Autocalibration, Predictive Models and Insurance Pricing

Arthur Charpentier

with Michel Denuit (UCL) & Julien Trufin (UCB)

ASTIN Reading Club (2023)

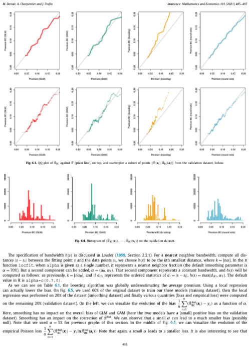
Autocalibration paper

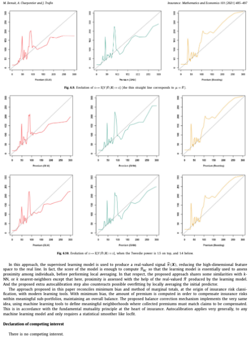
Denuit et al. (2021) *Autocalibration and Tweedie-dominance for insurance pricing with machine learning*, IME & ArXiv:2103.03635

Insurance: Mathematics and Economics

Autocalibration and Tweedie-dominance for insurance pricing with machine learning

Michel Denuit[a], Arthur Charpentier[b], Julien Trufin[c,*]





Autocalibration github

https://github.com/freakonometrics/autocalibration/

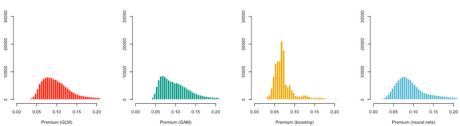
https://github.com/freakonometrics/autocalibration/

README.md

```
                kern="rect",deg=0,alpha=.05)
```

The figure below is the distribution of the premiums according to the four models

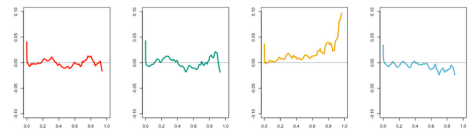
```
par(mfrow=c(1,4))
hist(valid_prime_glm, breaks = (0:(201+max(valid_prime_glm)))/200,xlim=c(0,.2),col=clrpal4[1],borde
     main="",xlab="Premium (GLM)",ylab="",ylim=c(0,37000))
hist(valid_prime_gam, breaks = (0:(201+max(valid_prime_gam)))/200,xlim=c(0,.2),col=clrpal4[2],borde
     main="",xlab="Premium (GAM)",ylab="",ylim=c(0,37000))
hist(valid_prime_bst, breaks = (0:(201+max(valid_prime_bst)))/200,xlim=c(0,.2),col=clrpal4[3],borde
     main="",xlab="Premium (boosting)",ylab="",ylim=c(0,37000))
hist(valid_prime_nn, breaks = (0:(201+max(valid_prime_nn)))/200,xlim=c(0,.2),col=clrpal4[4],border='
     main="",xlab="Premium (neural nets)",ylab="",ylim=c(0,37000))
```



The following graph is a visulisation of $s \mapsto \mathbb{E}[Y|\pi(\boldsymbol{X}) = s]$ where s is some premium level, for the three models π. If π is close to μ the curve should be close to the first diagonal.

README.md

```
vy_gam = predict(fit_loc_gam, newdata=vx)
vy_bst = predict(fit_loc_bst, newdata=vx)
vy_nn = predict(fit_loc_nn, newdata=vx)
plot(q_glm,vy_glm~vx,lwd=3,col=clrpal4[1],xlim=c(0,1),xlab="Premium quantile (GLM)",ylab="",ylim=c(-
abline(h=0,lwd=.4)
plot(q_gam,vy_gam~vx,lwd=3,col=clrpal4[2],xlim=c(0,1),xlab="Premium quantile (GAM)",ylab="",ylim=c(-
abline(h=0,lwd=.4)
plot(q_bst,vy_bst~vx,lwd=3,col=clrpal4[3],xlim=c(0,1),xlab="Premium quantile (boosting)",ylab="",yli
abline(h=0,lwd=.4)
plot(q_nn,vy_nn~vx,lwd=3,col=clrpal4[4],xlim=c(0,1),xlab="Premium quantile (neural nets)",ylab="",y
abline(h=0,lwd=.4)
```



Observe that $\mathbb{E}[Y|\pi(\boldsymbol{X}) = F_\pi^{-1}(u)] \geq F_\pi^{-1}(u)$ which reflects the local bias of the estimator π (except perhaps for very low risks). Other plots can be used to visualised using the following plots.

Since $\mathbb{E}[Y|\pi(\boldsymbol{X}) = F_\pi^{-1}(u)] \sim F_\pi^{-1}(u)$ if $\pi \sim \mu$, we can plot $\mathbb{E}[Y|\pi(\boldsymbol{X}) = F_\pi^{-1}(u)]/F_\pi^{-1}(u)$ which should be close to 1,

Warning and Preamble (1)

Important to define what probabilities are (see in French, in English, from the *Casualty Actuarial and Statistical Task Force* (CASTF) *Reading Club*)

"*Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means* ", Russell (1929), quoted in Bell (1945)

Very often, the "physical" probabilities receive an objective value only posterior on the basis of the law of large numbers, the empirical frequency converge towards the probability (frequentist theory of probabilities)

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(X_i \in A)}_{\text{(empirical) frequency}} \overset{\text{a.s.}}{\to} \underbrace{\mathbb{P}(X \in A)}_{\text{probability}} \text{ as } n \to \infty$$

Warning and Preamble (1)

But this approach is unable to make sense of the probability of a "single singular event", as noted by von Mises (1928, 1939).

"*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. We can say nothing about the probability of death of an individual even if we know his condition of life and health in detail. The phrase 'probability of death', when it refers to a single person, has no meaning for us at all*."

Warning and Preamble (2)

Econometrics is based on a probabilistic model, unlike most machine learning approaches, see Charpentier et al. (2018)

- in SVMs, the distance to the separation line is used as a score which can then be interpreted as a probability - Platt scaling, Platt et al. (1999) or isotonic regression Zadrozny and Elkan (2001, 2002) (see also Niculescu-Mizil and Caruana (2005) "good probabilities")
- in Neural Nets, Rumelhart et al. (1985), Rumelhart et al. (1986) Hertz et al. (1991) and Buntine and Weigend (1991) proposed to formalize back-propagation in a Bayesian context, taken up by MacKay (1992) and Neal (1992). State of the art in Neal (2012), more than 25 years ago (or more recently Neal (2012) Theodoridis (2015), Gal and Ghahramani (2016) and Goulet et al. (2021))

Motivation and Context

Ph.Demetri @PhDemetri · 6 avr. 2021

All this talk about XGboost prompted me to try it again on some toy datasets I have laying around.

The long and the short of it is: XGBoost results in a better AUC than my logistic regression (99.7 v 87) but XGB is so poorly calibrated it doesn't make sense to truest the probs

8 3 45

Ph.Demetri @PhDemetri · 6 avr. 2021

Calibration is a thing I never see people talk about in data science. We should really care about calibration more than we do.

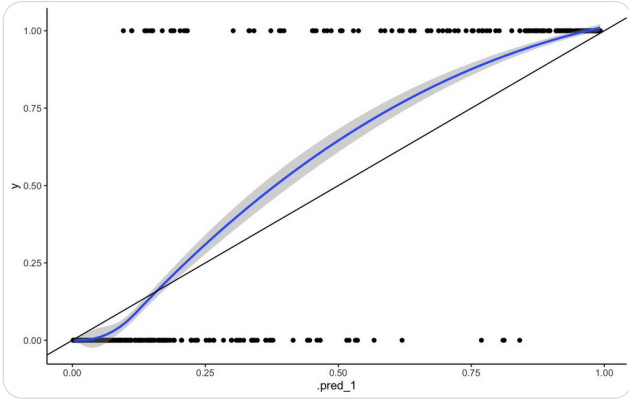
2 1 27

Motivation and Context

Ph.Demetri @PhDemetri · 6 avr. 2021

Calibration for the xgboost model.

Yuck



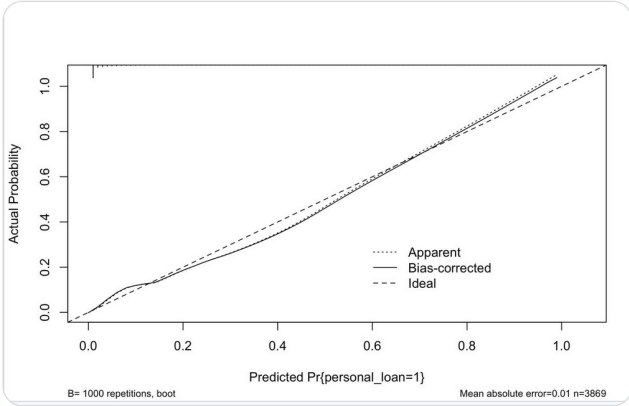
4 1 13

Motivation and Context

Ph.Demetri @PhDemetri · 6 avr. 2021

Calibration for the logistic regression

Not perfect, but a hell of a lot better, eh?



2 10

Calibration

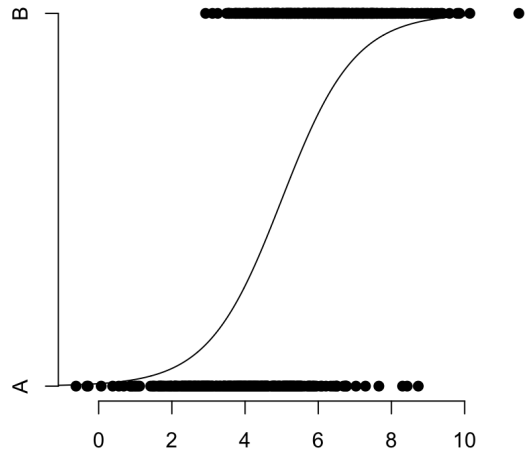
Dataset (x_i, y_i), $y_i \in \{A, B\}$,

$$\mathbb{P}[Y = \text{'A'}] = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

Convert into $y_i \in \{0, 1\}$,

$$m(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

(classical logistic regression)



Calibration

Given $p \in (0,1)$, let $\mathcal{X}_p \subset \mathbb{R}$ such that $\forall x \in \mathcal{X}_p$, $m(x) = p$

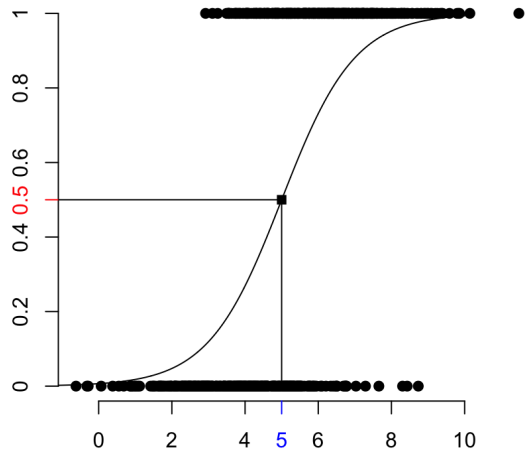
We want to compute

$$\mathbb{E}\big[Y \big| m(X) = p\big]$$

or a sample version

$$\frac{1}{\#\mathcal{V}_p} \sum_{i:x_i \in \mathcal{V}_p} y_i$$

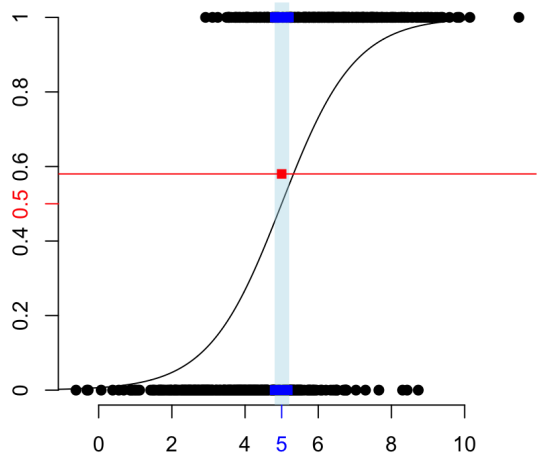
where $\mathcal{V}_p$ is a neighborhood of x



Calibration

Given $\mathcal{V}_p$ is a neighborhood of x

$$\frac{1}{\#\mathcal{V}_p} \sum_{i:x_i \in \mathcal{V}_p} y_i$$

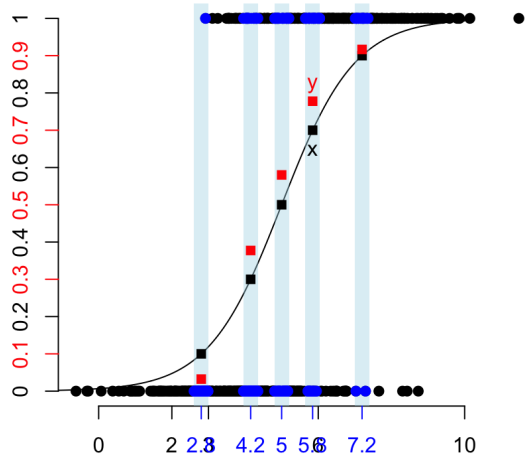


Calibration

for various $p \in (0,1)$, we can approximate

$$\mathbb{E}\big[Y \big| m(X) = p\big]$$

using averages on neighborhoods of $\mathcal{X}_p \subset \mathbb{R}$

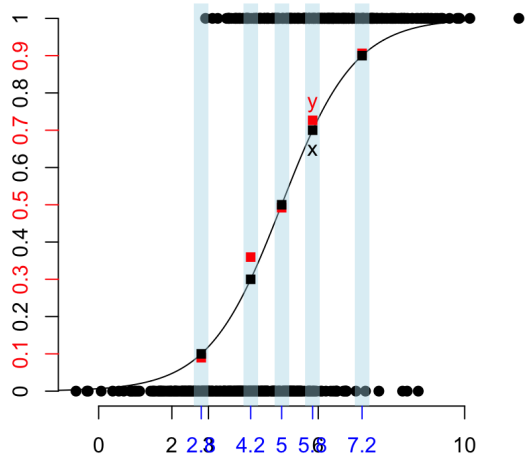


Calibration

We can also estimate

$$\mathbb{E}\big[Y\big|m(X)=p\big]$$

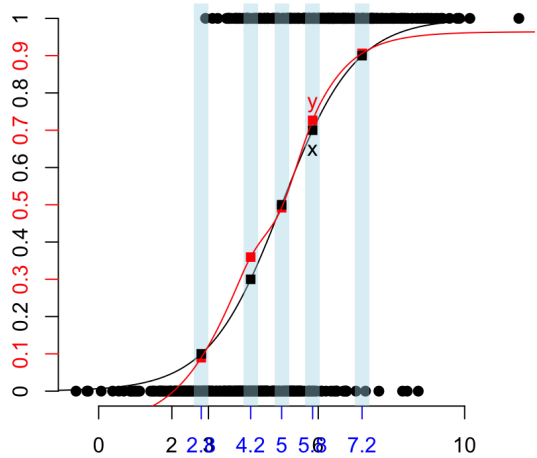
using a local regression of Y against $m(X)$,
on $\{m(x_i), y_i\}$



Calibration

Consider function

$$p \mapsto \mathbb{E}\big[Y \big| m(X) = p\big]$$

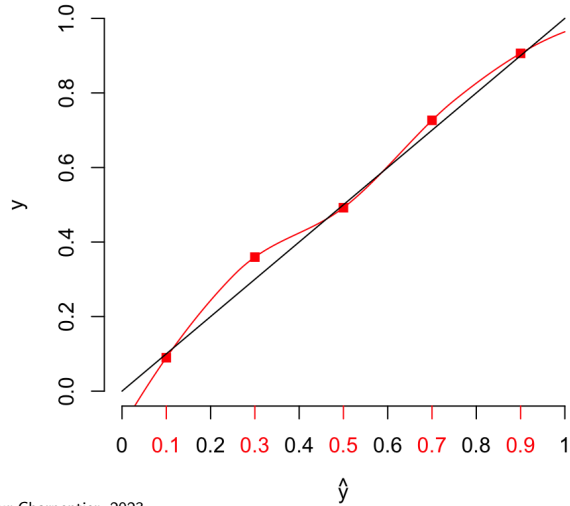


Calibration

If the model is well-calibrated

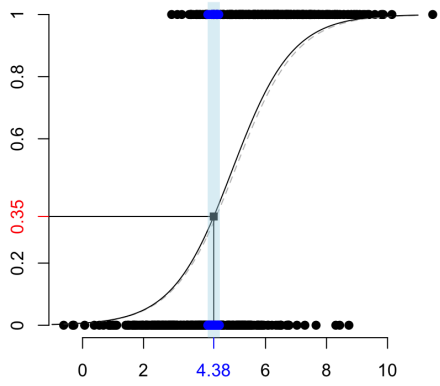
$$\mathbb{E}\big[Y\big|m(X)=p\big]\sim p$$

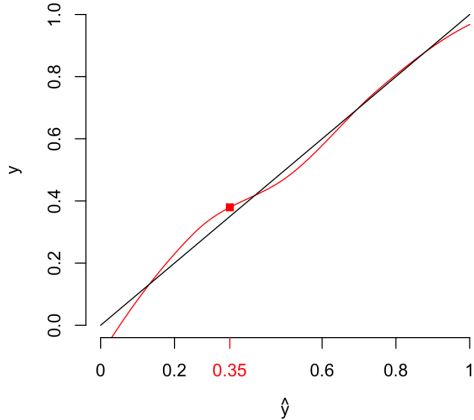
if not, m is "locally biased"



Calibration

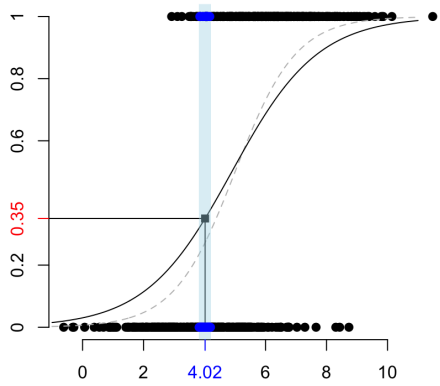
Let $\widehat{m}(x) = \dfrac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 x)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x)}$ denote the prediction of a logistic regression

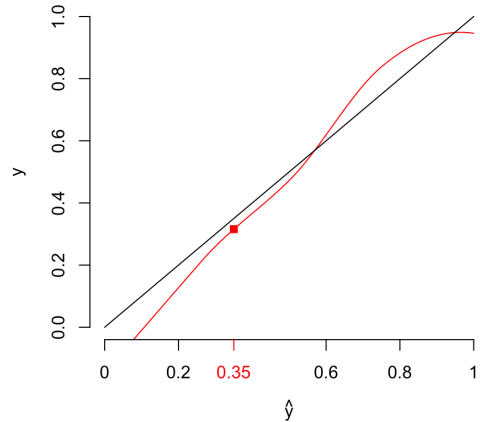




Calibration

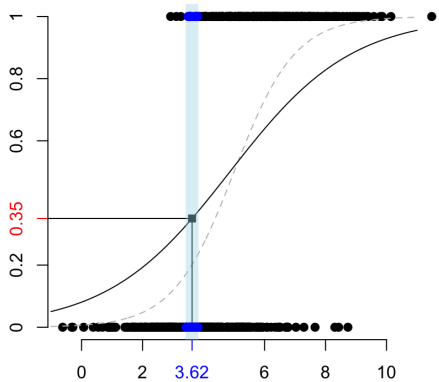
Consider a penalized Ridge logistic regression $\widehat{m}(x) = \dfrac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 x)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x)}$

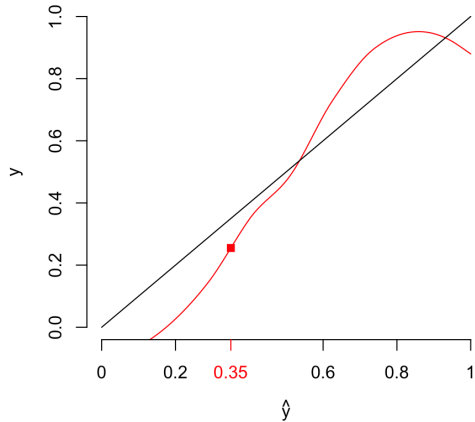




Calibration

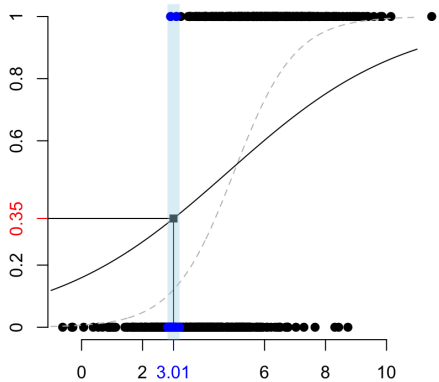
Consider a penalized Ridge logistic regression $\widehat{m}(x) = \dfrac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 x)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x)}$

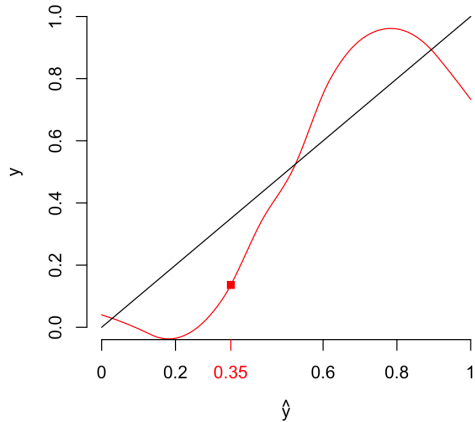




Calibration

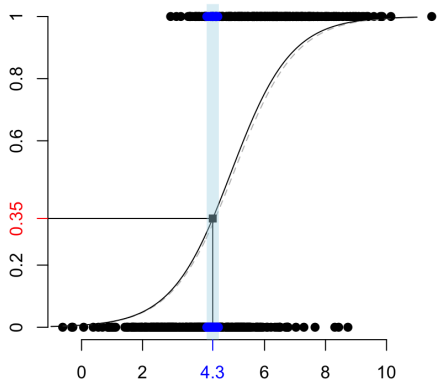
Consider a penalized Ridge logistic regression $\widehat{m}(x) = \dfrac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 x)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x)}$

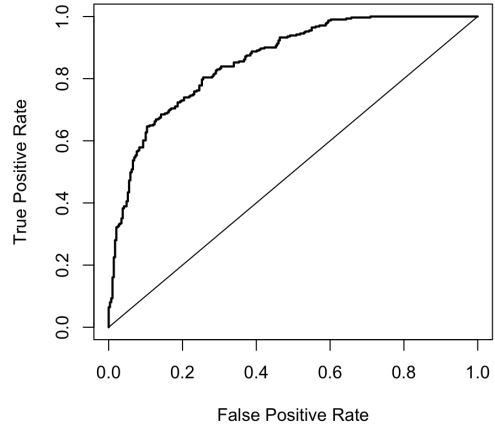




Calibration vs. Accuracy

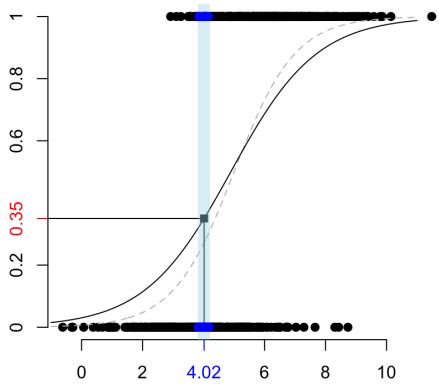
We can get ROC curves for those Ridge logistic regressions

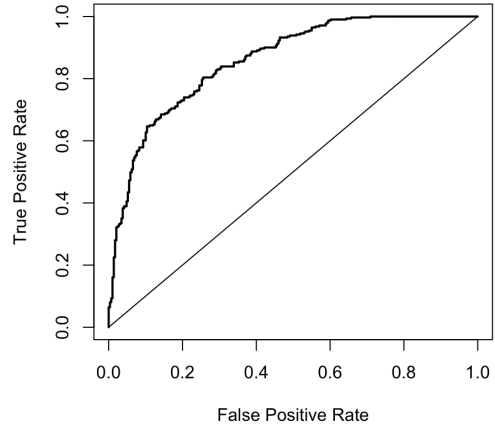




Calibration vs. Accuracy

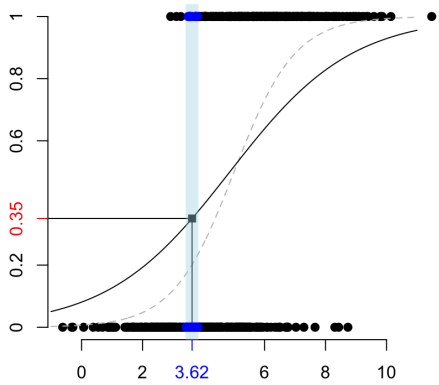
We can get ROC curves for those Ridge logistic regressions

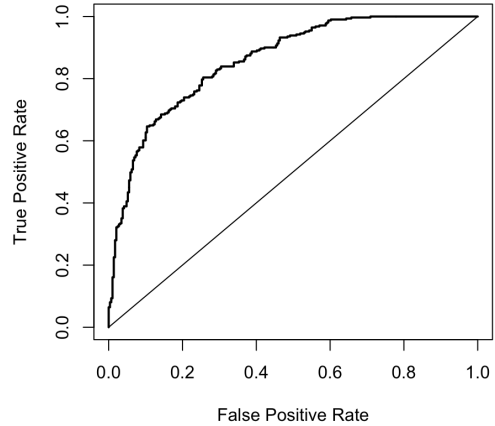




Calibration vs. Accuracy

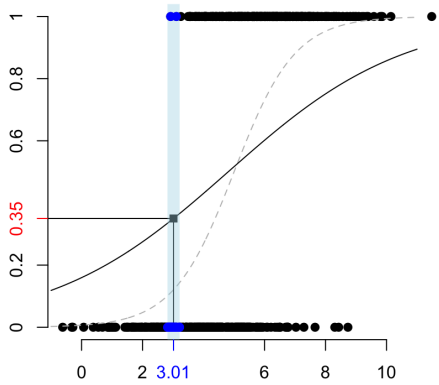
We can get ROC curves for those Ridge logistic regressions

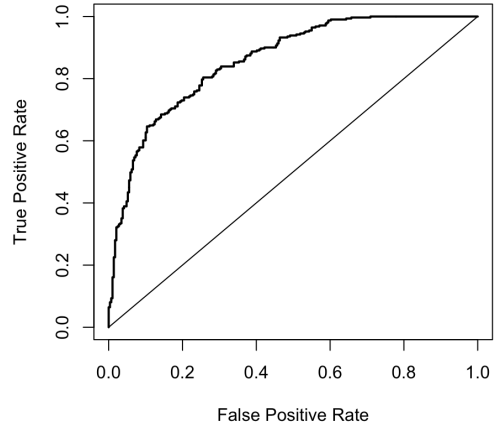




Calibration vs. Accuracy

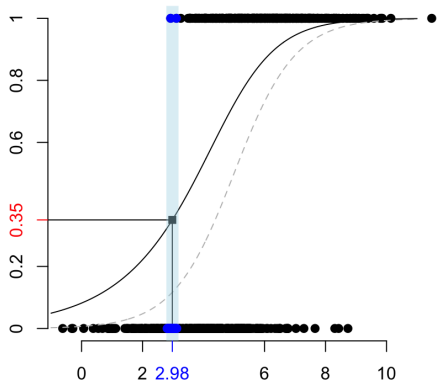
We can get ROC curves for those Ridge logistic regressions

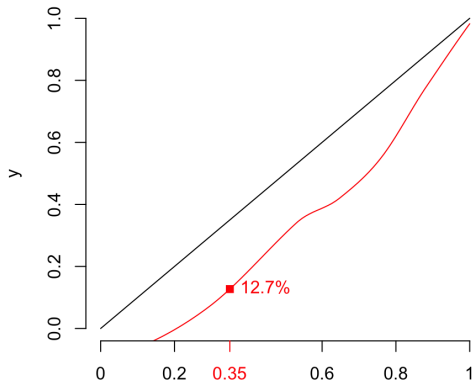




Calibration

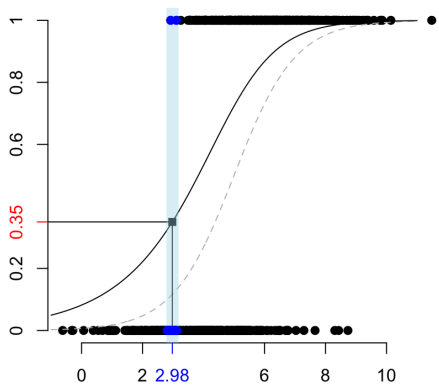
Consider the square root of the logistic regression $\widehat{m}(x) = \sqrt{\dfrac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 x)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x)}}$

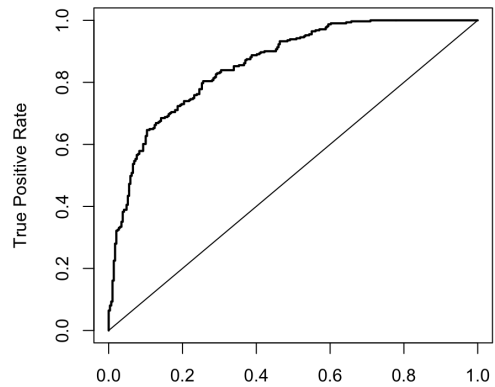




Calibration and Accuracy

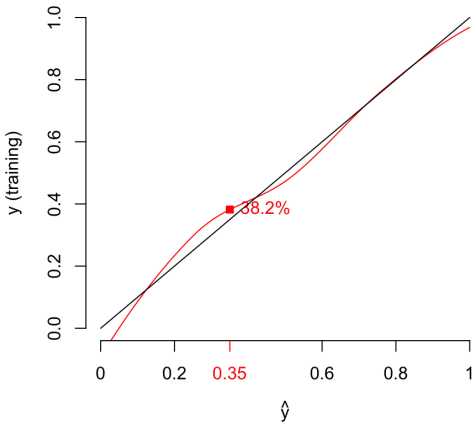
Consider the square root of the logistic regression $\widehat{m}(x) = \sqrt{\dfrac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 x)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x)}}$

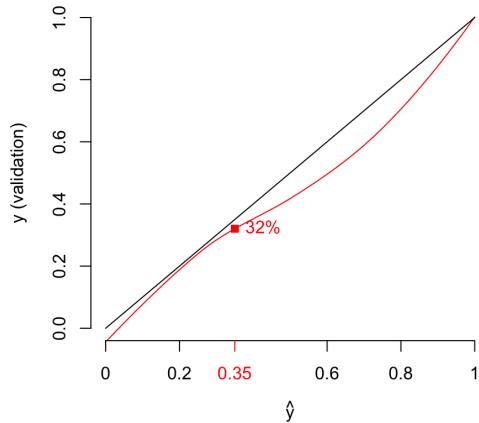




Training vs. Validation Datasets

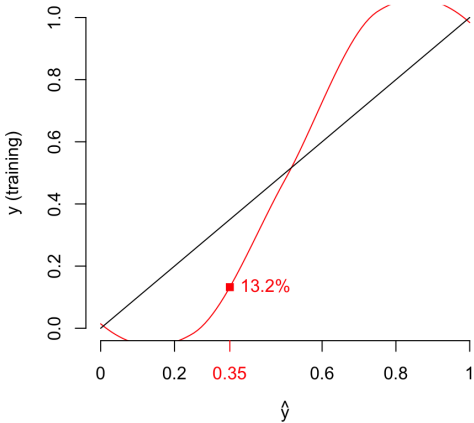
Consider a logistic regression, ← training dataset vs. validation dataset →

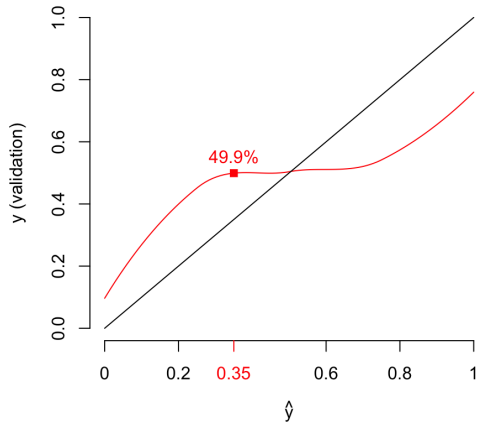




Training vs. Validation Datasets

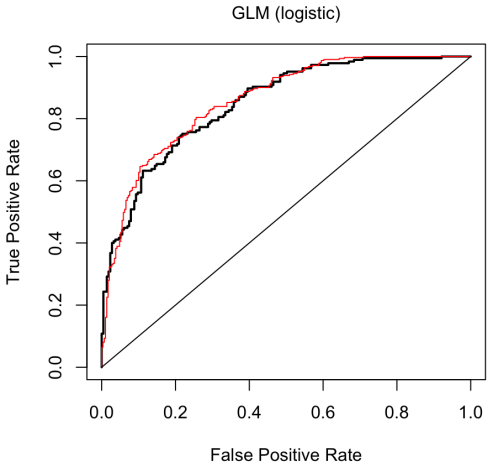
Consider a random forest, ← training dataset vs. validation dataset →

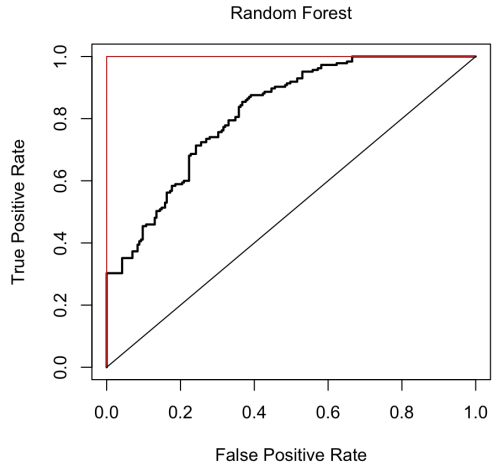




Training vs. Validation Datasets

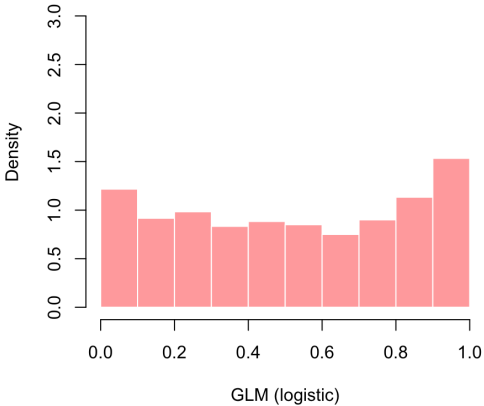
ROC curves, ← logistic regression vs. random forest →

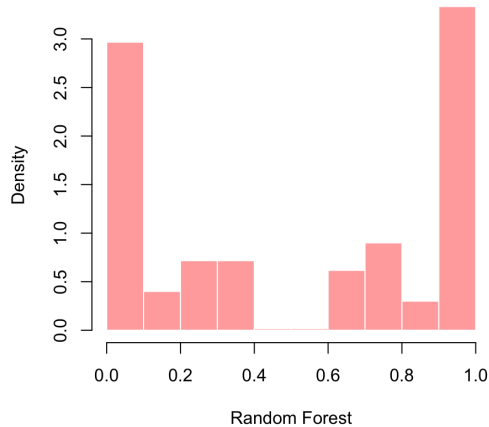




Training vs. Validation Datasets

Distribution of scores for the random forest (training dataset)



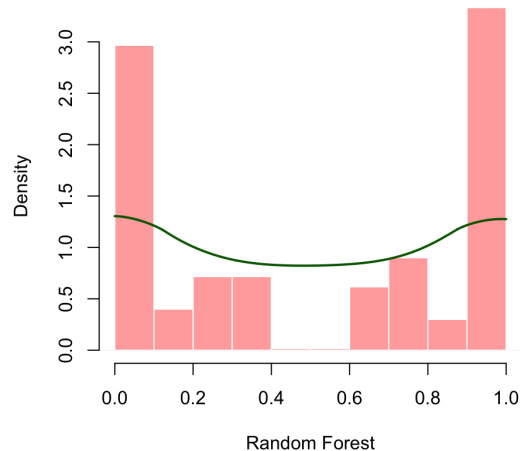


Training vs. Validation Datasets

Distribution of scores, ← logistic regression vs. random forest → (training dataset)

"*deep neural networks tend to be over-confident and poorly calibrated after training*", Wang et al. (2021)

"*Guo et al. (2017) have shown that modern neural networks are poorly calibrated and over-confident despite having better performance*", Müller et al. (2019)



Accuracy, Calibration & Bias

For classication problems, **calibration** measures how well your model's scores can be interpreted as probabilities. **Accuracy** measures how often your model produces correct answers.

"**Accuracy** *is a qualitative term referring to whether there is agreement between a measurement made on an object and its true (target or reference) value.* **Bias** *is a quantitative term describing the difference between the average of measurements made on the same object and its true value..* " Handbook of Statistical Methods

"**Well calibrated** *classifiers are probabilistic classifiers for which the output can be directly interpreted as a confidence level. For instance, a well calibrated (binary) classifier should classify the samples such that among the samples to which it gave a*[predicted probability] *value close to 0.8, approximately 80% actually belong to the positive class.* " scikit learn: Probability calibration

Calibration

"*Suppose that a forecaster sequentially assigns probabilities to events. He is* **well calibrated** *if, for example, of those events to which he assigns a probability 30 percent, the long-run proportion that actually occurs turns out to be 30 percent.* ", Dawid (1982), The Well-Calibrated Bayesian,

"*Out of all the times you said there was a 40 percent chance of rain, how often did rain actually occur? If, over the long run, it really did rain about 40 percent of the time, that means your forecasts were* **well calibrated** ", Silver (2012), The signal and the noise,

"*we desire that the estimated class probabilities are reflective of the true underlying probability of the sample* ", Kuhn and Johnson (2013) Applied Predictive Modeling

See Murphy and Epstein (1967), Roberts (1968), Gneiting and Raftery (2005) on ensemble methods for weather forecasting, or more generally Lichtenstein et al. (1977), Oakes (1985), Gneiting et al. (2007).

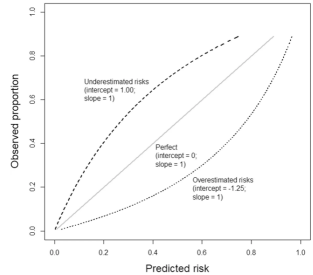
Calibration & Auto-calibration

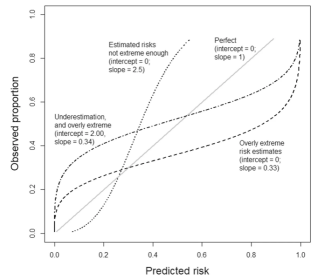
As explained in Van Calster et al. (2019), "*among patients with an estimated risk of 20%, we expect 20 in 100 to have or to develop the event*",

- 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.

Hosmer-Lemeshow test, from Hosmer Jr et al. (2013))(logistic regression), and Bier score, from Brier et al. (1950) and Murphy (1973)

Function ploted in psychological papers Keren (1991)



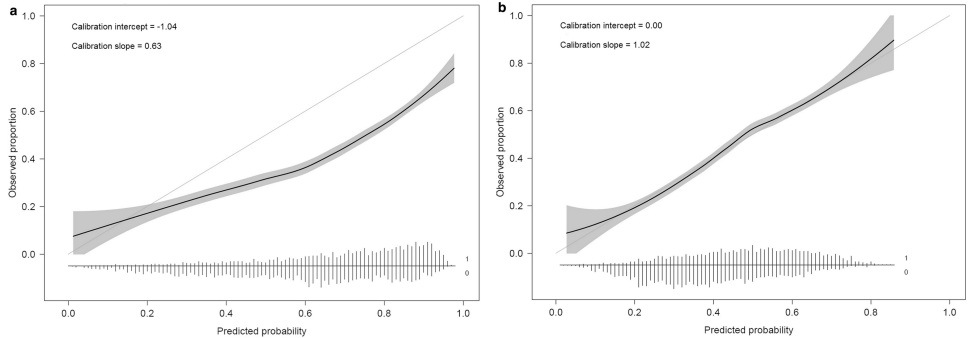


Calibration & Auto-calibration

Krüger and Ziegel (2021) Generic conditions for forecast dominance,

Definition 3.1 "*the forecast X of Y is an auto-calibrated forecast of Y if $\mathbb{E}(Y|X) = X$ almost surely* ", or $\mathbb{E}(Y|\widehat{Y} = y) = y, \ \forall y$

Van Calster et al. (2019) Calibration: the Achilles heel of predictive analytics



Calibration & Fairness

Consider a binary outcome Y, a binary protected attribute P, and a prediction $\widehat{Y}$.

Classical fairness concepts are related to independence ($\widehat{Y} \perp\!\!\!\perp P$) or separation ($\widehat{Y} \perp\!\!\!\perp P \mid Y$)

Based on sufficiency ($Y \perp\!\!\!\perp P \mid \widehat{Y}$), Sokolova et al. (2006) introduced a concept related to calibration (and later taken up by Kleinberg et al. (2016) and Zafar et al. (2017)).

We have calibration parity if

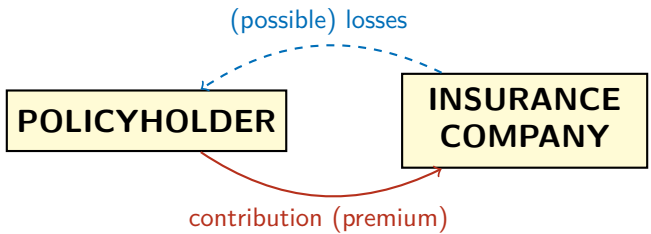
$$\mathbb{P}[Y = 1 | \widehat{Y} = y, P = 0] = \mathbb{P}[Y = 1 | \widehat{Y} = y, P = 1], \ \forall y \in (0, 1).$$

We have an fairness of good calibration if

$$\mathbb{P}[Y = 1 | \widehat{Y} = y, P = 0] = \mathbb{P}[Y = 1 | \widehat{Y} = y, P = 1] = y, \ \forall y \in (0, 1).$$

Insurance

"*Insurance is the contribution of the many to the misfortune of the few* "



The "*contribution* " is obtained using predictive models
(interpretability / black box / etc issues)

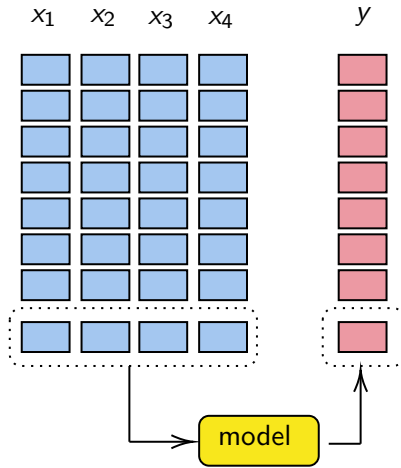
A model, $m : \boldsymbol{x} \mapsto y$

To train / estimate a model m,
we need a dataset, i.e. a collection
of observations $(\boldsymbol{x}_i, y_i)$

Usually y_i denotes the annual loss

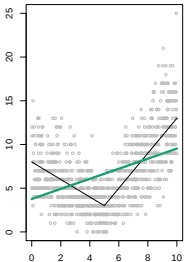
$$y_i = \sum_{j=1}^{n_i} z_{i,j}$$

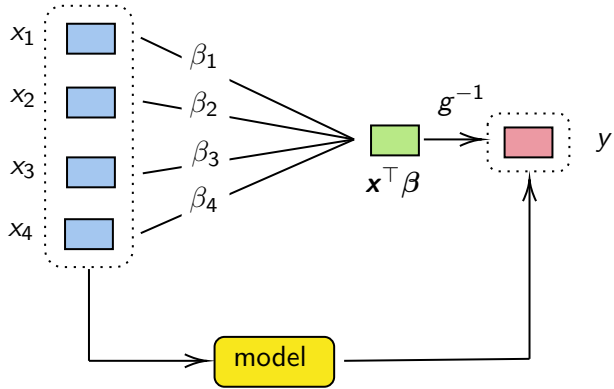
n is the annual frequency
z is the cost of a single claim
(see Tweedie models)
To illustrate, y is a counting variable



A model, $m : \boldsymbol{x} \mapsto y$ (GLM)

Poisson regression

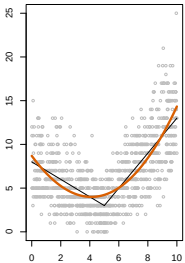


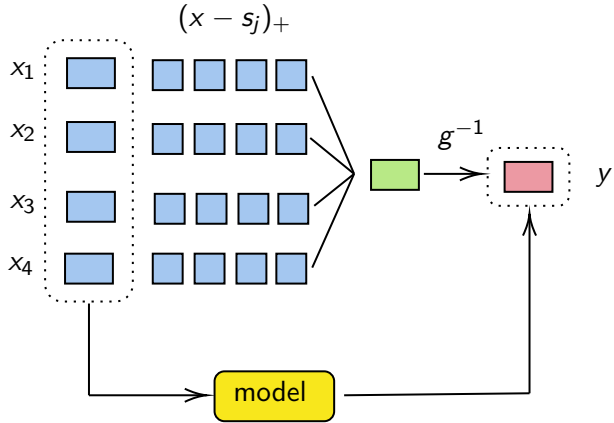


$$m(\boldsymbol{x}) = g^{-1}(\boldsymbol{x}^\top\boldsymbol{\beta}) = g^{-1}\left(\sum_{j=1}^{p}\beta_j x_j\right)$$

A model, $m : \boldsymbol{x} \mapsto y$ (GAM)

Poisson regression

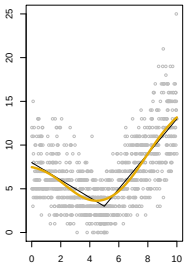


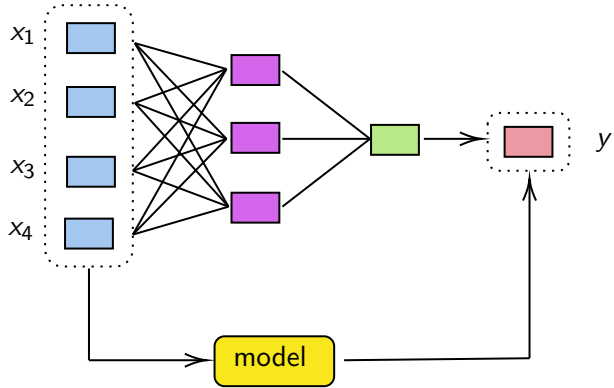


$$m(\boldsymbol{x}) = g^{-1}\left(\sum_{j=1}^{p} \beta_j \psi_j(x_j)\right), \text{ where } \psi_j(x) = \sum_{k=1}^{5} \alpha_{j,k}(x - s_{j,k})_+$$

A model, $m : \boldsymbol{x} \mapsto y$ (neural nets)

Poisson loss

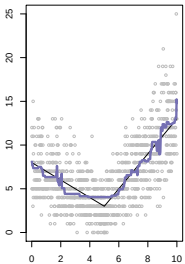


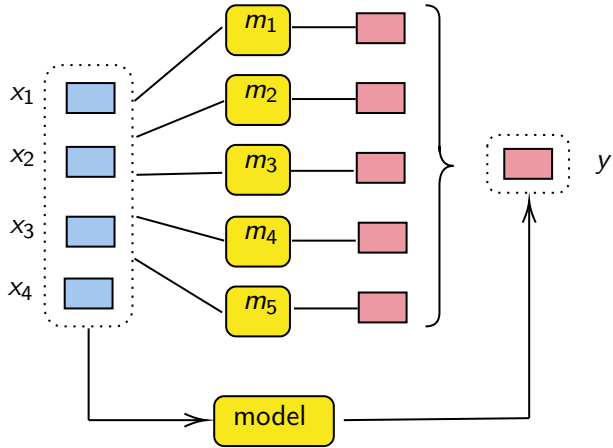


E.g. $m(\boldsymbol{x}) = \sum_{j=1}^{3} \omega_{1:j}\ h(\boldsymbol{x}^\top \boldsymbol{\omega}_{2:j})$

A model, $m : \boldsymbol{x} \mapsto y$ (ensemble parallel, bagging)

Random forest

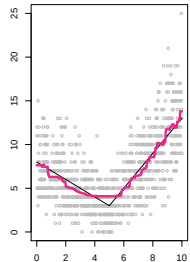


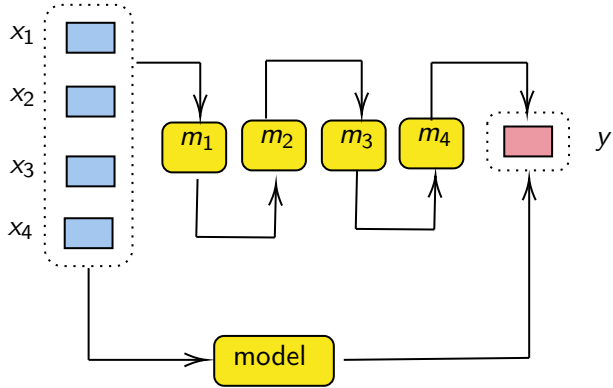


E.g. $m(\boldsymbol{x}) = \dfrac{1}{B}\sum_{b=1}^{B} m_b(\boldsymbol{x})$ where $m_b(\boldsymbol{x}) = \sum_{k=1}^{K} \alpha_{b,k}\mathbf{1}(\boldsymbol{x} \in \mathcal{X}_{b,k})$

A model, $m : \boldsymbol{x} \mapsto y$ (ensemble sequential, boosting)

Boosting





$m(\boldsymbol{x}) = \sum_{t=1}^{T} m_t(\boldsymbol{x})$ where m_t is a (weak) model on $y_i - m_{t-1}(\boldsymbol{x}_i)$

(residuals from the previous step) such as (not too deep) trees

Not a model, $\widehat{m}: \boldsymbol{x} \mapsto y$ (local regression)

To approximate $\mathbb{E}[Y]$ use

$$\widehat{m} = \underset{\mu \in \mathbb{R}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \frac{1}{n} [y_i - \mu]^2 \right\}$$

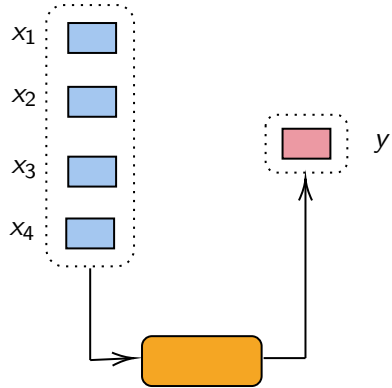
To approximate $\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$, use

$$\widehat{m}(\boldsymbol{x}) = \underset{\mu \in \mathbb{R}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \omega_i(\boldsymbol{x}) [y_i - \mu]^2 \right\}$$

where, see Loader (1999),

$$\omega_i(\boldsymbol{x}) \propto k_\alpha \left(\|\boldsymbol{x} - \boldsymbol{x}_i\| \right)$$

or k nearest neighbors indicator...

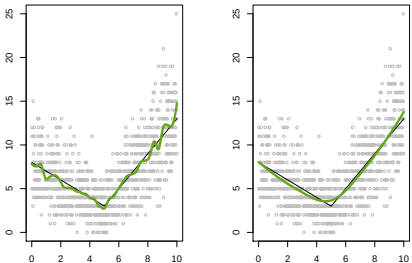


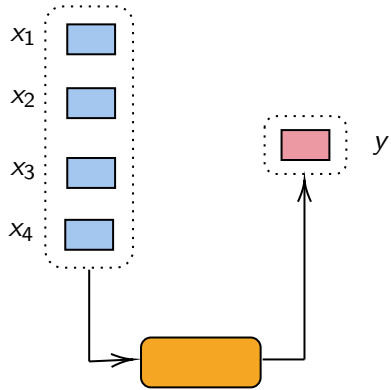
Behaves poorly in high dimension (but efficient in dimension 1)

Not a model, $\widehat{m} : \boldsymbol{x} \mapsto y$ (local regression)

$$\widehat{m}(\boldsymbol{x}) = \sum_{i=1}^{n} \omega_i(\boldsymbol{x})\, y_i, \text{ with } \sum_{i=1}^{n} \omega_i(\boldsymbol{x}) = 1$$

see `locfit` in R

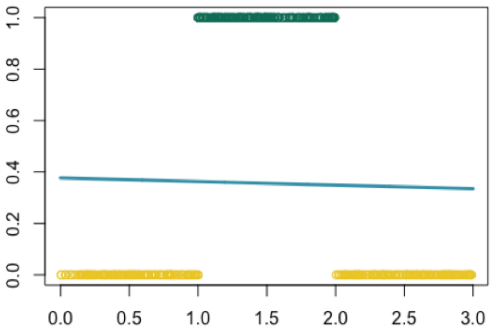


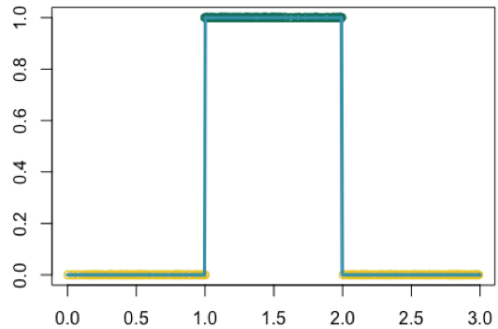


(depends on a bandwidth parameter α)
Provides a local estimate of $\mathbb{E}[Y|X = x]$ on a neighborhood of x

GLM, GAM & Neural Nets

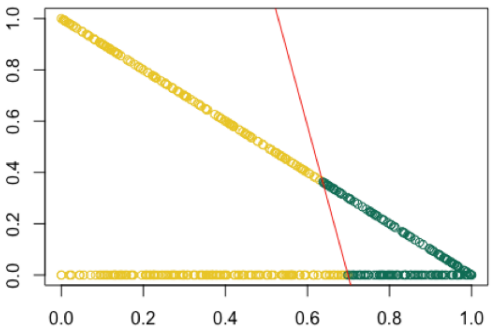
Consider some simple dataset $\{(x_i, y_i)\}$.
Fit a GLM and a GAM (linear, 3 knots)

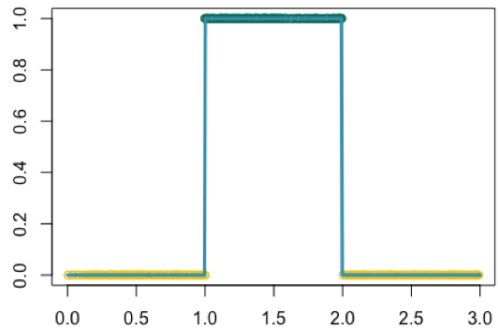




GLM, GAM & Neural Nets

Consider some simple dataset $\{(x_i, y_i)\}$.
Fit a GLM and a GAM (linear, 3 knots)

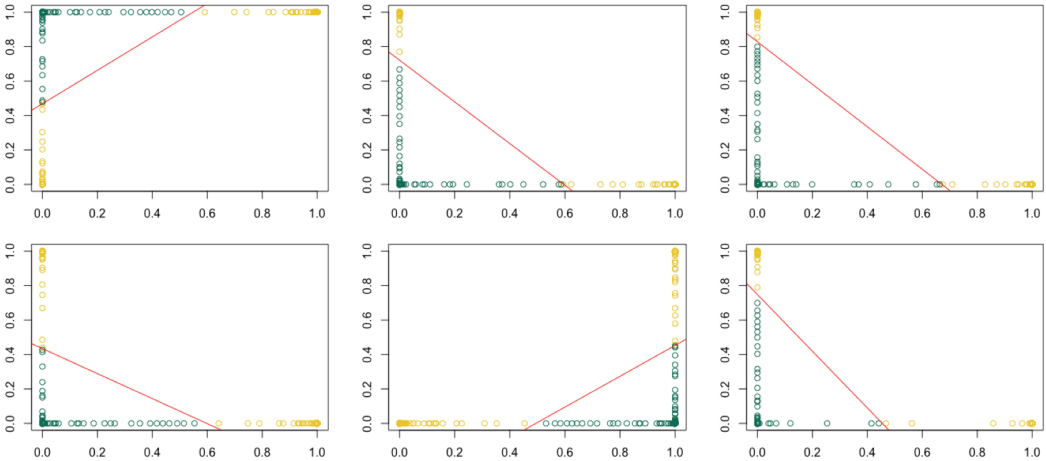




Two variables, $x_1 = x$ and $x_2 = (x - s)_+$

GLM, GAM & Neural Nets

Consider some simple dataset $\{(x_i, y_i)\}$.
Fit a Neural Net



GLM, Bias, & Economic Interpretation

For GLMs, $f(y_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\varphi} + c(y_i, \varphi)\right)$

The natural parameter for y_i : θ_i
Prediction for y_i : $\widehat{y}_i = \mu_i = \mathbb{E}(Y_i) = b'(\theta_i)$
Score associated with y_i : $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$
Link function : g such that $\eta_i = g(\mu_i) = g(b'(\theta_i))$

$$\log \mathcal{L}(\boldsymbol{\theta}, \varphi, \boldsymbol{y}) = \sum_{i=1}^{n} \log \mathcal{L}_i = \sum_{i=1}^{n} \left[\frac{y_i\theta_i - b(\theta_i)}{\varphi} + c(y_i, \varphi)\right]$$

First order conditions: $\frac{\partial \log \mathcal{L}_i}{\partial \beta_j} = \frac{\partial \log \mathcal{L}_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j}$

$$\frac{\partial \log \mathcal{L}_i}{\partial \theta_i} = \frac{y_i - \mu_i}{\varphi}, \quad \frac{\partial \theta_i}{\partial \mu_i} = \left(\frac{\partial \mu_i}{\partial \theta_i}\right)^{-1} = \frac{1}{b''(\theta_i)} = \frac{1}{V(\mu_i)}$$

$$\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \boldsymbol{x}_i^\top \boldsymbol{\beta}}{\partial \beta_j} = x_{i,j}, \quad \frac{\partial \mu_i}{\partial \eta_i} = (\frac{\partial \eta_i}{\partial \mu_i})^{-1} = (g'(\mu_i))^{-1}$$

GLM, Bias, & Economic Interpretation

With canonical link $g_\star = b'^{-1}$, i.e. $\eta_i = \theta_i$,

$$\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{0}, \quad \text{where } \widehat{\boldsymbol{y}} = \boldsymbol{\mu}$$

so, if there is an intercept, $\mathbf{1}^\top(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{0}$, i.e. $\sum_{i=1}^n y_i = \sum_{i=1}^n \widehat{y}_i$

which is the empirical version (training dataset) of $\mathbb{E}[Y] = \mathbb{E}[\widehat{Y}]$

(see logistic regression or Poisson with log-link)

But more generally, the first order condition is

$$\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{\Delta}(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{0},$$

where $\boldsymbol{W} = \text{diag}\big((V(\mu_i)g'(\mu_i)^2)^{-1}\big)$ and $\boldsymbol{\Delta} = \text{diag}\big(g'(\mu_i)\big)$.

But usually not an important issue in ML classification problems

GLM, Bias, & Economic Interpretation

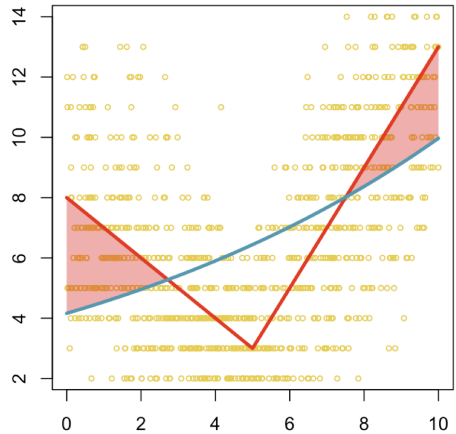
Model $\widehat{\pi}$ is globally unbiased if $\mathbb{E}[\widehat{\pi}(\boldsymbol{X})] = \mathbb{E}[Y]$, $\sum_{i=1}^{n} \widehat{y}_i = \sum_{i=1}^{n} y_i$

Model $\widehat{\pi}$ is locally unbiased if $\mathbb{E}[Y|\widehat{\pi}(\boldsymbol{X}) = s] = s$

GLM $\widehat{\pi}$ globally unbiased,
but possibly locally biased
Major economic impact

- $\widehat{\pi}(\boldsymbol{x}) < \mu(\boldsymbol{x})$
  attractive, but underpriced
- $\widehat{\pi}(\boldsymbol{x}) > \mu(\boldsymbol{x})$
  not attractive

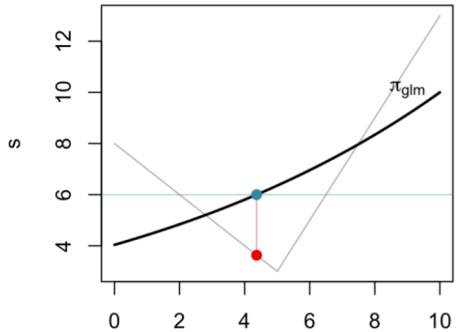
Natural idea: plot
$s \mapsto \mathbb{E}[Y|\widehat{\pi}(\boldsymbol{X}) = s]$ " $= \mu(\widehat{\pi}^{-1}(s))$ "

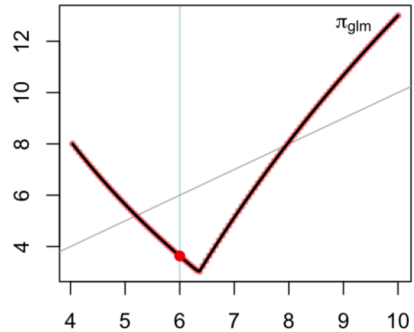


Computing $\mathbb{E}[Y|\widehat{\pi}(\boldsymbol{X}) = s]$

True model $\mu(x)$ and GLM model $\widehat{\pi}(x)$

- select s, e.g. $s = 6$
- compute $x = \widehat{\pi}^{-1}(s)$ ($\widehat{\pi}$ is strictly increasing), here $x = 4.2$
- compute $\mu(\widehat{\pi}^{-1}(s))$, here 3.8
- plot $(s, \mu(\widehat{\pi}^{-1}(s)))$



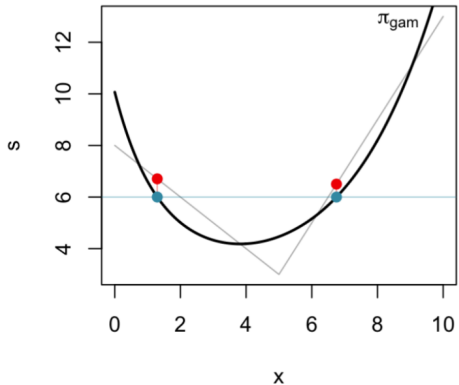


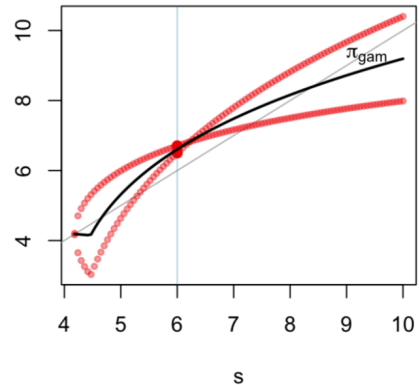
$s \mapsto \mathbb{E}[Y|\widehat{\pi}(X) = s] = \mu(\widehat{\pi}^{-1}(s))$ since $\mu(x) = \mathbb{E}[Y|X = x]$.

Computing $\mathbb{E}[Y|\widehat{\pi}(\boldsymbol{X}) = s]$

True model $\mu(x)$ and GAM model $\widehat{\pi}(x)$

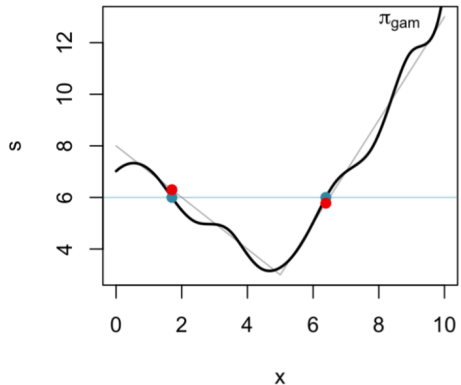
- select s, e.g. $s = 6$
- compute $\widehat{\mathcal{X}_s^{\pi}} = \{\boldsymbol{x} \in \mathcal{X}, \widehat{\pi}(\boldsymbol{x}) = s\}$, here $\{1.6; 6.7\}$
- compute $\mu(\widehat{\pi}^{-1}(s))$, $\{6.4, 6.2\}$ and its mean $\overline{\mu}(\widehat{\pi}^{-1}(s))$, 6.3
- plot $(s, \overline{\mu}(\widehat{\pi}^{-1}(s)))$

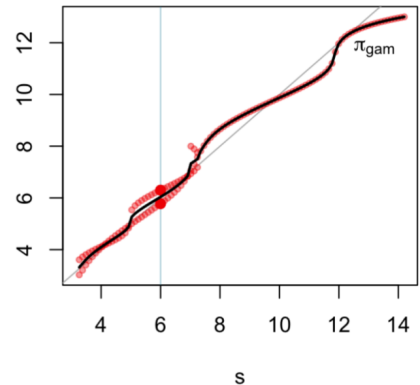




Computing $\mathbb{E}[Y|\widehat{\pi}(\boldsymbol{X}) = s]$

True model $\mu(x)$ and GAM model $\widehat{\pi}(x)$ (more degrees of freedom)
Plot $s \mapsto \mathbb{E}[Y|\widehat{\pi}(X) = s]$, should be close to the first diagonal





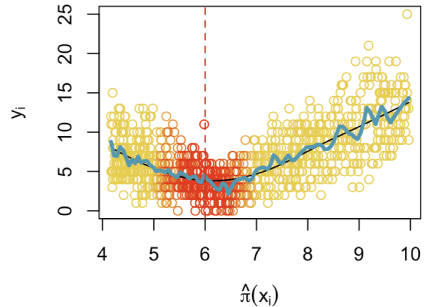
Seems locally unbiased...

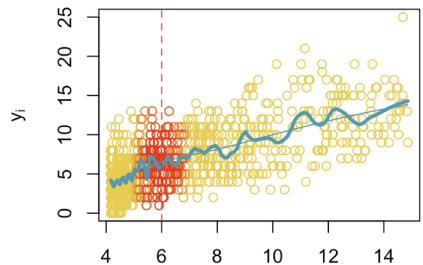
- impossible to get the figure on the left in higher dimension
- we used here μ but in practice, μ is unknown !

$\mathbb{E}[Y|\widehat{\pi}(\boldsymbol{X}) = s]$: empirical version

How to plot $s \mapsto \mathbb{E}[Y|\widehat{\pi}(X) = s]$?
but in real life, μ is unknown
Consider $\{(\widehat{\pi}(\boldsymbol{x}_i), y_i)\}$
and fit a local regression

- fit a model $\widehat{\pi}$
- estimate $\mathbb{E}[Y|\widehat{\pi}(X) = s]$
  local regression on $\{(\widehat{\pi}(\boldsymbol{x}_i), y_i)\}$
- local (multiplicative) correction
  $$\lambda_\alpha(s) = \frac{\mathbb{E}[Y|\widehat{\pi}(X) = s]}{s}$$
- correct $\widehat{\pi}$
  $$\widehat{\pi}_{BC}(\boldsymbol{x}) = \lambda_\alpha(\widehat{\pi}(\boldsymbol{x})) \cdot \widehat{\pi}(\boldsymbol{x})$$



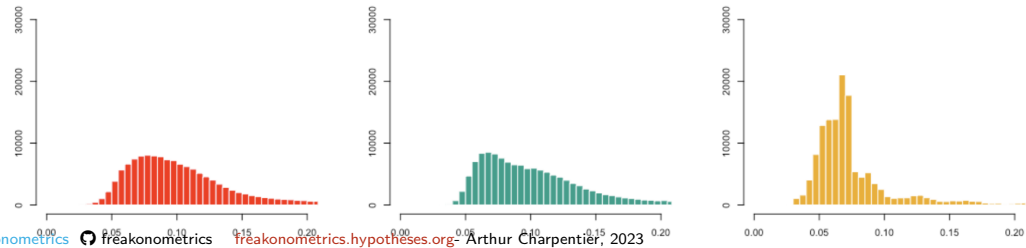


Application of a motor-insurance dataset

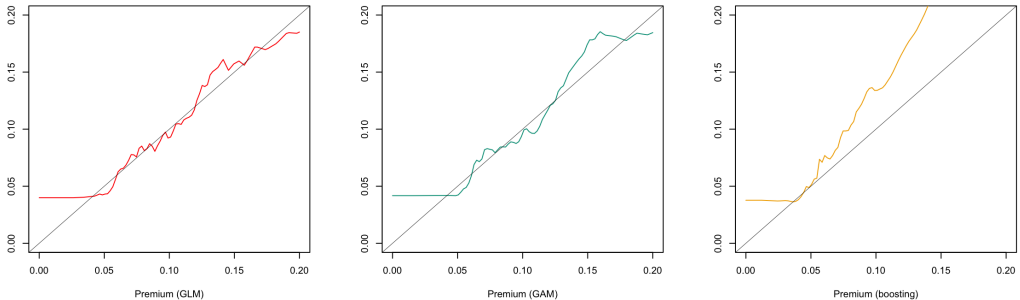
Here we focus only on claims (annual) frequency, corrected from the exposure, `freMTPL2freq` from `CASDataset` package,

	π^{glm}	π^{gam}	π^{bst}
average $\overline{\pi}$	0.1087	0.1092	0.0820
10% quantile	0.0605	0.0598	0.0498
90% quantile	0.1682	0.1713	0.1244

Table 1: Summary statistics on $\{\pi(\boldsymbol{x}_1), \cdots, \pi(\boldsymbol{x}_n)\}$, on the validation dataset (assuming an exposure of 1 to provide annualized predictions).

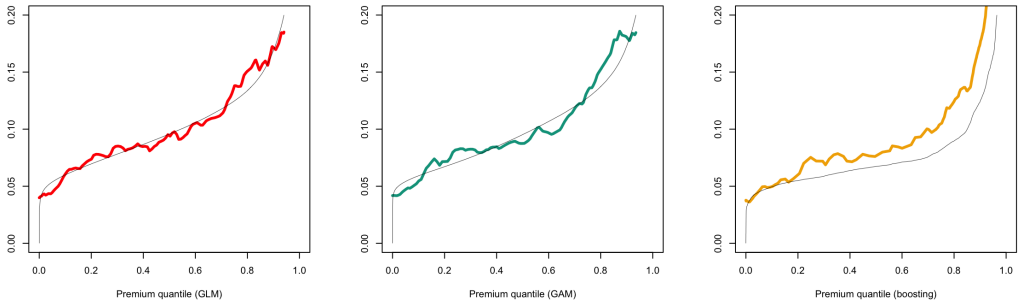


Application of a motor-insurance dataset



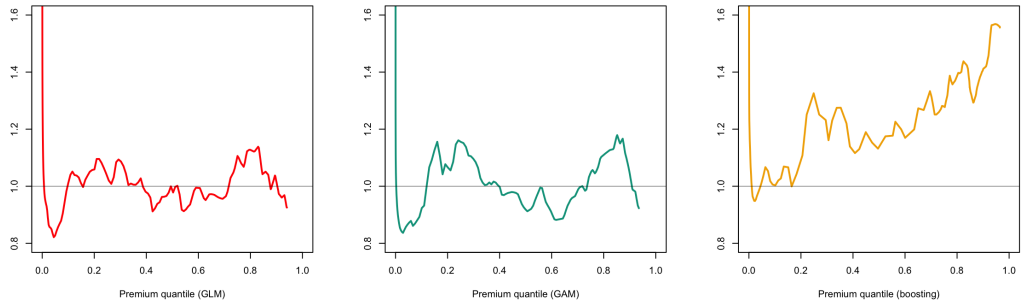
Evolution of $s \mapsto \mathbb{E}[Y|\widehat{\pi}(\boldsymbol{X}) = s]$

Application of a motor-insurance dataset



Evolution of $u \mapsto \mathbb{E}[Y|\widehat{\pi}(\boldsymbol{X}) = F_{\pi}^{-1}(u)]$

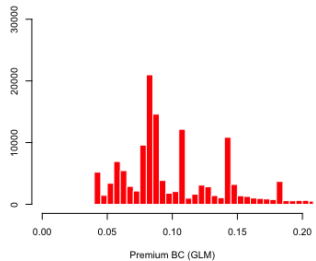
Application of a motor-insurance dataset

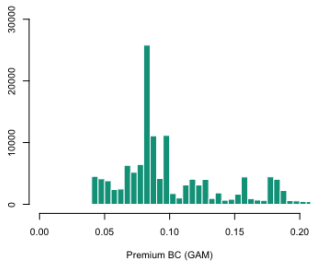


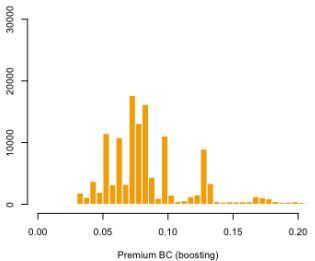
Multificative correction $\lambda_\alpha(u) = \mathbb{E}[Y|\widehat{\pi}(\boldsymbol{X}) = F^{-1}_{\widehat{\pi}}(u)]/F^{-1}_{\widehat{\pi}}(u)$

Application of a motor-insurance dataset

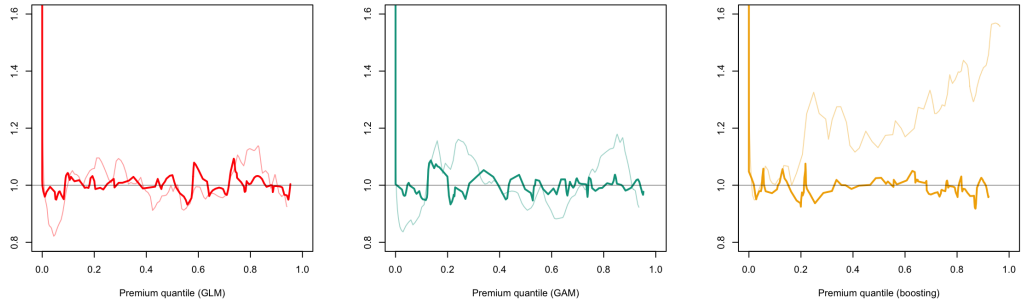
Multificative correction $\lambda_\alpha(u) = \mathbb{E}[Y|\widehat{\pi}(\boldsymbol{X}) = F^{-1}_{\widehat{\pi}}(u)]/F^{-1}_{\widehat{\pi}}(u)$







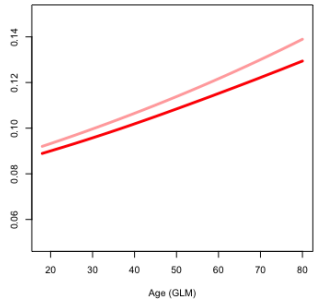
Application of a motor-insurance dataset

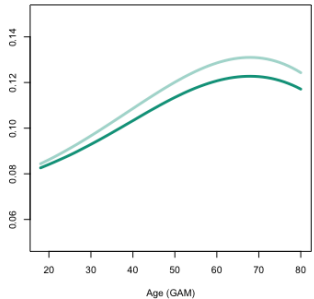


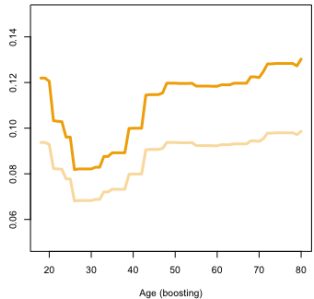
λ_α on the corrected model $\widehat{\pi}_{BC}$,

Application of a motor-insurance dataset

some partial dependence plot
(on the age of the driver)

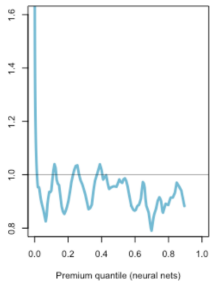


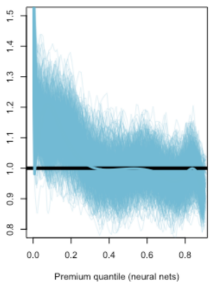


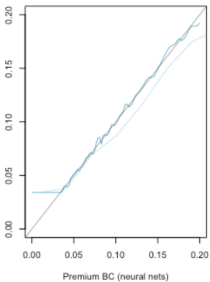


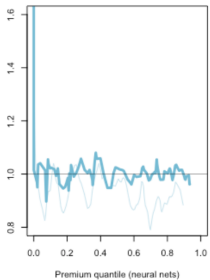
Using Neural Nets on motor-insurance dataset

We can also look at neural networks performance



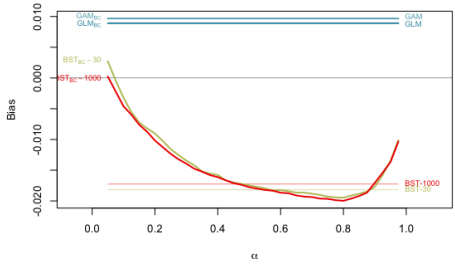


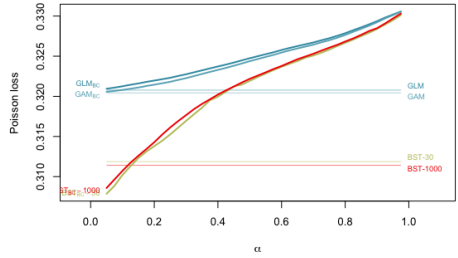


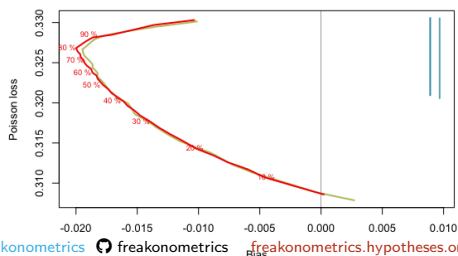


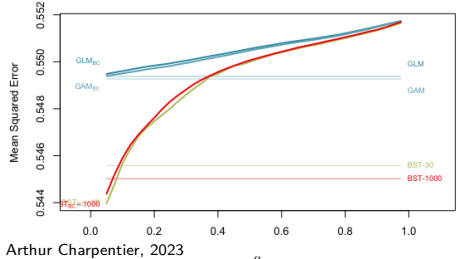
Reproducibility issues here

Choice of α

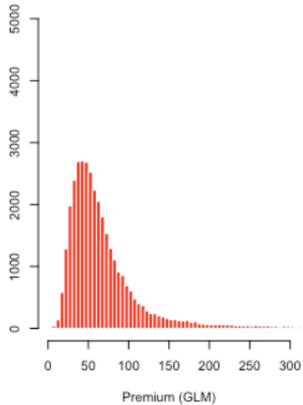


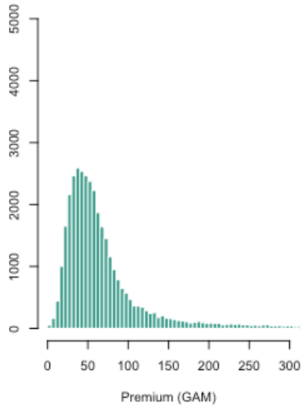


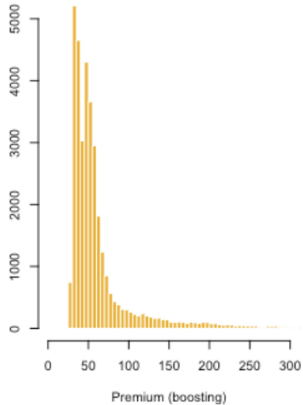




Tweedie model on motor-insurance dataset







Wrap-Up

- fit a model $\widehat{\pi}$ using an ML algorithm
- estimate $\mathbb{E}[Y|\widehat{\pi}(X)]$ with local regression on $\{(\widehat{\pi}(\boldsymbol{x}_i), y_i)\}$
- local (multiplicative) correction $\lambda_\alpha(s) = \dfrac{\mathbb{E}[Y|\widehat{\pi}(X) = s]}{s}$
- correct $\widehat{\pi}$ by setting $\widehat{\pi}_{BC}(\boldsymbol{x}) = \lambda_\alpha(\widehat{\pi}(\boldsymbol{x})) \cdot \widehat{\pi}(\boldsymbol{x})$

Denuit, M., Charpentier, A. & Trufin, J. (2021). Autocalibration and Tweedie-dominance for Insurance Pricing with Machine Learning, (IME)

References

Bell, E. T. (1945). *The development of mathematics*. Courier Corporation.

Brier, G. W. et al. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Buntine, W. L. and Weigend, A. S. (1991). Bayesian back-propagation. *Complex Systems*, 5.

Charpentier, A. (2023). *Insurance, biases, discrimination and fairness*. Springer Nature.

Charpentier, A., Flachaire, E., and Ly, A. (2018). Econometrics and machine learning. *Economics & Statistics*, 505(1):147–169.

Dawid, A. P. (1982). The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610.

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.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

References

Gneiting, T. and Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, 310(5746):248–249.

Goulet, J.-A., Nguyen, L. H., and Amiri, S. (2021). Tractable approximate gaussian inference for bayesian neural networks. *J. Mach. Learn. Res.*, 22:251–1.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the theory of neural computation*. CRC Press.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta psychologica*, 77(3):217–273.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv*, 1609.05807.

Krüger, F. and Ziegel, J. F. (2021). Generic conditions for forecast dominance. *Journal of Business & Economic Statistics*, 39(4):972–983.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

References

Lichtenstein, S., Fischhoff, B., and Phillips, L. D. (1977). Calibration of probabilities: The state of the art. *Decision making and change in human affairs*, pages 275–324.

MacKay, D. J. (1992). A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472.

Müller, R., Kornblith, S., and Hinton, G. E. (2019). When does label smoothing help? *Advances in neural information processing systems*, 32.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600.

Murphy, A. H. and Epstein, E. S. (1967). Verification of probabilistic predictions: A brief review. *Journal of Applied Meteorology and Climatology*, 6(5):748–755.

Neal, R. M. (1992). Bayesian training of backpropagation networks by the hybrid monte carlo method. Technical report, Citeseer.

Neal, R. M. (2012). *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

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.

References

Oakes, D. (1985). Self-calibrating priors do not exist. *Journal of the American Statistical Association*, 80(390):339–339.

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

Roberts, H. V. (1968). On the meaning of the probability of rain. In *first national conference on statistical meteorology*.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

Silver, N. (2012). *The signal and the noise: Why so many predictions fail-but some don't*. Penguin.

Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, pages 1015–1021. Springer.

Theodoridis, S. (2015). *Machine learning: a Bayesian and optimization perspective*. Academic press.

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.

References

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

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

Wang, D.-B., Feng, L., and Zhang, M.-L. (2021). Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. *Advances in Neural Information Processing Systems*, 34:11809–11820.

Zadrozny, B. and Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *ICML*, volume 1, pages 609–616. Citeseer.

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. *arXiv*, 1507.05259.