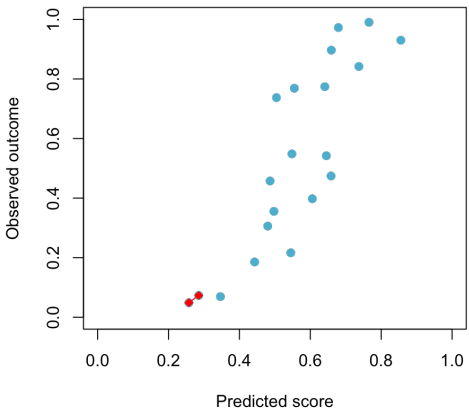
- **isotonic regression**, based on a sequential algorithm

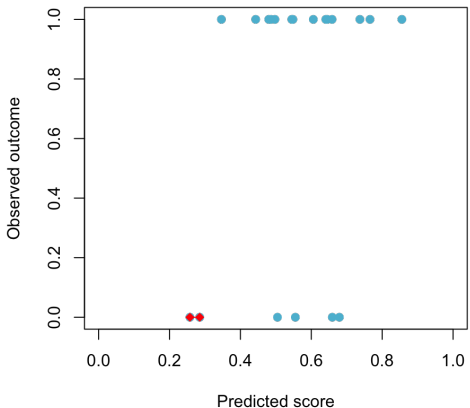




Calibration: Curve g

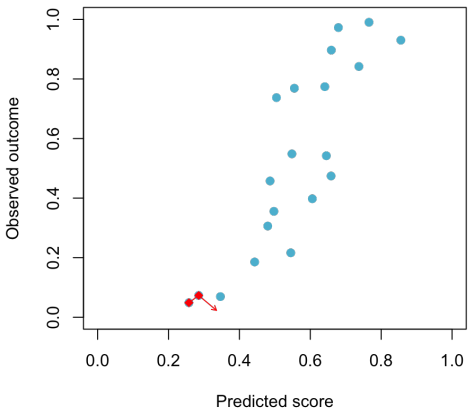
- **isotonic regression**, based on a sequential algorithm

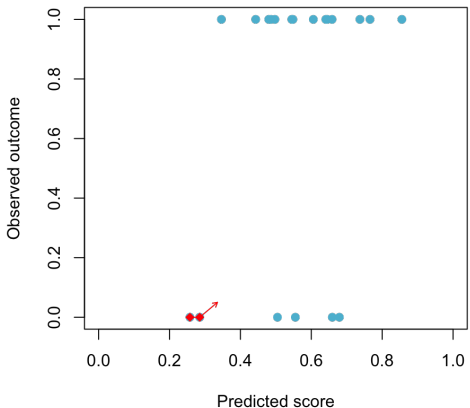




Calibration: Curve g

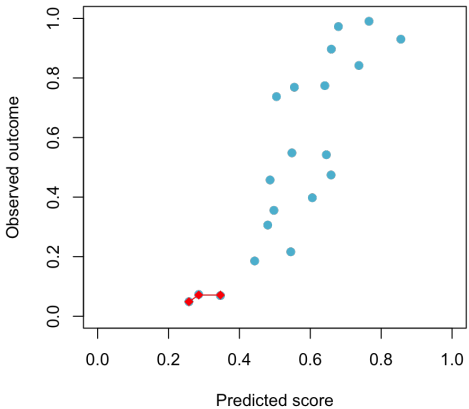
- **isotonic regression**, based on a sequential algorithm

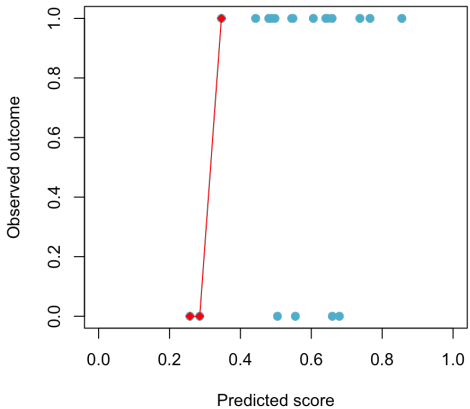




Calibration: Curve g

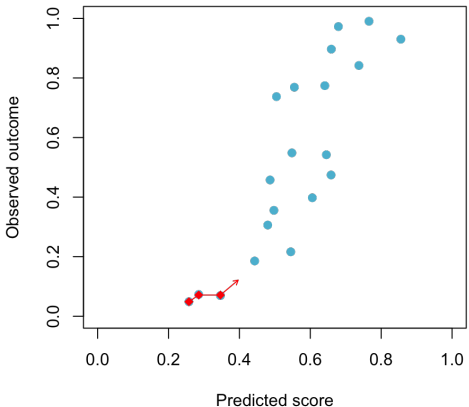
- **isotonic regression**, based on a sequential algorithm

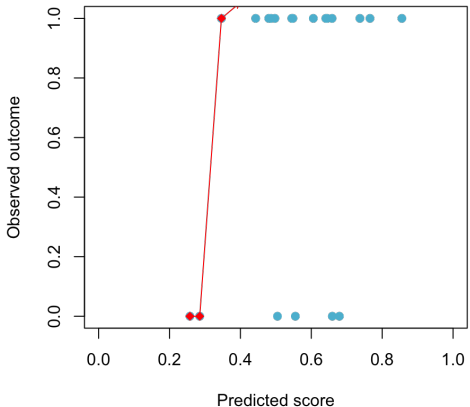




Calibration: Curve g

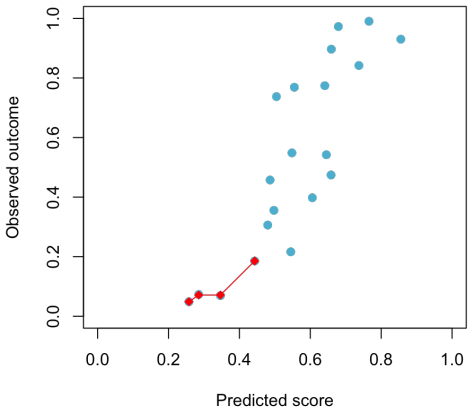
- **isotonic regression**, based on a sequential algorithm

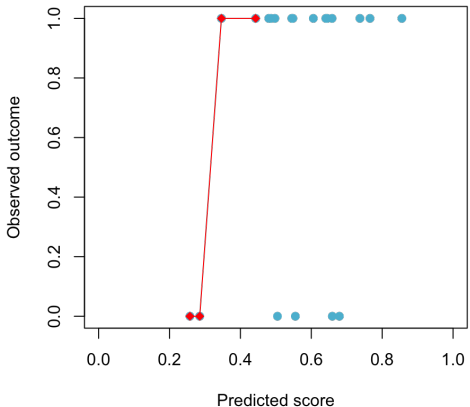




Calibration: Curve g

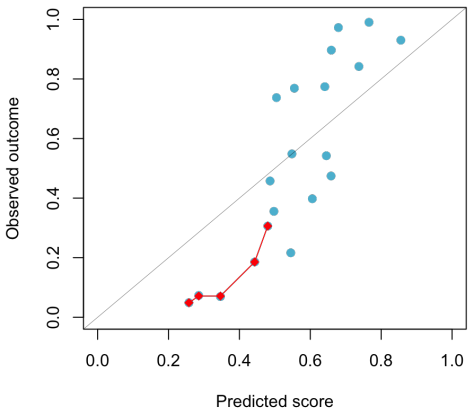
- **isotonic regression**, based on a sequential algorithm

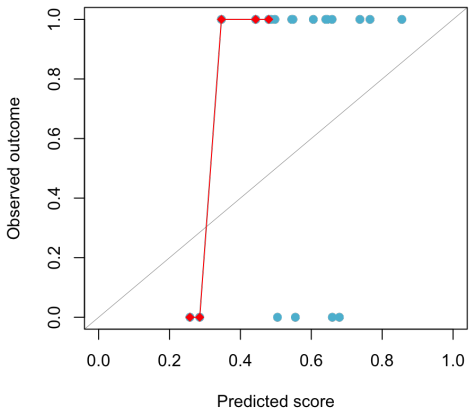




Calibration: Curve g

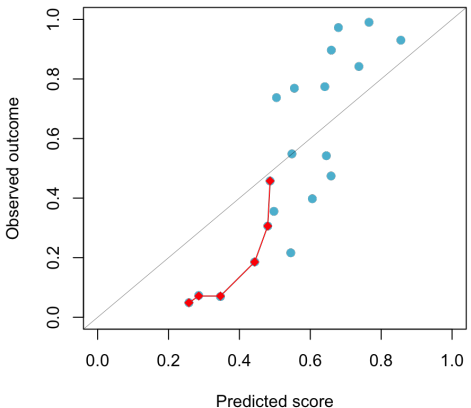
- **isotonic regression**, based on a sequential algorithm

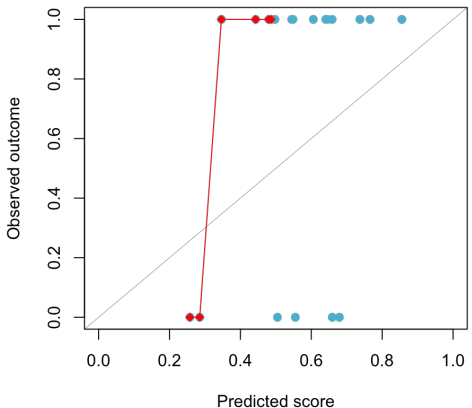




Calibration: Curve g

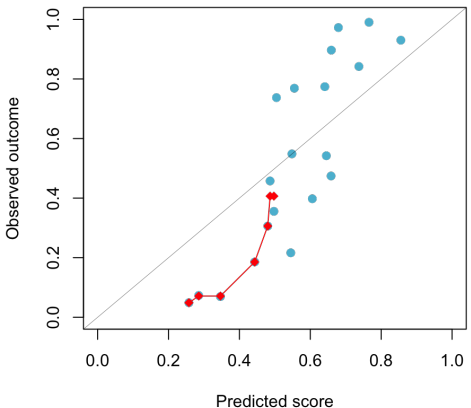
- **isotonic regression**, based on a sequential algorithm

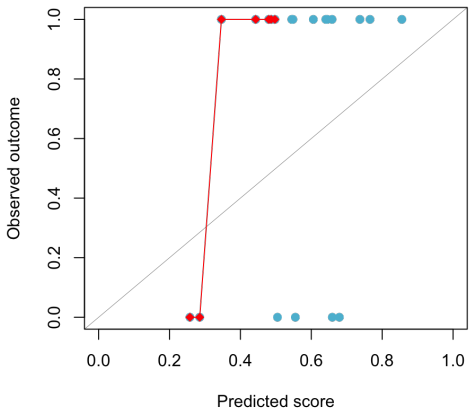




Calibration: Curve g

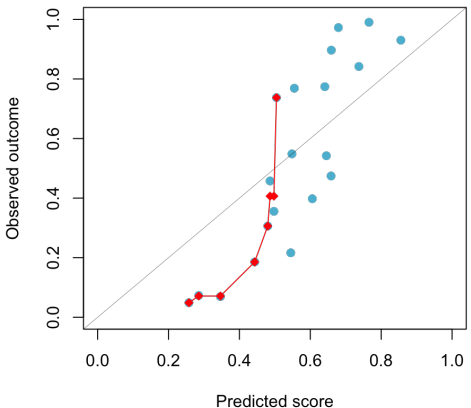
- **isotonic regression**, based on a sequential algorithm

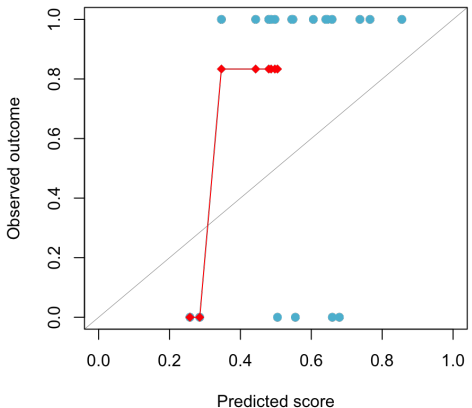




Calibration: Curve g

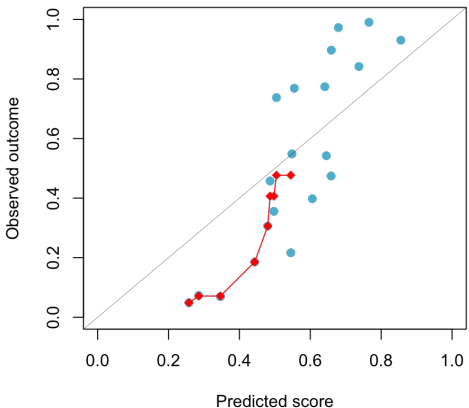
- **isotonic regression**, based on a sequential algorithm

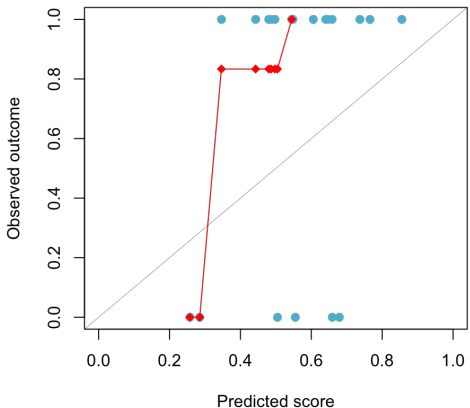




Calibration: Curve g

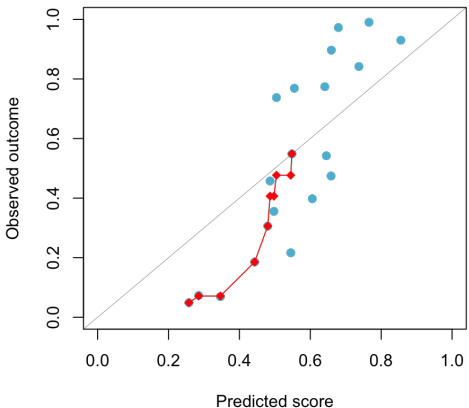
- **isotonic regression**, based on a sequential algorithm

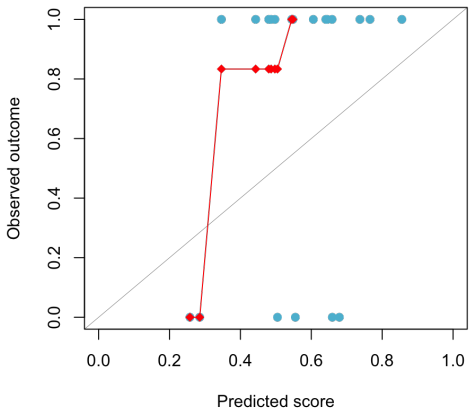




Calibration: Curve g

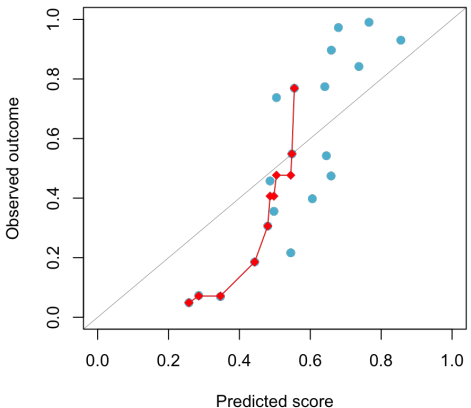
- **isotonic regression**, based on a sequential algorithm

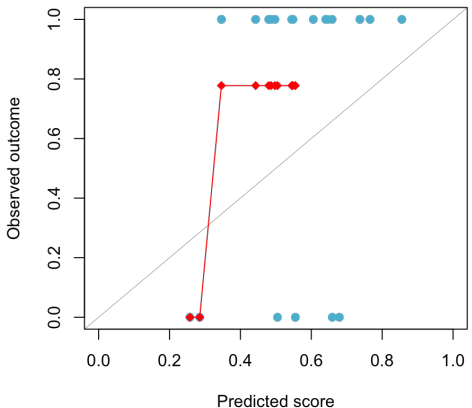




Calibration: Curve g

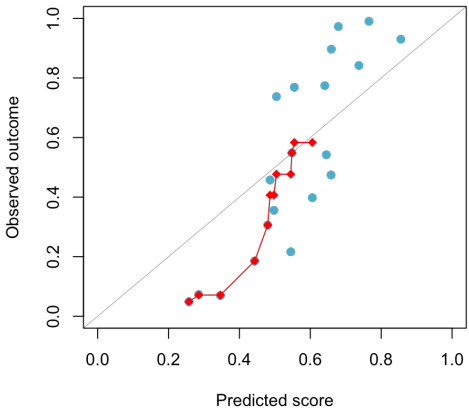
- **isotonic regression**, based on a sequential algorithm

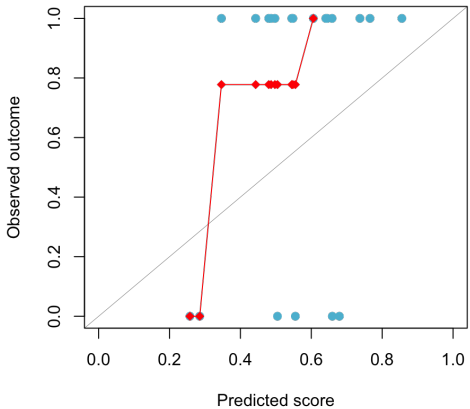




Calibration: Curve g

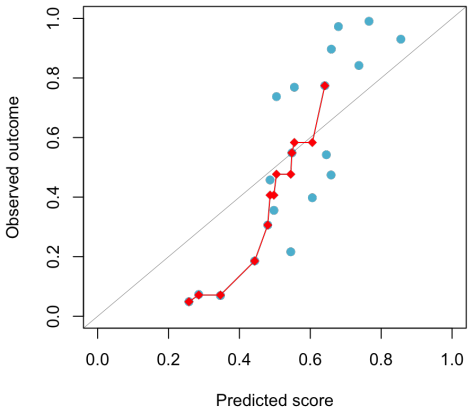
- **isotonic regression**, based on a sequential algorithm

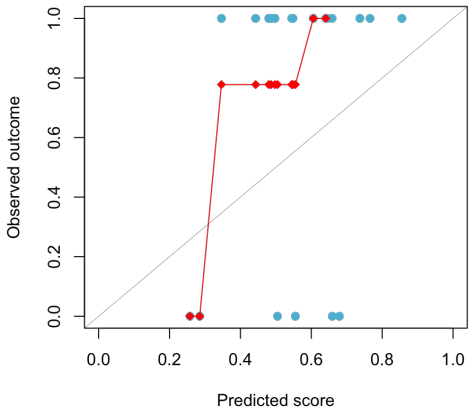




Calibration: Curve g

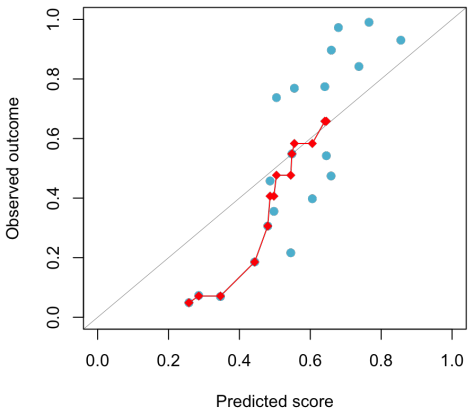
- **isotonic regression**, based on a sequential algorithm

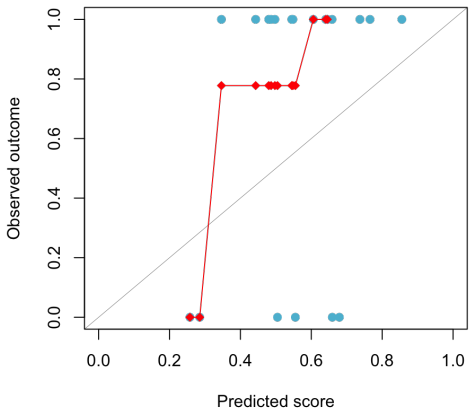




Calibration: Curve g

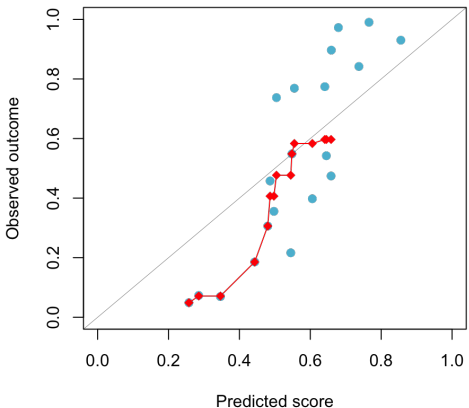
- **isotonic regression**, based on a sequential algorithm

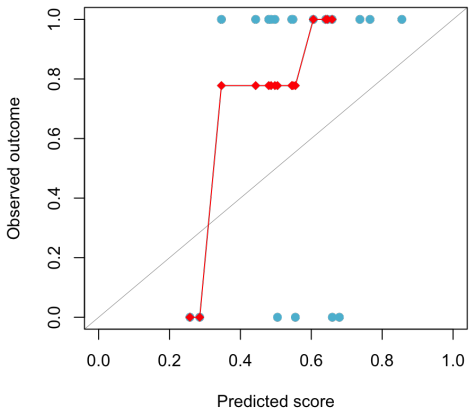




Calibration: Curve g

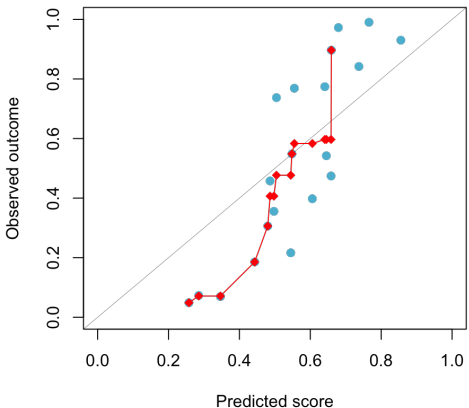
- **isotonic regression**, based on a sequential algorithm

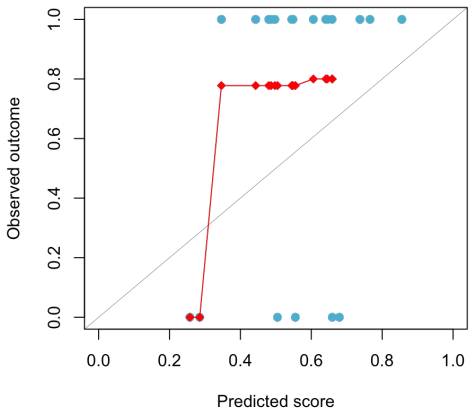




Calibration: Curve g

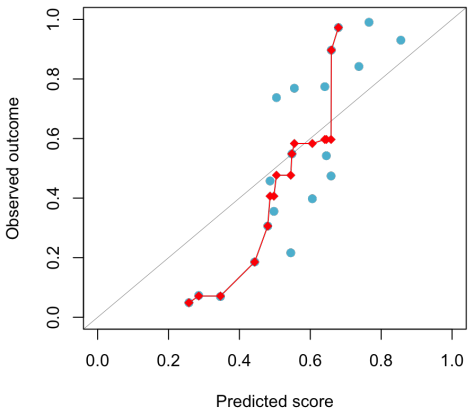
- **isotonic regression**, based on a sequential algorithm

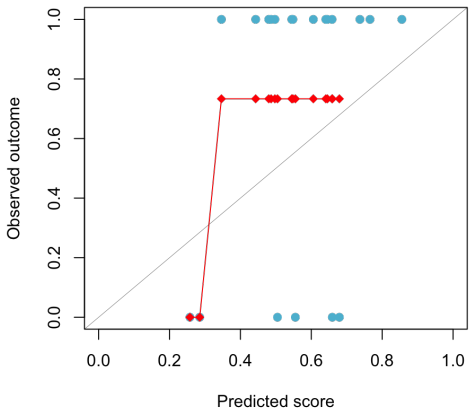




Calibration: Curve g

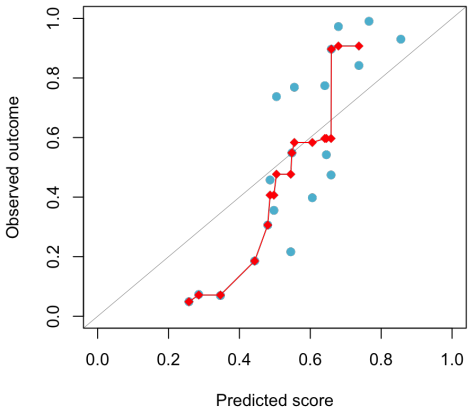
- **isotonic regression**, based on a sequential algorithm

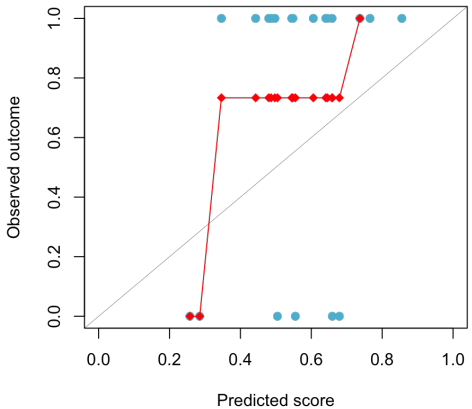




Calibration: Curve g

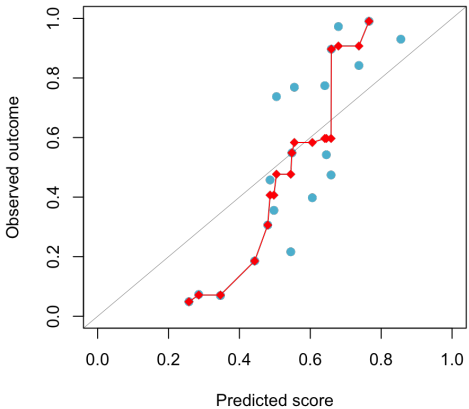
- **isotonic regression**, based on a sequential algorithm

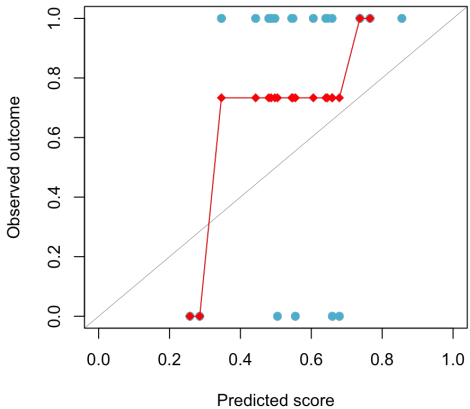




Calibration: Curve g

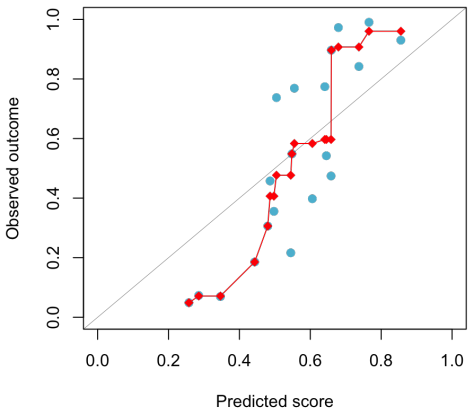
- **isotonic regression**, based on a sequential algorithm

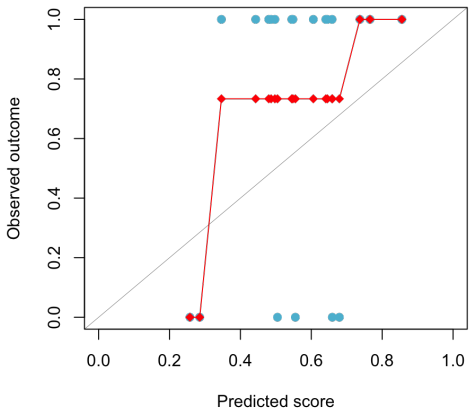




Calibration: Curve g

- **isotonic regression**, based on a sequential algorithm





Recalibration, $\widetilde{s}(\cdot) = \widehat{g}\big(\widehat{s}(\cdot)\big)$

If $g : p \mapsto \mathbb{E}(Y \mid \widehat{s}(\boldsymbol{X}) = p)$ if continuously increasing, recalibration is obtained using

$$\widetilde{s}(\boldsymbol{x}) := \mathbb{E}(Y \mid \widehat{s}(\boldsymbol{X}) = \widehat{s}(\boldsymbol{x})) = g\big(\widehat{s}(\boldsymbol{x})\big)$$

plugin any estimator $\widehat{g}$

see [Denuit et al., 2021].

- **quantile-based bins**, [Wilks, 1990]
- **local regression**, [Loader, 2006]
- **isotonic regression**, [Barlow and Brunk, 1972]
- **Platt (re)scalling**, from [Platt, 1999]
- **Beta regression**, from [Ferrari and Cribari-Neto, 2004]

Recalibration, $\widetilde{s}(\cdot) = \widehat{g}\big(\widehat{s}(\cdot)\big)$

Platt's scaling fits a logistic regression model to the output scores of a classifier:

$$\mathbb{P}(Y = 1 \mid \widehat{s}(\boldsymbol{X})) = \frac{1}{1 + \exp(A \cdot \widehat{s}(\boldsymbol{X}) + B)}$$

where $\widehat{s}(\cdot)$ is the uncalibrated score, and A, B are learned on a validation set. Parameters A, B are learned via logistic regression by minimizing the negative log-likelihood:

$$\log \mathcal{L}(A, B) = -\sum_{i=1}^{n} [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)]$$

where $\hat{p}_i = \dfrac{1}{1 + \exp(A \cdot \widehat{s}(\boldsymbol{x}_i) + B)}$. Then

$$\widetilde{s}(\boldsymbol{x}) = \frac{1}{1 + \exp(\widehat{A} \cdot \widehat{s}(\boldsymbol{X}) + \widehat{B})}$$

Recalibration, $\widetilde{s}(\cdot) = \widehat{g}\big(\widehat{s}(\cdot)\big)$

Beta regression is used to model variables that lie in the open interval $(0, 1)$.
Assumes that $Y \sim \mathcal{B}\text{eta}(\mu, \phi)$, where μ is the mean and ϕ is a precision parameter

$$f(y\,;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\,\Gamma((1-\mu)\phi)} \cdot y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},\ y \in (0,1).$$

Suppose that μ is function of $\widehat{s}(\boldsymbol{x})$

$$\text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1\widehat{s}(\boldsymbol{x}_i) = \widehat{s}(\boldsymbol{x}_i)^\top\boldsymbol{\beta}$$

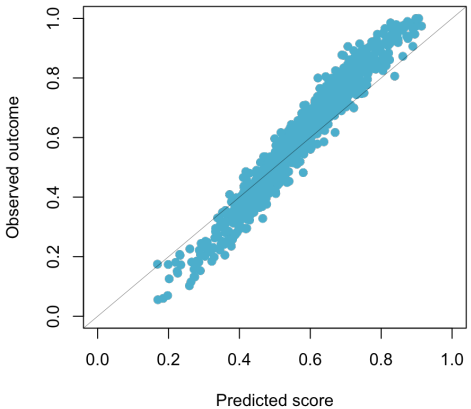
Log-likelihood ($\log\mathcal{L}(\beta,\phi)$) for n observations is:

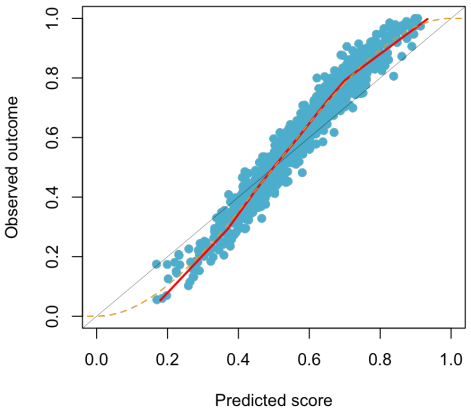
$$\sum_{i=1}^{n}\left[\log\Gamma(\phi) - \log\Gamma(\mu_i\phi) - \log\Gamma((1-\mu_i)\phi) + (\mu_i\phi-1)\log y_i + ((1-\mu_i)\phi-1)\log(1-y_i)\right]$$

Then $\widetilde{s}(\boldsymbol{x}) = \text{expit}(\widehat{\boldsymbol{s}}(\boldsymbol{x})\widehat{\boldsymbol{\beta}})$.

Recalibration, $\widetilde{s}(\cdot) = \widehat{g}\big(\widehat{s}(\cdot)\big)$

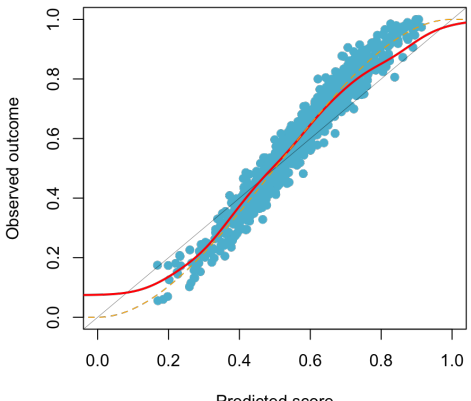
- on a larger validation sample, use **quantile based bins**, fit $\widehat{g}$

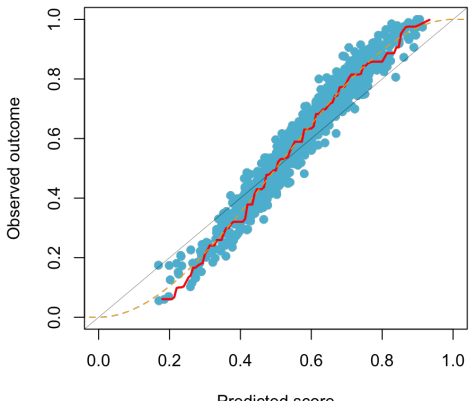




Recalibration, $\widetilde{s}(\cdot) = \widehat{g}\big(\widehat{s}(\cdot)\big)$

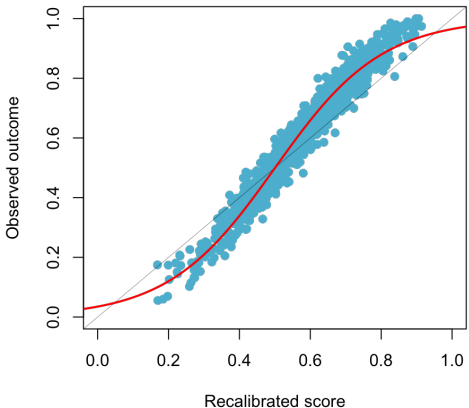
- **local regression** and **isotonic regression**

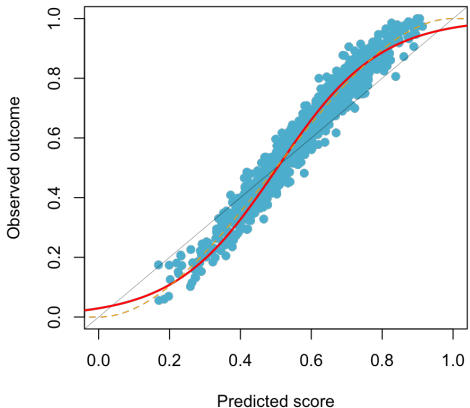




Recalibration, $\widetilde{s}(\cdot) = \widehat{g}\big(\widehat{s}(\cdot)\big)$

- **Platt (re)scalling** (logistic) and **Beta regression**



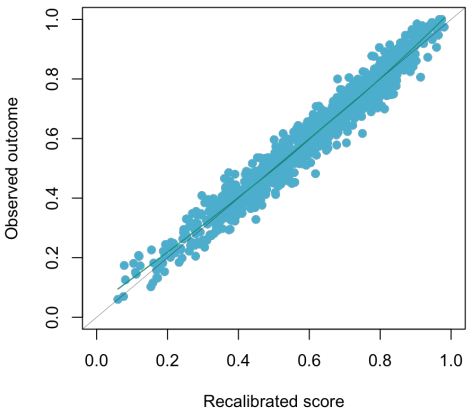


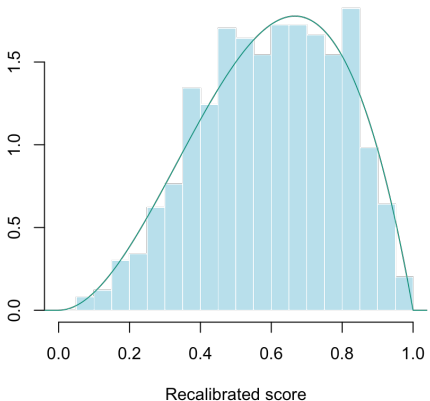
Recalibration, $\widetilde{s}(\cdot) = \widehat{g}\big(\widehat{s}(\cdot)\big)$

recalibrated score — initial score

- **quantile-based bins**

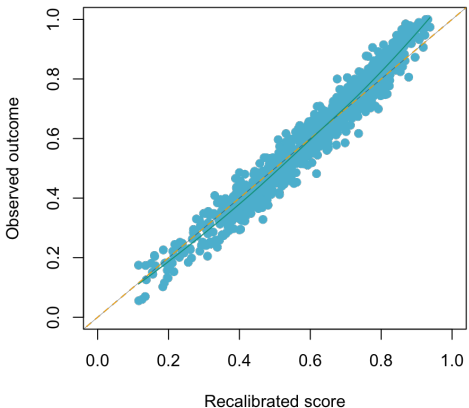
$$\widehat{s}(\boldsymbol{x}_i) \xrightarrow{\widehat{g}} \widetilde{s}(\boldsymbol{x}_i) = \widehat{g}\big(\widehat{s}(\boldsymbol{x}_i)\big)$$

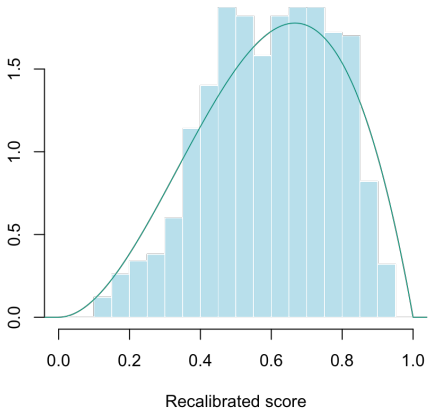




Recalibration, $\widetilde{s}(\cdot) = \widehat{g}\big(\widehat{s}(\cdot)\big)$

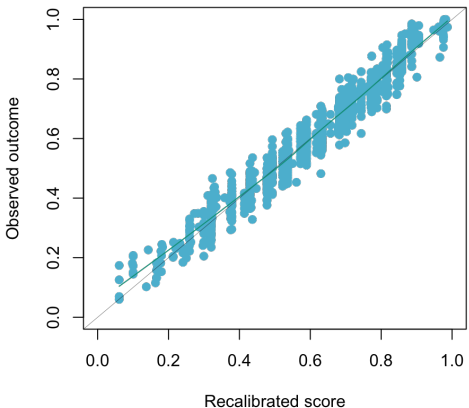
- **local regression**

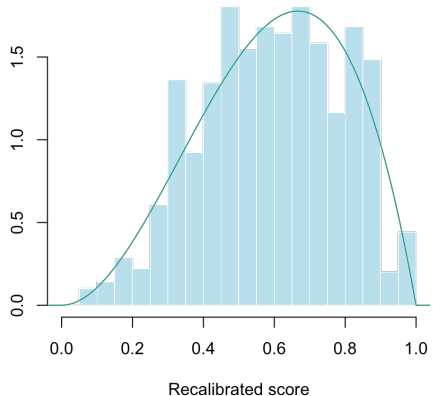




Recalibration, $\widetilde{s}(\cdot) = \widehat{g}\big(\widehat{s}(\cdot)\big)$

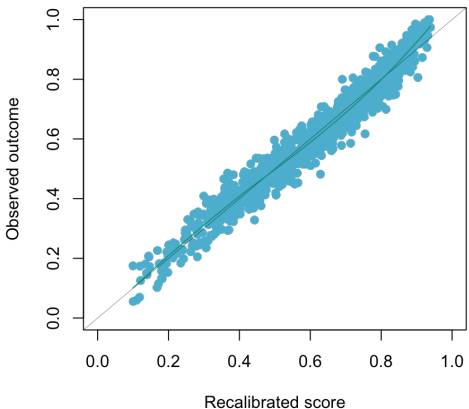
- **isotonic regression**

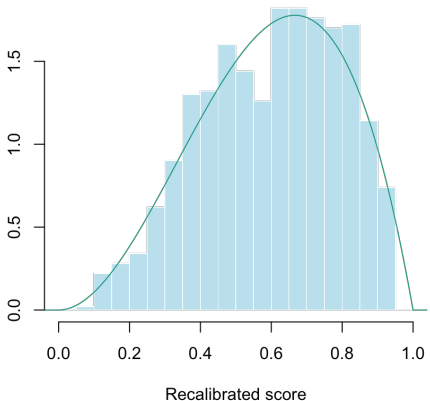




Recalibration, $\widetilde{s}(\cdot) = \widehat{g}\big(\widehat{s}(\cdot)\big)$

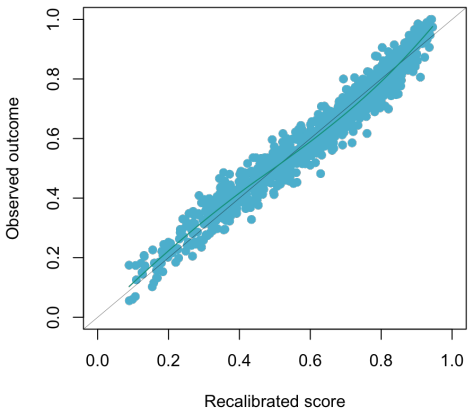
- **Platt (re)scalling**

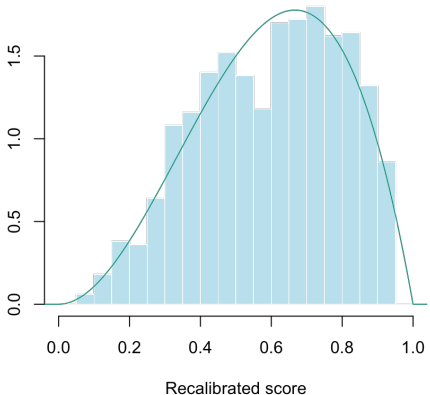




Recalibration, $\widetilde{s}(\cdot) = \widehat{g}\big(\widehat{s}(\cdot)\big)$

- **Beta regression**





Sidenote, Distance Between Distributions

- **Kullback–Leibler divergence** $KL(\mathbb{P} \parallel \mathbb{Q})$

$$KL(\mathbb{P} \parallel \mathbb{Q}) = \sum_i \mathbb{P}(i) \log \frac{\mathbb{P}(i)}{\mathbb{Q}(i)} \text{ or } \int_{\mathbb{R}} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x$$

for absolutely continuous distributions (with densities p and q)

- **Wasserstein or Cramér distance** $W_k(\mathbb{P}, \mathbb{Q})$ for $k \geq 1$,

$$W_k(\mathbb{P}, \mathbb{Q}) = \Big(\int_0^1 |F_p^{-1}(u) - F_q^{-1}(u)|^k \mathrm{d}u \Big)^{1/k}, \text{ where } \begin{cases} F_p(x) = \mathbb{P}((-\infty, x]) = \int_{-\infty}^{x} \mathrm{d}\mathbb{P}(t) \\ F_q(x) = \mathbb{Q}((-\infty, x]) = \int_{-\infty}^{x} \mathrm{d}\mathbb{Q}(t) \end{cases}$$

$$C_k(\mathbb{P}, \mathbb{Q}) = \Big(\int_{\mathbb{R}} |F_p(x) - F_q(x)|^k \mathrm{d}x \Big)^{1/k}.$$

Claims Frequency in Motor Insurance

As in [Denuit et al., 2021], consider claims (annual) frequency in motor insurance,

	$\widehat{s}^{\text{glm}}$	$\widehat{s}^{\text{gam}}$	$\widehat{s}^{\text{gbm}}$	$\widehat{s}^{\text{rf}}$
average $\widehat{s}(\mathbf{x})$'s	0.0875	0.0877	0.0875	0.0886
10% quantile	0.0295	0.0288	0.0508	0.0006
90% quantile	0.1589	0.1612	0.1321	0.4073

First desirable property, **global balance**, $\mathbb{E}(Y) = \mathbb{E}(\widehat{s}(\mathbf{X}))$

Another desirable property, **local balance**,

$$\mathbb{E}(Y \mid \widehat{s}(\mathbf{X})) = \mathbb{E}(\widehat{s}(\mathbf{X}) \mid \widehat{s}(\mathbf{X})) = \widehat{s}(\mathbf{X})$$

average loss

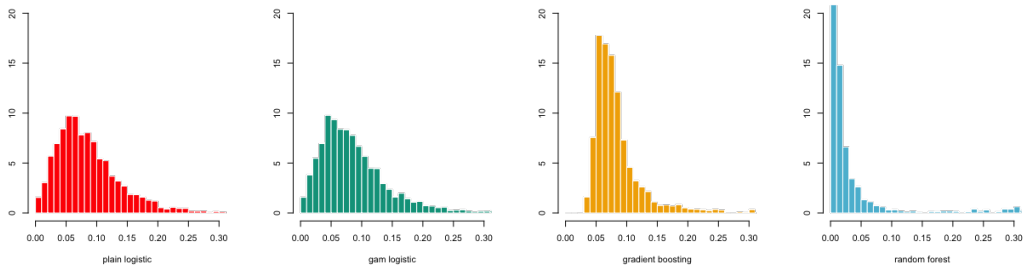
average premium

Calibration curve (for probabilistic scores) is defined as

$$g : \begin{cases} [0,1] \to [0,1] \\ p \mapsto g(p) := \mathbb{E}[Y \mid \widehat{s}(\mathbf{X}) = p] \end{cases}$$

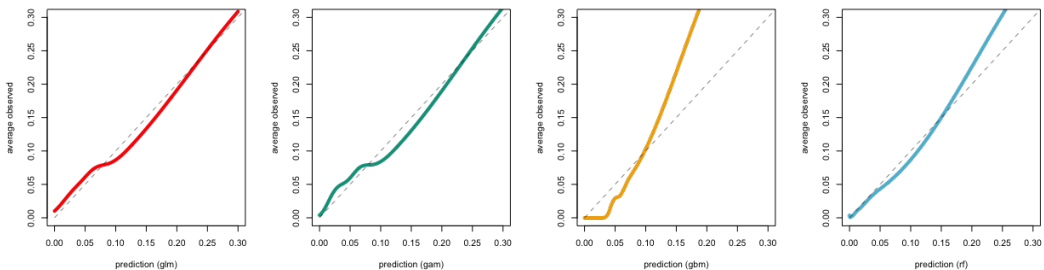
Claims Frequency in Motor Insurance

Distribution of $\widehat{s}(\mathbf{x}_i)$'s, for a plain logistic regression, GAM additive model, gradient boosting (sequential learning) and random forest.



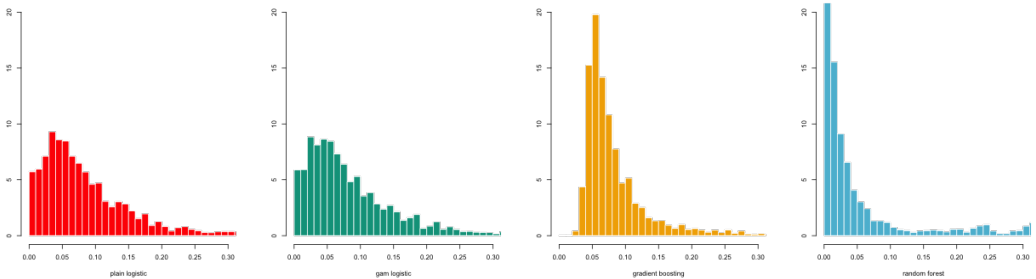
	$\widehat{s}^{\text{glm}}$	$\widehat{s}^{\text{gam}}$	$\widehat{s}^{\text{gbm}}$	$\widehat{s}^{\text{rf}}$
average $\widehat{s}(\mathbf{x})$'s	0.0875	0.0877	0.0875	0.0886
10% quantile	0.0295	0.0288	0.0508	0.0006
90% quantile	0.1589	0.1612	0.1321	0.4073

Claims Frequency in Motor Insurance



Evolution of the (estimated) calibration curve, $p \mapsto \mathbb{E}[Y|\widehat{s}(\mathbf{X}) = p]$
where $\widehat{s}(\mathbf{x})$ is a predicted claims (annual) frequency, for a plain logistic regression, GAM additive model, gradient boosting (sequential learning) and random forest.

Claims Frequency in Motor Insurance



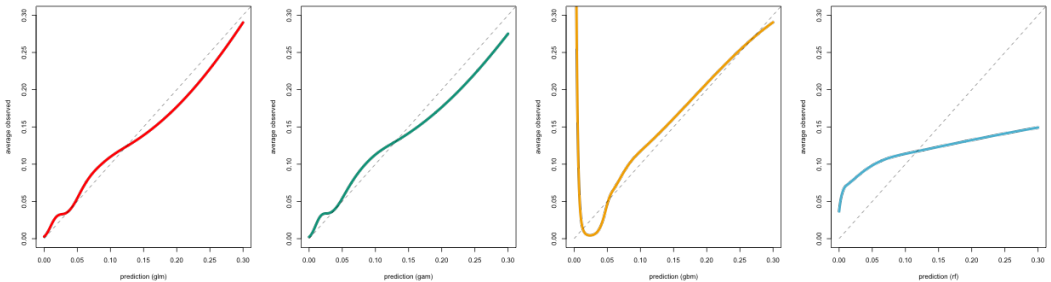
AUC

	$\widehat{s}^{\text{glm}}$	$\widehat{s}^{\text{gam}}$	$\widehat{s}^{\text{gbm}}$	$\widehat{s}^{\text{rf}}$
glm $\rightarrow$	**0.707**	**0.707**	**0.685**	**0.619**
gam $\rightarrow$	0.699	0.699	0.681	0.616
gbm $\rightarrow$	0.665	0.669	0.708	0.681
rf $\rightarrow$	0.642	0.646	0.726	0.822

inter quantile difference

	$\widehat{s}^{\text{glm}}$	$\widehat{s}^{\text{gam}}$	$\widehat{s}^{\text{gbm}}$	$\widehat{s}^{\text{rf}}$
glm $\rightarrow$	**0.159**	**0.163**	**0.099**	**0.306**
gam $\rightarrow$	0.157	0.159	0.095	0.305
gbm $\rightarrow$	0.135	0.143	0.089	0.296
rf $\rightarrow$	0.138	0.142	0.084	0.299

Claims Frequency in Motor Insurance



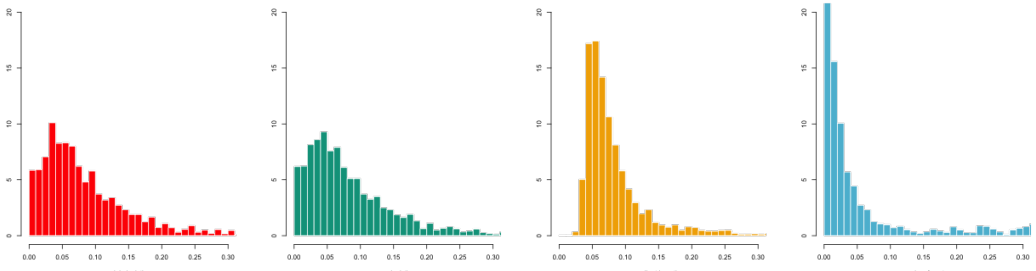
	$\widehat{s}^{\text{glm}}$	$\widehat{s}^{\text{gam}}$	$\widehat{s}^{\text{gbm}}$	$\widehat{s}^{\text{rf}}$
glm →	**0.14**	**0.16**	**0.36**	**12.99**
gam →	0.15	0.17	0.52	13.49
gbm →	0.40	0.52	0.30	9.69
rf →	0.46	0.48	0.92	1.91

un-calibration $\|\widehat{g} - Id\|_2^2$

	$\widehat{s}^{\text{glm}}$	$\widehat{s}^{\text{gam}}$	$\widehat{s}^{\text{gbm}}$	$\widehat{s}^{\text{rf}}$
glm →	**0.0482**	**0.0540**	**0.3435**	**1.0891**
gam →	0.0460	0.0400	0.3394	1.0977
gbm →	0.1389	0.1277	0.4567	0.8814
rf →	0.7133	0.7099	1.3091	0.1129

distance $KL(\mathbb{P}_p, \mathbb{P}_{\hat{p}})$

Claims Frequency in Motor Insurance



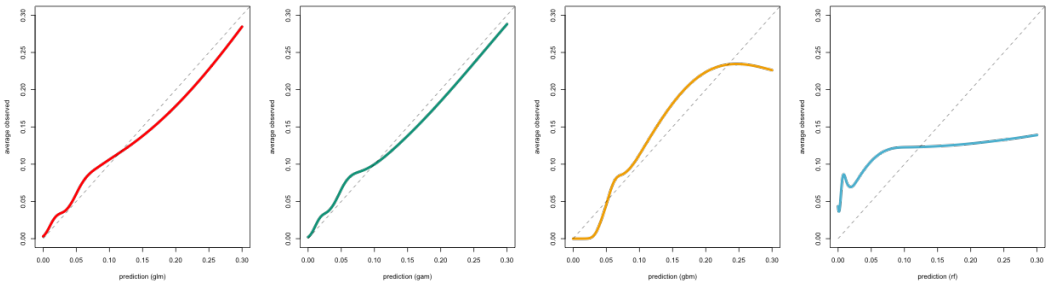
	$\widehat{s}^{\text{glm}}$	$\widehat{s}^{\text{gam}}$	$\widehat{s}^{\text{gbm}}$	$\widehat{s}^{\text{rf}}$
glm →	0.707	0.707	0.685	0.619
gam →	**0.699**	**0.699**	**0.681**	**0.616**
gbm →	0.665	0.669	0.708	0.681
rf →	0.642	0.646	0.726	0.822

AUC

	$\widehat{s}^{\text{glm}}$	$\widehat{s}^{\text{gam}}$	$\widehat{s}^{\text{gbm}}$	$\widehat{s}^{\text{rf}}$
glm →	0.159	0.163	0.099	0.306
gam →	**0.157**	**0.159**	**0.095**	**0.305**
gbm →	0.135	0.143	0.089	0.296
rf →	0.138	0.142	0.084	0.299

inter quantile difference

Claims Frequency in Motor Insurance



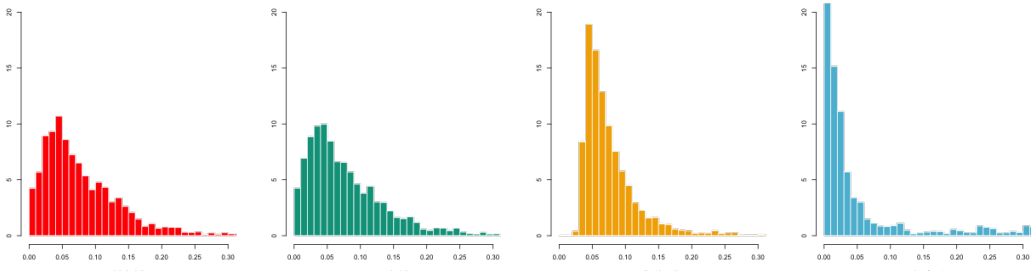
un-calibration $\|\widehat{g} - Id\|_2^2$

	$\widehat{s}^{\text{glm}}$	$\widehat{s}^{\text{gam}}$	$\widehat{s}^{\text{gbm}}$	$\widehat{s}^{\text{rf}}$
glm →	0.14	0.16	0.36	12.99
gam →	**0.15**	**0.17**	**0.52**	**13.49**
gbm →	0.40	0.52	0.30	9.69
rf →	0.46	0.48	0.92	1.91

distance $KL(\mathbb{P}_p, \mathbb{P}_{\widehat{p}})$

	$\widehat{s}^{\text{glm}}$	$\widehat{s}^{\text{gam}}$	$\widehat{s}^{\text{gbm}}$	$\widehat{s}^{\text{rf}}$
glm →	0.0482	0.0540	0.3435	1.0891
gam →	**0.0460**	**0.0400**	**0.3394**	**1.0977**
gbm →	0.1389	0.1277	0.4567	0.8814
rf →	0.7133	0.7099	1.3091	0.1129

Claims Frequency in Motor Insurance



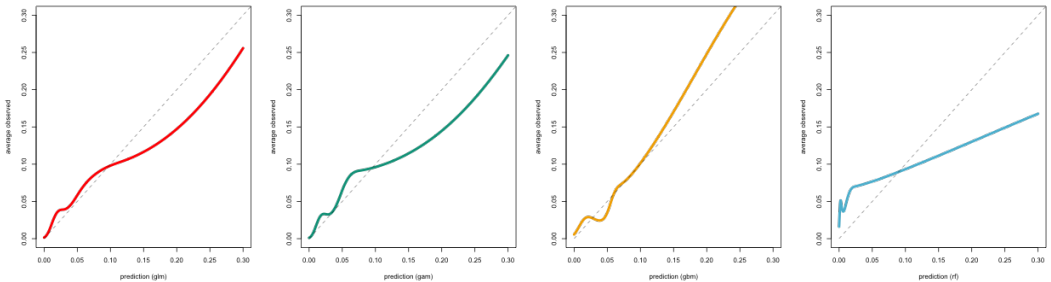
AUC

	$\widehat{s}^{\text{glm}}$	$\widehat{s}^{\text{gam}}$	$\widehat{s}^{\text{gbm}}$	$\widehat{s}^{\text{rf}}$
glm $\rightarrow$	0.707	0.707	0.685	0.619
gam $\rightarrow$	0.699	0.699	0.681	0.616
gbm $\rightarrow$	**0.665**	**0.669**	**0.708**	**0.681**
rf $\rightarrow$	0.642	0.646	0.726	0.822

inter quantile difference

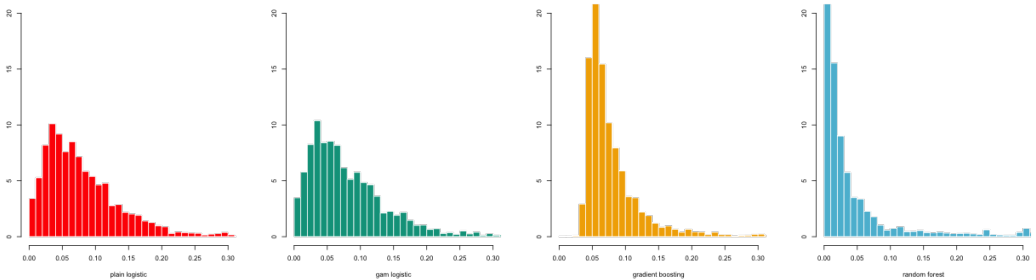
	$\widehat{s}^{\text{glm}}$	$\widehat{s}^{\text{gam}}$	$\widehat{s}^{\text{gbm}}$	$\widehat{s}^{\text{rf}}$
glm $\rightarrow$	0.159	0.163	0.099	0.306
gam $\rightarrow$	0.157	0.159	0.095	0.305
gbm $\rightarrow$	**0.135**	**0.143**	**0.089**	**0.296**
rf $\rightarrow$	0.138	0.142	0.084	0.299

Claims Frequency in Motor Insurance



	$\widehat{s}^{\text{glm}}$	$\widehat{s}^{\text{gam}}$	$\widehat{s}^{\text{gbm}}$	$\widehat{s}^{\text{rf}}$	$\widehat{s}^{\text{glm}}$	$\widehat{s}^{\text{gam}}$	$\widehat{s}^{\text{gbm}}$	$\widehat{s}^{\text{rf}}$
glm $\rightarrow$	0.14	0.16	0.36	12.99	0.0482	0.0540	0.3435	1.0891
gam $\rightarrow$	0.15	0.17	0.52	13.49	0.0460	0.0400	0.3394	1.0977
gbm $\rightarrow$	**0.40**	**0.52**	**0.30**	**9.69**	**0.1389**	**0.1277**	**0.4567**	**0.8814**
rf $\rightarrow$	0.46	0.48	0.92	1.91	0.7133	0.7099	1.3091	0.1129
	un-calibration $\|\widehat{g} - Id\|_2^2$				distance $KL(\mathbb{P}_p, \mathbb{P}_{\hat{p}})$			

Claims Frequency in Motor Insurance



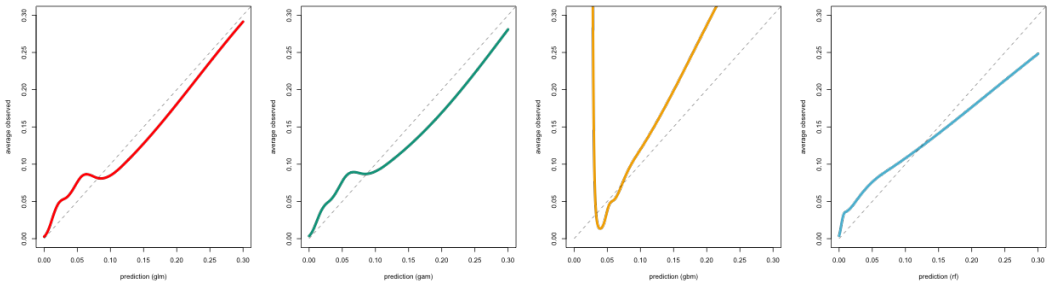
AUC

	$\widehat{s}^{\text{glm}}$	$\widehat{s}^{\text{gam}}$	$\widehat{s}^{\text{gbm}}$	$\widehat{s}^{\text{rf}}$
glm $\rightarrow$	0.707	0.707	0.685	0.619
gam $\rightarrow$	0.699	0.699	0.681	0.616
gbm $\rightarrow$	0.665	0.669	0.708	0.681
rf $\rightarrow$	**0.642**	**0.646**	**0.726**	**0.822**

inter quantile difference

$\widehat{s}^{\text{glm}}$	$\widehat{s}^{\text{gam}}$	$\widehat{s}^{\text{gbm}}$	$\widehat{s}^{\text{rf}}$
0.159	0.163	0.099	0.306
0.157	0.159	0.095	0.305
0.135	0.143	0.089	0.296
0.138	**0.142**	**0.084**	**0.299**

Claims Frequency in Motor Insurance



	$\widehat{s}^{\text{glm}}$	$\widehat{s}^{\text{gam}}$	$\widehat{s}^{\text{gbm}}$	$\widehat{s}^{\text{rf}}$
glm →	0.14	0.16	0.36	12.99
gam →	0.15	0.17	0.52	13.49
gbm →	0.40	0.52	0.30	9.69
rf →	**0.46**	**0.48**	**0.92**	**1.91**

un-calibration $\|\widehat{g} - Id\|_2^2$

	$\widehat{s}^{\text{glm}}$	$\widehat{s}^{\text{gam}}$	$\widehat{s}^{\text{gbm}}$	$\widehat{s}^{\text{rf}}$
glm →	0.0482	0.0540	0.3435	1.0891
gam →	0.0460	0.0400	0.3394	1.0977
gbm →	0.1389	0.1277	0.4567	0.8814
rf →	**0.7133**	**0.7099**	**1.3091**	**0.1129**

distance $KL(\mathbb{P}_p, \mathbb{P}_{\widehat{p}})$

References

Austin, P. C. and Steyerberg, E. W. (2019).
The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models.
Statistics in Medicine, 38:4051 – 4065.

Barlow, R. E. and Brunk, H. D. (1972).
The isotonic regression problem and its dual.
Journal of the American Statistical Association, 67(337):140–147.

Denuit, M., Charpentier, A., and Trufin, J. (2021).
Autocalibration and tweedie-dominance for insurance pricing with machine learning.
Insurance: Mathematics and Economics, 101:485–497.

Fernandes Machado, A., Charpentier, A., Flachaire, E., Gallic, E., and Hu, F. (2024a).
From uncertainty to precision: Enhancing binary classifier performance through calibration.
arXiv preprint arXiv:2402.07790.

References

Fernandes Machado, A., Charpentier, A., Flachaire, E., Gallic, E., and Hu, F. (2024b).
Post-calibration techniques: Balancing calibration and score distribution alignment.
Thirty-Eighth Annual Conference on Neural Information Processing Systems.

Fernandes Machado, A., Charpentier, A., Flachaire, E., Gallic, E., and Hu, F. (2024c).
Probabilistic scores of classifiers, calibration is not enough.
arXiv preprint arXiv:2408.03421.

Ferrari, S. and Cribari-Neto, F. (2004).
Beta regression for modelling rates and proportions.
Journal of applied statistics, 31(7):799–815.

Gourieroux, C. and Jasiak, J. (2015).
The econometrics of individual risk: credit, insurance, and marketing.
Princeton University Press.

References

Kruskal, J. B. (1964).
Nonmetric multidimensional scaling: a numerical method.
Psychometrika, 29(2):115–129.

Kumar, A., Liang, P. S., and Ma, T. (2019).
Verified uncertainty calibration.
In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Loader, C. (2006).
Local regression and likelihood.
Springer.

Nadaraya, E. A. (1964).
On estimating regression.
Theory of Probability & Its Applications, 9(1):141–142.

References

Niculescu-Mizil, A. and Caruana, R. (2005).
Predicting good probabilities with supervised learning.
In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632.

Pakdaman Naeini, M., Cooper, G., and Hauskrecht, M. (2015).
Obtaining well calibrated probabilities using bayesian binning.
Proceedings of the AAAI Conference on Artificial Intelligence, 29(1):2901–2907.

Platt, J. (1999).
Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.
Advances in large margin classifiers, 10(3):61–74.

Reichenbach, H. (1971).
The theory of probability.
University of California Press.

References


Sanders, F. (1963).
On subjective probability forecasting.
Journal of Applied Meteorology and Climatology, 2(2):191–201.


Schervish, M. J. (1989).
A General Method for Comparing Probability Assessors.
The Annals of Statistics, 17(4):1856–1879.


Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., and Steyerberg, E. W. (2019).
Calibration: the achilles heel of predictive analytics.
BMC medicine, 17(1):1–7.


von Mises, R. (1928).
Wahrscheinlichkeit Statistik und Wahrheit.
Springer.

References


von Mises, R. (1939).
Probability, statistics and truth.
Macmillan.


Watson, G. S. (1964).
Smooth regression analysis.
Sankhyā: The Indian Journal of Statistics, Series A, pages 359–372.


Wilks, D. S. (1990).
On the combination of forecast probabilities for consecutive precipitation periods.
Weather and Forecasting, 5(4):640–650.


Wüthrich, M. V. and Ziegel, J. (2024).
Isotonic recalibration under a low signal-to-noise ratio.
Scandinavian Actuarial Journal, 2024(3):279–299.