Model Selection I and II

Lauren J Beesley

Postdoctoral Research Fellow in Biostatistics at UM

BDSI Summer Program
(Some slides stolen from Bhramar)
Introduction

- Note: We are not trying to find the “correct” model
  - “All models are wrong, but some are useful” – George Box, 1979

- Want to find a “good” model (whatever that means)
Introduction

- Inferential tools assume that the model was specified ahead of time. They are invalid if the model was chosen based on data.
  - E.g. in clinical trials, models usually specified ahead of time
  - Generally, data-adaptive model selection often biases coefficients away from zero and residual variance towards zero. **False positives and overfitting!**

- Easiest solution is not to carry out model selection based on data, but often unavoidable
Overstretching your data to create a buzz...

Fig. 1. The relations of Africanized and stingless (meliponine) bee abundances on flowering *Melochia villosa*. The dashed line is a quadratic polynomial (given by $y = -0.516 + 1.08x - 0.023x^2$) which gave the best fit to the points (7).

Bee Careful of Stinging rebuke

Curve-Fitting

The rather fanciful curve-fitting of Roubik (Reports, 15 Sept., p. 1030, Fig. 1) has prompted me to propose an alternative interpretation of his data (see below).

ROBERT M. HAZEN

So it fits the data...what about generalizability?
Model selection strategy may depend on inferential objective:

- Prediction
- Estimation
- Hypothesis Testing
- Interpretation as risk factors
- Discovery of biomarkers
- Testing treatment effect
- Causal interpretation of a coefficient
- Identification of sets of important predictors/variables

Model selection refers to both (1) model structure/type (2) included predictors
Cox Regression

Generalized Linear Models (GLMs)
- Logistic
- Poisson
- Negative Binomial
- Multinomial
- Proportional Odds

CART (Trees)

Linear Regression

Multivariate Regression Models
  e.g. Multistate models

Latent Variable Models

Discriminant Analysis

Mixed Modeling

Penalized Regression

Quantile Regression

Non-parametric Regression

Principal Components Regression

Support Vector Machines

Unsupervised Learning

Latent Variable Models

Principal Components Regression

DATA
Which predictors to include?

Usually faced with problem of selecting subset of $p$ possible predictors to include in model.

- Have to balance conflicting objectives
  - Predictive Accuracy versus Model Parsimony
- Ideal: determine single best subset of predictors
  - But no single definition of “best”
- Different algorithms will produce different "best" subsets
- Problems magnified by correlation among predictors
When model not pre-specified (like in your projects),
Get to know your data and your problem!

### Understanding the problem

**Clarify your scientific question**
- What do you want to know?
- Why?

### Some issues to consider
- How did you select your subjects?
- How were the data collected?
- Are observations independent?
- Potential sources of confounding

### Getting to know your data

#### Response Variables
- Distribution
- Associations with predictors
- Outliers

#### Predictors
- Distributions
- Relationships with other predictors
- Outliers
- Collinearity?
- Small categories?

#### Some issues to consider
- Missing data?
- Evidence of “strange” values
Prostate Cancer Example

- Prostate-Specific Antigen (PSA) is a protein produced by the prostate
  - Values change over time
  - Increase in PSA is a potential sign of prostate cancer

- Consider a dataset consisting of 4544 men newly-diagnosed with prostate cancer

- Measure their PSA at diagnosis along with a lot of other variables

- **Goal:** Identify factors related to PSA levels at prostate cancer diagnosis.

- Why are we studying this?
  - Baseline PSA levels are related to prognosis in prostate cancer patients
  - It is a convenient example
  - You will have a better reason

## Exploratory Analysis

- Quick look at our data

<table>
<thead>
<tr>
<th>comorbidity</th>
<th>pni</th>
<th>gleason</th>
<th>age_decade</th>
<th>radiation</th>
<th>stage</th>
<th>PSA</th>
<th>caucasian</th>
<th>glandvol</th>
<th>txyear</th>
<th>group</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2599</td>
<td>Min.</td>
<td>6</td>
<td>2038</td>
<td>Min.</td>
<td>4.00</td>
<td>Min.</td>
<td>0.0000</td>
<td>1</td>
<td>1:1205</td>
</tr>
<tr>
<td>1</td>
<td>527</td>
<td>1st Qu.</td>
<td>7</td>
<td>1513</td>
<td>1st Qu.</td>
<td>5.600</td>
<td>1st Qu.</td>
<td>0.0000</td>
<td>2</td>
<td>2:1578</td>
</tr>
<tr>
<td>2</td>
<td>354</td>
<td>Median</td>
<td>7.5</td>
<td>521</td>
<td>Median</td>
<td>6.100</td>
<td>Median</td>
<td>0.0000</td>
<td>3</td>
<td>3:1761</td>
</tr>
<tr>
<td>3+</td>
<td>136</td>
<td>Mean</td>
<td>8</td>
<td>245</td>
<td>Mean</td>
<td>6.122</td>
<td>Mean</td>
<td>0.1706</td>
<td>NA's</td>
<td>928</td>
</tr>
<tr>
<td>NA's: 928</td>
<td></td>
<td>3rd Qu.</td>
<td>9</td>
<td>205</td>
<td>3rd Qu.</td>
<td>6.700</td>
<td>3rd Qu.</td>
<td>0.0000</td>
<td>Max.</td>
<td>1.0000</td>
</tr>
<tr>
<td>Max.</td>
<td>1.000</td>
<td>NA's</td>
<td>22</td>
<td>8.328</td>
<td>Max.</td>
<td>1.000</td>
<td>Max</td>
<td>1.0000</td>
<td>NA's</td>
<td>94</td>
</tr>
<tr>
<td>NA's: 94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stage</td>
<td>PSA</td>
<td>caucasian</td>
<td>glandvol</td>
<td>txyear</td>
<td>group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>3117</td>
<td>Min.</td>
<td>0.100</td>
<td>Min.</td>
<td>0.0000</td>
<td>4.00</td>
<td>Group</td>
<td>1:1205</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2</td>
<td>1354</td>
<td>1st Qu.</td>
<td>4.400</td>
<td>1st Qu.</td>
<td>1.0000</td>
<td>30.00</td>
<td>Group</td>
<td>2:1578</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T3</td>
<td>60</td>
<td>Median</td>
<td>6.100</td>
<td>Median</td>
<td>1.0000</td>
<td>40.00</td>
<td>Group</td>
<td>3:1761</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA's: 13</td>
<td></td>
<td>Mean</td>
<td>8.403</td>
<td>Mean</td>
<td>0.9228</td>
<td>41.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3rd Qu.</td>
<td>9.000</td>
<td>3rd Qu.</td>
<td>1.0000</td>
<td>49.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Max.</td>
<td>219.000</td>
<td>Max.</td>
<td>1.0000</td>
<td>265.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>NA's</td>
<td>4</td>
<td>NA's</td>
<td>905</td>
<td>1574</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Exploratory Analysis

• Quick look at our data

<table>
<thead>
<tr>
<th>comorbidity</th>
<th>pni</th>
<th>gleason</th>
<th>age_decade</th>
<th>radiation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2599 Min. : 0.000 6 : 2038 Min. : 4.000 Min. : 0.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>527 1st Qu.: 0.000 7 : 1513 1st Qu.: 5.600 1st Qu.: 0.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>354 Median : 0.000 7.5 : 521 Median : 6.100 Median : 0.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3+</td>
<td>136 Mean : 0.249 8 : 245 Mean : 6.122 Mean : 0.1706</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA's: 928</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd Qu.: 0.000 9 : 205 3rd Qu.: 6.700 3rd Qu.: 0.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max. : 1.000 NA's: 22 Max. : 8.328 Max. : 1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>stage</th>
<th>psa</th>
<th>caucasian</th>
<th>glandvol</th>
<th>txyear</th>
<th>group</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>3117 Min. : 0.100 Min. : 0.0000 Min. : 4.000 Group 1:1205</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2</td>
<td>1354 1st Qu.: 4.400 1st Qu.: 1.0000 1st Qu.: 30.00 Group 2:1578</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T3</td>
<td>60 Median : 6.100 Median : 1.0000 Median : 40.00 Group 3:1761</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA's: 13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean : 8.403 Mean : 0.9228 Mean : 41.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd Qu.: 9.000 3rd Qu.: 1.0000 3rd Qu.: 49.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max. : 219.000 Max. : 1.0000 Max. : 265.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA's :4 NA's :905 NA's :1574</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• We will ignore these for now (complete case analysis)
Baseline PSA

Histogram of PSA

Histogram of log(PSA)

Boxplot of log(PSA)

PSA Trend over Time

Years Since Start of Study
Quantile-Quantile Plots

- Is log-PSA “normal-ish?”

Normal Q-Q Plot of log(PSA)

https://kanbanize.com/blog/normal-gaussian-distribution-over-cycle-time/
Associations with Predictors

For example,

Boxplots of $\log(\text{PSA})$ by Gleason Score

Plot of $\log(\text{PSA})$ by Age

Plot of $\log(\text{PSA})$ by Gland Volume
## Linear Regression Model Fit

- Starting point: propose a “reasonable” model

**Linear Regression of log(PSA):**

| Coefficients: | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------|----------|------------|---------|----------|
| (Intercept)   | 1.709698 | 0.137941   | 12.394  | < 2e-16  *** |
| gleason7      | 0.159534 | 0.032033   | 4.980   | 6.87e-07 *** |
| gleason7.5    | 0.330396 | 0.045686   | 7.232   | 6.64e-13 *** |
| gleason8      | 0.438873 | 0.063317   | 6.931   | 5.53e-12 *** |
| gleason9      | 0.715239 | 0.066903   | 10.691  | < 2e-16  *** |
| age_decade    | 0.006566 | 0.018936   | 0.347   | 0.728811 |
| radiation     | 0.202324 | 0.036616   | 5.526   | 3.69e-08 *** |
| stageT2       | -0.114469| 0.032287   | -3.545  | 0.000401 *** |
| stageT3       | 0.104128 | 0.142540   | 0.731   | 0.465153 |
| caucasian     | -0.145573| 0.047275   | -3.079  | 0.002102 ** |
| pni           | 0.115990 | 0.033280   | 3.485   | 0.000502 *** |
| comorbidity1  | -0.056721| 0.039098   | -1.451  | 0.146999 |
| comorbidity2  | -0.068202| 0.045458   | -1.500  | 0.133677 |
| comorbidity3+ | -0.154585| 0.069095   | -2.237  | 0.025373 * |
| glandvol      | 0.005832 | 0.000712   | 8.191   | 4.45e-16 *** |
| txyearGroup2  | -0.180386| 0.074366   | -2.426  | 0.015364 * |
| txyearGroup3  | -0.311743| 0.075141   | -4.149  | 3.48e-05 *** |

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

- Are the model assumptions reasonably met?
  - Residual diagnostics
  - Knowledge about problem

- Are some subjects particularly “influential”?
  - Leverage, Cook’s D

- Multicollinearity?
  - Variance Inflation Factors (VIF)
  - Partial correlations

- Which covariates should I include and how?
  - Variable selection
  - Knowledge about problem
Evaluating Standardized Residuals

- **Linearity** and **constant variance** (homoscedasticity)

Might improve **non-constant variance** and **linearity** issues by replacing gland volume with log(gland volume) in the model.
Assessing Distributional Assumptions: QQ Plots

- Studentized residuals (“deleted” residuals standardized by their estimated standard errors) should be roughly t-distributed.

Does removing some outliers make sense?
- Normality “least important” of the assumptions (tail of distribution)
Evaluating Multi-Collinearity

- Strongly correlated predictors → Inflated standard errors of parameters
- Compare standard errors to theoretical minimum standard errors
- The variance inflation factor for the $k^{th}$ predictor is

$$VIF_k = \frac{1}{1 - R_k^2}$$

where $R_k^2$ is the $R^2$ value for a regression of the $k^{th}$ predictor on other predictors

Rule of Thumb:
- VIF = 1: No correlation
- VIF > 4: Investigate
- VIF > 10: Serious issues
And iterate….

- Fixing one problem may make another problem more visible
Comparing Models

• Suppose you have multiple different models to choose from. How to decide?

• Lots of different methods to compare
  • Based on your analytical goal
  • Look at multiple different metrics

• Some popular model comparison methods
  • P-value based methods (Forward/backward/stepwise selection)
  • Adjusted R2, Likelihood Ratio testing
  • Information criteria (AIC/BIC)
  • Prediction/Cross-Validation
  • ROC/AUC Analysis (next time)
P-Value Based Model Selection

• For nested models

• Want to select best subset of covariates
• Suppose we compare all possible subsets
  • With 10 predictors, \(2^{10}\) or 1024 models to evaluate
  • Compare models using some metric
  • Computation can get tricky
• Leaps and Bounds
  • search through smaller model space
Backward Elimination

(1) Start with all $p$ predictors.

(2) Remove the least significant predictor with $p >$ pre-determined threshold $= \alpha^*$

(3) Re-fit model and go to step 2.

(4) Stop when $p$-values for all predictors retained in model are less than $\alpha^*$

- This threshold is typically not set at 5% but at 10-20%
- Popular alternative: use another metric such as AIC to choose what to remove
Forward Selection

(1) Start with intercept only model.

(2) For all potential predictors check p-values if they are added to the model, choose the one with lowest p-value (< $\alpha^*$).

(3) Continue until no new predictor can be added.

- Variables entered at earlier steps may lose significance as new predictors are added.
Stepwise Regression

- Each step a variable can be added or removed, bidirectional.
- This can be carried out in a number of ways.
- At each step of forward selection you check whether one or more predictors can be removed without increasing the residual sum of squares “too much”.

**Drawbacks for these methods**

- No guarantee of optimal model
- So much unaccounted-for multiple testing and the p-values are dubious at best
- No direct connection to the application context (prediction, estimation)
- Tends to overstate the effect of predictors retained in the model.
- Trouble with highly correlated predictors
- Sometimes predictors only significant in presence of other predictors
Application to PSA Modeling

- Methods may give slightly or very different model fits

| Coefficients:                                   | Estimate | Std. Error | t value | Pr(>|t|) |
|------------------------------------------------|----------|------------|---------|----------|
| (Intercept)                                    | 0.90254  | 0.15857    | 5.692   | 1.43e-08 *** |
| as.factor(gleason)7                            | 0.16368  | 0.03199    | 5.116   | 3.40e-07 *** |
| as.factor(gleason)7.5                          | 0.33155  | 0.04533    | 7.314   | 3.66e-13 *** |
| as.factor(gleason)8                            | 0.44210  | 0.06270    | 7.051   | 2.40e-12 *** |
| as.factor(gleason)9                            | 0.71246  | 0.06499    | 10.715  | < 2e-16 *** |
| radiation                                      | 0.21516  | 0.03493    | 6.159   | 8.73e-10 *** |
| stageT2                                        | -0.11482 | 0.03222    | -3.539  | 0.000411 *** |
| stageT3                                        | -0.10239 | 0.14242    | 0.719   | 0.472257 |
| caucasian                                      | -0.14878 | 0.04713    | -3.157  | 0.001617 ** |
| pni                                            | 0.11754  | 0.03325    | 3.535   | 0.000417 *** |
| comorbidity1                                   | -0.06052 | 0.03889    | -1.556  | 0.119829 |
| comorbidity2                                   | -0.06973 | 0.04494    | -1.552  | 0.120870 |
| comorbidity3+                                  | -0.15429 | 0.06872    | -2.245  | 0.024856 * |
| log(glandvol)                                   | 0.29972  | 0.03549    | 8.445   | < 2e-16 *** |
| txyeargroupGroup 2                             | -0.18167 | 0.07433    | -2.444  | 0.014606 * |
| txyeargroupGroup 3                             | -0.31940 | 0.07507    | -4.254  | 2.19e-05 *** |

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 1

<table>
<thead>
<tr>
<th>Backward Elimination</th>
</tr>
</thead>
</table>

| Coefficients:                                   | Estimate | Std. Error | t value | Pr(>|t|) |
|------------------------------------------------|----------|------------|---------|----------|
| (Intercept)                                    | 0.879012 | 0.177618   | 4.949   | 8.06e-07 *** |
| as.factor(gleason)7                            | 0.162981 | 0.032086   | 5.080   | 4.12e-07 *** |
| as.factor(gleason)7.5                          | 0.329916 | 0.045680   | 7.222   | 7.12e-13 *** |
| as.factor(gleason)8                            | 0.439521 | 0.063321   | 6.941   | 5.16e-12 *** |
| as.factor(gleason)9                            | 0.710471 | 0.068522   | 10.628  | < 2e-16 *** |
| age_decade                                     | 0.005579 | 0.018960   | 0.294   | 0.768581 |
| radiation                                      | 0.212048 | 0.036505   | 5.809   | 7.26e-09 *** |
| stageT2                                        | -0.114582| 0.032286   | -3.549  | 0.000395 *** |
| stageT3                                        | 0.103845 | 0.142538   | 0.729   | 0.466364 |
| caucasian                                      | -0.149883| 0.047286   | -3.170  | 0.001548 ** |
| pni                                            | 0.117144 | 0.033285   | 3.519   | 0.000442 *** |
| comorbidity1                                   | -0.061719| 0.039112   | -1.578  | 0.114720 |
| comorbidity2                                   | -0.071722| 0.045453   | -1.578  | 0.114732 |
| comorbidity3+                                  | -0.156372| 0.069096   | -2.363  | 0.023731 * |
| log(glandvol)                                   | 0.297492 | 0.036301   | 8.195   | 4.30e-16 *** |
| txyeargroupGroup 2                             | -0.181290| 0.074360   | -2.438  | 0.014850 * |
| txyeargroupGroup 3                             | -0.319092| 0.075097   | -4.249  | 2.24e-05 *** |

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 1
Adjusted $R^2$

$R^2 = \text{corr}(Y, \hat{Y})^2$ for linear regression

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Pseudo $R^2$ measures for logistic regression, Cox and Snell $R^2$, Nagelkerke $R^2$

$$R^2 = 1 - \left( \frac{L(0)}{L(\hat{\beta})} \right)^{2/n}$$

Likelihood of model with only intercept: $L(0)$
Likelihood evaluated at MLE: $L(\hat{\beta})$

These values increase for larger models. Will pick the larger model.

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}$$

Adjusted $R^2$ penalizes for larger models.
Larger adjusted $R^2$ is better!
Likelihood ratio testing

For nested models
Test whether some parameters can be set to zero:

(1) Fit both models: full model and reduced model
(2) Calculate likelihood using two estimated parameters

\[ LRT = -2 \log\text{lik}(\beta^{\hat{\text{reduced}}}) + 2 \log\text{lik}(\beta^{\hat{\text{full}}}) \]

\[ LRT \sim \text{Chi-squared with} \]
\[ df = \text{number of parameters being set to zero} \]
AIC/BIC

- Measure goodness of fit
- Akaike Information Criterion, Bayes Information Criteria (smaller is better)
  - $-2 \times \text{maximized log likelihood} + 2p$ : AIC
  - $-2 \times \text{maximized log likelihood} + p \log(n)$ : BIC
- For small data sets a correction is needed for AIC, namely $\text{AIC}_c$
- BIC gives more parsimonious models
- Often used to compare non-nested models
- Often good to use both and compare
Comparing Models for PSA Example

Compare

(1) full model (2) model only including treatment and Gleason
(3) full model + extra nonsense covariates

<table>
<thead>
<tr>
<th></th>
<th>Full Model</th>
<th>Reduced Model</th>
<th>Full Model + Extra</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictors</td>
<td>16</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.17</td>
<td>0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>AIC</td>
<td>3961.7</td>
<td>4079.8</td>
<td>3967.4</td>
</tr>
<tr>
<td>BIC</td>
<td>4063.5</td>
<td>4119.4</td>
<td>4086.2</td>
</tr>
<tr>
<td>MSE</td>
<td>0.377</td>
<td>0.401</td>
<td>0.378</td>
</tr>
</tbody>
</table>

LRT (Full vs. Reduced): $p < 0.001$   LRT (Extra vs. Full): $p = 0.96$
Prediction Measures

• Previously, we were evaluating how well the model fits our data
• Often, our goal is prediction!
  • Risk prediction models
  • Precision medicine (e.g. which treatment will be best for the patient)
  • Weather, stock market prices, etc.

• Model that fits our data best may not predict future/new data the best

Some approaches to evaluate prediction abilities
• PRESS
• Mallow’s Cp
• Cross-Validation
PRESS

**PRESS** (prediction sum of squares) \[ \text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2 \]

1. Remove the \(i\)th observation
2. Re-fit model, re-estimate parameters
3. Predict the \(i\)th observation with this model: \(\hat{y}_{i,-i}\)
4. Calculate the residual. Do it for each observation (or a random subset).

The model structure leading to smallest value of PRESS is preferred.
aka take-one-out cross validation

**Prediction R\(^2\)**

\[ R_{pred}^2 = 1 - \frac{\text{PRESS}}{\text{SSY}} \]

Measures ability to predict future responses
Mallow’s Cp

- Combines bias and variance of the predicted Y

\[
\Gamma_p = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left \{ [E(\widehat{Y}_i - Y_i)]^2 + Var(\widehat{Y}_i) \right \}
\]

\[
C_p \equiv \widehat{\Gamma}_p = p + \left \{ \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_{full}} - 1 \right \} (n - p)
\]

- Helps strike a balance between including
  - enough covariates to avoid underfitting
  - not too many that we over-fit the data
- Want values near p
- Cannot use to evaluate “full” model (Cp always = p)
Cross Validation/Data Splitting:

Data divided into two parts: **test data and training data.**

- **Training Data**
  - Exploratory Analysis
  - Model selection
  - Fitting the model

- **(Independent) Test Data**
  - Evaluating the model

- Quantify how well model predicts test data set

- Provides a more realistic estimate of the predictive power of a model

- Test data could be part of your main dataset or external dataset
K-fold Cross Validation

(1) Split the data into k subsets of equal size.
(2) Estimate/fit model based on all subsets except one.
(3) Use the left out subset to test your model by calculating a metric of your choice.
(4) Average the metrics across the subsets to get an estimate of the cross-validation error.
Some cross-validation metrics

\[ \text{Error}_i = y_i - \hat{y}_i \]

Mean Squared Prediction Error:

\[ \text{MSPE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]

Mean Absolute Percentage Error (MAPE):

\[ \text{MAPE} = \left( \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \right) \times 100 \]
Comparing Models for PSA Example

Compare

1. full model
2. model only including treatment and Gleason
3. full model + extra nonsense covariates

<table>
<thead>
<tr>
<th></th>
<th>Full Model</th>
<th>Reduced Model</th>
<th>Full Model + Extra</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictors</td>
<td>16</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>PRESS</td>
<td>807.8</td>
<td>852.1</td>
<td>809.9</td>
</tr>
<tr>
<td>Mallow's Cp</td>
<td>13.2</td>
<td>134.78</td>
<td>-</td>
</tr>
<tr>
<td>10-Fold CV MSPE</td>
<td>0.382</td>
<td>0.403</td>
<td>0.383</td>
</tr>
</tbody>
</table>
General Guidelines

• Numerical criteria are useful, but don't rely too heavily on them
  • A lot of model selection is judgment calls and balanced opposing forces

• All models are wrong, some less wrong.

• Be guided by background knowledge of relationships whenever possible
  • Use information from the data AND your knowledge of the problem
  • Model may fit well but unmeasured confounding/selection biases could create problems

• Follow Occam’s Razor principle: beauty in simplicity, parsimony, succinctness

• For GLMs, there are really two parts to model selection: link function & variable selection/modeling
Some Alternatives to “Standard” Regression Models
Penalized Regression Models
An Alternative Method: Penalization

• Rather than directly choosing a subset of predictors to include in the model, can use **penalization** methods
  • Involve fitting full regression model with a penalty term
  • Penalizes more complicated models
  • Add some bias in exchange for smaller standard errors

\[
\begin{align*}
\text{No penalization} & \quad l(\beta) \\
\text{LASSO} & \quad l(\beta) + \lambda \sum_{k=1}^{p} |\beta_k| \\
\text{Ridge} & \quad l(\beta) + \lambda \sum_{k=1}^{p} \beta_k^2 \\
\text{Elastic Net} & \quad l(\beta) + \lambda \sum_{k=1}^{p} \beta_k^2 + (1 - \lambda) \sum_{k=1}^{p} |\beta_k|
\end{align*}
\]

(LASSO = Least Absolute Shrinkage and Selection Operator)
Tuning Parameter

• These methods all involve a tuning parameter, which controls how much you penalize.

• When the tuning parameter = 0, you get linear regression (ridge and LASSO)

• When the tuning parameter increases, parameters shrink toward zero

• Bias increases and variance decreases as the tuning parameter increases.

• You will **center** and **scale** the predictors before doing applying penalization
LASSO shrinks parameters exactly to zero

Can touch the contour ellipse for the first time at a corner of the square, corresponding to a zero coefficient.

In ridge there are no corners for the contour to hit, zero solutions will rarely result.
Ridge vs. LASSO in PSA Example

- Ridge versus LASSO in a sample dataset: note coefficients go to zero as lambda increases.
Choosing a Tuning Parameter (PSA Example)
## Comparing Betas (PSA Example)

<table>
<thead>
<tr>
<th>LASSO</th>
<th>Ridge</th>
<th>Standard GLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>(Intercept)</td>
<td>(Intercept)</td>
</tr>
<tr>
<td>1.58726</td>
<td>1.51606</td>
<td>1.709698</td>
</tr>
<tr>
<td>comorbidity1</td>
<td>comorbidity1</td>
<td>comorbidity1</td>
</tr>
<tr>
<td>.</td>
<td>0.00243</td>
<td>-0.056721</td>
</tr>
<tr>
<td>comorbidity2</td>
<td>comorbidity2</td>
<td>comorbidity2</td>
</tr>
<tr>
<td>.</td>
<td>-0.01718</td>
<td>-0.068203</td>
</tr>
<tr>
<td>comorbidity3+</td>
<td>comorbidity3+</td>
<td>comorbidity3+</td>
</tr>
<tr>
<td>.</td>
<td>-0.03890</td>
<td>-0.154585</td>
</tr>
<tr>
<td>pni</td>
<td>pni</td>
<td>pni</td>
</tr>
<tr>
<td>0.08993</td>
<td>0.08160</td>
<td>0.115990</td>
</tr>
<tr>
<td>gleason7</td>
<td>gleason7</td>
<td>gleason7</td>
</tr>
<tr>
<td>0.12119</td>
<td>0.01154</td>
<td>0.159534</td>
</tr>
<tr>
<td>gleason7.5</td>
<td>gleason7.5</td>
<td>gleason7.5</td>
</tr>
<tr>
<td>0.16885</td>
<td>0.09229</td>
<td>0.330396</td>
</tr>
<tr>
<td>gleason8</td>
<td>gleason8</td>
<td>gleason8</td>
</tr>
<tr>
<td>0.40656</td>
<td>0.13298</td>
<td>0.438873</td>
</tr>
<tr>
<td>age_decade</td>
<td>age_decade</td>
<td>age_decade</td>
</tr>
<tr>
<td>0.19946</td>
<td>0.02981</td>
<td>0.006566</td>
</tr>
<tr>
<td>radiation</td>
<td>radiation</td>
<td>radiation</td>
</tr>
<tr>
<td>-0.03228</td>
<td>0.12372</td>
<td>0.202324</td>
</tr>
<tr>
<td>stageT2</td>
<td>stageT2</td>
<td>stageT2</td>
</tr>
<tr>
<td>.</td>
<td>-0.01429</td>
<td>-0.114469</td>
</tr>
<tr>
<td>stageT3</td>
<td>stageT3</td>
<td>stageT3</td>
</tr>
<tr>
<td>.</td>
<td>0.17870</td>
<td>0.104128</td>
</tr>
<tr>
<td>caucasian</td>
<td>caucasian</td>
<td>caucasian</td>
</tr>
<tr>
<td>0.00396</td>
<td>-0.06523</td>
<td>-0.145573</td>
</tr>
<tr>
<td>glandvol</td>
<td>glandvol</td>
<td>glandvol</td>
</tr>
<tr>
<td>0.02476</td>
<td>0.00241</td>
<td>0.005832</td>
</tr>
<tr>
<td>txyeargroupGroup 2</td>
<td>txyeargroupGroup 2</td>
<td>txyeargroupGroup 2</td>
</tr>
<tr>
<td>0.06255</td>
<td>-0.04979</td>
<td>-0.180386</td>
</tr>
<tr>
<td>txyeargroupGroup 3</td>
<td>txyeargroupGroup 3</td>
<td>txyeargroupGroup 3</td>
</tr>
<tr>
<td>.</td>
<td>-0.311743</td>
<td>-0.311743</td>
</tr>
</tbody>
</table>
More on Elastic Net (Zou and Hastie, 2005)

- LASSO does not do very well for a correlated set of predictors and when p much larger than n.

- If there is a group of predictors with high pairwise correlation, LASSO tends to select only one from the group and does not care which one it is.

- Prediction performance of LASSO not satisfactory with highly correlated set of predictors, and elastic net dominated by ridge.

\[ \hat{\beta}_{ENET} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \right\} \]

- Ridge and LASSO are special cases
Elastic Net

- Combination of LASSO and Ridge penalties

\[ \hat{\beta}_{\text{ENET}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij} \beta_j)^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \right\} \]

- ENET beats LASSO in presence of collinearity in terms of prediction error
- Produces larger models than LASSO
- Produces sparse models with good prediction accuracy.
Estimating Standard Errors

• Penalization methods give coefficients but not standard errors

• Often, people will choose variables by LASSO and then go back to usual regression for inference. This is **WRONG**.

• Inference post-selection is hard for these penalization methods.

• There are methods in the literature for doing this
  • Based on asymptotic results
  • Based on bootstrap methods
Classification and Regression Trees (CART)
Machine Learning

- Completely different in flavor than classical parametric statistical inference
- Often borrows ideas from computer science and engineering
  - uncertainty is de-emphasized
  - more algorithmic than stochastic
- Learn from the data as opposed to cast the data into a structured model
- Goal is often prediction of new data
Supervised Learning Set-up

• Output measurement $Y$ (also called class label, response, dependent variable, target).

• Vector of $p$ input measurements $X$ (aka predictors, covariates, regressors)

• We have training data $(y_1, x_1), \ldots, (y_n, x_n)$. These are observations (instances) of these measurements.

• On the basis of the training data we would like to
  • Accurately predict unseen test cases
  • Understand which inputs affect the output and how
  • Assess the quality of our predictions and inferences
General Tree-based methods

• **Prediction**, classification and assessment of variable importance are critical questions in statistical inference.

• Recursive partitioning
  feature space (e.g. space spanned by all predictors) is split into regions containing observations with similar response values.

**A Simple Regression Tree Example:**

(1) Separates data into $X>0$, $X\leq 0$
(2) Separates data into $X>5$, $0<X\leq 5$
Classification vs. Regression

• **Classification Tree:** When Y (outcome) is binary/unordered categorical
  
  Want to assign each subject to a category Y=k
  
  Terminal nodes result in classifications
  
  Error assessment through misclassification cost.

• **Regression Tree:** Y is continuous or ordered discrete values.
  
  Prediction error measured by squared or relative absolute difference between observed and predicted values.
• Classification Tree: 3 class labels, two predictors, partition X space (feature space) into rectangular sets

FIGURE 1 | Partitions (left) and decision tree structure (right) for a classification tree model with three classes labeled 1, 2, and 3. At each intermediate node, a case goes to the left child node if and only if the condition is satisfied. The predicted class is given beneath each leaf node.

Loh et al, 2011
• Regression Tree: Break up covariate space based on outcome mean

- **Mean Y = 5.9**
  - n = 602 100%

- **Mean Y = 5.9**
  - n = 595 99%

- **Mean Y = 5.2**
  - n = 314 52%
  - educ >= 0.5
    - < 0.5
      - sbp < 144
        - trig >= 106
          - < 106
            - sodium >= 2312
              - < 2312
                - Mean Y = 5.9
                  - n = 284 4%
                - Mean Y = 10
                  - n = 7 1%
              - Mean Y = 5.6
                - n = 202 34%
                - trig >= 176
                  - < 176
                    - Mean Y = 4.5
                      - n = 55 9%
                    - Mean Y = 4
                      - n = 147 24%
                  - > 176
                    - Mean Y = 5.8
                      - n = 12 2%
                    - Mean Y = 12
                      - n = 9 1%
    - >= 106
      - Mean Y = 6.8
        - n = 281 47%

- **Mean Y = 5.3**
  - n = 99 16%

- **Mean Y = 6.6**
  - n = 281 47%

- **Mean Y = 6**
  - n = 9 1%

- **Mean Y = 4**
  - n = 5 1%

- **Mean Y = 10**
  - n = 33 5%

- **Mean Y = 5.8**
  - n = 12 2%

- **Mean Y = 12**
  - n = 9 1%

- **Mean Y = 4.5**
  - n = 55 9%

- **Mean Y = 10**
  - n = 7 1%

- **Mean Y = 5.8**
  - n = 12 2%

- **Mean Y = 9.9**
  - n = 9 1%

- **Mean Y = 5.9**
  - n = 314 52%

- **Mean Y = 5.9**
  - n = 595 99%

- **Mean Y = 9.6**

- **Mean Y = 6.6**
  - n = 281 47%

- **Mean Y = 164**

- **Mean Y = 6.5**
  - n = 274 46%

- **Mean Y = 9**
  - n = 9 1%

- **Mean Y = 10**
  - n = 17 3%

- **Mean Y = 8.5**
  - n = 21 3%

- **Mean Y = 8.9**
  - n = 26 4%

- **Mean Y = 7**
  - n = 17 3%

- **Mean Y = 7**
  - n = 17 3%

- **Mean Y = 9**
  - n = 17 3%

- **Mean Y = 10**
  - n = 17 3%

- **Mean Y = 14**
  - n = 8 1%

- **Mean Y = 13**
  - n = 7 1%

- **Mean Y = 15**
  - n = 1 0.3%

- **Mean Y = 16**
  - n = 1 0.3%

- **Mean Y = 17**
  - n = 1 0.3%

- **Mean Y = 18**
  - n = 1 0.3%

- **Mean Y = 19**
  - n = 1 0.3%
Classification and Regression Trees (CART)

• Breiman, Friedman, Olshen and Stone (1984), proposed this 30 years ago.

• Feature space recursively partitioned into rectangular areas such that observations with similar responses are grouped together.

• When you stop, you provide a common prediction for Y for subjects in the same group.
Distinction from GLMs (e.g. linear regression)

• Non-linear and even non-monotone associations are identified
• Can capture complex variable relationships

Why use trees?
• often yield relatively simple and easy to comprehend models.
• frequently more accurate than parametric tools.
• method can sift through any number of variables.
• can separate relevant from irrelevant predictors.
• no/fewer prior assumptions on data structure
• “pretty” pictures can give insight into relative importance of variables
The rise of CART

Tons of publications, use in biomedical applications

Why?

• Availability of huge data sets requiring analysis
• Need to automate or accelerate and improve analysis process
• Rising interest in data mining
• New software and documentation make techniques accessible to researchers
• Next generation CART techniques appear to be even better than former
Growing a Tree

1. Fix a predictor in X
2. Fix a cut-point for the predictor, \( c \)
3. Compute measure of the quality of the split
   - E.g. the impurity (homogeneity) of the daughter nodes/leaves
   - E.g. test statistic for difference between daughter nodes
4. Repeat for all cut-points and all predictors
5. Choose best split using some metric
   - E.g. producing the best in terms of impurity, largest test statistic,
6. Repeat for each daughter node

Grow very large tree (believed to overfit the data)
Keep growing until you have nodes of a certain size or impurity
Pruning the Tree

- Pruning
  - Take the maximal tree (radically overfit).
  - Prune branches from the large tree
  - Pruning at a node means deleting all of its descendants/leaves

- Challenge is how to prune
  - which branch to cut?
  - Point is to find a subtree that is most “predictive” of the outcome and least vulnerable to the noise in the data

- Cost-complexity pruning
  - External validation, internal cross-validation
Drawbacks of CART

Drawbacks

- **MODEST ACCURACY**
  - current methods, such as ensemble classifiers often have 30% lower error rates than CART.

- **INSTABILITY**
  - if we change the data a little, tree picture can change a lot

Some alternatives (ensemble methods)

- Bagging
- Boosting
Bagging and Boosting

Obtaining Bootstrap Sample:
- Sample with replacement from training data \((y_1, x_1), \ldots, (y_n, x_n)\) to get dataset of same size
- Do it \(B\) times to get \(B\) bootstrap samples of data

**Bagging** (Breiman 1996): Fit many large trees to bootstrap resampled versions of the training data, and classify by majority vote.

**Boosting** (Freund & Schapire 1996): Fit many large or small trees to reweighted versions of the training data. Classify by weighted majority vote.
- Weights related to prediction error for subject
Visualization of Bagging and Boosting

- **single**: complete training set
- **bagging**: random sampling with replacement
- **boosting**: random sampling with replacement over weighted data

Visualization of Bagging and Boosting

Generally, boosting > bagging > single tree

https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/
Random Forests and Out of Bag Prediction

- For each tree, generate a bootstrap sample of the data.
- The bootstrap sample is used to grow the tree.
- The remaining data are said to be “out-of-bag”
- The out-of-bag (oob) data can serve as a test set for the tree grown on the bootstrap sample.

- For each subject, get classification in out-of-bag trees.
- For each case, the RF prediction is either correct or incorrect
  - Average over the subjects within each class to get a classwise oob error rate
  - Average over all subjects to get an overall oob error rate
Prostate Example

- Classification tree for I(treated with radiation)

fit = rpart(radiation~AGE_CAT+log(psa)+stage+caucasian+pni+comorbidity+
log(glandvol) + txyeargroup, data = data, method = 'class', control = list(cp = 0.001))

Cp = Complexity parameter

X-val Relative Error: Measure of relative prediction error from cross-validation
Prostate Example

Who is classified as being treated with radiation?

- RADIATION
- SURGERY

Note: We might get more “sensible” groupings by considering an ensemble method like bagging/boosting.
Prostate Example

• Regression tree for log(PSA)

Pruned Tree

- radiatio < 0.5
- pni < 0.5
- stage = T1,T2

fit = rpart(log(PSA)~AGE_CAT+radiation+stage+caucasian+pni+comorbidity+
log(glandvol) + txyeargroup, data = data, method = ‘anova’, control = list(cp = 0.001))
An additional topic:
Evaluating Risk Prediction Models with ROC Curves
Sensitivity and Specificity

- Sensitivity = \( \frac{a}{a+b} = P(\text{Test Positive | Diseased}) \)
- Specificity = \( \frac{d}{c+d} = P(\text{Test Negative | Not Diseased}) \)
- Can also estimate for continuous risk predictors (tests)
Sensitivity/Specificity for Continuous Scores

- Want to know sensitivity/specificity of continuous score $X$ for disease status

- Consider different thresholds, $c$ where $X > c$ is a positive test
  \[
  \text{Specificity} = P(X < c \mid D = 0) \\
  \text{Sensitivity} = P(X > c \mid D = 1)
  \]

- Can estimate these quantities for different values of $c$
  - Gives curve of sensitivity and specificity values depending on $c$
ROC Curves

TP = Sensitivity vs. FP = 1-Specificity

Measure of Discrimination

Note: figure is misleading, “good” depends on your problem

AUC = Area under ROC Curve

Higher AUC = better discrimination
Prostate Data

- Model whether subjects assigned to radiation vs. surgery based on covariates

**Model 1**: All covariates in full model

**Model 2**: Gleason Only

```r
glm(formula = radiation ~ comorbidity + pni + gleason + age_decade + stage + caucasian + log(glandvol) + txyeargroup + log( PSA), family = "binomial", data = dataCC)
```
A “Real” Example using External Validation

- External validation better measure of model quality
- Compare online prediction calculators for 5-year survival for patients with Oropharyngeal cancer with observed UM data

AUC just using cancer stage: 0.70

Beesley et al. 2019 “Individualized Survival Prediction for Patients with Oropharyngeal Cancer in the Human Papillomavirus Era”
Exploring Calibration

- Calibration of online calculators with observed survival probabilities
Exploring Risk Stratification

- How well do the calculators stratify patients by risk?
Questions?