Prediction intervals and calibration

Quantiles, coverage, conformal prediction and calibration

A prediction is one number. A decision usually needs to know how far off that number may be. This note covers two ways of saying so, and how to check them on data the model has not seen: an interval for a numerical outcome, and a probability for a binary one. Throughout, the numerical example is a simulation, so that the right answers are known: x is uniform on $[0, 10]$, and

$$Y = 3 + x + 2\sin x + (0.5 + 0.35\,x)\,\varepsilon, \qquad \varepsilon \sim N(0, 1),$$

so the noise grows with x. There are 1,000 points to fit on and 2,000 fresh points to check on.

Two kinds of interval

At a new input x_0 the outcome is $Y_0 = f^\star(x_0) + \varepsilon_0$. A **confidence interval** is for $f^\star(x_0)$, the average outcome at x_0: it describes how far the fitted model may be from the regression function, and it narrows as the training sample grows. A **prediction interval** is for Y_0 itself, so it must also cover ε_0. The bias–variance decomposition at a point says why it cannot narrow to nothing:

$$\mathbb{E}\big[(Y_0 - \hat{f}(x_0))^2\big] = \sigma^2 + \text{bias}^2 + \text{variance},$$

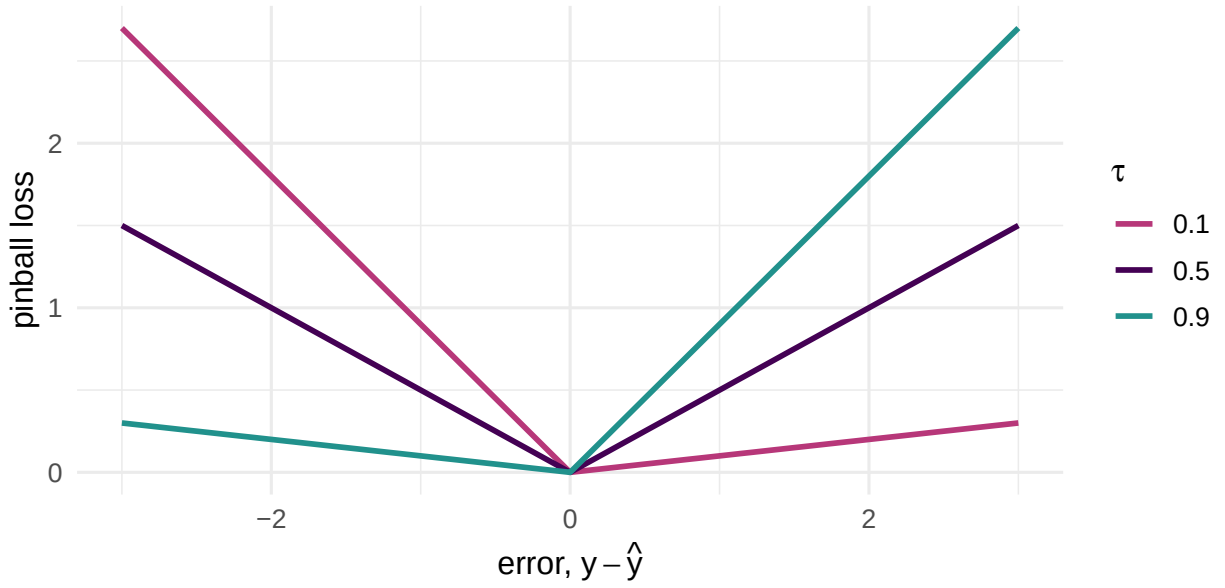
and more data shrink the variance, never σ^2. For a linear model with normal noise of constant variance, R gives both: `predict(fit, newdata, interval = "confidence")` and `interval = "prediction"`. This note is about the second kind, without assuming normal noise of constant variance.

Quantiles and the pinball loss

Squared error makes the conditional mean the best prediction, and absolute error the conditional median. The **pinball loss** makes it a conditional quantile. For an outcome y and a prediction $\hat{y}$,

$$\rho_\tau(y - \hat{y}) = (y - \hat{y})\,(\tau - \mathbf{1}\{y < \hat{y}\}),$$

which charges τ for each unit of underprediction and $1 - \tau$ for each unit of overprediction; with $\tau = \frac{1}{2}$ it is half the absolute error. As a function of the error $y - \hat{y}$ it is two lines that meet at zero: slope τ to the right, where the prediction is too low, and slope $-(1 - \tau)$ to the left, where it is too high. With $\tau = 0.9$ a shortfall costs nine times an overshoot of the same size, so the best prediction is high.



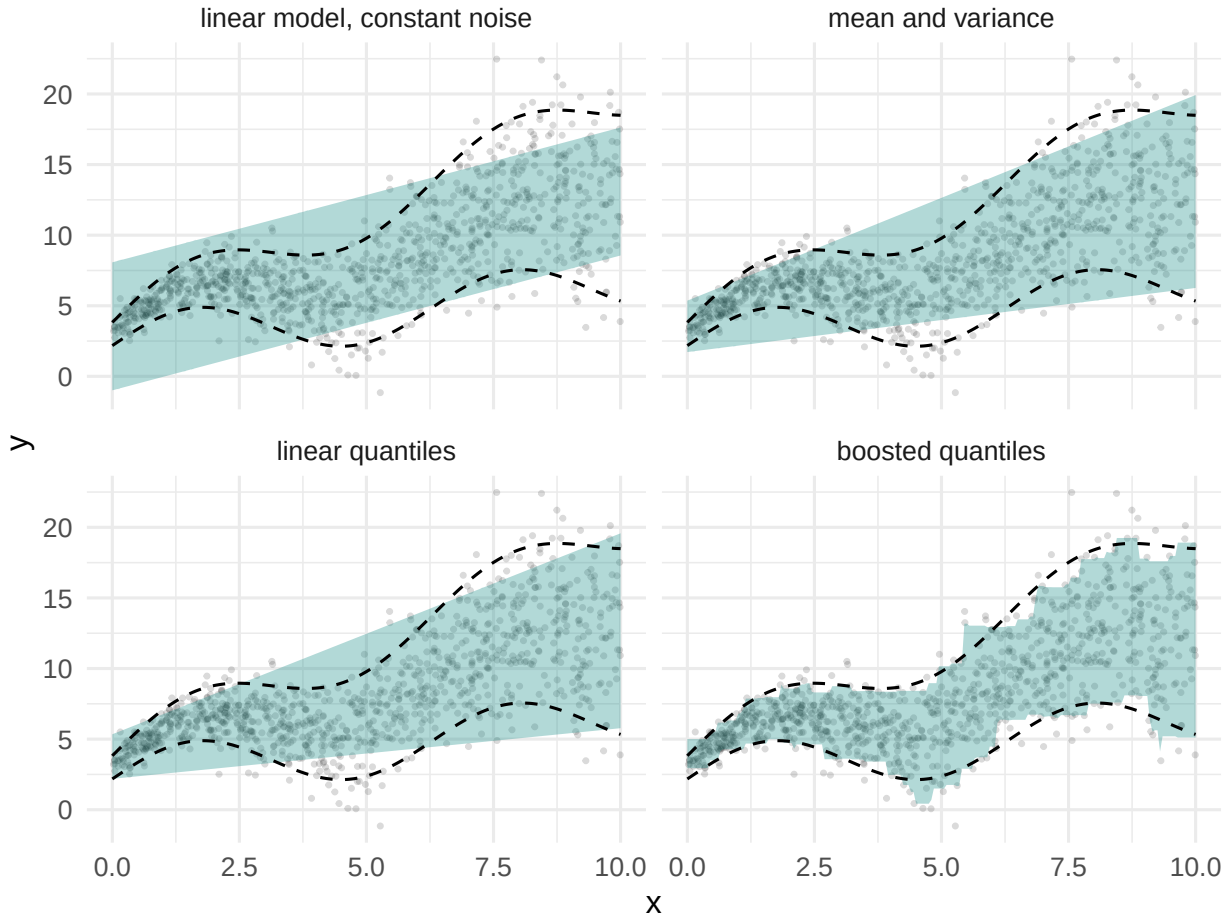
If Y given $X = x$ has a continuous distribution with a finite mean, the conditional τ-quantile $q_\tau(x)$ minimizes the expected loss $\mathbb{E}[\rho_\tau(Y - \hat{y}) \mid X = x]$ over $\hat{y}$, uniquely unless the distribution function is flat at τ (the quantile optimality theorem, stated in the lecture).

Fitting a model with the pinball loss is **quantile regression**: minimize $\frac{1}{n}\sum_i \rho_\tau(y_i - f(x_i))$ over a class of functions f, a line or a sum of trees, just as least squares minimizes the average squared error.

An interval from two quantiles

A 90% interval from two quantile fits is $\hat{C}(x) = [\hat{q}_{0.05}(x), \hat{q}_{0.95}(x)]$. The true quantiles would cover Y with probability 0.9 at every x; estimated quantiles aim at that, and miss by however much the fits are wrong. The figure compares four ways to make a 90% interval on the simulated data; the dashed lines are the true 5% and 95% quantiles.

- **Linear model, constant noise**: least squares, with R's prediction interval, which assumes normal noise of constant variance. It widens slightly toward the ends of the range, with the uncertainty in the fitted line.
- **Mean and variance**: least squares for the mean, then a second regression, of the absolute residuals on x, for the spread; the interval is the mean $\pm 1.645\times$ the spread (the absolute value of a normal error has mean $\sqrt{2/\pi}$ times its standard deviation, so the second fit is scaled by $\sqrt{\pi/2}$).
- **Linear quantiles**: two quantile regressions on a line, with `quantreg::rq()`.
- **Boosted quantiles**: two boosted-tree fits with the pinball loss, xgboost's `objective = "reg:quantileerror"`.



Checking an interval

Check a claimed 90% on data that played no part in the fit: the share of fresh outcomes inside their intervals (the **coverage**), the mean width, and both again within ranges of x. The linear model's interval is close to 90% overall and far from it within ranges, too wide where the noise is small and too narrow where it is large; covering is not enough on its own either, since widening an interval never lowers its coverage.

method	all	x in [0, 2]	(2, 4]	(4, 6]	(6, 8]	(8, 10]	mean width
linear model, constant noise	88%	100%	98%	86%	82%	78%	9.1
mean and variance	89%	93%	93%	83%	87%	90%	8.7
linear quantiles	88%	92%	93%	82%	85%	90%	8.6
boosted quantiles	87%	84%	87%	94%	79%	89%	7.3

The true quantiles have mean width 7.5 on these points. Coverage by range is a diagnostic, not a certificate: a range can be right on average while a smaller group within it is not. Could a method promise 90% at every x? **Marginal coverage** averages over X, $P(Y \in \hat{C}(X)) \geq 0.9$; **conditional coverage** asks for $P(Y \in \hat{C}(X) \mid X = x) \geq 0.9$ at every x. Barber, Candès, Ramdas and Tibshirani

(2021) show that if a procedure guarantees the second in finite samples for every distribution, then wherever the inputs are continuous its intervals are infinitely long on average (their Proposition 2.2).[1] Guarantees for groups, such as each range of x, and results that rest on assumptions about the distribution, or hold as n grows, remain possible. No method does best on every distribution, and none can promise coverage at every x without saying something about the distribution.

Conformalized quantile regression

An interval from two quantile fits covers as well as the fits are right, and the fits do not say how right they are. Checked on the points that fitted them, they look better than they are, just as training error underestimates the risk. **Conformalized quantile regression** (CQR; Romano, Patterson and Candès 2019, section 4) measures how far the fits miss on rows they did not see, and moves both ends of the interval by that amount. For a $1-\alpha$ interval, here $\alpha = 0.1$:

1. Split the training rows in two. On the first part, fit a lower and an upper quantile, $\hat{q}_{\text{lo}}$ and $\hat{q}_{\text{hi}}$ (here the 5% and 95% quantiles), by any method. Keep the other m rows, the **calibration set**, for the next step.
2. On each calibration row, compute the **score**

$$E_i = \max\left\{\hat{q}_{\text{lo}}(x_i) - y_i,\ y_i - \hat{q}_{\text{hi}}(x_i)\right\}$$

   (their equation 6). If y_i is above the interval, E_i is how far above the upper end; if below, how far below the lower end; if inside, $E_i \le 0$, minus the distance to the nearer end.
3. Let Q be the k-th smallest of the m scores, with $k = \lceil (1-\alpha)(m+1) \rceil$; in their equation 8, the $(1-\alpha)(1+1/m)$-th empirical quantile. (If $k > m$, take $Q = \infty$.)
4. Report

$$\hat{C}(x) = \left[\hat{q}_{\text{lo}}(x) - Q,\ \hat{q}_{\text{hi}}(x) + Q\right]$$

   (their equation 7). If the fits were too narrow, $Q > 0$ and the interval widens; if too wide, $Q < 0$ and it narrows.

Proof sketch. A new row (X, Y) is inside $\hat{C}(X)$ exactly when its own score, $E = \max\{\hat{q}_{\text{lo}}(X) - Y,\ Y - \hat{q}_{\text{hi}}(X)\}$, is at most Q. Suppose all the rows, including the new one, are independent draws from one distribution.[2] The fits used neither the calibration rows nor the new row, so given the rows they were fitted on, the $m+1$ scores of these rows are independent draws from one distribution too. With no ties, the new score's rank among them is then equally likely to be any of $1, \ldots, m+1$, and the new score is at most the k-th smallest of the other m exactly when its rank is at most k. So

$$P(Y \in \hat{C}(X)) = \frac{k}{m+1}, \qquad 1-\alpha \ \le\ \frac{k}{m+1}\ <\ 1-\alpha+\frac{1}{m+1},$$

and with ties the probability is at least $k/(m+1)$. This is their Theorem 1: coverage at least $1-\alpha$, and at most $1-\alpha+1/(m+1)$ when the scores are almost surely distinct, as they are when continuous. Nothing is assumed about the quantile fits. A poor fit gives a wide or badly shaped

[1] Their statement is about every x that is not an atom of the inputs' distribution, a value with positive probability; continuous inputs have none.

[2] Romano, Patterson and Candès, and Lei et al., state their results for exchangeable rows, a weaker condition: the joint distribution of the rows is unchanged when they are reordered. Independent draws from one distribution are exchangeable; draws without replacement from a finite population are exchangeable but not independent. The argument here uses only exchangeability. The Beta distribution in the next paragraph uses independence.

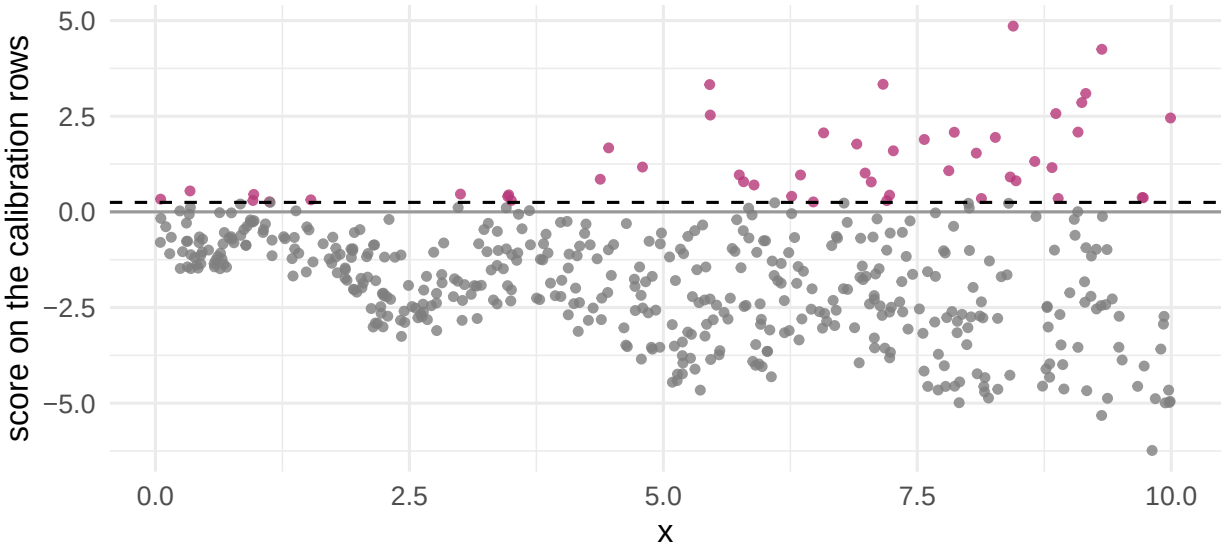
interval, not a wrong overall level. The argument needs fits that did not use the calibration rows in any way, and that includes tuning: if the rows that set Q also chose the fits' settings, the setting chosen is one whose scores there happen to be small, and coverage falls below $1-\alpha$. In the lecture's simulation, choosing among 50 settings on 500 calibration rows cost about a percentage point at most; with fewer rows or more candidates it costs more. Use separate rows to fit, tune and calibrate, or tune by cross-validation within the fitting rows.

Scope of the guarantee. First, the probability is over the calibration rows and the new row together. For one fixed calibration set, the coverage of future rows is the chance that a new score is at most that set's Q. If the rows are independent draws from one distribution and the scores have a continuous distribution, this chance is the k-th smallest of m independent uniform draws, which has the Beta$(k, m+1-k)$ distribution: mean $k/(m+1)$, standard deviation about $\sqrt{\alpha(1-\alpha)/m}$, 1.3 percentage points when $m = 500$. Second, the guarantee is marginal. The same Q is added at every x, so CQR corrects the overall level and not the shape: where the fits are too narrow in one range and too wide in another, the conformalized interval is too.

Conditional coverage as n grows. If the quantile fits are consistent, approaching the true conditional quantiles as the fitting and calibration sets grow, then the CQR interval approaches $[q_{\alpha/2}(x), q_{1-\alpha/2}(x)]$ and its coverage given $X = x$ tends to $1-\alpha$, except on a set of x whose probability tends to zero (Sesia and Candès 2020, Theorem 1 and its corollary, for independent rows and scores with a density bounded away from zero near zero). This does not contradict the impossibility result: consistency is an assumption about the distribution and the method. How fast the fits converge depends on how smooth the quantiles are and on the number of inputs. The same limits apply to checking coverage near x, which needs rows near x: 2,000 fresh rows in 5 bins for each input give 400 rows a bin with one input, 16 with three, under one with five, and with 16 rows the standard error of a bin's coverage near 90% is 7.5 percentage points.

Split conformal prediction is a special case. If both quantile fits are the same fit of the mean, $\hat{q}_{\text{lo}} = \hat{q}_{\text{hi}} = \hat{f}$, the score is $\max\{\hat{f}(x_i) - y_i,\ y_i - \hat{f}(x_i)\} = |y_i - \hat{f}(x_i)|$ and the interval is $\hat{f}(x) \pm Q$, the same width at every x (Lei et al. 2018, section 2.2).

On the simulated data: boosted 5% and 95% fits, with the settings above, on the first 500 training points, and scores on the other 500, so $m = 500$ and $k = 451$. The figure shows the scores; the dashed line is Q, and the points above it are the outcomes the interval still misses.



Here $Q = 0.25$. The two fits cover 92% of their own 500 training points but only 87% of the calibration rows, so CQR widens each end by Q. The scores spread out above zero for x beyond 6, where the fits are too narrow, and one Q cannot fix that. On the fresh points, against split conformal prediction around a least squares line fitted and calibrated on the same halves:

method	all	x in [0, 2]	(2, 4]	(4, 6]	(6, 8]	(8, 10]	mean width
boosted quantiles, 500 points	87%	85%	92%	93%	80%	86%	7.2
conformalized, 500 + 500	90%	92%	94%	94%	82%	87%	7.7
split conformal, least squares, 500 + 500	87%	100%	97%	83%	81%	75%	8.7

Overall, CQR covers 90% of the fresh points and split conformal 87%. Theorem 1 is about the average over calibration sets: for one set of 500, the coverage varies with a standard deviation of about 1.3 percentage points (above), and measuring it on 2,000 fresh points adds a standard error of about 0.7. Split conformal has one width everywhere, so it covers too much where the noise is small and too little where it is large; CQR keeps the shape of its quantile fits, including their error beyond $x = 6$.

Ways to make an interval

Method	How	Assumes	Can claim
Linear model, normal noise	$\hat{f}(x)$ plus or minus a multiple of the estimated noise sd	linear mean, constant variance, normal noise	90% at every x, if the model is right
Mean and variance	a model for the mean, another for the spread	both models right, normal noise	about 90% at every x, if right
Quantile regression	two fits with the pinball loss	the quantile model right	90% at every x as n grows, if right

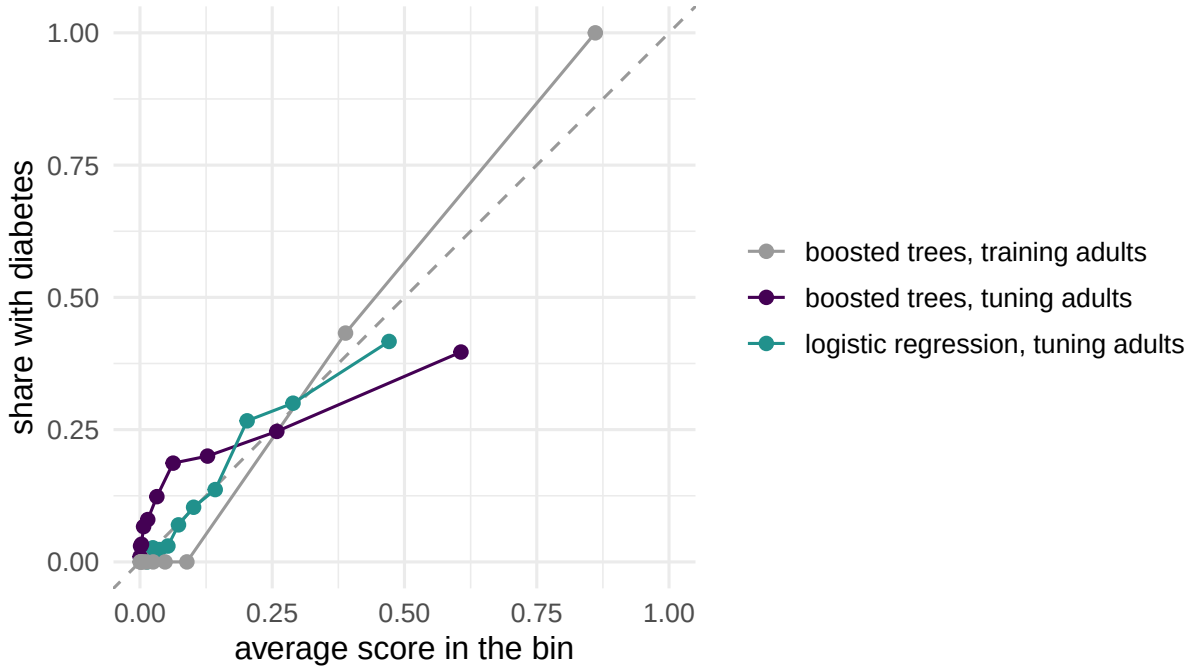
Method	How	Assumes	Can claim
Split conformal prediction	fit $\hat{f}$ on part of the data; on the rest, find the residual size d that 90% of residuals stay within (with a small correction); report $\hat{f}(x) \pm d$	independent rows from one distribution	at least 90% marginally, in finite samples, with any model; the same width at every x
Conformalized quantile regression	fit two quantiles on part of the data; on the rest, find how far to widen (or narrow) them so that 90% of outcomes fall inside	independent rows from one distribution	at least 90% marginally, in finite samples; widths that follow the quantile fits; 90% at every x as n grows, if the fits are consistent

The first three rows rest on a model being right; the two conformal methods (Lei et al. 2018, section 2.2; Romano, Patterson and Candès 2019, Theorem 1) need only independent rows from one distribution, and in return their guarantee is marginal.

Calibration

The same idea for a binary outcome: a stated probability should match a frequency. Of 500 adults given probability 0.2 of having diabetes by a model, about 100 should have it. Probabilities are acted on as probabilities: a risk told to a patient, a premium set from an expected cost, a threshold set from the costs of errors. If adults given 0.2 in fact have the condition 40% of the time, a threshold of 0.25 leaves them unreferred although referring them costs less on average. A score $\hat{p}$ is **calibrated** if $P\big(Y = 1 \mid \hat{p}(X)\big) = \hat{p}(X)$: it conditions on the score's value, not on x. A **reliability diagram** checks this on fresh data: sort the units by their score into bins of equal size, and plot each bin's share of 1s against its average score.

Here the outcome is diabetes, as reported by adults in the NHANES survey (data note; 14% of the training adults). Two scores use nine measurements (age, sex, BMI, pulse, systolic blood pressure, total and HDL cholesterol, the income-to-poverty ratio and physical activity); they are fitted on 4,000 adults and checked on 3,000 others: logistic regression, and boosted trees with xgboost's defaults (depth 6, learning rate 0.3, 100 iterations). The boosted score is too confident on fresh adults. Checked on the adults who fitted it (grey), it looks too cautious at the top instead, which is why the check needs fresh data.



- **Calibration and sharpness.** The base rate, $\hat{p}(x) = 0.14$ for every adult, is calibrated, and it does not separate those with diabetes from the rest at all. A useful score is calibrated and also sharp, with values near 0 and 1 where it can; Gneiting, Balabdaoui and Raftery (2007) make this principle precise for forecasts of a continuous outcome: as sharp as possible, subject to calibration.
- **Calibration and the threshold.** If $\hat{p}$ is calibrated, the optimal threshold theorem applies with $\hat{p}(X)$ as the input: when a false positive costs c_{FP} and a false negative c_{FN}, predicting 1 when $\hat{p}(x) > c_{FP}/(c_{FP} + c_{FN})$ is optimal among rules based on that score. It need not be optimal among rules that use all of x. Without calibration, that threshold need not be the optimal cut on $\hat{p}$: a miscalibrated score can still give the right decisions (with p the true class probability, $\hat{p} = 0.25 + 0.5p$ does at the threshold 0.5), but in general it does not.
- **Recalibration.** A score can be repaired on rows not used to fit it. One method: a logistic regression of the label on $\operatorname{logit} \hat{p} = \log \frac{\hat{p}}{1-\hat{p}}$, giving $\tilde{p} = 1/(1 + e^{-(a+b \operatorname{logit} \hat{p})})$. For $b > 0$ the map from $\hat{p}$ to $\tilde{p}$ is increasing, so the new score ranks the adults as the old one did (its AUC is unchanged), with new probabilities. Fitted on the first 1,500 tuning adults and checked on the other 1,500:

	log loss on the other 1,500 adults
boosted trees	0.375
recalibrated	0.329

The fitted slope is $b = 0.39$: two adults whose boosted log-odds differ by 1 differ by 0.39 after recalibration, in the same order. Van Calster et al. (2019) describe the same checks and repairs for risk models in medicine, where a miscalibrated risk can change a treatment decision.

Calibration within groups. A score can be right on average and still too high for one group and too low for another, with the errors cancelling. Used with one threshold for everyone, such a

score refers the first group for tests more often, and the second less often, than their risk warrants: it discriminates, even when group membership is not one of its inputs. Calibration within groups is one fairness criterion among several, and they conflict. Kleinberg, Mullainathan and Raghavan (2017) show that when the groups' base rates differ, a score that is calibrated within groups cannot also be balanced, meaning that its average is the same in each group among those with the outcome, and the same in each group among those without, unless it predicts perfectly. For a rule that thresholds a score, Chouldechova (2017) shows that with different base rates, equal positive predictive values across groups rule out equal false positive and false negative rates (for a rule whose positive predictive value is strictly between 0 and 1). **Multicalibration** (Hébert-Johnson, Kim, Reingold and Rothblum 2018, definition 2.6) asks for calibration within every group in a collection named in advance: within each group, among the units with score near v, the share of 1s should be within a tolerance of v, for every v. Their algorithm (section 3.2) repeatedly finds a group and a range of scores where the average score and the share of 1s differ by more than the tolerance, and shifts the scores there by the difference. The groups must be large enough to check, and a group not in the collection gets no guarantee. On these data the logistic score above, which leaves race and ethnicity out, is right on average on the tuning adults (an average score of 14.1% against 13.7% with diabetes), but for White adults it averages 14.8% against 10.6% with diabetes, a gap of about 5 standard errors; for the survey's other groups the gaps are within about two standard errors. Matching averages, overall or in a group, would not show calibration: errors at different scores can cancel, which is why multicalibration asks for calibration at every range of scores within each group. In the lecture, shifts learned on the tuning adults shrink the gaps for White and Black adults on the test adults and widen the gap for Mexican American adults: with a few hundred adults in a cell, the shifts partly fit noise. Applying them also needs each person's group when the score is used.

Further reading

- Barber, R. F., Candès, E. J., Ramdas, A. and Tibshirani, R. J. (2021). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA* 10(2), 455–482. doi:10.1093/imaiai/iaaa017. Section 1 defines marginal and conditional coverage; Proposition 2.2 is the impossibility.
- Chouldechova, A. (2017). Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data* 5(2), 153–163. doi:10.1089/big.2016.0047. Equal positive predictive values and equal error rates cannot hold together when base rates differ.
- Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B* 69(2), 243–268. doi:10.1111/j.1467-9868.2007.00587.x. The principle of a sharp forecast subject to calibration, for forecasts that are whole distributions.
- Hébert-Johnson, Ú., Kim, M. P., Reingold, O. and Rothblum, G. N. (2018). Multicalibration: calibration for the (computationally-identifiable) masses. *Proceedings of the 35th International Conference on Machine Learning*, PMLR 80, 1939–1948. proceedings.mlr.press/v80/hebert-johnson18a.html. Definition 2.6 and the algorithm of section 3.2 (numbered as in the arXiv version, 1711.08513).
- James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R*, 2nd edition. Springer. Section 3.2.2, its last question, on predictions: confidence against prediction intervals for a linear model. statlearning.com
- Kleinberg, J., Mullainathan, S. and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. *8th Innovations in Theoretical Computer Science Conference*

(ITCS 2017), LIPIcs 67, 43:1–43:23. doi:10.4230/LIPIcs.ITCS.2017.43. Calibration within groups and balance conditions on the scores, compatible only with perfect prediction or equal base rates.

- Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J. and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association* 113(523), 1094–1111. doi:10.1080/01621459.2017.1307116. Section 2.2, split conformal prediction.
- National Center for Health Statistics, National Health and Nutrition Examination Survey, 2009–2012, as shipped in the R package NHANES: Pruim, R., *NHANES: Data from the US National Health and Nutrition Examination Study*, version 2.1.4, CRAN. The calibration examples; data note.
- Romano, Y., Patterson, E. and Candès, E. J. (2019). Conformalized quantile regression. *Advances in Neural Information Processing Systems* 32, 3543–3553. NeurIPS proceedings. Section 4, equations 6 to 8 and Theorem 1: the method above.
- Sesia, M. and Candès, E. J. (2020). A comparison of some conformal quantile regression methods. *Stat* 9(1), e261. doi:10.1002/sta4.261. Theorem 1 and its corollary (numbered as in arXiv:1909.05433): asymptotic conditional coverage of CQR when the quantile fits are consistent.
- 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, article 230. doi:10.1186/s12916-019-1466-7. Levels of calibration, how to check them, and an example of two heart-disease risk models in which the one with the slightly higher AUC overestimated risk and, at a 20% threshold, flagged about twice as many men.