# Table of Contents

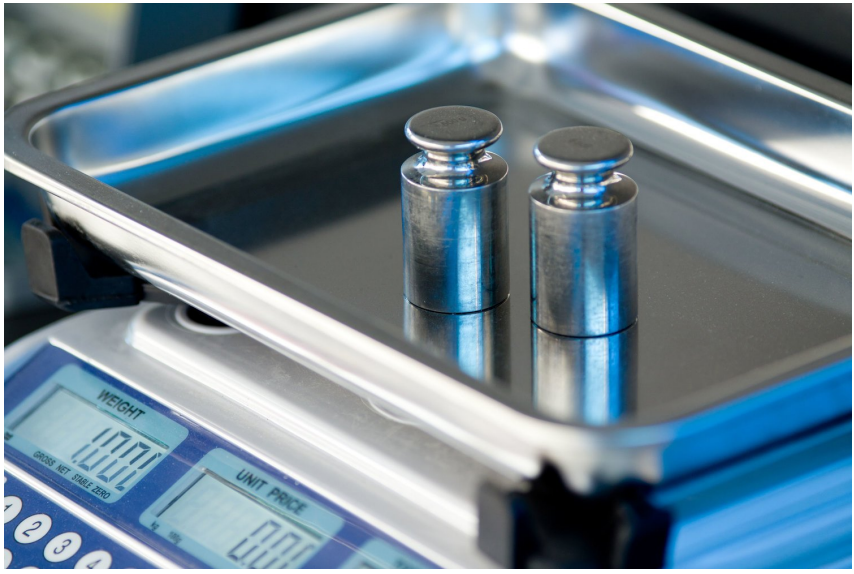



Figure: Scale calibration
source: https://ssec-epos.co.uk/hardware/weighing-scale-calibration-and-service/

# Motivation for Calibration

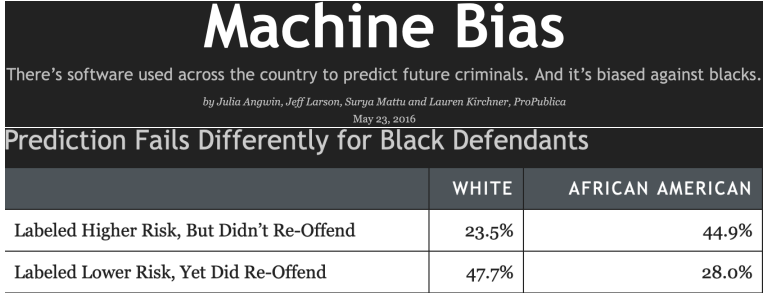
- Is the model fit for its prediction task?
- How well does the predictions align with observations?
- Detect bias and discrimination.

Bias can result in bad news.

# Motivation for Calibration

- Is the model fit for its prediction task?
- How well does the predictions align with observations?
- Detect bias and discrimination.

Bias can result in bad news.



## Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

*by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica*
May 23, 2016

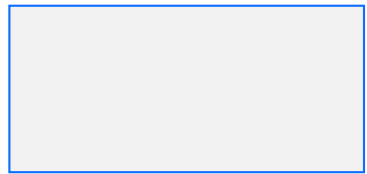
### Prediction Fails Differently for Black Defendants

| | WHITE | AFRICAN AMERICAN |
|---|---|---|
| Labeled Higher Risk, But Didn't Re-Offend | 23.5% | 44.9% |
| Labeled Lower Risk, Yet Did Re-Offend | 47.7% | 28.0% |

Figure: ProPublica article on COMPAS.

# Measuring Model Calibration

Distance



Is $m(\boldsymbol{x})$ calibrated?

# Measuring Model Calibration

Distance

**Strict**
**Identification Function** $V$

Is $m(\boldsymbol{x})$ calibrated?

# 2 Notions of Calibration

## Definition

Given a feature–response pair $(\boldsymbol{X}, Y)$, the model $m(\boldsymbol{X})$ is **conditionally calibrated** for the functional $T$ if

$$m(\boldsymbol{X}) = T(Y|\boldsymbol{X}) \qquad \text{almost surely.}$$

The model $m(\boldsymbol{X})$ is **auto-calibrated** for $T$ if

$$m(\boldsymbol{X}) = T(Y|m(\boldsymbol{X})) \qquad \text{almost surely.}$$

## Note

- Calibration uncovers the use of information.
- Conditional calibrated models use information ideally and predict **oracle function** $\boldsymbol{x} \to T(Y|\boldsymbol{X} = \boldsymbol{x})$.
- The trivial model $m(\boldsymbol{X}) = T(Y)$ is auto-calibrated, but a.s. not conditionally calibrated.
- **Unconditional calibration** does not have such a definition in terms of $T$ and $m(x)$.

# Identification Functions

## Definition

Let $\mathcal{F}$ be a class of probability distributions where the functional $T$ is defined on. A **strict $\mathcal{F}$-identification function** for $T$ is a function $V(z, y)$ in a forecast $z$ and an observation $y$ such that

$$\int V(z, y)\, \mathrm{d}F(y) = 0 \quad \Longleftrightarrow \quad z \in T(F) \qquad \text{for all } z \in \mathbb{R},\ F \in \mathcal{F}\,. \tag{22}$$

If only the implication $\Longleftarrow$ in (22) holds, then $V$ is just called an $\mathcal{F}$-identification function for $T$. If $T$ admits a strict $\mathcal{F}$-identification function, it is **identifiable** on $\mathcal{F}$.

## Canonical strict identification functions

| Functional | Strict Identification Function | Domain of $y$, $z$ |
|---|---|---|
| expectation $\mathbb{E}[Y]$ | $V(z, y) = z - y$ | $\mathbb{R}$ |
| $\alpha$-expectile | $V(z, y) = 2\lvert\mathbb{1}\{z \geq y\} - \alpha\rvert(z - y)$ | $\mathbb{R}$ |
| median $F_Y^{-1}(0.5)$ | $V(z, y) = \mathbb{1}\{z \geq y\} - 1/2$ | $\mathbb{R}$ |
| $\alpha$-quantile $F_Y^{-1}(\alpha)$ | $V(z, y) = \mathbb{1}\{z \geq y\} - \alpha$ | $\mathbb{R}$ |

# Osbands's Principle for Scoring Functions

Assume $\mathbb{P}(Y = a) = p$ and $\mathbb{P}(Y = b) = 1 - p$. Then, for any $S$ consistent for $T$ and smooth in its first argument, the expected score $\epsilon(c) = pS(c, a) + (1 - p)S(c, b)$ is minimised at $c = t$ fulfilling

$$\epsilon'(t) = pS_{(1)}(t, a) + (1 - p)S_{(1)}(t, b) = 0\,.$$

Furthermore, (22) implies

$$pV(t, a) + (1 - p)V(t, b) = 0\,.$$

Combinining them gives for all pairwise distinct $a, b$ and $t \in D$

$$S_{(1)}(t, a)/V(t, a) = S_{(1)}(t, b)/V(t, b) = \text{function in } t \text{ alone}$$

and we can write, for some function $h : \mathrm{D} \to D$

$$S_{(1)}(z, y) = h(z)V(z, y)\,. \tag{23}$$

# Osbands's Principle for Identification Functions

## Characterisation[7]

*Subject to mild regularity conditions on $\mathcal{F}$ and $V$:*
If $V$ is a strict identification function for $T : \mathcal{F} \to \mathrm{A} \subseteq \mathbb{R}^k$, then

$$\{h(z)V(z,y) | h : \mathrm{A} \to \mathbb{R}^{k,k}, \det h(z) \neq 0 \text{ for all } z \in \mathrm{A}\} \tag{24}$$

is the entire class of strict identification functions for $T$.

## Identifiability and elicitability

*Under some richness assumptions on the class $\mathcal{F}$ and continuity assumptions on $T$:*
For one-dimentional functionals $T$, identifiability and elicitability are equivalent.

[7] T. Dimitriadis, T. Fissler and J. Ziegel. "Osband's Principle for Identification Functions". arxiv:2208.07685

# Identification Functions and Calibration

Let $V$ be any strict $\mathcal{F}$-identification function for $T$.

## Conditional calibration

Suppose that $\mathcal{F}$ contains the conditional distributions $F_{Y|\boldsymbol{X}=\boldsymbol{x}}$ for almost all $\boldsymbol{x} \in \mathcal{X}$. Application of (22) to these conditional distributions yields that $m(\boldsymbol{x}) \in T(Y|\boldsymbol{X} = \boldsymbol{x})$ if and only if $\int V(m(\boldsymbol{x}), y)\, \mathrm{d}F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y) = 0$. This shows that $m$ is conditionally calibrated for $T$ if and only if

$$\mathbb{E}[V(m(\boldsymbol{X}), Y)|\boldsymbol{X}] = 0 \qquad \text{almost surely.} \tag{25}$$

## Auto-Calibration

Suppose the conditional distributions $F_{Y|m(\boldsymbol{X})=z}$ are in $\mathcal{F}$ for almost all $z \in \mathbb{R}$. Then $m$ is auto-calibrated for $T$ if and only if

$$\mathbb{E}[V(m(\boldsymbol{X}), Y)|m(\boldsymbol{X})] = 0 \qquad \text{almost surely.} \tag{26}$$

## Note

By the tower property of the conditional expectation, conditional calibration implies auto-calibration for identifiable functionals with a sufficiently rich class $\mathcal{F}$.

# Unconditional Calibration

**Definition**

Let $V$ be any strict identification function for $T$. We say that $m(\boldsymbol{X})$ is **unconditionally calibrated** for $T$ relative to $V$ if $\mathbb{E}[V(m(\boldsymbol{X}), Y)] = 0$.

Unless $m(\boldsymbol{X})$ is constant, and in stark contrast to conditional calibration and auto-calibration, the notion of unconditional calibration depends on the choice of the identification function $V$ used.

# Assessing Calibration: Overview

| Notion | Definition | Check |
|---|---|---|
| conditional calibration | $m(\boldsymbol{X}) = T(Y\|\boldsymbol{X})$ | $\mathbb{E}[V(m(\boldsymbol{X}), Y)\|\boldsymbol{X}] = 0 \quad a.s.$ |
| auto-calibration | $m(\boldsymbol{X}) = T(Y\|m(\boldsymbol{X}))$ | $\mathbb{E}[V(m(\boldsymbol{X}), Y)\|m(\boldsymbol{X})] = 0 \quad a.s.$ |
| unconditional calibration | $\mathbb{E}[V(m(\boldsymbol{X}), Y)] = 0$ | $\mathbb{E}[V(m(\boldsymbol{X}), Y)] = 0$ |

Table: Types of calibration for an identifiable functional $T$ with strict identification function $V$.

# Assessing Calibration: Test Functions

How to measure *good* calibration on a sample level?

## Test functions

$$\mathbb{E}[V(m(\boldsymbol{X}), Y)|\boldsymbol{X}] = 0 \quad a.s.$$

is equivalent to

$$\mathbb{E}[\varphi(\boldsymbol{X})V(m(\boldsymbol{X}), Y)] = 0 \quad \text{for } \textbf{all} \text{ (measurable) test functions } \varphi\colon \mathcal{X} \to \mathbb{R}. \tag{27}$$

Similarly, auto-calibration is equivalent to

$$\mathbb{E}[\varphi(m(\boldsymbol{X}))V(m(\boldsymbol{X}), Y)] = 0 \quad \text{for } \textbf{all} \text{ (measurable) test functions } \varphi\colon \mathbb{R} \to \mathbb{R}. \tag{28}$$

Here, $\varphi$ is a univariate function (only).

# Practical Considerations

- Check for calibration on the training as well as on the test set.
- Quantify calibration by making a choice for the test function $\varphi$ and report $\overline{V}_\varphi(m; D) = \frac{1}{n}\sum_{i=1}^{n} \varphi(\boldsymbol{x}_i) V(m(\boldsymbol{x}_i), y_i)$. Consider at least
  - $\varphi(\boldsymbol{x}) = 1$
  - all projections to single components of the feature vector $\boldsymbol{x}$
  - $\phi(\boldsymbol{x}) = m(\boldsymbol{x})$ to assesses auto-calibration.

  Continuous features can be binned, e.g., $\varphi(\boldsymbol{x}) = \mathbb{1}\{\text{lower} \leq x^1 < \text{upper}\}$.
- Assess calibration visually:
  - Plot the generalised residuals $V(m(\boldsymbol{x}_i), y_i)$ versus $\varphi(\boldsymbol{x}_i)$ for the above choices of test functions $\varphi$. Average of the generalised residuals should be around 0 for all values of the test function.
  - Another possibility is to plot the values of $\overline{V}_\varphi(m; D)$ for different $\varphi$, for instance projections to single feature columns.
  - A reliability diagram assesses auto-calibration: a graph of the mapping $m(\boldsymbol{x}) \to T(Y|m(\boldsymbol{x}))$, see [5]. $T(Y|m(\boldsymbol{x}))$ can be estimated by isotonic regression of $y_i$ against $m(x_i)$. For an auto-calibrated model, the graph is the diagonal line.

# Unconditional Calibration in Numbers

Would we have made profit or loss (on test set) on the portfolio?

Note: Ideally neither loss nor profit, i.e. *balanced*.

$n_{\text{test}} = 20504$

| | $\frac{1}{n}\sum_i m(\boldsymbol{x}_i) - y_i$ | $p$-value of $t$-test |
|---|---|---|
| Trivial | $\mathbf{-24}$ | $9.5 \times 10^{-1}$ |
| GLM Gamma | $-1207$ | $8.8 \times 10^{-4}$ |
| GLM Poisson | $125$ | $7.3 \times 10^{-1}$ |
| XGBoost | $-2044$ | $1.4 \times 10^{-8}$ |

$\Rightarrow$ **unconditional calibration:** $\mathbb{E}[m(\boldsymbol{X}) - Y] \approx 0$

# Unconditional Calibration in Numbers

Would we have made profit or loss (on test set) on the portfolio?
Note: Ideally neither loss nor profit, i.e. *balanced*.
$n_{\text{test}} = 20504$

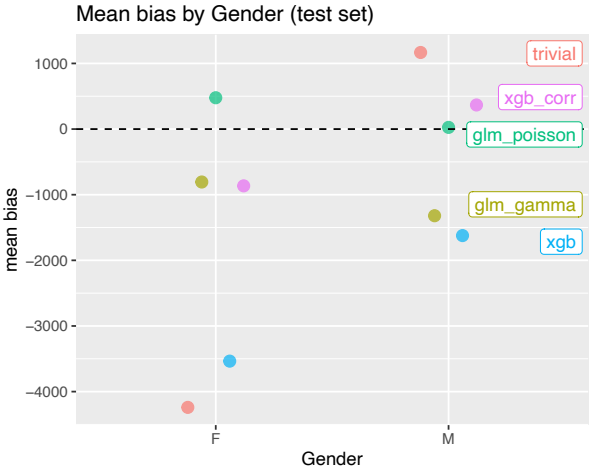
| | $\frac{1}{n}\sum_i m(\boldsymbol{x}_i) - y_i$ | $p$-value of $t$-test |
|---|---|---|
| Trivial | $\mathbf{-24}$ | $9.5 \times 10^{-1}$ |
| GLM Gamma | $-1207$ | $8.8 \times 10^{-4}$ |
| GLM Poisson | $125$ | $7.3 \times 10^{-1}$ |
| XGBoost | $-2044$ | $1.4 \times 10^{-8}$ |
| XGBoost corr | $96$ | $7.9 \times 10^{-1}$ |

Recalibrate XGBoost by a multiplicative constant (on training set).

$\Rightarrow$ **unconditional calibration:** $\mathbb{E}[m(\boldsymbol{X}) - Y] \approx 0$

# Calibration Conditional on Gender

## Is there a gender bias in the models?



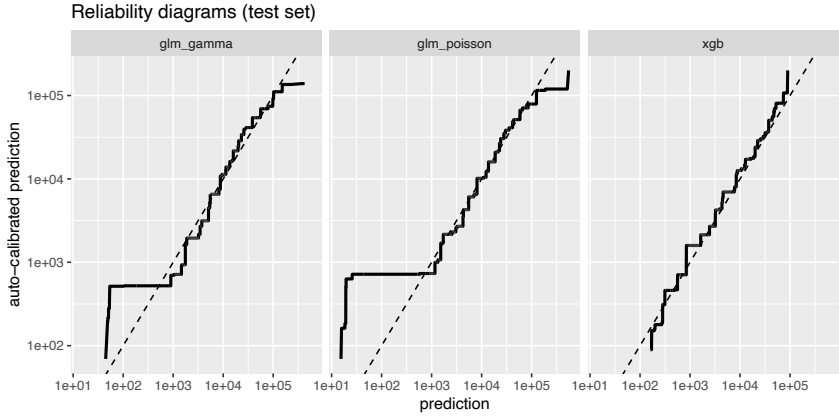

| model | $\frac{1}{n}\sum_{i \in subset} m(\boldsymbol{x})_i - y_i$ bias F | bias M |
|---|---|---|
| Trivial | −4240 | 1167 |
| GLM Gamma | −807 | −1320 |
| GLM Poisson | **477** | **26** |
| XGBoost | −3536 | −1623 |
| XGBoost corr | −865 | 367 |

$\Rightarrow$ **conditional calibration:**
$\mathbb{E}[m(\boldsymbol{X}) - Y|\boldsymbol{X}] \approx 0$

# Auto-Calibration

## Are policies with same (actuarial) price self-financing?

**Reliability diagram**: Estimate $\mathbb{E}[Y|m(\boldsymbol{X})]$ via isotonic regression (PAV) and plot vs $m(\boldsymbol{X})$.




$\Rightarrow$ **auto-calibration:** $\mathbb{E}[m(\boldsymbol{X}) - Y|m(\boldsymbol{X})] \approx 0$

# Exercises

## Exercise 14

Find a strict identification $V(z_1, z_2, y) : \mathbb{R}^3 \to \mathbb{R}^2$ for the pair (mean, variance).

## Exercise 15

Plot a reliability diagram for your model(s). Hint: The R package monotone might help.

## Exercise 16

What are the differences between assessing calibration and the analysis of residuals with ordinary least squares.

# Table of Contents