Model assessment

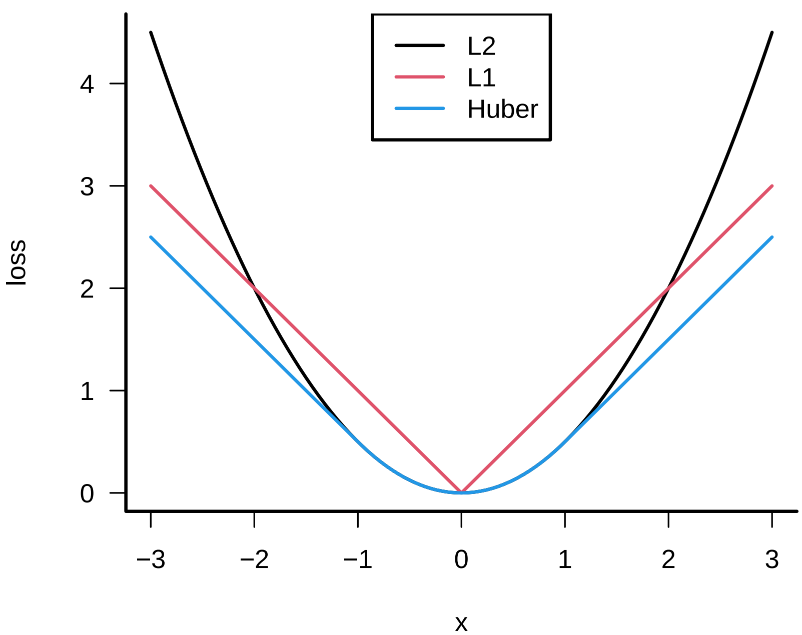
22 Mar 2023

Table of contents

loss functions (regression/quantitative outcome)

- continuous: L2, L1, **Huber loss**:

```
par(las = 1, bty = "l", lwd = 2)
huber <- function(x, d) ifelse(abs(x)<d, x^2/2, d*abs(x)-d/2)
curve(x^2/2, from = -3, to = 3, ylab = "loss")
curve(abs(x), add = TRUE, col = 2)
curve(huber(x, 1), add = TRUE, col = 4)
legend("top", c("L2", "L1", "Huber"), col = c(1, 2, 4), lty = 1)
```



loss functions (classification)

- 0-1
- **deviance**: $-2 \sum I(G = k) \log \hat{p}_k$ = -2 × log-likelihood
- deviance generalizes to other distributions

a short rant about categorical loss functions

- 0-1 scoring dichotomizes prematurely
- leads to lots of confusing discussion about balancing data sets
- lots of discussion of what to do about imbalanced data sets (SMOTE etc.) (Chawla et al. 2002; van den Goorbergh et al. 2022)
- when **should** we balance?
  - when we have to use 0-1 scoring for some technical reason
  - when we have too **much** data (downsampling, i.e., throw away majority class)
- (cf. discussion of variable selection)

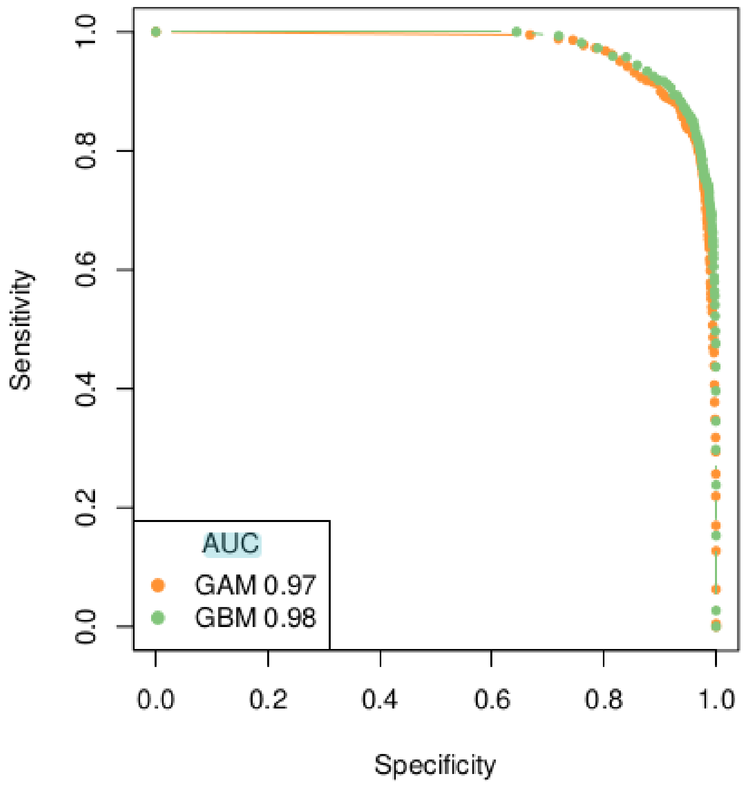
Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "SMOTE: Synthetic Minority Over-Sampling Technique." *Journal of Artificial Intelligence Research* 16 (June): 321–57. https://doi.org/10.1613/jair.953.

van den Goorbergh, Ruben, Maarten van Smeden, Dirk Timmerman, and Ben Van Calster. 2022. "The Harm of Class Imbalance Corrections for Risk Prediction Models: Illustration and Simulation Using Logistic Regression." *Journal of the American Medical Informatics Association*, June, ocac093. https://doi.org/10.1093/jamia/ocac093.

from loss functions to model quality metrics

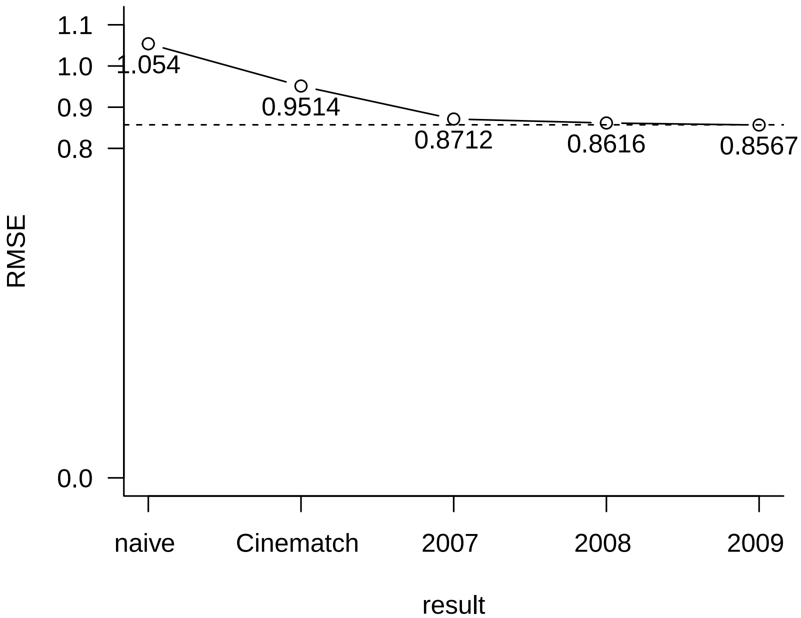
- categorical predictors:
- accuracy (total fraction correct); same problems as 0-1 classification
- AUC (area under the curve)
  - may be problematic in terms of implied misclassification costs? (Hand 2009)

Hand, David J. 2009. "Measuring Classifier Performance: A Coherent Alternative to the Area Under the ROC Curve." *Machine Learning* 77 (1): 103–23. https://doi.org/10.1007/s10994-009-5119-5.



quality metrics

- some combination of loss functions per point
- scaled for **interpretability**
  - how good is good enough?
  - how much difference in model predictions matters?
  - e.g. Netflix prize
- R^2
- MSE $\rightarrow$ RMSE $\rightarrow$ scaled RMSE (or mean-squared log error?)
- always a **business** or **scientific** decision (*value of information*)



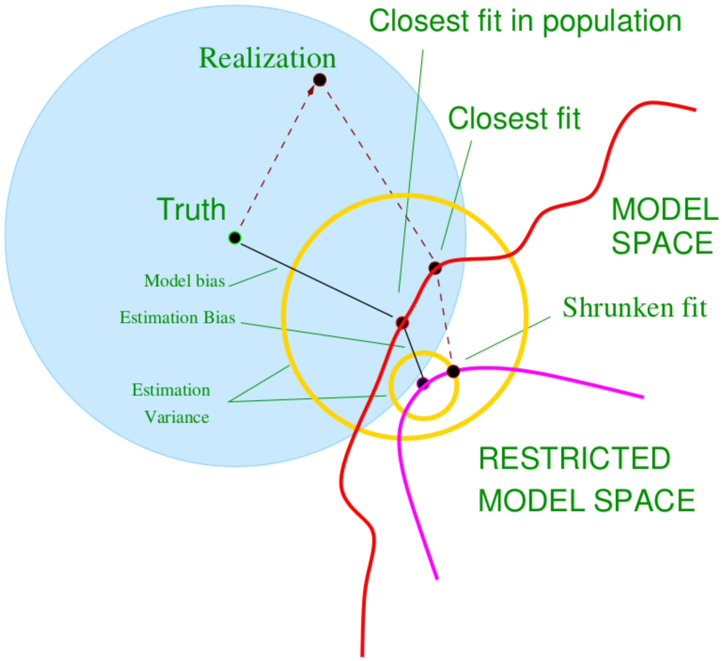
what error are we trying to estimate?

- training error (within-sample): average error within sample
- test error (generalization error), Err_T: expected prediction error for a **fixed** training sample
- **expected** prediction error: test error *averaged over training sets* = $E[\text{Err}_T]$

bias/variance etc.

$$E[f(x_0) - x_0^\top \beta^*]^2 + E[x_0^\top \beta^* - E x_0^\top \hat{\beta}_\alpha]^2$$

- estimation bias = 0 for linear regression etc., positive for ridge etc. **given correct model**



train/validate/test

- training to estimate parameters
- validation: select models/tune hyperparameters
- test: evaluate; **must be independent**, don't snoop!
- select models based on **estimated test error**: only need to get relative values right

within- (training) and out-of-sample (test) error

- within-sample: R^2
- out-of-sample: adjusted R^2 (scaled by $n-p$), PRESS (predicted out-of-sample error) = LOOCV SSQ, AIC ($-2\log L + 2p$), Mallows' C_p ($\frac{\text{RSS}+2p\hat{\sigma}^2}{n}$). AIC and C_p equivalent for Gaussian models. AIC asymptotically → LOOCV for linear models.
- finite-size corrections for AIC (AICc): more conservative for smaller samples

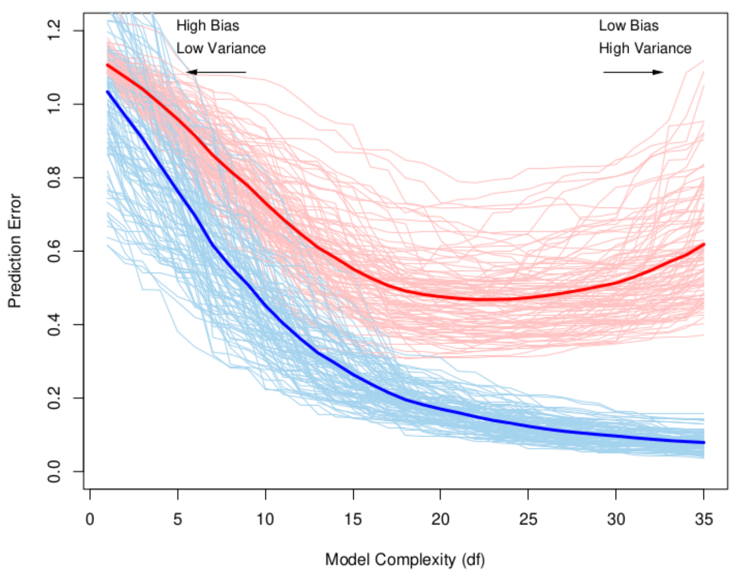
- ESL gives a weird/unusual (scaled-by-N) definition of AIC
- GCV, AUC
- BIC: $-2 \log L + (\log N)p$; higher penalty for $N > e^2$ (almost always)
  - derived from a **Laplace approximation** to the **Bayes factor** (quadratic approx; $\approx$ multivariate normal posterior) given equal priors on models
  - (¿ what is N for non-iid data ?)
- BIC is **consistent**, AIC is **predictive** (Yang 2005); M-closed vs M-complete vs M-open (Clarke, Clarke, and Yu 2014)

Yang, Yuhong. 2005. "Can the Strengths of AIC and BIC Be Shared? A Conflict Between Model Identification and Regression Estimation." *Biometrika* 92 (4): 937–50. https://doi.org/10.1093/biomet/92.4.937.

Clarke, Bertrand, Jennifer Clarke, and Chi Wai Yu. 2014. "Statistical Problem Classes and Their Links to Information Theory." *Econometric Reviews* 33 (1-4): 337–71. https://doi.org/10.1080/07474938.2013.807190.

effective number of parameters

- (generalized, penalized) linear models: trace(Hat)
- additive-error models: $\sum(\text{Cov}(\hat{y}_i, y)/\sigma_\epsilon^2)$



cross-validation

- typical used for **hyperparameter tuning** (e.g. ridge/lasso/spline penalty, elasticnet α)

- LOOCV
  - sometimes easy/closed-form solution
  - expensive otherwise
- *k*-fold

bias vs variance in CV

- more folds = smaller folds = larger training sets
- training error decreases with training set size (i.e. **decreasing** bias in error estimate)
- high variance because training sets are highly correlated (i.e., we're estimating Err_T)

one-standard-error rule

- account for uncertainty in cross-validation error estimate; choose a slightly more **parsimonious** (i.e. higher penalty/lower complexity) model than min-CV
- ¿ not on strong foundations ? "Occam's razor": is there a general trend toward overoptimism?
- CrossValidated q.
- Chen and Yang (2021)

Chen, Yuchen, and Yuhong Yang. 2021. "The One Standard Error Rule for Model Selection: Does It Work?" *Stats* 4 (4): 868–92. https://doi.org/10.3390/stats4040051.

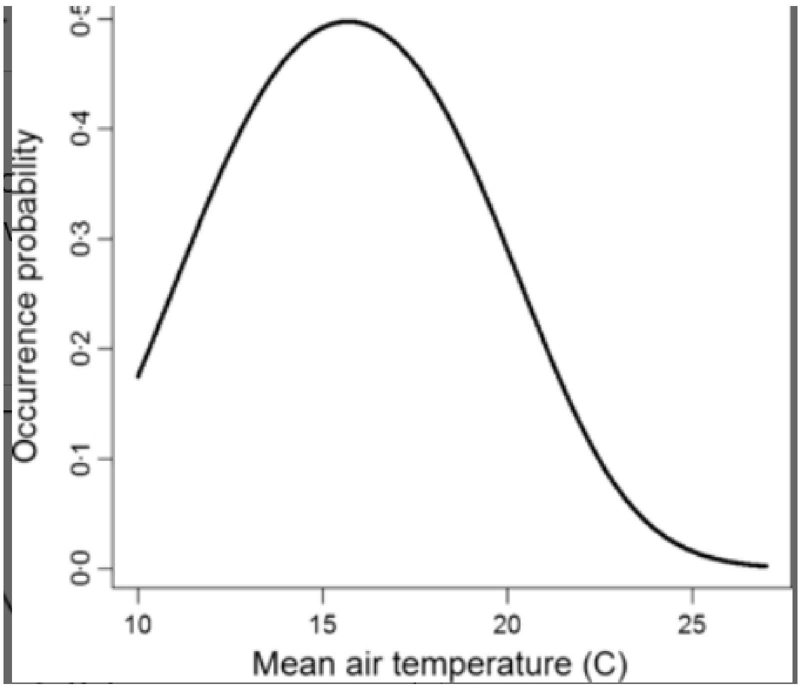
data leakage

- inadmissible predictors (e.g. palliative care in Cygu et al. (2023))
- ESL § 7.10.2
  - example: screen predictors first
  - then use CV to tune the model
  - the **full** 'training' sequence must be done on every CV training set
    - can do **unsupervised** model reduction (i.e., not looking at predictions), e.g. select PCA components or high-variance predictors

Cygu, Steve, Hsien Seow, Jonathan Dushoff, and Benjamin M. Bolker. 2023. "Comparing Machine Learning Approaches to Incorporate Time-Varying Covariates in Predicting Cancer Survival Time." *Scientific Reports* 13 (1): 1370. https://doi.org/10.1038/s41598-023-28393-7.

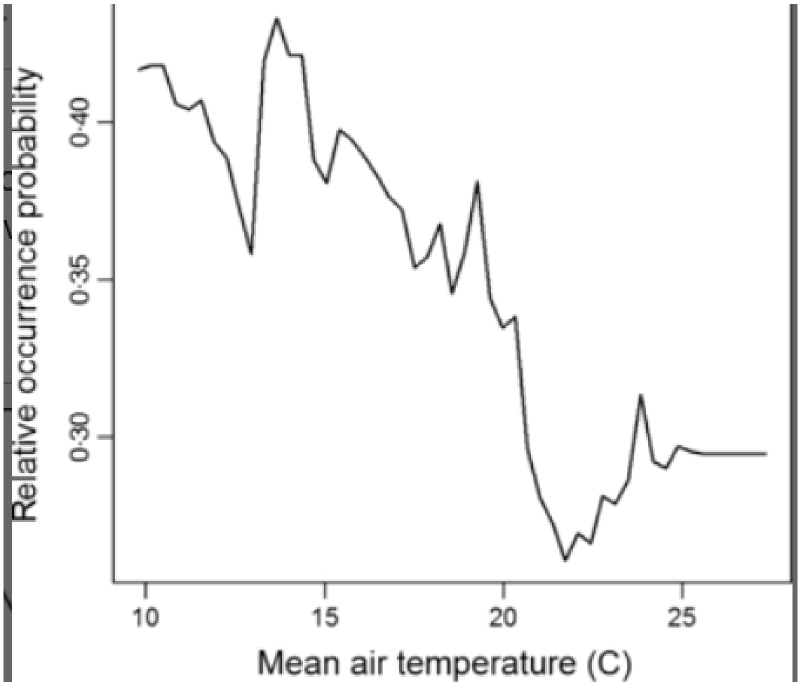
dependent data

- blocking factors: patient, space, time, etc. (Wenger and Olden 2012; Bussola et al. 2020)



Wenger, Seth J., and Julian D. Olden. 2012. "Assessing Transferability of Ecological Models: An Underappreciated Aspect of Statistical Validation." *Methods in Ecology and Evolution* 3 (2): 260–67. https://doi.org/10.1111/j.2041-210X.2011.00170.x.

Bussola, Nicole, Alessia Marcolini, Valerio Maggio, Giuseppe Jurman, and Cesare Furlanello. 2020. "AI Slipping on Tiles: Data Leakage in Digital Pathology." arXiv. https://doi.org/10.48550/arXiv.1909.06539.



solutions

- consider admissibility of predictors
- stratify CV folds
- organize spatially blocked or buffered test/train splits (Roberts et al. 2017; Valavi et al. 2019; Milà et al. 2022)
- account for blocking/correlation in the model (mixed models, spatial correlation models …?)

Roberts, David R., Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, et al. 2017. "Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure." *Ecography* 40 (8): 913–29. https://doi.org/10.1111/ecog.02881.

Valavi, Roozbeh, Jane Elith, José J. Lahoz-Monfort, and Gurutzeta Guillera-Arroita. 2019. "blockCV: An r Package for Generating Spatially or Environmentally Separated Folds for k-Fold Cross-Validation of Species Distribution Models." *Methods in Ecology and Evolution* 10 (2): 225–32. https://doi.org/10.1111/2041-210X.13107.

Milà, Carles, Jorge Mateu, Edzer Pebesma, and Hanna Meyer. 2022. "Nearest Neighbour Distance Matching Leave-One-Out Cross-Validation for Map Validation." *Methods in Ecology and Evolution* 13 (6): 1304–16. https://doi.org/10.1111/2041-210X.13851.

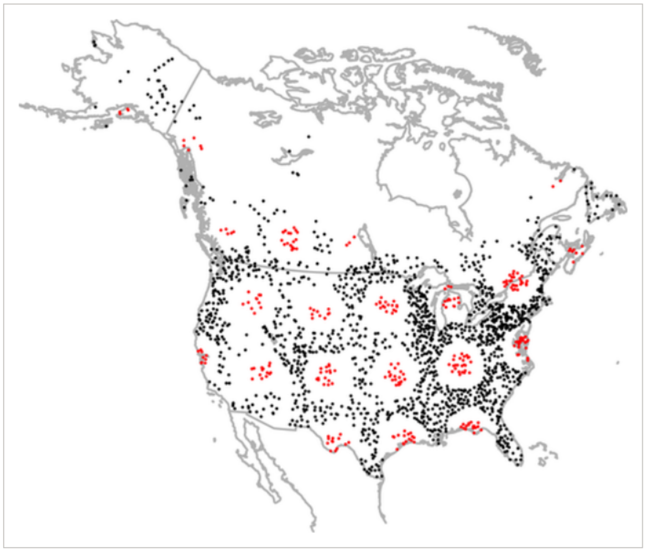


Figure 2 Open in figure viewer | PowerPoint

Map of the BBS routes used in this analysis. Black points are training routes; red ones are test routes. The training and test routes are separated by a 150-km buffer to minimize spatial autocorrelation across the two partitions.

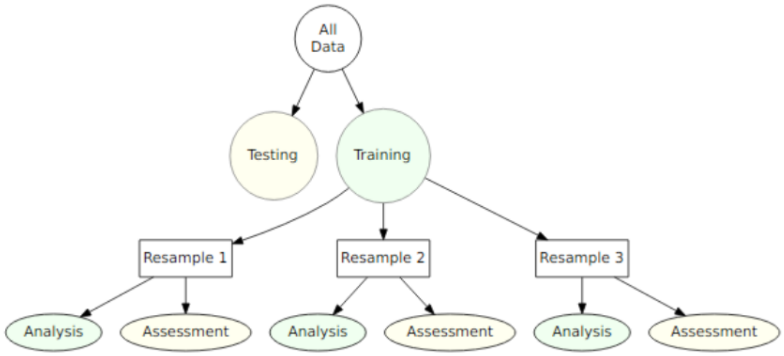
bootstrapping

- we can also use bootstrapping
- $\widehat{\text{Err}}_{\text{boot}} = \frac{1}{B}\frac{1}{N}\sum_b \sum_i L(y_i, \hat{f}^{*b}(x_i))$
  (applies bootstrap estimate to *training data*: overlap of $\phi_b = \approx 1 - e^{-1}$
- **leave-one-out bootstrap error**: $\widehat{\text{Err}}^{(1)}_{\text{boot}}$: only average $L(i)$ over bootstrap replicates not containing i
- bias correction for smallness of bootstrap set: $(1-\phi_b) \cdot$ sample error $+ \phi_b \cdot \widehat{\text{Err}}^{(1)}_{\text{boot}}$

nested cross-validation

Kuhn (2017)

Kuhn, Max. 2017. "Nested Resampling with Rsample." *Applied Predictive Modeling*. http://appliedpredictivemodeling.com/blog/2017/9/2/njdc83d01pzysvvlgik02t5qnaljnd.



> Nested resampling does an additional layer of resampling that separates the tuning activities from the process used to estimate the efficacy of the model. … For example, if 10-fold cross-validation is used on the outside and 5-fold cross-validation on the inside, a total of 500 models will be fit. The parameter tuning will be conducted 10 times and the best parameters are determined from the average of the 5 assessment sets.
>
> Once the tuning results are complete, a model is fit to each of the outer resampling splits using the best parameter associated with that resample. The average of the outer method's assessment sets are a unbiased estimate of the model.

- maybe overkill for practical purposes (Wainer and Cawley 2021) ?

from model assessment to uncertainty estimation

- RMSE is the *average* inaccuracy; use it as a standard error?
- conformal prediction (Shafer and Vovk 2008)
- jackknife (Barber et al. 2021; Efron and Gong 1983)
  - R_i^{LOO} is the leave-one-out residual for point i
  - jackknife pred interval: quantiles of $\hat{\mu}(X_{n+1}) \pm R_i^{\text{LOO}}$

Wainer, Jacques, and Gavin Cawley. 2021. "Nested Cross-Validation When Selecting Classifiers Is Overzealous for Most Practical Applications." *Expert Systems with Applications* 182 (November): 115222. https://doi.org/10.1016/j.eswa.2021.115222.

Shafer, Glenn, and Vladimir Vovk. 2008. "A Tutorial on Conformal Prediction." *Journal of Machine Learning Research* 9: 371–421.

Barber, Rina Foygel, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. 2021. "Predictive Inference with the Jackknife+." *The Annals of Statistics* 49 (1): 486–507. https://doi.org/10.1214/20-AOS1965.

Efron, Bradley, and Gail Gong. 1983. "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation." *The American Statistician* 37 (1): 36–48. https://doi.org/10.1080/00031305.1983.10483087.

- jackknife+: quantiles of $\hat{\mu}_{-i}(X_{n+1}) \pm R_i^{\text{LOO}}$
- **or** K-fold CV+ intervals (Taquet 2021)

Taquet, Vianney. 2021. "With MAPIE, Uncertainties Are Back in Machine Learning !" *Medium*. https://towardsdatascience.com/with-mapie-uncertainties-are-back-in-machine-learning-882d5c17fdc3.

coverage

- a measure of the accuracy of confidence intervals
- do $(1-\alpha)$ CIs include the true value a fraction α of the time?
- $\approx$ accuracy of model assessment
- MAPIE

calibration

- for categorical prediction
- do predicted probabilities match observed probabilities (e.g. fraction of positives)?
- (Guo et al. 2017; Minderer et al. 2021)

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. "On Calibration of Modern Neural Networks." In *Proceedings of the 34th International Conference on Machine Learning*, 1321–30. PMLR. https://proceedings.mlr.press/v70/guo17a.html.

Minderer, Matthias, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. "Revisiting the Calibration of Modern Neural Networks." In *Advances in Neural Information Processing Systems*, 34:15682–94. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2021/hash/8420d359404024567b5aefda1231af24-Abstract.html.

Journal of the American Medical Informatics Association, 2022, Vol. 00, No. 0

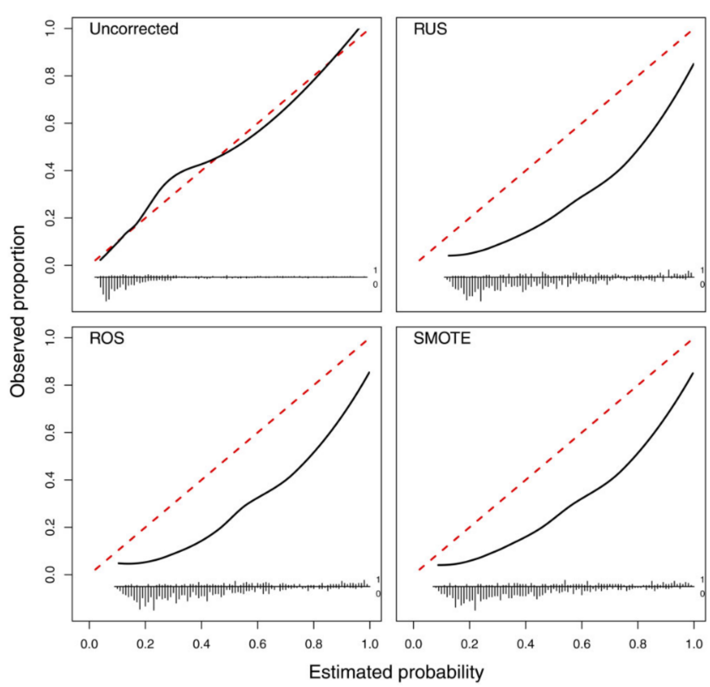


Figure 2. Flexible calibration curves on the test set for the Ridge models to diagnose ovarian cancer.

Model/parameter interpretation

goals

- wanting **pure** prediction is very unusual
- evaluate effects of variables on predictions
- tell a story/interpret results
- prioritize data collections
- counterfactuals (**not** causal inference!)
    - **danger**: strong assumptions of representative sampling etc.
    - effects of correlated predictors
- *conditional* vs *marginal* effects
    - "marginal" as in "marginal probability"
        - any nonlinearity makes $E(f(\beta)) \neq f(E(\beta))$
    - "marginal" as in "marginal effect" (**partial derivative** of predictions wrt predictors) ["average" marginal effects]

by variable: p-values

- de-emphasized/impractical
- usually parameter-specific
- usually model-dependent (although permutation test)
- theory difficult under penalization,
- measure clarity, not effect size
- usually messed up by penalization, hyperparameter tuning, etc..
- multiple-comparisons testing
    - **false discovery rate** (Benjamini-Hochberg)
        - rank p-values
        - critical value $(i/m)Q$
        - all p-values < crit value are significant
- **high-dimensional inference**: e.g.
    - asymptotic assumptions
    - requires sparsity (e.g. $\log(p)/\sqrt{n} \to 0$

- may require a bound on the smallest non-zero parameter

by variable: "relevance"

- "relevance"
- single CART: average improvement (decrease of squared loss) over splits that use variable v
- boosted trees: average over trees
- (splits importance between strongly correlated predictors …)
- do we have to worry about overfitting (training vs testing)?

permutation measures

- random forests: permute jth variable in OOB samples, compare accuracy
- can be generalized, but expensive
- more even: correlated variables can be substituted

partial dependence

- S = focal variable, C = complement (all other variables)
- average dependence: $f_S(X_S) = E_{X_C} f(X_S, X_C)$
- $\to \bar{f}_S(X_S) = \frac{1}{N} \sum f(X_S, x_{iC})$
- compare with **individual conditional** expectations
  - plot predictions for observations i while modifying X_S
  - could plot effects at **model centre** ('average individual')

Shapley values

- Game theoretic
- *average additive* contributions/decreases in loss rate

- … over all possible combinations of previously included variables
- fast algorithm for trees
- challenges … Kumar et al. (2020)

See Burzykowski (2020) for more details …

Kumar, I. Elizabeth, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. 2020. "Problems with Shapley-value-based Explanations as Feature Importance Measures." In *Proceedings of the 37th International Conference on Machine Learning*, 5491–5500. PMLR. http://proceedings.mlr.press/v119/kumar20e/kumar20e.pdf.

Burzykowski, Przemyslaw Biecek and Tomasz. 2020. *Explanatory Model Analysis*.