Statistics for Radiomics

Shouhao Zhou, PhD
The University of Texas MD Anderson Cancer Center

Radiomics Certificate, AAPM 2018
Nashville, TN

Outline

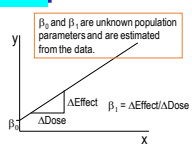
1. General statistical modeling concepts and approaches
   - Illustrated in linear regression
   - Objective: Explanatory vs Predictive
2. A statistical predictive feature selection approach for Radiomics
3. Machine learning and Statistics
   - Cox regression, cross-validation, LASSO, Bootstrap resampling, … …

Estimation of Parameters in Linear Model

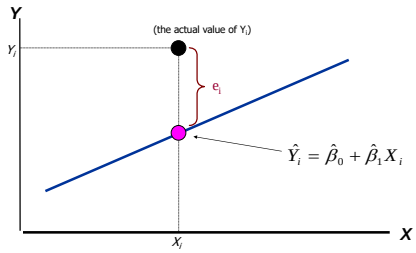
- The linear model

$$y = \beta_0 + \beta_1 x + e$$

y = dependent variable
x = independent variable
β_0 = y-intercept
β_1 = slope of the line
e = error term



Linear Regression - Model



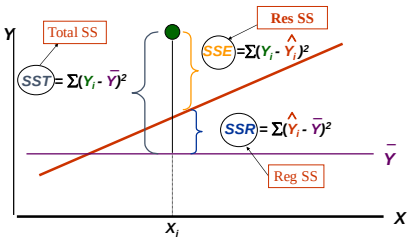
Objective is to find the estimates for β_0 and β_1 that minimize the sum of the squared distance from the points to the line.

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 x_i \right] \right)^2$$

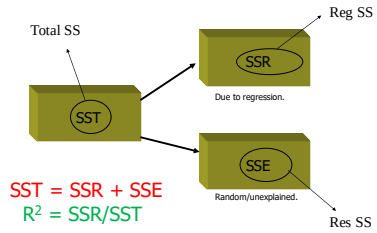
A good line is one that minimizes the sum of squared differences between the points and the line (Res SS, aka SSE).

SS = "sums of squares"

Linear Regression – Sources of Variation



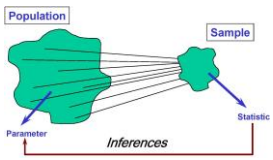
Linear Regression - Variation



SST = SSR + SSE

R^2 = SSR/SST

Overfitting



- Modeling techniques tend to overfit the data.
- Multiple regression:
  - Every time you add a variable to the regression, the model's **R^2** goes up.
  - Naïve interpretation: every additional predictive variable helps to explain yet more of the target's variance.
    - But that can't be true!
  - Error on the dataset used to *fit* the model can be misleading
    - Doesn't predict future performance.
  - Too much complexity can diminish model's accuracy on future data.
    - Sometimes called the Bias-Variance Tradeoff.

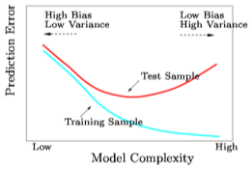
Choosing the Optimal Predictive Model

- The model containing all the predictors will always have the smallest RSS and the largest R^2, since these quantities are related to the training error.
- We wish to choose a model with **low test error**, not a model with low training error.
  - Training error is usually a poor estimate of test error.
- RSS and R^2 are not suitable for selecting the *best* predictive model among a collection of models with different numbers of predictors.

Estimating Prediction Error

Two approaches:

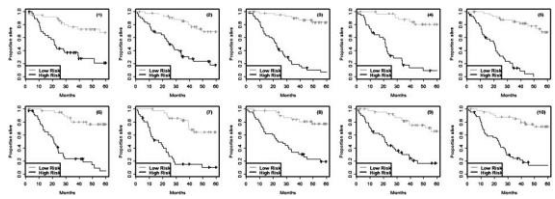
1. Estimate the test error *directly*, by using either a validation set approach or a cross-validation approach.
2. Estimate the test error *indirectly*, by making an adjustment to the training error to account for the bias due to overfitting. (e.g. AIC)



Validation and Overfitting

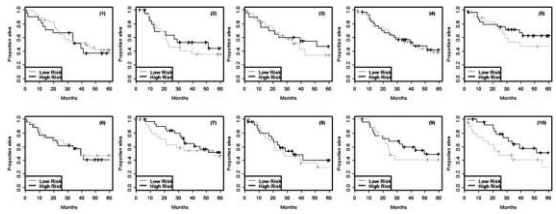
- To assess predictive performance, entire data can be separated into two pieces:
  - Training sample
    - Model fitting -> Minimize training error
  - Test sample
    - Model evaluation -> Compare Test Error

Kaplan–Meier curve in training sample (n=129, p=5K)



Subramanian and Simon, JNCI 102, 2010

K-M curve in test sample

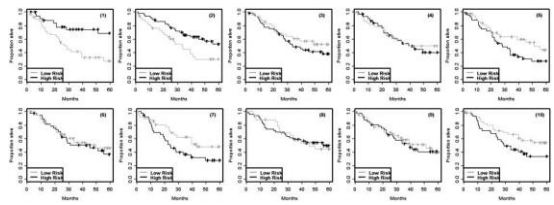


Subramanian and Simon, JNCI 102, 2010

Cross-Validation (CV, Stone 1974)

- Partition the full data set **D** into a prespecified **K** approximately equal parts $\mathbf{D_1}$, $\mathbf{D_2}$, $\mathbf{D_3}$, …, $\mathbf{D_K}$
- For each k, form a training set $\mathbf{T_k = D - D_k}$ for model fitting. Compute the validation set error or the CV error for subset $\mathbf{D_k}$ under consideration.
- Select the *model* for which the resulting estimated test error is smallest.
  - A direct estimate of the test error (prediction error)
  - Leave-one-out Cross-validation (LOO-CV) is a special case when **K=n**
  - Can be used in a wider range of model selection tasks

10-fold Cross-validated K-M curve



Simon et al., Briefings in Bioinfo 12, 2011

Bias Correction Approaches

- To assess prediction error, we can use other approaches to correct the bias in estimating the training error:
  - AIC (Akaike information criterion)
  - Mallow's C_p (equivalent to AIC for linear regression)
- Avoid partitioning the data
- These techniques adjust the training error for the model size, and can be used to select among a set of models with different numbers of variables.
- A small value of C_p and AIC indicates a low error, and thus a better model.

Mallow's C_p (Mallow 1973)

- For a fitted OLS model containing ***d*** predictors, the $\mathbf{C_p}$ estimate of test mean squared error:

$$C_p = \frac{1}{n}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right)$$

where $\hat{\boldsymbol{\sigma}}^2$ is an estimate of the variance of the error **ε** associated with each response measurement.

- Here, a penalty is added to the training RSS (Residual SS) in order to adjust for the fact that the training error tends to underestimate the test error.

Akaike Information Criterion (AIC, Akaike 1973)

- Defined for a large class of models fit by maximum likelihood.

$$\mathrm{AIC} = -2\log L + 2 \cdot d$$

where *L* is the maximized value of the likelihood function for the estimated model.

- In the case of the linear model with Gaussian errors, MLE and OLS are the same things; thus, C_p and AIC are equivalent.
- Stone (1977) proved that in linear regression the error term of both LOO-CV and AIC in estimating the test error are in the same order of $o(n^{-1})$.
  - When sample size **n** -> Inf, the error -> 0.

LASSO (Tibshirani 1996)

- Recall that the OLS fitting procedure estimates the beta coefficients using the values that minimize:

$$\text{RSS} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

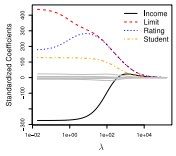
- LASSO (least absolute shrinkage and selection operator) is similar to OLS, except that the coefficients are estimated by minimizing a slightly different quantity:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

where $\lambda \geq 0$ is a tuning parameter, to be determined separately.

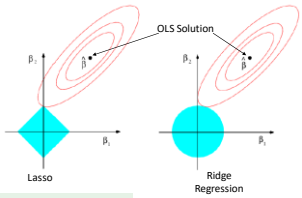
LASSO

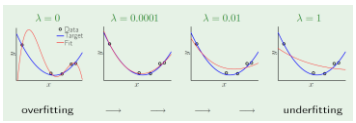
- Note that $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage.
- The idea of penalizing by the sum-of-squares of the parameters is also used in neural networks, where it is known as *weight decay*.
- Lasso uses an ℓ_1 penalty, which has the effect of forcing some of the coefficients to be exactly equal to zero when the tuning parameter λ is sufficiently large. Thus, the lasso performs variable/feature selection.
- As λ increases, the standardized ridge regression coefficients shrinks towards zero.
- Thus, when λ is extremely large, then all of the LASSO coefficient estimates are basically zero; this corresponds to the *null model* that contains no predictors.



Shrinkage estimators

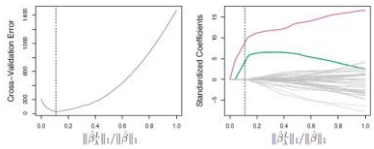
The lasso and ridge regression coefficient estimates are given by the first point at which an ellipse contacts the constraint region.





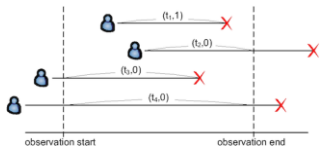
Selection of the Tuning Parameter λ

- Select a grid of potential values; use cross-validation to estimate the error rate on test data (for each value of λ) and select the value that gives the smallest error rate.
- The model is re-fit using all of the variable observations and the selected value of the tuning parameter λ.



Survival Analysis

- In many biomedical studies, the primary endpoint is time until an event occurs (e.g. death, recurrence, new symptoms, etc.)
- Data are typically subject to right *censoring* when a study ends before the event occurs.



- Often we assume censoring is noninformative, i.e., patients who are censored have the same underlying survival curve after their censoring time as patients who are not censored.

30

Description of a probability distribution:

1. Distribution function
$$F(t) = \Pr(T \le t)$$
2. Density function
$$f(t) = \frac{\partial F(t)}{\partial t}$$
3. Survival function
$$S(t) = \Pr(T > t) = 1 - F(t)$$
4. Hazard function
$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T \le t + \Delta t \mid T \ge t)}{\Delta t} = \frac{f(t)}{S(t)}$$
5. Cumulative hazard function
$$\Lambda(t) = \int_0^t \lambda(u)du$$

Cox PH Regression Model

- Proportional hazards assumption
- No distributional assumptions on $\lambda_0(t)$
- Linear effects for numeric covariates
- Generalized Linear Model

$\lambda(t; X) = \lambda_0(t) \exp[\beta_1 X_1 + ... + \beta_p X_p]$
- $\lambda_0(t)$ is baseline hazard function
- relates hazard function to covariates

For a single binary covariate, X_1:
$X_1 = 0 \Rightarrow \lambda(t; 0) = \lambda_0(t)$
$X_1 = 1 \Rightarrow \lambda(t; 1) = \lambda_0(t)\exp(\beta_1)$

The ratio of hazard functions,
$HR = \lambda(t; 1)/ \lambda(t; 0) = \exp(\beta_1)$
Or, more generally
$HR = \lambda(t; X_0+\Delta X)/ \lambda(t; X_0) = \exp(\beta_1 \Delta X)$

Challenges in Modeling for Radiomics

- Large *p* small *n* problem
- Data splitting into training and test sets would reduce the size of the already small training set that is used for the development of the risk prediction model thus increasing problems of instability.

Guideline:
- take the information on survival and censored times into consideration.
- binary classification highly depends on the survival threshold used to define the two classes. A slight change of the threshold can lead to very different prediction accuracy and interpretation.
- binary modeling approach can result in loss of efficiency and potential bias in high dimensional settings.
- the choice of threshold affected the predictions.

33

Procedure

- Pre-screening
  - Identify robust and/or correlated radiomics features
- Feature Selection
  - multivariate Cox proportional hazards regression model with LASSO-penalty
  - CV/GIC to determine the tuning parameters
  - Boostrap/CV to determine the most frequently selected features
- Model fitting and Evaluation
  - Comparison between the best clinical model and best clinical+imaging model
- Validation
  - based on independent test sample

34

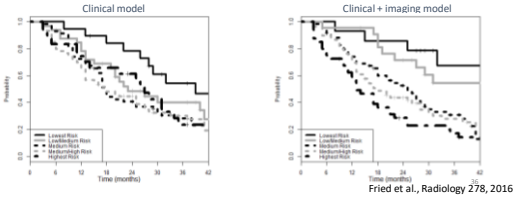
Other measures for Model Evaluation

- Cross-validated K-M curves
- Harrell's Concordance index (C-index)
  - Fraction of all pairs of subjects whose predicted survival times are correctly ordered
  - Cross-validated C-index for predictive assessment
- Time-dependent ROC curves
  - For every specific time t, ROC(t) was plotted as sensitivity(t) versus 1-specificity(t) for all values of the risk score cutoff used to define binary classes
  - Optimal cutoff can be determined based Youden's index
- Integrated Brier scores
  - Integrating out over time the score functions which measure the discrepancies between true disease status and predictive risk scores

35

Risk classification based on Statistical Modeling

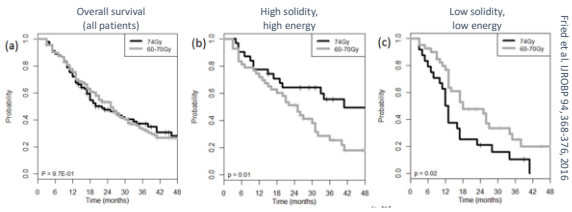
- 195 Patients, stage III NSCLC w/ definitive XRT
- 11 conventional prognostic factors
- MIM PETedge: Semi-automated delineation
- 47 Quantitative Image Features (QIFs) [IBEX]
- Clustering to try to identify multiple risk groups



Fried et al., Radiology 278, 2016

A case when machine learning techniques could be very useful

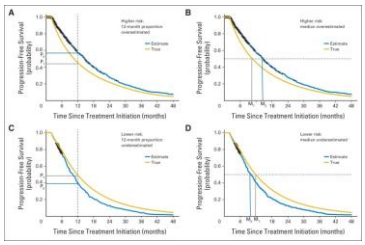
- RTOG 0617 showed no benefit (possible harm) in dose escalation for stage III NSCLC patients
- What if there are sub-groups of patients that would benefit?



Fried et al. IJROBP 94, 368-376, 2016

Impact of Informative Censoring in Survival Analysis

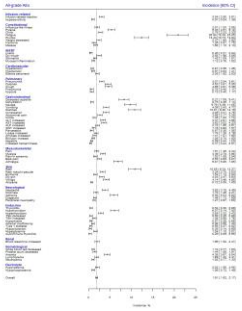
- An event is said to be censored if the exact time at which the event occurs is not observed.
- Often we assume censoring is noninformative, i.e., patients who are censored have the same underlying survival curve after their censoring time as patients who are not censored.
- In reality, informative censoring for PFS could arise when patients are censored for initiation of an effective anticancer treatment before the protocol-defined progression.



Campigotto and Weller, *JCO* **32:** 3068-74, 2014

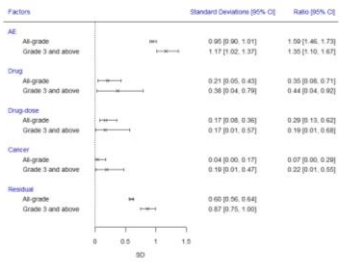
Informative Censoring

- Background: programmed cell death protein 1 (PD-1) and programmed death-ligand 1 (PD-L1) inhibitors have been increasingly used in cancer therapy.
- Objective: to understand toxicity profile of treatment-related AEs of PD-1/PD-L1 inhibitors
- Approach: to conduct a systematic review and meta-analysis of treatment-related AEs of single-agent PD-1 and PD-L1 inhibitors
- Method: to develop and apply an innovative Bayesian approach for deriving exact inference identical to results using individual-level patient data.
- Result: without correcting the informative censoring, the overall average incidence of all-grade AE would have been 2.65%, an over-estimation of 0.74% or by nearly 40%.

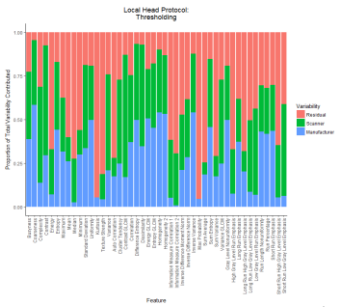


Wang et al. under review

Variance Decomposition



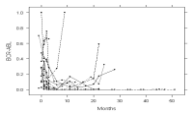
Variance Decomposition



Ger et al. Under review

Estimation of Longitudinal Biomarker Trajectory

- 635 CML patients received TKI in first/second line treatment.
- 6035 measurements of BCR-ABL expression levels.
- Patients are instructed to schedule follow-up visits approximately at 3, 6, 9, 12, 18, 24 months, and yearly thereafter.
- Real total number of visits varies (Max:23). Real patient visiting times scatter through all the time.



- Bayesian Semi-parametric beta regression
- Estimation of subject-specific Time-dependent effect

$$f(Y_{i,t} = y_{i,t}|\boldsymbol{\mu}, \boldsymbol{\phi}) \sim \texttt{Beta}(\mu_{i,t}\phi_{i,t}, (1-\mu_{i,t})\phi_{i,t})$$
$$\texttt{Link}^{\mu}(\mu_{i,t}) = \texttt{logit}^{-1}(\mu_{i,t}) = r_{i,\mu}(t) + \Sigma_j s_{\mu,k}(X_{\mu,j,i}),$$
$$\texttt{Link}^{\phi}(\phi_{i,t}) = \texttt{exp}(\phi_{i,t}) = r_{i,\phi}(t) + \Sigma_j s_{\phi,k}(X_{\phi,j,i}).$$

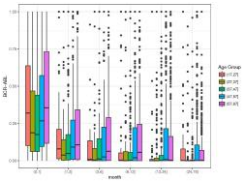
Term specification:

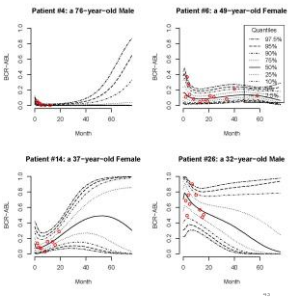
$$r_{i,\mu}(t) = \sum_k \mathbf{u}_{T,i,k,\mu} h_k(t)$$
$$\Sigma_j s_{\mu,k}(X_{\mu,j,i}) = \alpha_{i,\mu} + \alpha_{\mathbb{S},\mu}\texttt{Sex}_i + \sum_k u_{A,k,\mu} h_k(\texttt{Age}_i)$$
$$r_{i,\phi}(t) = \sum_k \mathbf{u}_{T,k,\phi} h_k(t)$$
$$\Sigma_j s_{\phi,j}(X_{\phi,j,i}) = \alpha_{i,\phi} + \alpha_{\mathbb{S},\phi}\texttt{Sex}_i$$

Patient-Specific Time trend

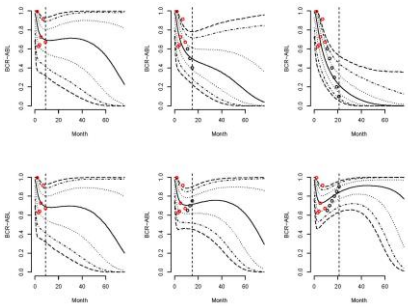
$$\mathbf{u}_{T,i,k,\mu} \sim \mathbb{N}(\mathbf{u}_{T,k,\mu}, \sigma^2_{T,k,\mu})$$

Zhou et al. Under review





Real-time Prediction



2 Cultures in Data Science

Machine Learning

- Look for generalizable predictive patterns
- Choose a predictive algorithm by relying on its empirical capabilities
- Focus on computational methodology
- Result driven

Traditional Statistics

- Draw population inferences from a sample
- Choose a model that incorporates current knowledge of the system
- Focus on math/probability methodology
- Process driven

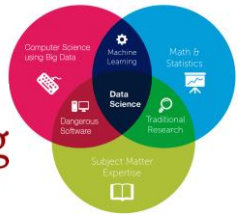
Breiman, *Statist.Sci.* **16:** 199-231, 2001
Blei and Smyth, *PNAS* **114:** 8689-92, 2017
Bzdok et al, *Nature Methods* **15:** 233-34, 2018

Glossary

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = $1,000,000	large grant= $50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

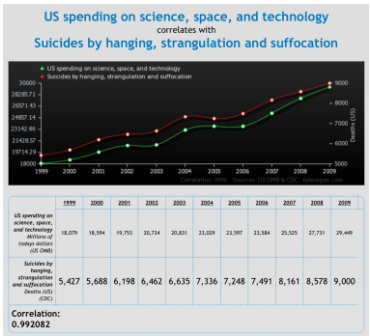
Source: Rob Tibshirani's personal website

An Era of Data Science



Seeing scientific applications turn into methodological advances is always a joy, at least for those of us who care about advancing the science of data, concurrent with advancing science through data.

Xiao-Li Meng, Harvard U



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
US spending on science, space, and technology Millions of today dollars (US OMB)	18,079	18,594	19,753	20,734	20,831	23,029	23,597	23,584	25,525	27,731	29,449
Suicides by hanging, strangulation and suffocation Deaths (US) (CDC)	5,427	5,688	6,198	6,462	6,635	7,336	7,248	7,491	8,161	8,578	9,000

Correlation: 0.992082

Acknowledgements

• Biostatistics
Don Berry, PhD
Xuelin Huang, PhD
• Epigenetics
Claudio Aldaz, PhD
• Emergency Medicine
Sai-Ching Yeung, MD PhD FACP
• Endocrine Neoplasia
Ramona Dadu, MD
• Experimental Therapeutics
Waldemar Priebe, PhD
• Imaging Physics
Kyle Jones, PhD
• Lymphoma
Michael Wang, MD
• Stem Cell Transplantation
Katy Rezvani, MD PhD

• Leukemia
Kapil Bhalla, MD FACP
Jorge Cortes, MD
• Neurosurgery
Frederick Lang, MD FACS
Amy Heimberger, MD
• Pediatrics
Shulin Li, PhD
• Radiation Physics
Laurence Court, PhD
Stephen Kry, PhD

• GRA
Shuangshuang Fu
Nancy Qi
Bo Zhao

• 5P30CA016672
• P50CA100632
• P50CA136411
• P50CA168505
• R01CA210250
• R01CA214526
• R01CA214749
• R01EB026291
• **R21CA216572**
• R21CA202104
• HHSN261201500018

• CPRIT
• Gateway
• BMS
• MDACC Moonshot