2020-05-12 to 15: Frank Harrell - Regression Modeling Strategies

NOTE: In addition to any relevant papers on the given topics, I include the section in Frank Harrell's Regression Modeling Strategies (RMS) textbook where these ideas are discussed. They are referenced as (Chapter.Section).

References to BBR, refer to publicly available slides to Biostatistics for Biomedical Research course given by Frank Harrell, http://hbiostat.org/doc/bbr.pdf.

Statistical Insight	Comments	Reference	Category
Careful model specification is MUCH more important than variable selection	Most important take-home lesson of the course		
There is no reason for log-rank test to exist anymore. It's same as cox model without adjusting for covariates		(1.1)	Survival
If you do multiplicity adjustment you must also adjust the coefficient estimates. In general, multiplicity adjustments shouldn't be used.		(1.1)	Modeling Strategies
• Classification vs Prediction (most popular blog): • Lists many issues with dichotomizing continuous outcome. Preferred approach is to use continuous outcome and then develop classification rules on basis of estimated probabilities • Even if classification is used, "proportion classified correctly (PPC)" is a poor accuracy measure.	• Increased power and precision with continuous outcome • PPC is very sensitive to relative frequencies of outcome • Many other problems • Why do we use dichotomous BOR as outcome? At least should use ordinal outcome??	• (1.3) • https://www.fharrell.com/post/classification/ • Valerii Fedorov, Frank Mannino, and Rongmei Zhang. "Consequences of Dichotomization". In: Pharm Stat 8 (2009), pp. 50{61. doi: 10.1002/pst.331. url: http://dx.doi.org/10.1002/pst.331 (cit. on pp. 1-6, 2-13). ◦ Optimal cutpoint depends on unknown parameters ◦ Should only entertain dichotomization when "estimating a value of the cumulative distribution and when the assumed model is very different from the true model" ◦ nice graphics	Modeling Strategies

• Avoid drawing ROC curves (it is misleading). Never saw an example where an ROC curve provided any insight on a paper.
• AUC/C-index can be useful, but only as a descriptive measure, NOT to compare models. There are better descriptive measures as well.
	To make an optimal decision you need to know all relevant data about an individual (used to estimate the probability of an outcome), and the utility (cost, loss function) of making each decision. Sensitivity and specificity do not provide this information. For example, if one estimated that the probability of a disease given age, sex, and symptoms is 0.1 and the "cost" of a false positive equaled the "cost" of a false negative, one would act as if the person does not have the disease. Given other utilities, one would make different decisions. If the utilities are unknown, one gives the best estimate of the probability of the outcome to the decision maker and let her incorporate her own unspoken utilities in making an optimum decision for her.
	Besides the fact that cutoffs do not apply to individuals, only to groups, individual decision making does not utilize sensitivity and specificity. For an individual we can compute Prob(Y = 1|X = x); we don't care about Prob(Y = 1|X > c), and an individual having X = x would be quite puzzled if she were given Prob(X > c|future unknown Y) when she already knows X = x so X is no longer a random variable.
	Even when group decision making is needed, sensitivity and specificity can be bypassed. For mass marketing, for example, one can rank order individuals by the estimated probability of buying the product, to create a lift curve. This is then used to target the k most likely buyers where k is chosen to meet total program cost constraints.
		• (1.3)
		• Andrew J. Vickers. "Decision Analysis for the Evaluation of Diagnostic Tests, Prediction Models, and Molecular Markers". In: Am Statistician 62.4 (2008), pp. 314{320 (cit. on p. 1-7).
		  ◦ limitations of accuracy metrics
		  ◦ incorporating clinical consequences
		  ◦ nice example of calculation of expected outcome
		  ◦ drawbacks of conventional decision analysis, especially because of the difficulty of eliciting the expected harm of a missed diagnosis
		  ◦ use of a threshold on the probability of disease for taking some action
		  ◦ decision curve
		  ◦ has other good references to decision analysis
		• William M. Briggs and Russell Zaretzki. "The Skill Plot: A Graphical Technique for Evaluating Continuous Diagnostic Tests (with Discussion)". In: Biometrics 64 (2008), pp. 250{261 (cit. on p. 1-7).
		  ◦ "statistics such as the AUC are not especially relevant to someone who must make a decision about a particular x c. ... ROC curves lack or obscure several quantities that are necessary for evaluating the operational effectiveness of diagnostic tests. ... ROC curves were first used to check how radio receivers (like radar receivers) operated over a range of frequencies. ... This is not how most ROC curves are used now, particularly in medicine. The receiver of a diagnostic measurement ... wants to make a decision based on some x c, and is not especially interested in how well he would have done had he used some different cutoff."
		  ◦ In the discussion David Hand states "when integrating to yield the overall AUC measure, it is necessary to decide what weight to give each value in the integration. The AUC implicitly does this using a weighting derived empirically from the data. This is nonsensical. The relative importance of misclassifying a case as a noncase, compared to the reverse, cannot come from the data itself. It must come externally, from considerations of the severity one attaches to the different kinds of misclassications."
		  ◦ See Lin, Kvam, Lu Stat in Med 28:798-813; 2009
		• Mitchell H. Gail and Ruth M. Pfeiffer. "On Criteria for Evaluating Models of Absolute Risk". In: Biostatistics 6.2 (2005), pp. 227{239 (cit. on p. 1-7).
		• Robert Bordley. "Statistical Decision making without Math". In: Chance 20.3 (2007), pp. 39{44 (cit. on p. 1-7).
		• Juanjuan Fan and Richard A. Levine."To Amnio or Not to Amnio: That Is the Decision for Bayes". In: Chance 20.3 (2007), pp. 26{32 (cit. on p. 1-7).
		• Tilmann Gneiting and Adrian E. Raftery. "Strictly Proper Scoring Rules, Prediction, and Estimation". In: J Am Stat Assoc 102 (2007), pp. 359{378 (cit. on p. 1-7).
		  ◦ wonderful review article except missing references from Scandanavian and German medical decision making literature
		• fharrell.com/post/addvalue
		• https://www.fharrell.com/post/mlconfusion/
		• https://www.fharrell.com/post/class-damage/
			Performance Metrics

Statistical models vs Machine learning (e.g. random forests, bagging, boosting, SVM, neural networks, deep learning)	• Machine learning techniques can require 200 events per candidate predictor (binary Y) as opposed to logistic regression which would require ~ 20 events • Statistical models allow for nonlinearity, interactions • Prefer machine learning if signal:noise ratio is large and outcome doesn't have strong component of randomness • Machine learning techniques assume interaction effects are likely to be as strong as main effects as opposed to statistical models where you can limit the number of interactions investigated	• (2.1) • (2.5) • Tjeerd van der Ploeg, Peter C. Austin, and Ewout W. Steyerberg.\Modern Modelling Techniques Are Data Hungry: A Simulation Study for Predicting Dichotomous Endpoints." In: BMC medical research methodology 14.1 (Dec. 2014), pp. 137+. issn: 1471-2288. doi: 10.1186/1471-2288-14-137. pmid: 25532820. url: http://dx.doi.org/10.1186/1471-2288-14-137 (cit. on pp. 2-4, 2-40, 4-18).	Machine Learning

• Multiplicative interaction effects from linear model can be misleading • Problem modeling for interactions if main effects aren't correctly specified (e.g. missing non-linear terms) • Should pre-specify interactions as much as possible, big opportunity for noisy results	• Don't test interactions by themselves, instead can do LRT checking all interactions in one shot ◦ Much more reliable • Misspecified model will yield false interactions • Shouldn't search for all possible 2-way interactions. Can have very high Chi-square in training set but then Chi square of 0 in test set.	• (2.3) • (2.7.2) • http://yiqingxu.org/papers/english/2018_HMX_interaction/main.pdf • Min Zhang et al. "Interaction Analysis under Misspecication of Main Effects: Some Common Mistakes and Simple Solutions". In: Statistics in Medicine n/a.n/a (2020). eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.8505. issn: 1097-0258. doi: 10.1002/sim.8505. url: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8505	Interactions

Categorizing continuous predictors is a disaster	• Amongst MANY other issues, the "optimal" cutpoint does not replicate over studies
	• Why do we use dichotomous BOR? At least we should keep it as ordinal?
	  ◦ https://discourse.datamethods.org/t/responder-analysis-loser-x-4/1262
		• (2.4)
		• Patrick Royston, Douglas G. Altman, and Willi Sauerbrei. "Dichotomizing Continuous Predictors in Multiple Regression: A Bad Idea". In: Stat Med 25 (2006), pp. 127{141. doi: 10.1002/sim.2331. url: http://dx.doi.org/10.1002/sim.2331 (cit. on p. 2-13).
		  ◦ Destruction of statistical inference when cutpoints are chosen using the response variable; varying effect estimates when change cutpoints
		  ◦ Difficult to interpret effects when dichotomize
		  ◦ Nice plot showing effect of categorization
		• Many other citations
			Modeling Strategies

Don't use polynomials to fit non-linear data	Does not fit adequately logarithmic functions or threshold effects. Has unwanted peaks and valleys		Non-linear Data

• Suggests using Restricted Cubic Splines to fit non-linear data
  ◦ Location of knots is not important in most situations (except where there's a "takeoff" in the function)
  ◦ Should place knots where data exist - fixed quantiles of predictor's marginal distribution
  ◦ Fit depends more on the choice of k (the number of knots)
  ◦ Should allow for 3 - 5 knots for each continuous predictor, if you can afford the df.
	• Smoothing splines (penalized splines) are useful for estimating trends in residual plots
	  ◦ Smoothing splines contain a lot of parameters, generally prefers regression splines
	  ▪ Burdensome (due to many tuning parameters), harder to program and describe. Regression splines typically perform just as well
	• In practice, can use Y to rank all covariates to determine which covariates to spend df on, in terms of number of knots. Not a problem of overfitting because a strong correlation with Y does not imply lack of linearity, it just prioritizes which covariates we are willing to invest in.
	• Interactions will be misleading if main effects are not properly modeled (e.g. modeled as linear instead of non-linear)
	• Should fit interactions of the form: X1f(X2) and X2g(X1)
		• (2.4)
		• F. E. Harrell, K. L. Lee, and B. G. Pollock. "Regression Models in Clinical Studies: Determining Relationships between Predictors and Response". In: J Nat Cancer Inst 80 (1988), pp. 1198{1202 (cit. on p. 2-32).
		• C. J. Stone. "Comment: Generalized Additive Models". In: Stat Sci 1 (1986), pp. 312{314 (cit. on p. 2-28).
		• S. Durrleman and R. Simon. "Flexible Regression Models with Cubic Splines". In: Stat Med 8 (1989), pp. 551{561 (cit. on p. 2-28).
		• Min Zhang et al. "Interaction Analysis under Misspecification of Main Effects: Some Common Mistakes and Simple Solutions". In: Statistics in Medicine n/a.n/a (2020). eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.8505. issn: 1097-0258. doi: 10.1002/sim.8505. url: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8505 (visited on 02/27/2020) (cit. on p. 2-47).
			Non-linear Data

• Rarely expect linearity and therefore *need* to allow flexibility for any *strong* predictor not known to predict linearly
  ◦ Can use Y to determine which parameters are "strong" to best determine how to spend df. Not a problem of information leakage since can also hurt performance
  ◦ In longitudinal studies, spaghetti plots should NOT be used to determine non-linearity (phantom df) rather just to look for abnormalities in data
	• Just because parameter is more associated with Y doesn't imply it is non-linear, it just means we care more about making mistake of linearity assumption. Therefore it's ok to use Y to help decide which covariates to spend extra df on to allow flexibililty
	  ◦ Compute partial ChiSquared for testing association of each predictor with response, adjusted for all other predictors
	  ▪ For categorical, subtract the degrees of freedom from Chisquare score
	  ▪ Should still incorporate into CV loop, although it's a conservative estimate so not terrible to leave out
	• In high dimensions, typically people don't allow for flexibility.
	  ◦ Better approach is Ridge Regression or Bayesian with Horseshoe Prior. NOT elastic net/Lasso
		• (4.1)	Non-linear Data

• Ridge regression performs better than elastic net/ lasso. NO method can reliably perform variable selection with -omics data.
  ◦ Can simplify final model by approximating it
	• Lasso has problems with scaling variables before analysis. The adaptive lasso tries to fix this (not sure if implemented in glmnet?).
	• When a predictor is represented by nonlinear basis functions, the scaling recommendations in the literature are not sensible. Also a problem for categorical predictors.
	• Problems with elastic net:
	  1. Default assigns equal penalties to L1 and L2 - should at least optimize alpha, although very difficult to optimize 2 tuning parameters simultaneously
	  2. Very small probability of identifying correct variables
	  a. Downfall of Lasso is that it tries to do variable selection
	  3. Ridge regression has better predictions than elastic net/lasso
	• Can approximate full model by treating full model as gold standard, then find subset of predictors which has highest R2 predicting Y_hat from full model
		• (2.5)
		• (5.5)
		• http://www.birs.ca/events/2019/5-day-workshops/19w5198/videos/watch/201906041303-Harrell.html
		  ◦ This is a talk from Frank Harrell I found browsing his website where he seems to cover this topic
		  ◦ Slides: http://hbiostat.org/talks/stratos19.pdf
		• https://discourse.datamethods.org/t/choosing-penalty-parameters-for-elastic-net/922
		  ◦ A blog post where Frank Harrell describes that elastic net is typically poor choose with binary outcome unless if there is a VERY large sample size
			Machine Learning

• Should use Multiple Imputation (MI), NOT complete case analysis. When proportion missing is <.03 it is ok to do complete case analysis, just to save analyst time, can also inset median in this scenario (otherwise don't).
  ◦ If MCAR, one should delete cases missing Y, otherwise don't (see Sybil paper)
	• Imputation doesn't help gain what was lost, but prevents you from losing what you've already measured
	• Uses Predictive Mean Matching (PMM) where missing value is replaced by value from subject with closest match (the *donor*, determined by closest prediction score from MI model)
	• Both MICE and aregImpute use chained equation approach
	  ◦ transcan performs single imputation
	• MI was developed as an approximation to full Bayesian model
	  ◦ Bayesian model treats missings as unknown parameters and provides exact inference and correct measures of uncertainty
	• Need to include imputation into validation cycle: e.g. imputation, validation, imputation, validation
	  ◦ Very long process, people are working on shortcuts
			Missing Data

- MI model should be at least as big as prediction model, but can be bigger
  - Should include Y
  - Interactions should also be imputed (instead of just multiplying the imputed main effect values) - Skeptical, but has paper which supports this
  - Number of imputations should be equal to max(5, 100*f) where f is proportion of observations with *any* variables missing
    - If very large sample size, can have less imputations
  - Should use PMM with weighted selection of donors
  - Use t distribution for tests and CI instead of Gaussian
  - Variables with the most NAs should be imputed first
- Both MICE and aregImpute (Hmisc package) perform MI, although aregImpute is faster (and preferred)
- Even if extreme amounts of data are missing, one should still use MI, since alternative (i.e. complete case analysis) is worse
- Bayesian approach is best (implemented with brms package)
  - Can also do "posterior stacking" by running typical MI algorithm and run Bayesian analysis for each completed dataset and draw thousands of samples from posterior distribution of the parameters. Then pool all of the posterior draws and do posterior inference as usual with no special correction required

- Chapter 3
- Soeun Kim, Catherine A. Sugar, and Thomas R. Belin. "Evaluating Model-Based Imputation Methods for Missing Covariates in Regression Models with Interactions". In: Stat Med 34.11 (May 2015), pp. 1876{1888. issn: 02776715. doi: 10.1002/sim.6435. url: http://dx.doi.org/10.1002/sim.6435 (cit. on p. 3-10).
- Sybil L. Crawford, Sharon L. Tennstedt, and John B. McKinlay. "A Comparison of Analytic Methods for Non-
Random Missingness of Outcome Data". In: J Clin Epi 48 (1995), pp. 209{219 (cit. on pp. 3-4, 4-46).
- Karel G. M. Moons et al. "Using the Outcome for Imputation of Missing Predictor Values Was Preferred". In: J Clin Epi 59 (2006), pp. 1092{1101. doi: 10.1016/j.jclinepi.2006.01.009. url: http://dx.doi.org/10.1016/j.jclinepi.2006.01.009 (cit. on p. 3-13).
  - Use of outcome variable
  - Excellent graphical summaries of simulations
- Ian R. White, Patrick Royston, and Angela M. Wood. "Multiple Imputation Using Chained Equations: Issues and Guidance for Practice". In: Stat Med 30.4 (2011), pp. 377{399 (cit. on pp. 3-1, 3-10, 3-15, 3-19).
  - Practical guidance for the use of multiple imputation using chained equations
  - MICE
  - Imputation models for different types of target variables
  - PMM choosing at random from among a few closest matches
  - Choosing number of multiple imputations by a reproducibility argument, suggesting 100f imputations when f is the fraction of cases that are incomplete
- K. J. Janssen et al. "Missing Covariate Data in Medical Research: To Impute Is Better than to Ignore". In: J Clin Epi 63 (2010), pp. 721{727 (cit. on p. 3-20).
- Paul Madley-Dowd et al. "The Proportion of Missing Data Should Not Be Used to Guide Decisions on Multiple Imputation". In: Journal of Clinical Epidemiology 110 (June 1, 2019), pp. 63{73. issn: 0895-4356. doi: 10.1016/j.jclinepi.2019.02.016.
  - url: http://www.sciencedirect.com/science/article/pii/S0895435618308710 (visited on 09/14/2019) (cit. on p. 3-20).
- https://cran.r-project.org/web/packages/brms/vignettes/brms_missings.html
  - Vignette demonstrating Bayesian approach to missing data via brms
  - http://hbiostat.org/R/rms/blrm.html

• Should have p < m /15, where p = # parameters in model and m = effective sample size
• NEVER use univariable screening (graphs, crosstabs, etc.) to choose predictors
	• Effective sample size is defined as n (total sample size for Continuous outcome)
	  ◦ min (n1, n2) for Binary outcome
	  ◦ Number of events for Survival outcome
	  ▪ Censored observations *slightly* boost effective sample size
		• (4.4)
		• Guo-Wen Sun, Thomas L. Shook, and Gregory L. Kay. "Inappropriate Use of Bivariable Analysis to Screen Risk Factors for Use in Multivariable Analysis". In: J Clin Epi 49 (1996), pp. 907{916 (cit. on p. 4-19).
			Feature Reduction

• Unless n >> p, model is unlikely to validate, there necessary to first do data reduction
  ◦ Data reduction MUST be masked from Y
  ◦ Eliminate variables with narrow distributions
  ◦ Can use PCA (unsupervised) or variable clustering and then summarize each cluster with its 1st principal component
  ◦ Can do "redundant analysis" removing variables that can be predicted well from others
  ◦ Univariable screening is the WORST of all approaches
  ▪ Worse than False Positives are the high numbers of false negatives
  ▪ Works like forward stepwise regression, but worse since once gene is rejected it doesn't have chance of being allowed back in
  ▪ Assumes genes work in trivial way instead of complex framework /pathways
	• Likes sparsePCA to help in interpretability
	• Redundant analysis implemented by "redun" function in Hmisc package
	  ◦ Related to Principal Variables, but faster
	• Look into "transcan" function for data reduction via transformation and imputation
	• "varclus" in Hmisc package helps cluster the variables
	• Prefers data reduction to penalized regression, because more stable, although tricky because it's a 2-step procedure as opposed to penalized regression which does both in one step.
		• (4.7)
		• George P. McCabe. "Principal Variables". In: Technometrics 26 (1984), pp. 137{144 (cit. on p. 4-27).
		• Jian Guo et al. "Principal Component Analysis with Sparse Fused Loadings". In: J Comp Graph Stat 19.4 (2011), pp. 930{946 (cit. on p. 4-28).
		  ◦ Incorporates blocking structure in the variables
		  ◦ Selects different variables for different components
		  ◦ Encourages loadings of highly correlated variables to have same magnitude, which aids in interpretation
			Feature Reduction

• When comparing models, 1st dismiss any model with poor calibration (unless if goal is to rank order) and only afterwards compare discrimination
  ◦ Discrimination can be measured by R2, model Chi_Square, Somers D, Spearmans p, or AUC (see comments)
• Shouldn't compare models on measure used to optimize either model
	• Problems with AUC or other rank measures
	  ◦ Rank measures will not give enough credit to extreme predictions, therefore model Chi_Square and R2 are preferred
	  ◦ Gold standard is LR Chi_Square, Deviance, R2, examine extremes of distribution of Y_hat
	  ◦ Test of difference between models in AUC does not have enough power
	  ◦ AUC should only be descriptive tool, not to choose model
		• (4.10)
		• N. J. D. Nagelkerke. "A Note on a General Denition of the Coecient of Determination". In: Biometrika 78 (1991), pp. 691{692 (cit. on p. 4-41).
			Performance Metrics

• Summary of Strategy for Developing Predictive Models:
  1. See Greenland paper
  2. In most cases use MI to impute X based on Y
  3. Specify df willing to spend for each predictor
  a. More for important predictors
  4. Do data reduction if needed (pre-transformations, combinations), or penalized estimation
  5. Check distributional assumptions and choose different model if needed
  6. Validate via bootstrap or repeated CV
  a. Generally prefers bootstrap except when n < p (see comments)
  7. When doing effect estimation, typically don't penalize main variable of interest
	• Prefers bootstrap over repeated CV because it's more computationally efficient, easier, and performs just as well,
	  ◦ CV is less precise than bootstrap
	  ◦ Bootstrap has advantage of estimating optimism of final whole sample fit without any holdout data, as opposed to CV
	• Validation process using bootstrap:
		• (4.11 -4 .12)
		• (5.6)
		• (5.3)
		• Sander Greenland. "When Should Epidemiologic Regressions Use Random Coefficients?" In: Biometrics 56 (2000), pp. 915{921. doi: 10.1111/j.0006- 341X.2000.00915.x. url: http://dx.doi.org/10.1111/j.0006-341X.2000.00915.x (cit. on pp. 4-10, 4-43).
		  ◦ Use of statistics in epidemiology is largely primitive
		  ◦ Stepwise variable selection on confounders leaves important confounders uncontrolled
		  ◦ Composition matrix
		  ◦ Example with far too many signicant predictors with many regression coeffcients absurdly inflated when overt
		  ◦ Lack of evidence for dietary effects mediated through constituents
		  ◦ Shrinkage instead of variable selection
		  ◦ Larger effect on confidence interval width than on point estimates with variable selection
		  ◦ Uncertainty about variance of random effects is just uncertainty about prior opinion
		  ◦ Estimation of variance is pointless
		  ◦ Instead the analysis should be repeated using different values
		  ◦ "If one feels compelled to estimate Tau^2, I would recommend giving it a proper prior concentrated amount contextually reasonable values"
		  ◦ Claim about ordinary MLE being unbiased is misleading because it assumes the model is correct and is the only model entertained
		  ◦ Shrinkage towards compositional model
		  ◦ "Models need to be complex to capture uncertainty about the relations...an honest uncertainty assessment requires parameters for all effects that we know may be present. This advice is implicit in an antiparsimony principle often attributed to L. J. Savage 'All models should be as big as an elephant (see Draper, 1995)'".
		  ◦ See also gus06per.
		• Frank E. Harrell et al. "Development of a Clinical Prediction Model for an Ordinal Outcome: The World Health Organization ARI Multicentre Study of Clinical Signs and Etiologic Agents of Pneumonia, Sepsis, and Meningitis in Young Infants". In: Stat Med 17 (1998), pp. 909{944. url: http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0258(19980430)17:8%3C909::AID SIM753%3E3.0.CO;2-O/abstract (cit. on pp. 4-21, 4-46).
			Modeling Strategies

• Describing Fitted Model
  ◦ Use partial effect plots to summarize each continuous predictor instead of OR/HR of IQR, unless if relationship is linear
  ◦ Use nomograms
	• Partial effects plots are easy to implement, but more complicated if there's interactions (in which case 2 curves are needed).
	  ◦ Hold other variables to median (or model if categorical)
		• (5.1)
		• Juha Karvanen and Frank E. Harrell. "Visualizing Covariates in Proportional Hazards Model". In: Stat Med 28 (2009), pp. 1957{1966 (cit. on p. 5-2).
		• https://stats.stackexchange.com/questions/155430/clarifications-regarding-reading-a-nomogram
			Visualization

• Error Measures
  ◦ For binary Y, use Brier Score
• Discrimination measures
  ◦ Use R2 or R2_adj
• Look at calibration slope and intercept
  ◦ Can also look at max calibration error, mean calibration error, and 0.9 quantile of calibration error
	• Model Y vs Y_hat
	  ◦ Perfect calibration would have intercept = 0 and slope = 1
		• Ben Van Calster et al. "A Calibration Hierarchy for Risk Models Was Dened: From Utopia to Empirical Data". In: J Clin Epi 74 (June 2016), pp. 167{176. issn: 08954356. doi: 10.1016/j.jclinepi.2015.12.005. url: http://dx.doi.org/10.1016/j.jclinepi.2015.12.005 (cit. on p. 5-4).	Performance Metrics

• Model Validation
  ◦ DON'T split data (once)
  ◦ Should report arbitrariness of selected "important variables" across repeated sampling (if variable selection is used)
  ◦ Generally, better not to do external validation and instead include all data available, and include country/time, etc. as a predictor. Unless if very large sample size
  ▪ Instead do rigorous internal validation
• Can use adjusted R2 for validating OLS model
	• If model was fully pre-specified, External validation tests the model. BUT if "final model" was fit via machine learning, feature selection, etc. the model developed via training data is only an "example model" and therefore the external validation is only an "example validation".
	• Resampling reveals the volatility of model selection process
	  ◦ Should report ranks of variables selected across all repeats
		• (5.2 - 5.3)
		• (5.4)
		• Bradley Efron and Balasubramanian Narasimhan. "The Automatic Construction of Bootstrap Confidence Intervals". In: Journal of Computational and Graphical Statistics 0.0 (Jan. 14, 2020). eprint: https://doi.org/10.1080/10618600.2020.1714633, pp. 1{12. issn: 1061-8600. doi: 10.1080/10618600.2020.1714633. url: https://doi.org/10.1080/10618600.2020.1714633 (visited on 03/13/2020) (cit. on p. 5-10).
		  ◦ Contains information about bootstrap CI and latest R functions
		• BBR 10.11
			Validation

• Validation process using bootstrap:
  ◦ Derive model from each bootstrap sample and apply it to original sample
  ◦ Estimate performance on original sample
  ◦ Estimate performance on each bootstrap sample as well
  ◦ Estimated Optimism = difference between average performance on original sample to performance on bs sample
  ◦ Avg. Optimism = Average the estimated optimism across all bs samples
  ◦ Final performance metric = Original performance metric calculated on full (training) data set (predicted from final model) - Avg. Optimism
	• Prefers bootstrap over repeated CV because it's more computationally efficient, easier, and performs just as well,
	  ◦ CV is less precise than bootstrap
	  ▪ Need to do about 50 repeats of 10-fold CV to ensure adequate precision
	  ◦ Bootstrap has advantage of estimating optimism of final whole sample fit without any holdout data, as opposed to CV
	• Modifications to bootstrap are not worth the trouble.
	  ◦ Standard "0.632" Bootstrap performs better for small n
		• (5.3)
		• B. Efron. "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation". In: J Am Stat Assoc 78 (1983), pp. 316{331 (cit. on pp. 5-17, 5-20, 5-21).
		  ◦ Suggested need at least 200 models to get an average that is adequate, i.e., 20 repeats of 10-fold cv
		• Bradley Efron and Robert Tibshirani. "Improvements on Cross-Validation: The .632+ Bootstrap Method". In: J Am Stat Assoc 92 (1997), pp. 548 -560 (cit. on p. 5-20).
		• J. C. van Houwelingen and S. le Cessie.\Predictive Value of Statistical Models". In: Stat Med 9 (1990), pp. 1303-1325 (cit. on pp. 2-29, 4-21, 5-17, 5-21, 5-22).
			Validation

• In repeated measures data, adjust for baseline as covariate in model - DO NOT include as first outcome
• Do not use change from baseline as a response variable
• Prefers to treat time as continuous with spline rather than categorical
	• See Modeling Baseline in Repeated Measures Studies for more detailed discussion regarding modeling baseline.
	• Baseline cannot be a response to treatment, therefore baseline and response cannot be modeled in an integrated framework
	• Baseline tends to be best predictor of outcome post-randomization and keeping baseline on LHS will increase precision of estimated treatment effect
	• Any prognostic factors correlated with outcome will also be correlated with baseline.
	  ◦ This reduces the need to enter a large number of prognostic variables into model
	• Also, many studies have inclusion/exclusion criteria that include cutoffs on baseline and therefore baseline comes from a truncated distribution and shouldn't be modeled with same distributional shape as other timepoints.
	• Splines take up less df than categorical time
	  ◦ Splines also smoother and therefore information from earlier timepoints can still be used to inform later timepoints as opposed to if time was categorical
		• (7.2)
		• BBR 14.4
		  ◦ https://www.youtube.com/watch?v=1r5iNqtVGuE
		
		Discussion regarding whether to include baseline as predictor or outcome in model: The 1st two papers below are rebuttals to the 3rd and 4th papers.
			Longitudinal Data

- Michael G. Kenward, Ian R. White, and James R. Carpener.\Should Baseline Be a Covariate or Dependent Variable
in Analyses of Change from Baseline in Clinical Trials? (Letter to the Editor)". In: Stat Med 29 (2010), pp. 1455{
1456
    - This is a strong rebuke of the Liu paper (see below), supporting the inclusion of baseline as predictor in the model
- Stephen Senn. \Change from Baseline and Analysis of Covariance Revisited". In: Stat Med 25 (2006), pp. 4334{
4344 (cit. on p. 7-5).
    - Shows that claims that in a 2-arm study it is not true that ANCOVA requires the population means at baseline to be identical
    - Refutes some claims of Liang and Zeger (see below)
    - problems with counterfactuals
    - temporal additivity ("amounts to supposing that despite the fact that groups are different at baseline they would show the same evolution over time")
    - Causal additivity is difficult to design trials for which simple analysis of change scores is unbiased, ANCOVA is biased, and a causal interpretation can be given
    - temporally and logically, a "baseline cannot be a "response to treatment", so baseline and response cannot be modeled in an integrated framework as Laird and Ware's model has been used"
    - One should focus clearly on `outcomes' as being the only values that can be influenced by treatment and
examine critically any schemes that assume that these are linked in some rigid and deterministic view to `baseline' values. An alternative tradition sees a baseline as being merely one of a number of measurements capable of improving predictions of outcomes and models it in this way."
    - "You cannot establish necessary conditions for an estimator to be valid by nominating a model and seeing what the model implies unless the model is universally agreed to be impeccable. On the contrary it is appropriate to start with the estimator and see
what assumptions are implied by valid conclusions." - this is in distinction to Liang and Zeger (see below)
- Guanghan F. Liu et al.\Should Baseline Be a Covariate or Dependent Variable in Analyses of Change from Baseline
in Clinical Trials?" In: Stat Med 28 (2009), pp. 2509-2530
    - This is a paper suggesting to use baseline as an outcome (and not predictor)
- Kung-Yee Liang and Scott L. Zeger.\Longitudinal Data Analysis of Continuous and Discrete Responses for Pre-Post
Designs". In: Sankhya 62 (2000), pp. 134{148 (cit. on p. 7-5).
    - This is the first paper (in this discussion) which is in support of using baseline as outcome (and not predictor)
    - Makes an error in assuming the baseline variable will have the same univariate distribution as the response except for a shift
    - baseline may have for example a truncated distribution based on a trial's inclusion criteria
    - if correlation between baseline and response is zero, ANCOVA will be twice as efficient as simple analysis of change scores
    - if correlation is one they may be equally efficient
- J. Martin Bland and Douglas G. Altman. "Comparisons against Baseline within Randomised Groups Are Often Used and Can Be Highly Misleading". In: Trials 12.1 (Dec. 2011), p. 264. doi: 10.1186/1745- 6215- 12- 264. url: https://doi.org/10.1186/1745-6215-12-264 (cit. on pp. 13-2, 14-9).

• Prefers GLS to mixed effects model
• Does not recommend GEE
  ◦ Mixed model is better than GEE
	• Random Effects are not robust for non-continuous outcomes - "just used for convenience"
	  ◦ In linear model, they're robust and work well, but not for non-linear model (e.g. logistic)
	  ▪ Requires marginalization/integration to get estimates you need, very tricky
	• With random effects, Induced correlation structure for Y is unrealistic
	  ◦ Can add to model specifications to get more versatile structure instead of Compound Symmetry, which is default
	• Random effects are numerically demanding
	• Random effects cause computing time to go up
	• Random effects requires complex approximations for distributions of test statistics
	• Random effects are good for clustering, not for repeated measures
	• Generalized Least Squares (a.k.a Growth Curve Models) allows for each subject to have separate covariance matrix, although typically we assume a diagonal matrix
	• GEE is not as robust as GLS or Mixed models because it needs a large sample to work
	  ◦ Also not robust to non-random dropout as opposed to mixed effects and GLS where you can have missing at random (GEE needs missing completely at random)
	  ◦ GEE requires using MI to fill in missing visits as opposed to GLS or mixed effects models
	  ◦ GEE also relies on large sample size
		• (7.3)
		• Joseph C. Gardiner, Zhehui Luo, and Lee A. Roman. "Fixed Effects, Random Effects and GEE: What Are the Differences?" In: Stat Med 28 (2009), pp. 221{239 (cit. on p. 7-9).
		  ◦ Nice comparison of models; econometrics
		  ◦ Different use of the term "fixed effects model"
		• See (7.8) for Bayesian Proportional Odds Random Effects model approach
			Longitudinal Data

Bayesian Logistic Regression	Good examples of using brms package and rms package to implement Bayesian Logistic Regression	• (8.11) • (9.8)	Bayesian

• Ordinal Logistic Regression can be applied to continuous Y - Can have as many "categories" as observations
  ◦ Proportional Odds model
  ◦ Continuation Ratio Model (same as discrete proportional hazards model)
	• PO model assumes parallel odds (intercepts can be based on category of Y but not slopes
	• Has advantages including robustness and freedom of distributional assumptions for Y conditional on any given set of predictors
	  ◦ Don't need to select transformations
	  ◦ Can work well even for extremely skewed data, based only on ranks
	• If interested in estimating quantiles, quantile regression is good. If interested in estimating the mean, linear model is good. But ordinal regression is good at estimating anything.
		• (10.1)
		• See Chapter 11 for case study comparing ordinal logistic regression to OLS and quantile regression
			Ordinal Logistic Regression