Model Uncertainty, Penalization, and Parsimony

Frank E Harrell Jr
Professor of Biostatistics and Statistics
Division of Biostatistics and Epidemiology
Department of Health Evaluation Sciences
University of Virginia School of Medicine
Charlottesville VA 22908 USA

fharrell@virginia.edu
http://hesweb1.med.virginia.edu/biostat

JOHNS HOPKINS UNIVERSITY
DEPARTMENT OF BIOSTATISTICS
17 JANUARY 2001

Outline

- Nature and complexity of models
- Problems with data–guided model selection
- Need for a monotonic model selection process
- Biased estimators (shrinkage)
- How does one choose a penalty?
- Can the data tell us the optimum penalty?
- Advantages of penalized estimation
- Achieving parsimony by approximating the “best” model
- Summary

Nature and Complexity of Models

- Response = anatomy + physiology + pathology + genetics + quality of medical care + compliance + medical decisions + society + environment + personal wealth + . . .
- **Model**: a *current* approximation to complex relationships
- Model/variable selection implies that there is some likelihood of a "true" model (some pre–specified variables have zero association with Y)
- Bayesian Information Criterion (BIC) "assumes that a true model exists and is low-dimensional"[6]
- Akaike Information Criterion (AIC) allows the complexity of the model to grow with the availability of information

Data–Guided Modeling Problems

- General: using data to guide modeling process, with no structure $\longrightarrow$overfitting, failure to validate, nonsense models
- If stepwise variable selection were invented today and submitted to a statistical journal the paper would be rejected
- Problem 1: data not capable of discerning some things (e.g., selecting from among collinear predictors)
- Problem 2: analysts have short memories
  - Run stepwise variable selection
  - Tear off last page of printout
  - Show this page to client, including R^2, standard errors
- Altman&Andersen[1]: If properly account for variability of variables selected (e.g., bootstrap),

length of confidence intervals of $\hat{S}(t|X)$ could be 60% longer than naive estimates

- Nearly unbiased estimate of σ^2 in OLS regression after stepwise: SSE / $(n-$ # **candidate** d.f.)
- Ye [37]: "generalized degrees of freedom" (GDF) for any "data mining" or model selection procedure based on least squares
  - Example: 20 candidate predictors, $n = 22$, forward stepwise: GDF=14.1
  - Example: CART, 10 candidate predictors, $n = 100$, 19 nodes: GDF=76

Monotonic Modeling Process

- Enter variables in order dictated by subject matter experts or cost
- Enter principal components in order of total variance explained (if can estimate component coefficients with little error!)
- Pre–specify full model and examine effects of increasing amount of penalization
- Any strategy controlled by a monotonic process (especially by a single parameter) will have
  - Less variability
  - Stopping rule easier to specify

Biased Estimation — Background

- Ex: clinical trial with 10 treatments, apparently best treatment is # 7
- Can test whether $\mu_7 = \mu_i$ for pre–specified i using Bonferroni inequality
- $\bar{Y}_7$ is a poor estimator of μ_7!
- Much lower MSE estimator obtained by shrinking all treatment means toward grand mean[12]
- By building in bias we protect outselves in how estimates will be used
- $\rightarrow$prognostication of very low or high risk patients improved by shrinkage; prediction for randomly chosen patient not improved

- Quadratic penalized likelihood (Verweij&van Houwelingen[32])
  $\log L - \frac{1}{2}\lambda\beta' P\beta$
- P is flexible
  - Can penalize categorical variables properly (ignore s.d.)
  - Can keep simple components (linear, additive) unpenalized
    - penalize nonlinear terms[15, 16],
    - interaction terms
    - terms representing departure from constant slopes assumptions (e.g., proportional odds)[19]
  - If sufficient number of terms of each type, can find best penalty for different term types
    - simple, nonlinear, linear interaction, nonlinear interaction

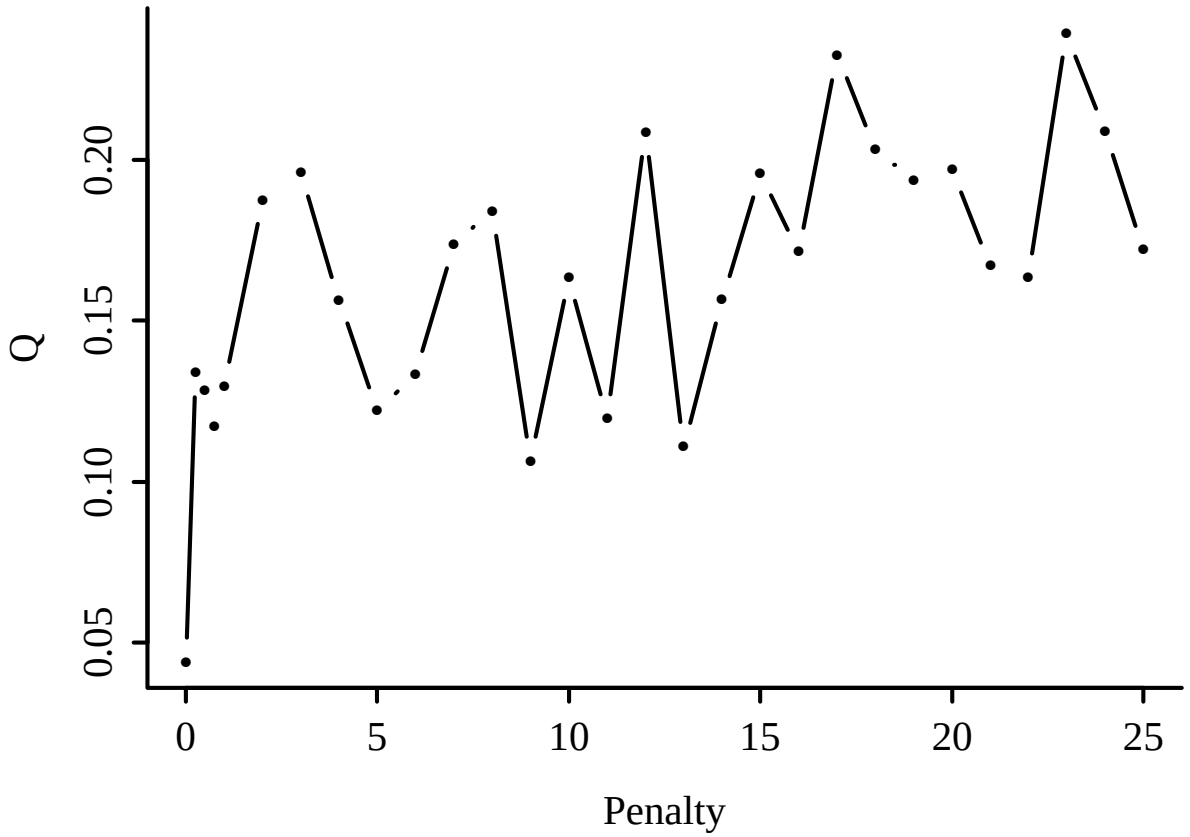
How to Choose λ?

- Specify d.f. to allow, solve for λ (later)
- Can't choose λ by naive 10–fold cross–validation
- Example: $n = 175$, binary logistic model, 92 events

  5 random normal predictors, linear with $\beta = 1, .5, .25, .125, 0$

  Expanded into restricted cublic splines with 4 knots; Total $p = 15$
- Cross–validated Q statistic (proportional to deviance; measures discrimination + calibration accuracy[a])

[a] logarithmic probability scoring rule



10–fold cross–validation of Q statistic in sample of $n = 175$ with 15 predictor d.f. Much smoother results could have been obtained by using the same splits for different values of the penalty.

- Based on corrected AIC (small sample correction for AIC)[21]
- Effective AIC = (likelihood ratio χ^2 for penalized model, but ignoring the penalty function) $-2\times$ effective d.f.
- Effective d.f.= trace[$I(\hat{\beta}^P)V(\hat{\beta}^P)$]
  $I =$ information matrix ignoring penalty function
  $V =$ inverse of information matrix including penalty[15]
- Need to demonstrate that there is minimal uncertainty about λ so that double bootstrap not needed;
  Limited experience repeating AIC trace for multiple re–samples is encouraging.
- Example: same training sample ($n = 175$), 10,000 observation validation sample

- If use BIC, optimum penalty nearly ∞

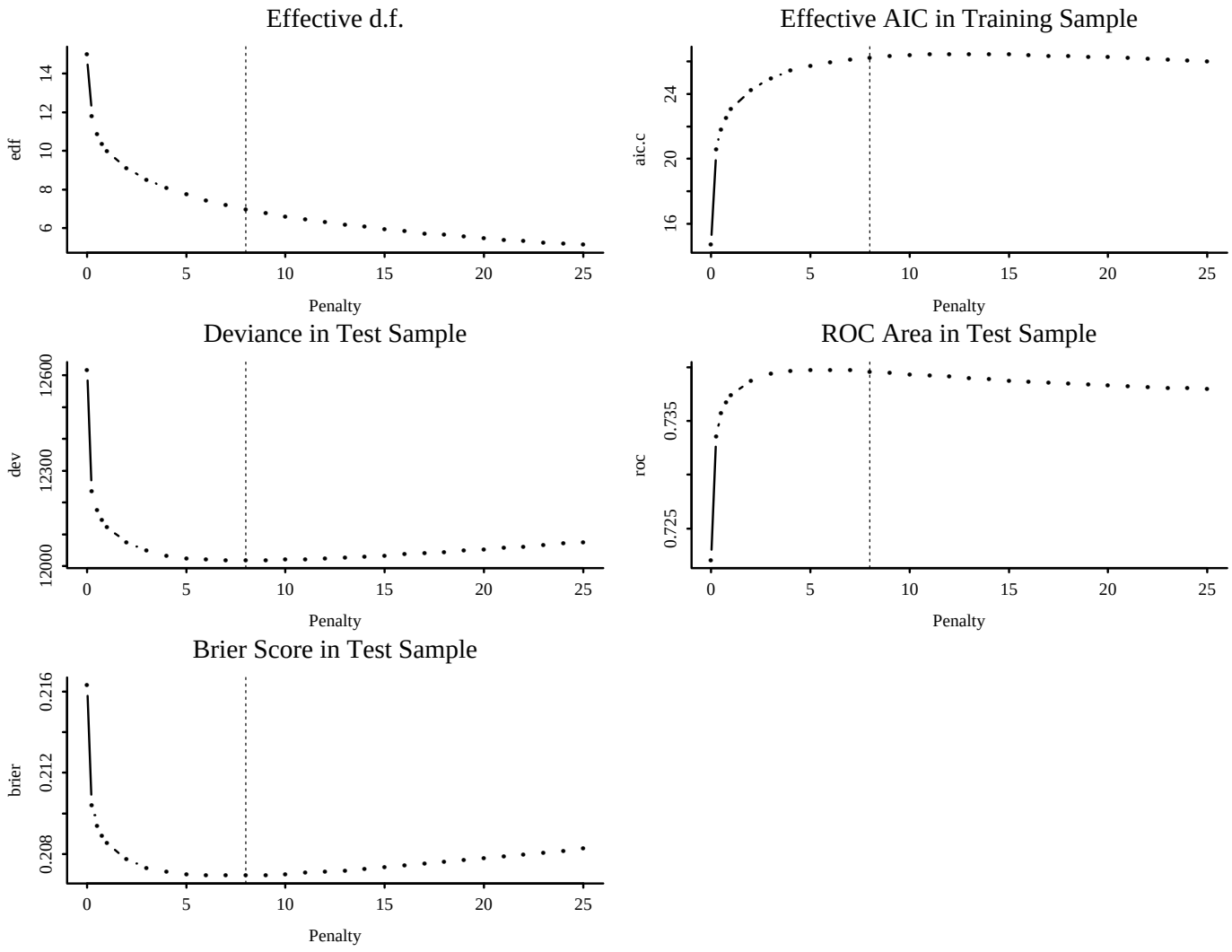
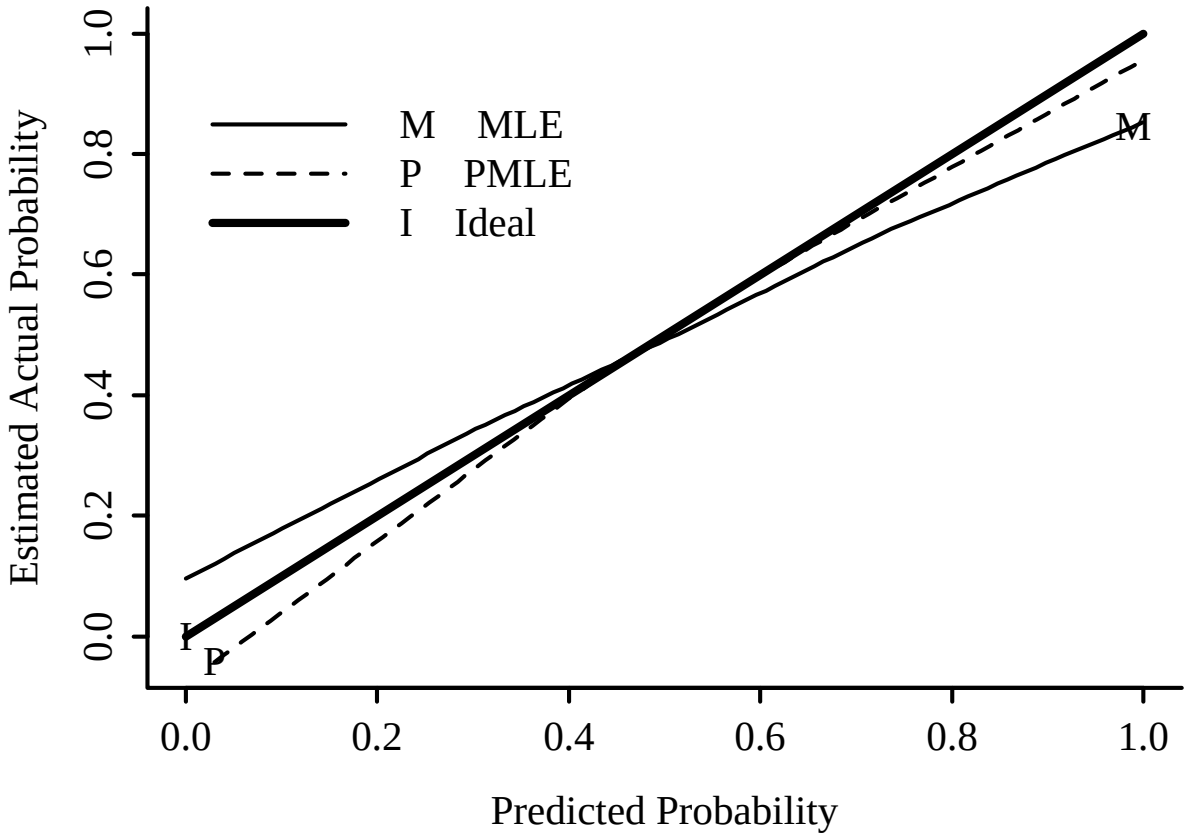


Figure 1: *Choosing best penality parameter by examining effective corrected AIC in $n = 175$ training sample. Upper left panel shows effective d.f. corresponding to each penalty. Dotted vertical lines are drawn at optimal penalty for $n = 10,000$ observation validation sample. Effective d.f. for penalty with best corrected effective AIC is 6.185 = # events / 14.9. D.f. for penalty with best validation deviance is 6.98, corresponding to an event ratio of 13.2. The Brier score[3] is the mean squared error in predicting binary responses with probabilities. It is a measure of calibration + discrimination.*

Shrinkage Improves Calibration Validation

- Previous example focused on $\uparrow$ discrimination
- Often get larger improvement in calibration accuracy
- Makes predictions more conservative for extreme subjects
- Calibration curve estimation: nonparametric regression on $(\hat{P}_i, Y_i)$[18]
- Useful index of calibration error:
  $\bar{E}$ =avg. $|\hat{P}_i - \hat{P}_i^c|$,
  $\hat{P}^c$ = calibrated $\hat{P}$
- In example, $\bar{E} = 0.051$ using MLE, 0.015 using PMLE



Nonparametric calibration plot in $n = 10,000$ observation test sample, for ordinary ($\bar{E} = 0.051$) and penalized MLE ($\bar{E} = 0.015$)

- Dropping insignificant terms to arrive at parsimonious model is an illusion
  - Low probability of selecting correct model if one exists
  - "Phantom" parameters still inflate variances (model uncertainty)
  - Parameters overestimated, F, χ^2 stats. destroyed
  - Subject matter knowledge ignored, interpretation messy
- Tibshirani's *lasso*[29] does penalization and variable selection[a] and has good performance; computationally difficult
- Breiman's *garrotte*[5] is an alternative

[a] The penalized likelihood requires that $\sum |\beta_i| < k$ for some choice of the penality parameter k; This is similar to using a weird Bayesian prior with a spike at $\beta = 0$.

- Another possibility is growing large trees using recursive partitioning
  - Different consumers could prune tree to different levels
  - Basic problem: usual regression trees have poor predictive discrimination when they don’t overfit
- Simplifying model during development is irrevocable
- Different consumers require differing degrees of simplicity
- Set of variables available when predictions are sought may not be the set envisioned by the analyst

Model Approximation — Methods

- "Gold standard" model is model with all pre–specified terms, with possible deletion of a block of variables having $P = 0.8$ for the pooled test
- This "gold standard" model is penalized to have maximum likely forecast accuracy
- Let $\hat{Z} = X\hat{\beta}$ be the linear predictor from this full penalized model (e.g., log odds, log hazard, expected value)
- Approximate $\hat{Z}$ from a set of candidate variables
  - Original set with original complexity
  - Original set, simplified (fewer knots, linear, fewer interactions)
  - Subset of original variables
  - Alternate "surrogate" variables

- Stepwise regression
- Recursive partitioning
- Both OK since $V[\hat{Z}|X] = 0$
- Stop simplifying when approximation error becomes unacceptable
  - R^2 in predicting $\hat{Z}$
  - Avg. absolute prediction error
  - 0.95 quantile of absolute pred. error
- Less uncertainty in $\hat{\hat{Z}}$ than in variable selection using Y
- But conditional on variables selected, in OLS coefficients of approx. model are identical to those of model re-fit against Y (see `http://hesweb1.med.virginia.edu/biostat/presentations/dilemmas.pdf`)

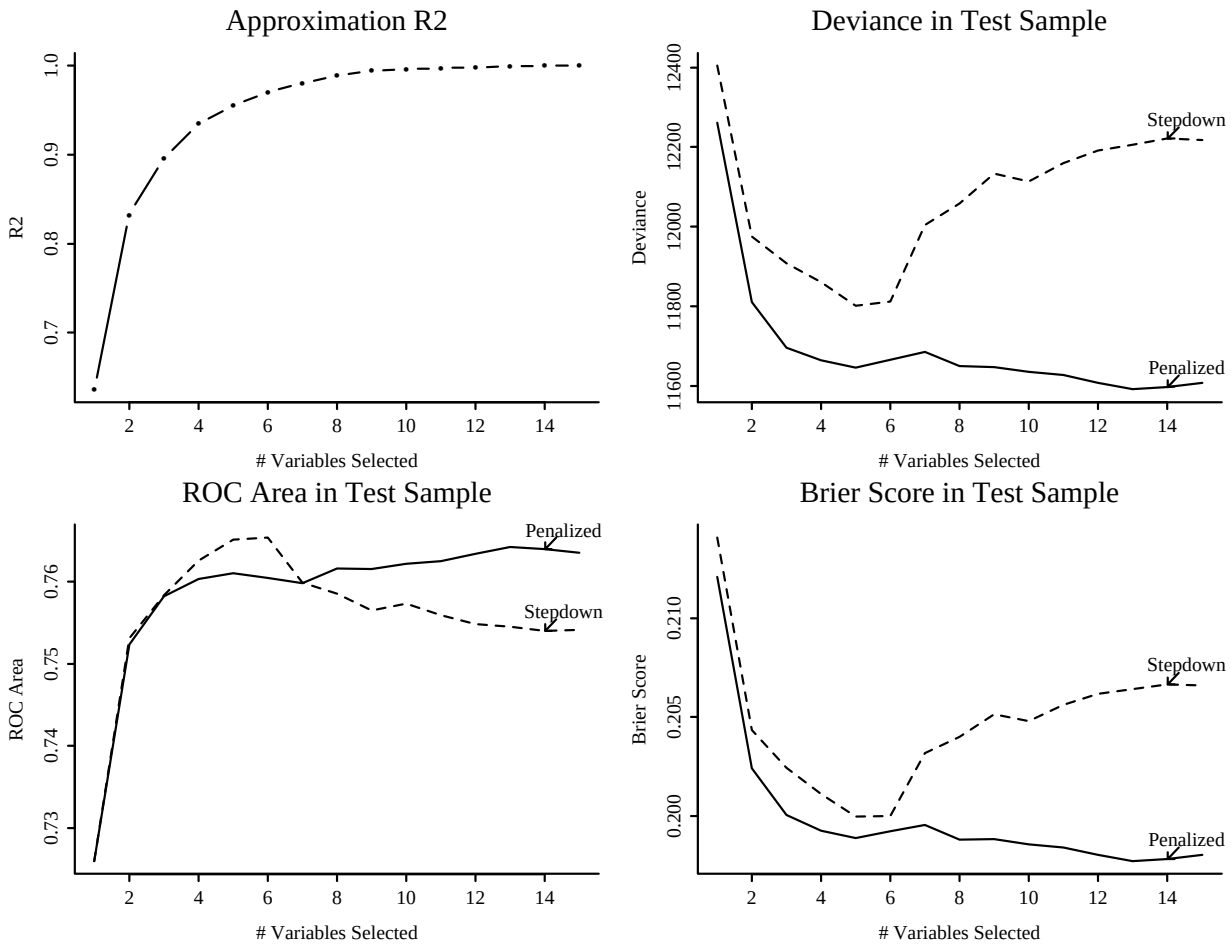
- Training sample: $n = 250,\ p = 15$ predictors $N(0, 1),\ \rho = 0.4$, 119 events
- $16\beta = 8\ \ 4\ \ 4\ \ 3\ \ 2\ \ 2\ \ 2\ \ 1\ \ 1\ \ .5\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0$
- Penalty with best AIC_c = 29.1, effective d.f.=7.06 = 119/16.9
- Backward stepdown against Y chose variables

  1 2 8 7 6 12 4 14 15 11 13 3 9 5 10

  (stop with 1 2 8 if use $\alpha = 0.05$)
- Backward stepdown against $\hat{Z}$ chose

  1 2 7 8 11 14 12 6 13 10 9 5 3 15 4

  (stop with desired complexity)
- Validation sample: $n = 10,000$



The upper left panel is the R^2 with which the model approximates the full penalized model. Other panels depict the forecast accuracy of two approaches in the 10,000 observation test sample. The backward stepdown procedure is a usual binary logistic stepwise variable selection using ordinary MLE. When all 15 variables are included this corresponds to the full unpenalized fit. The various approximations are to the full penalized model by predicting $\hat{Z}$ from a subset of X. All model components are linear and additive.

Summary

- Known simple models are uncommon in clinical biostatistics
- Simple models can be useful if derived sensibly
- Modeling process needs to be monotonic
- Biased estimators are better than unbiased ones
- Penalty parameters can be chosen from the data (effective AIC)
- Parsimony of various degrees can be achieved by approximating the “best” validating (penalized) model
- Completely empirical approaches to model selection are beset with problems of model uncertainty
- Shrinkage uses the full model; $\downarrow$ uncertainty

Abstract

Biostatistical modeling is often a contest between bias and variance. A good modeler tries to reduce bias in predicting patient responses by having the model incorporate all of the potentially relevant predictor variables and by adding additional terms to allow for non–linearity and non–additivity. Judging by the extent to which stepwise modeling and other data mining techniques are used, most practitioners seem to be very willing to let the data drive the modeling process, and then to pretend that the model was pre–specified in order to make statistical inferences using traditional variance formulas, etc. This process has the apparent advantage of development of "small" models that at face value seem to have parameter estimates with small variances. But overfitting is a common result of either data mining or of fitting complex pre–specified models, and variances of estimates, if computed correctly by accounting for model uncertainty, can become large; likewise measures of predictive discrimination shrink once one accounts for model searching. J Ye (JASA 93:120, 1998) has developed a very useful way to quantifying the degree of "data dredging" done in developing a model.

Pre–specifying the fullest reasonable model and using penalization (shrinkage) to downweight parameter estimates is a promising solution to the bias–variance trade–off, especially when more complex portions of the model receive greater shrinkage. See for example PJM Verweij and HC van Houwelingen[32] and FE Harrell *et al.*[19]. Of concern is the choice of the penalty parameter. Some simulated examples showing advantages of "effective AIC" over 10–fold cross–validation are presented.

When full but penalized model fits are used to reduce model uncertainty, complexity of the final model may be a hindrance to its clinical use. Harrell *et al.* (op cite) proposed that the final "gold standard" model be approximated to any desired accuracy using stepwise regression or recursive partitioning. In that way, an entire sequence of approximate models, inheriting the shrinkage of the full model, can be presented to clinical users having different needs. Some examples of this more linear approach to parsimony are presented.

References

[1] D. G. Altman and P. K. Andersen. Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine*, 8:771–783, 1989.

[2] P. K. Andersen, J. P. Klein, and M. Zhang. Testing for centre effects in multi-centre survival studies: A monte carlo comparison of fixed and random effects tests. *Statistics in Medicine*, 18:1489–1500, 1999.

[3] G. Blattenberger and F. Lad. Separating the Brier score into calibration and refinement components: A graphical exposition. *American Statistician*, 39:26–32, 1985.

[4] M. Blettner and W. Sauerbrei. Influence of model-building strategies on the results of a case-control study. *Statistics in Medicine*, 12:1325–1338, 1993.

[5] L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37:373–384, 1995.

[6] S. T. Buckland, K. P. Burnham, and N. H. Augustin. Model selection: An integral part of inference. *Biometrics*, 53:603–618, 1997.

[7] C. Chatfield. Model uncertainty, data mining and statistical inference (with discussion). *Journal of the Royal Statistical Society A*, 158:419–466, 1995.

[8] J. B. Copas. Regression, prediction and shrinkage (with discussion). *Journal of the Royal Statistical Society B*, 45:311–354, 1983.

[9] J. B. Copas. Cross-validation shrinkage of regression predictors. *Journal of the Royal Statistical Society B*, 49:175–183, 1987.

[10] D. Draper. Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society B*, 57:45–97, 1995.

[11] B. Efron and C. Morris. Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70:311–319, 1975.

[12] B. Efron and C. Morris. Stein's paradox in statistics. *Scientific American*, 236:119–127, 1977.

[13] J. J. Faraway. The cost of data analysis. *Journal of Computational and Graphical Statistics*, 1:213–229, 1992.

[14] L. S. Freedman and D. Pee. Return to a note on screening regression equations. *American Statistician*, 43:279–282, 1989.

[15] R. J. Gray. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87:942–951, 1992.

[16] R. J. Gray. Spline-based tests in survival analysis. *Biometrics*, 50:640–652, 1994.

[17] S. Greenland. Second-stage least squares versus penalized quasi-likelihood for fitting hierarchical models in epidemiologic analyses. *Statistics in Medicine*, 16:515–526, 1997.

[18] F. E. Harrell, K. L. Lee, and D. B. Mark. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15:361–387, 1996.

[19] F. E. Harrell, P. A. Margolis, S. Gove, K. E. Mason, E. K. Mulholland, D. Lehmann, L. Muhe, S. Gatchalian, and H. F. Eichenwald. 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. *Statistics in Medicine*, 17:909–944, 1998.

[20] T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *Applied Statistics*, 23:73–102, 1995.

[21] C. M. Hurvich and C. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76:297–307, 1989.

[22] le Cessie, S. and van Houwelingen, J. C. Ridge estimators in logistic regression. *Applied Statistics*, 41:191–201, 1992.

[23] J. K. Lindsey and B. Jones. Choosing among generalized linear models applied to medical data. *Statistics in Medicine*, 17:59–68, 1998.

[24] A. N. Phillips, S. G. Thompson, and S. J. Pocock. Prognostic scores for detecting a high risk group: Estimating the sensitivity when applied to new data. *Statistics in Medicine*, 9:1189–1198, 1990.

[25] P. Royston and D. G. Altman. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling (disc: P453-467). *ApplStat*, 43:429–453, 1994.

[26] D. J. Sargent and J. S. Hodges. A hierarchical model method for subgroup analysis of time-to-event data in the Cox regression setting. Presented at the Joint Statistical Meetings, Chicago, 1996.

[27] W. Sauerbrei and M. Schumacher. A bootstrap resampling procedure for model building: Application to the Cox regression model. *Statistics in Medicine*, 11:2093–2109, 1992.

[28] P. E. Scott and G. Campbell. Interpretation of subgroup analyses in medical device clinical trials. *Drug Information Journal*, 32:213–220, 1998.

[29] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58:267–288, 1996.

[30] R. Tibshirani. The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16:385–395, 1997.

[31] H. C. van Houwelingen and J. Thorogood. Construction, validation and updating of a prognostic model for kidney graft survival. *Statistics in Medicine*, 14:1999–2008, 1995.

[32] P. Verweij and H. C. van Houwelingen. Penalized likelihood in Cox regression. *Statistics in Medicine*, 13:2427–2436, 1994.

[33] P. J. M. Verweij and H. C. V. Houwelingen. Cross-validation in survival analysis. *Statistics in Medicine*, 12:2305–2314, 1993.

[34] P. J. M. Verweij and H. C. van Houwelingen. Time-dependent effects of fixed covariates in Cox regression. *Biometrics*, 51:1550–1556, 1995.

[35] C. T. Volinsky, D. Madigan, A. E. Raftery, and R. A. Kronmal. Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke. *Applied Statistics*, 46:433–448, 1997.

[36] Y. Wang and J. M. G. Taylor. Inference for smooth curves in longitudinal data with application to an aids clinical trial. *Statistics in Medicine*, 14:1205–1218, 1995.

[37] J. Ye. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93:120–131, 1998.

Thanks to Ewout Steyerberg of Erasmus University for comments which led to improvements in this presentation.

```
#This code requires the Hmisc and Design libraries
# See U. Virginia web page or lib.stat.cmu.edu
#See help file for lrm for more simulation/penalization

#Use 10-fold cross-validation to estimate predictive acc
#logistic models with various penalties
#
store()
options(digits=3)
set.seed(123)
n <- 175
nval <- 10000
nt <- n + nval
x1 <- rnorm(nt)
x2 <- rnorm(nt)
x3 <- rnorm(nt)
x4 <- rnorm(nt)
x5 <- rnorm(nt)
logit <- x1+.5*x2+.25*x3+.125*x4
y  <- ifelse(runif(nt) < plogis(logit), 1, 0)
f  <- lrm(y ~ rcs(x1,4)+rcs(x2,4)+rcs(x3,4)+rcs(x4,4)+rc
               x=T,y=T, subset=1:n)

new.data <- data.frame(x1,x2,x3,x4,x5,y)[-(1:n),]
Xnew <- predict(f, new.data, type="x", incl.non.slopes=F
Ynew <- new.data$y

penalties <- c(0,.25,.5,.75,1:25)

pt <- pentrace(f, penalties)
# Use pentrace(f, 40, method='optimize') to find best pe
#  (40 = starting value)
aic.c <- pt$results.all[,'aic.c']
edf   <- pt$results.all[,'df']

index <- matrix(NA, nrow=length(penalties), ncol=9,
           dimnames=list(format(penalties),
          c("Dxy","R2","Intercept","Slope","Emax","D","U
```

```
dev <- roc <- brier <- single(length(penalties))

evaltest <- function(cof,w=1:(length(cof)-1)) {
  pred <- plogis(cof[1] + (Xnew[,w,drop=F] %*% cof[-1]))
  C.index <- somers2(pred, Ynew)["C"]; names(C.index) <-
  Brier   <- mean((pred-Ynew)^2)
  Deviance<- -2*sum( Ynew*log(pred) + (1-Ynew)*log(1-pre
  c(deviance=Deviance, roc=C.index, brier=Brier)
}

i <- 0
set.seed(143)
for(penlty in penalties) {

  cat(penlty, "")
  i <- i+1
  if(penlty==0) {
    g <- f
    X <- f$x
    Y <- f$y
    penalty.matrix <- diag(diag(var(X)))   # save time -
  } else g <- lrm(Y ~ X, penalty=penlty,
                  penalty.matrix=penalty.matrix, x=T,y=T
  val <- validate(g, method="cross", B=10)
  index[i,] <- val[,"index.corrected"]
  w <- evaltest(g$coef)
  dev[i] <- w[1]; roc[i] <- w[2]; brier[i] <- w[3]
}

stores(aic.c, edf, index, dev, roc, brier)

#ps.slide('crossval.penalty.Q',type=3,hor=F,las=1,height
setps(crossval.penalty.Q)
plot(penalties, index[,'Q'], xlab='Penalty',
     ylab='Q', type='b')
dev.off()

setps(examine.test, h=6, pointsize=12, toplines=1)
par(mfrow=c(3,2))
Penalty <- penalties
best <- penalties[dev==min(dev)]
```

```
w <- function() invisible(abline(v=best, lty=2, lwd=1))

plot(Penalty, edf,   type='b', main='Effective d.f.'); w
plot(Penalty, aic.c, type='b',
     main='Effective AIC in Training Sample');  w()
plot(Penalty, dev,   type='b', main='Deviance in Test Sa
plot(Penalty, roc,   type='b', main='ROC Area in Test Sa
plot(Penalty, brier, type='b', main='Brier Score in Test
dev.off()

#Assess calibration accuracy in test sample
pred <- plogis(f$coef[1] + (Xnew %*% f$coef[-1]))
val.prob(pred, Ynew, group=T)

g <- update(f, penalty=penalties[aic.c==max(aic.c)],
            penalty.matrix=penalty.matrix, x=F, y=F)
predp <- plogis(g$coef[1] + (Xnew %*% g$coef[-1]))
val.prob(pred, Ynew, group=T)

z <- list('MLE'=wtd.loess.noiter(pred,Ynew,type='eval'),
          'PMLE'=wtd.loess.noiter(predp,Ynew,type='eval'
          'Ideal'=list(x=c(0,1),y=c(0,1)))
setps(calibration.test)
labcurve(z, lty=c(1,3,1), lwd=c(2,2,4),
         keys=c('M','P','I'), method='on top',
         xlab='Predicted Probability',
         ylab='Estimated Actual Probability', pl=T)
dev.off()


#Model approximation - simulate a new training and test

# Function to generate n p-variate normal variates with
# and covariance matrix S
# Slight modification of function written by Bill Venabl
mvrnorm <- function(n, p = 1, u = rep(0, p), S = diag(p)
  Z <- matrix(rnorm(n * p), p, n)
  t(u + t(chol(S)) %*% Z)
}

n <- 250
```

```
nval <- 10000
nt <- n + nval

# Generate multivariate normal covariables for nt subjec
# Assume equal correlations of rho=.4, independent subje
rho <- .4
set.seed(19)
X <- mvrnorm(nt, p=15, S=diag(rep(1-rho,15))+rho)

x1 <- X[,1]
x2 <- X[,2]
x3 <- X[,3]
x4 <- X[,4]
x5 <- X[,5]
x6 <- X[,6]
x7 <- X[,7]
x8 <- X[,8]
x9 <- X[,9]
x10<- X[,10]
x11<- X[,11]
x12<- X[,12]
x13<- X[,13]
x14<- X[,14]
x15<- X[,15]

logit <- .25*(2*x1+x2+x3+.75*x4+.5*x5+.5*x6+.5*x7+.25*x8
              .25*x9+.125*x10)
set.seed(149)
y  <- ifelse(runif(nt) < plogis(logit), 1, 0)

f  <- lrm(y ~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13
                x=T,y=T, subset=1:n)
best <- pentrace(f, 60, method='optimize')
pentrace(f, c(0,5,10,15,20,30,40,50,60,70,90))
fp <- update(f, penalty=best$penalty)

new.data <- data.frame(x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,
                       x11,x12,x13,x14,x15,y)[-(1:n),]
Xnew <- predict(f, new.data, type="x", incl.non.slopes=F
Ynew <- new.data$y
```

```
fastbw(f) # found following order of variables:
ovb <- c(1,2,8,7,6,12,4,14,15,11,13,3,9,5,10)

z <- predict(fp)
h <- ols(z ~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+
         subset=1:n, sigma=1)
fastbw(h, aics=1000)  # found following order of variabl
ov <- c(1,2,7,8,11,14,12,6,13,10,9,5,3,15,4)
rsq <- deva <- roca <- briera <- devb <- rocb <- brierb
for(i in 1:15) {
  fa <- lm.fit.qr.bare(X[1:n,ov[1:i],drop=F], z) # in Hm
  rsq[i] <- fa$rsquared
  w <- evaltest(fa$coef, ov[1:i])
  deva[i] <- w[1]; roca[i] <- w[2]; briera[i] <- w[3]

  fb <- lrm.fit(X[1:n,ovb[1:i],drop=F], y[1:n])
  w <- evaltest(fb$coef, ovb[1:i])
  devb[i] <- w[1]; rocb[i] <- w[2]; brierb[i] <- w[3]
}
stores(rsq, deva, roca, briera, devb, rocb, brierb)


setps(approx.test, h=6, pointsize=12, toplines=1)
par(mfrow=c(2,2))

plot(1:15, rsq, type='b', xlab='# Variables Selected',
     ylab='R2', main='Approximation R2')

pl <- function(y1,y2,ylab)
  invisible(labcurve(list('Penalized'=list(1:15, y1),
                          'Stepdown'=list(1:15, y2)),
                     xlab='# Variables Selected', ylab=y
                     lty=c(1,3), pl=T, method='arrow'))

pl(deva, devb, 'Deviance'); title('Deviance in Test Samp
pl(roca, rocb, 'ROC Area'); title('ROC Area in Test Samp
pl(briera, brierb, 'Brier Score'); title('Brier Score in

dev.off()
```