

Model Assessment and Selection for Prediction - Part 2

UC Irvine - ISI BUDS 2023

Presented July 26, 2023

Daniel L. Gillen
Chancellor's Professor and Chair
Department of Statistics
University of California, Irvine

Ex: King County Birth
Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and
BIC)

Estimation of the
extra-sample error

Cross-validation
Bootstrap methods

Summary

Ex: King County birth weight data

Model selection and coefficient shrinkage

- ▶ In many prediction situations there are a large number of inputs, X
- ▶ While it may be the case that $f(X) = X^T \beta$ appropriately describes the underlying mechanisms, it is always the case that we have a finite training sample size, n
- ▶ Prediction accuracy:
 - ▶ least squares estimates may have low bias, but in 'small'-sample settings can exhibit large variability
 - ▶ we could sacrifice a little bias to reduce variation and achieve better overall predictive accuracy

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation
Bootstrap methods

Summary

Ex: King County birth weight data

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation
Bootstrap methods

Summary

Model selection and coefficient shrinkage

- ▶ Another issue is interpretation:
 - ▶ with a large number of predictors, it may be hard to conceptualize 'holding everything else constant'
 - ▶ may be desirable to restrict attention to a smaller subset of variables which exhibit the strongest effects

Ex: King County birth weight data

King County birth data

- ▶ As an example, let's consider data on child birth weights for children born in King County, WA in 2001
- ▶ The dataset contains information on a sample of $n=2,500$ births from 2001
- ▶ The data was originally obtained to determine if a new state program ('First Steps') to educate women on proper nutrition during pregnancy was associated with greater birth weight
- ▶ The key outcome variable of interest is birth weight
 - ▶ Birth weight ranges from 255g to 5,175g
 - ▶ 5.1% of babies (127) were born at *low birth weight* ($< 2,500\text{g}$)
- ▶ A total of 15 potential predictor variables are available for investigation

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation
Bootstrap methods

Summary

Ex: King County birth weight data

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation
Bootstrap methods

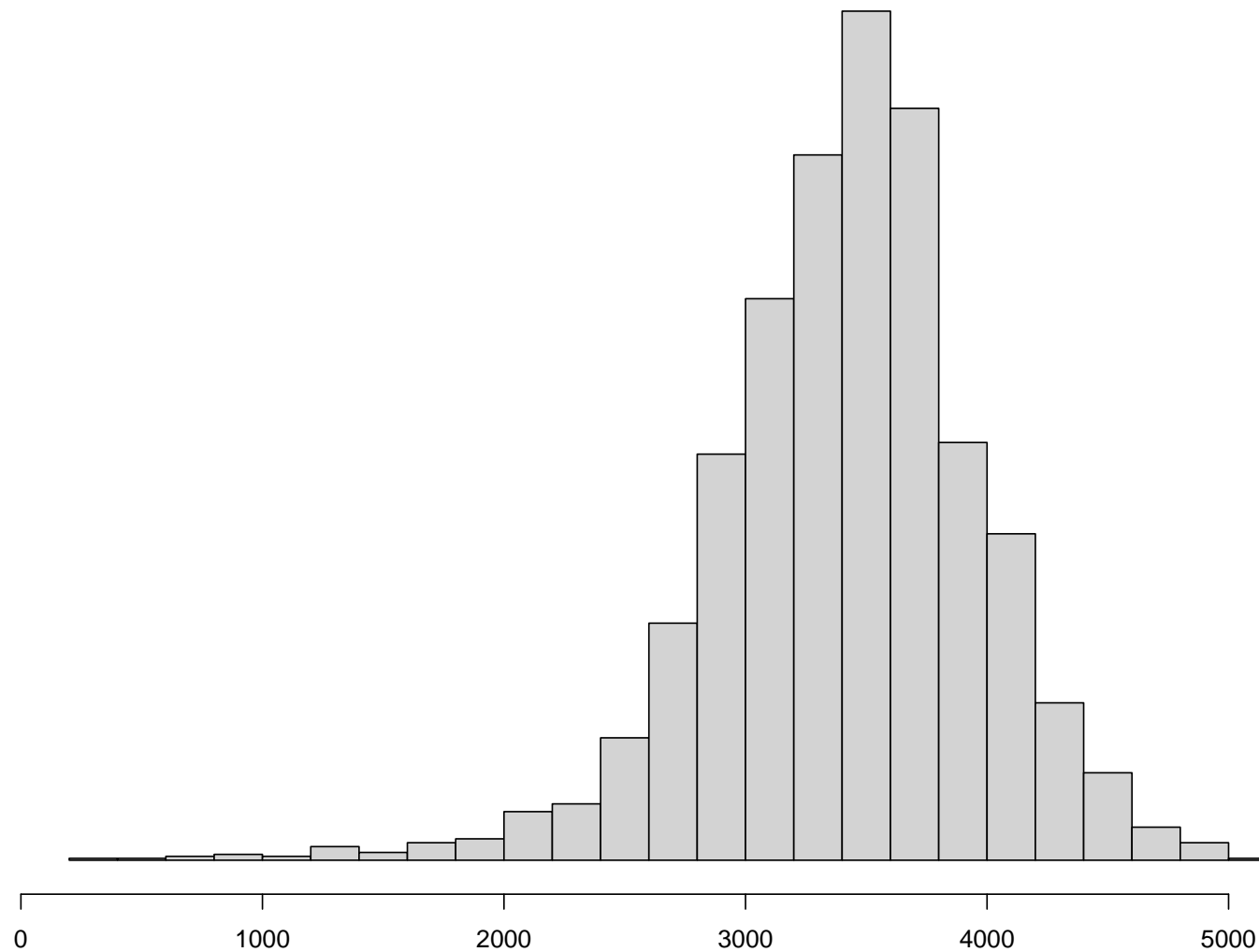
Summary

Complete variable listing

"gender"	M = male, F = female baby
"plural"	1 = singleton, 2 = twin, 3 = triplet
"age"	mother's age in years
"race"	race categories (for mother)
"parity"	number of previous live born infants
"married"	Y = yes, N = no
"bwt"	birth weight in grams
"smokeN"	number of cigarettes smoked per day during pregnancy
"drinkN"	number of alcoholic drinks per week during pregnancy
"firststep"	1 = participant in program; 0 = did not participate
"welfare"	1 = participant in public assistance program; 0 = did not
"smoker"	Y = yes, N = no, U = unknown
"drinker"	Y = yes, N = no, U = unknown
"wpre"	mother's weight in pounds prior to pregnancy
"wgain"	mother's weight gain in pounds during pregnancy
"education"	highest grade completed (add 12 + 1 / year of college)
"gestation"	weeks from last menses to birth of child

Distribution of birth weights from the King County data

Birth weight, grams



Ex: King County Birth Weight Data

- Best subsets regression
- Ridge regression
- Simulation study (AIC and BIC)

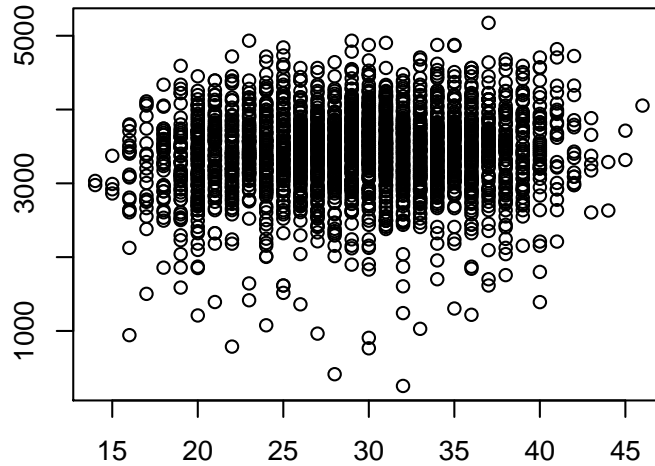
Estimation of the extra-sample error

- Cross-validation
- Bootstrap methods

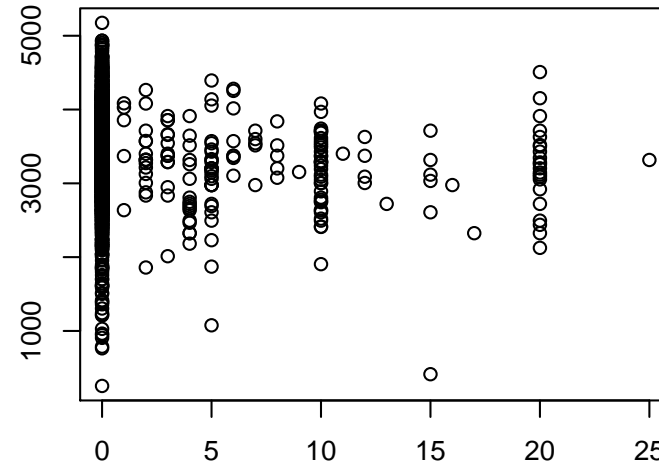
Summary

Selected scatterplots from the King County data

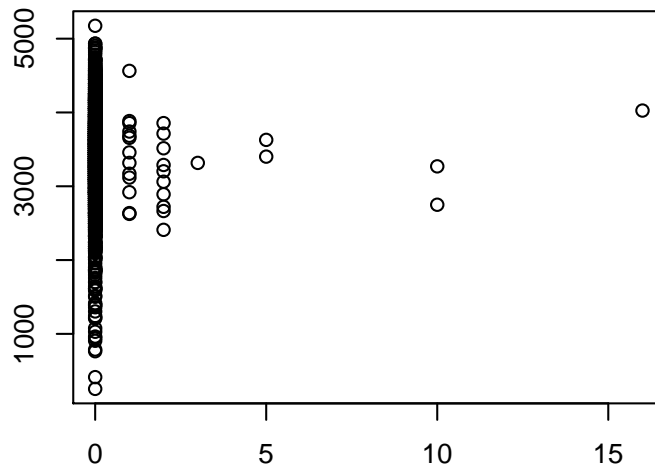
Mother's age, years



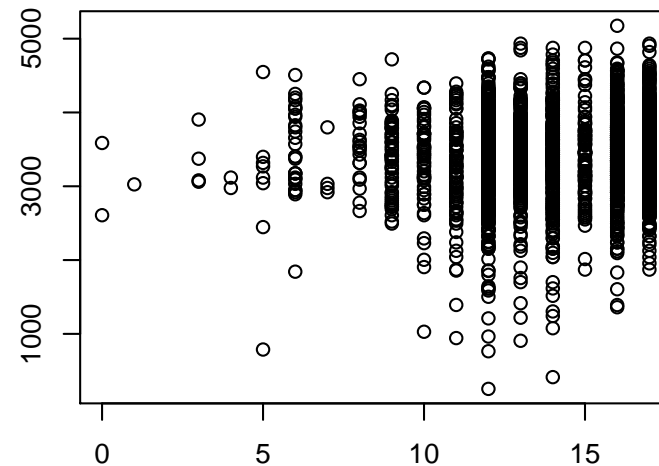
Cigarettes smoked per day



Alcoholic drinks per week



Highest grade completed



Ex: King County Birth Weight Data

- Best subsets regression
- Ridge regression
- Simulation study (AIC and BIC)

Estimation of the extra-sample error

- Cross-validation
- Bootstrap methods

Summary

Ex: King County birth weight data

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation
Bootstrap methods

Summary

Subset selection vs. shrinkage

- ▶ Rather than attempting to fit and report a model which includes all the potential predictors, we can consider two strategies
 - ▶ subset selection
 - ▶ shrinkage methods

Ex: King County birth weight data

Ex: King County Birth Weight Data

Best subsets regression

Ridge regression

Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

Subset selection

- ▶ Here we retain only a subset of variables
 - ▶ the remaining variables essentially have their β coefficients set to zero
- ▶ Various strategies exist for ‘choosing’ the variables to keep (or throw out)
 - ▶ best subset selection
 - ▶ stepwise strategies

Ex: King County birth weight data

Best subsets regression

- ▶ Suppose X consists of p components; X_1, \dots, X_p
- ▶ For each $k \in \{1, \dots, p\}$, find the subset of k variables which results in the smallest residual sums of squares
 - ▶ other criteria include Mallows's C_p , R^2 and adjusted R^2
- ▶ Can quickly become computationally intensive when p gets large
- ▶ In \mathbb{R} , code is implemented in the `leaps` package

Ex: King County Birth Weight Data

Best subsets regression

Ridge regression

Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

Ex: King County birth weight data

Best subsets regression in R

```
library(leaps)

## Model with only the intercept
##
fit0 <- lm(bwt ~ 1, data=weight)

## Perform best subsets analysis
##
## maxModel: a model which includes all the variables you wish to
##            entertain
## nvmax:    maximum number of variables for the subset selection
## nbest:    specify, for any given k, the number of the best models
##            are to be returned
##
maxModel <- as.formula(bwt ~ gender + age + race + parity + married
                      + smokeN + drinkN + firststep + welfare
                      + smoker + drinker + wpre + education )

bestSub <- summary(regsubsets(maxModel, data=weight,
                             nvmax=17, nbest=10))
```

Ex: King County Birth Weight Data

Best subsets regression

Ridge regression

Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

Ex: King County birth weight data

Ex: King County Birth Weight Data

Best subsets regression

Ridge regression

Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

Best subsets regression in R

```
names(bestSub)
[1] "which" "rsq" "rss" "adjr2" "cp" "bic"
     "outmat" "obj"

## 'results' contains the subset size, k, and the residual
## sum of squares

results <- c(0, sum((weight$bwt- fitted(fit0))^2))
results <- rbind(results,
                 cbind(apply(bestSub$which, 1, sum)-1, bestSub$rss))

##
##### Look at minimum residual sums of squares
##
minRSS <- tapply(results[,2], results[,1], FUN=min)
```

Best subset selection for the King County 2001 birth weight data

Ex: King County Birth Weight Data

Best subsets regression

Ridge regression

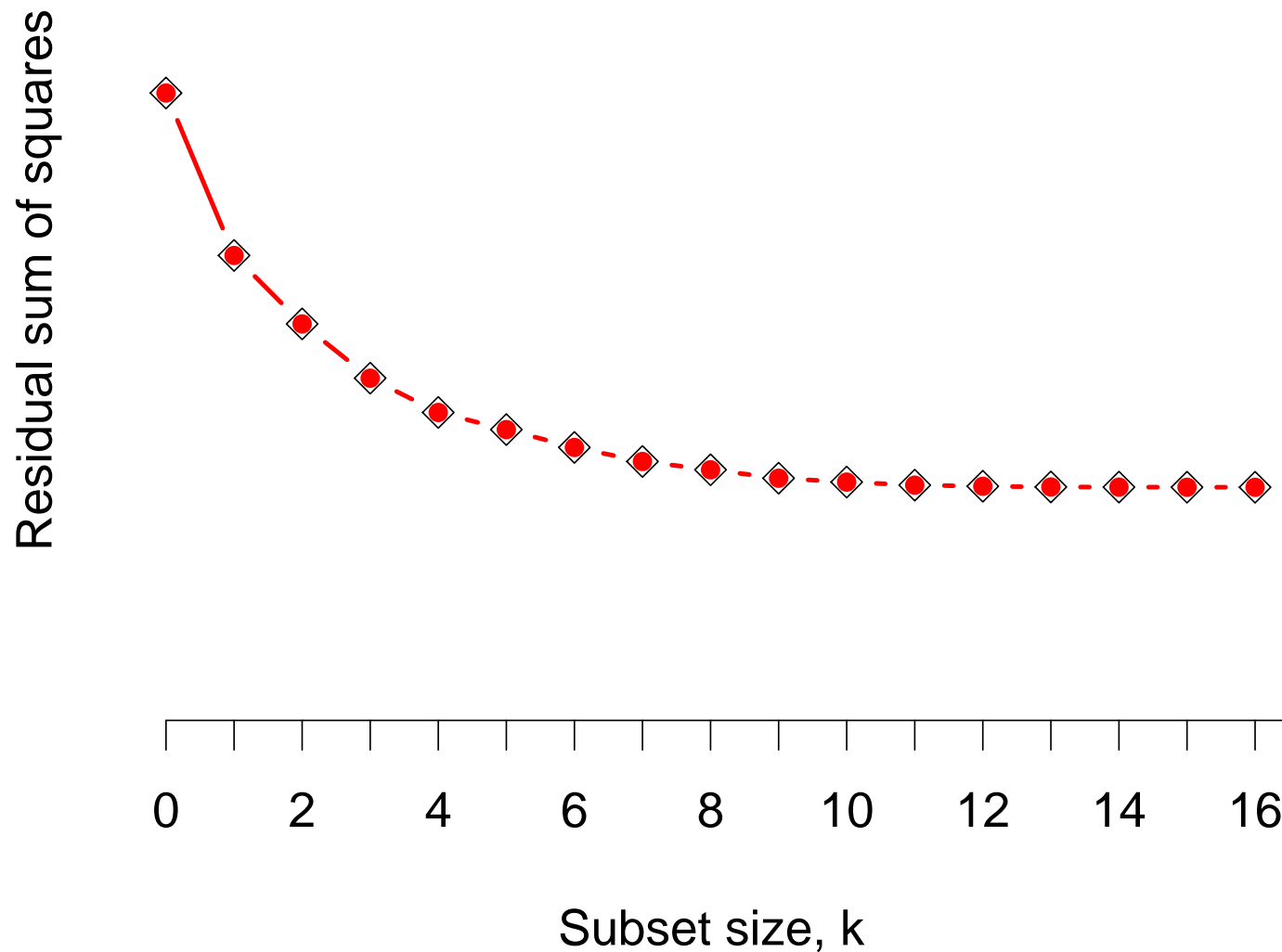
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary



Ex: King County birth weight data

Ex: King County Birth Weight Data

Best subsets regression

Ridge regression

Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

Best subsets regression

- ▶ The best-subset curve is necessarily decreasing, so it cannot be used as a criteria for choosing k
- ▶ Typically choose a model which minimizes an estimate of the EPE
 - ▶ Mallows's C_p , AIC, BIC, cross-validation

Ex: King County birth weight data

Ex: King County Birth Weight Data

Best subsets regression

Ridge regression

Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

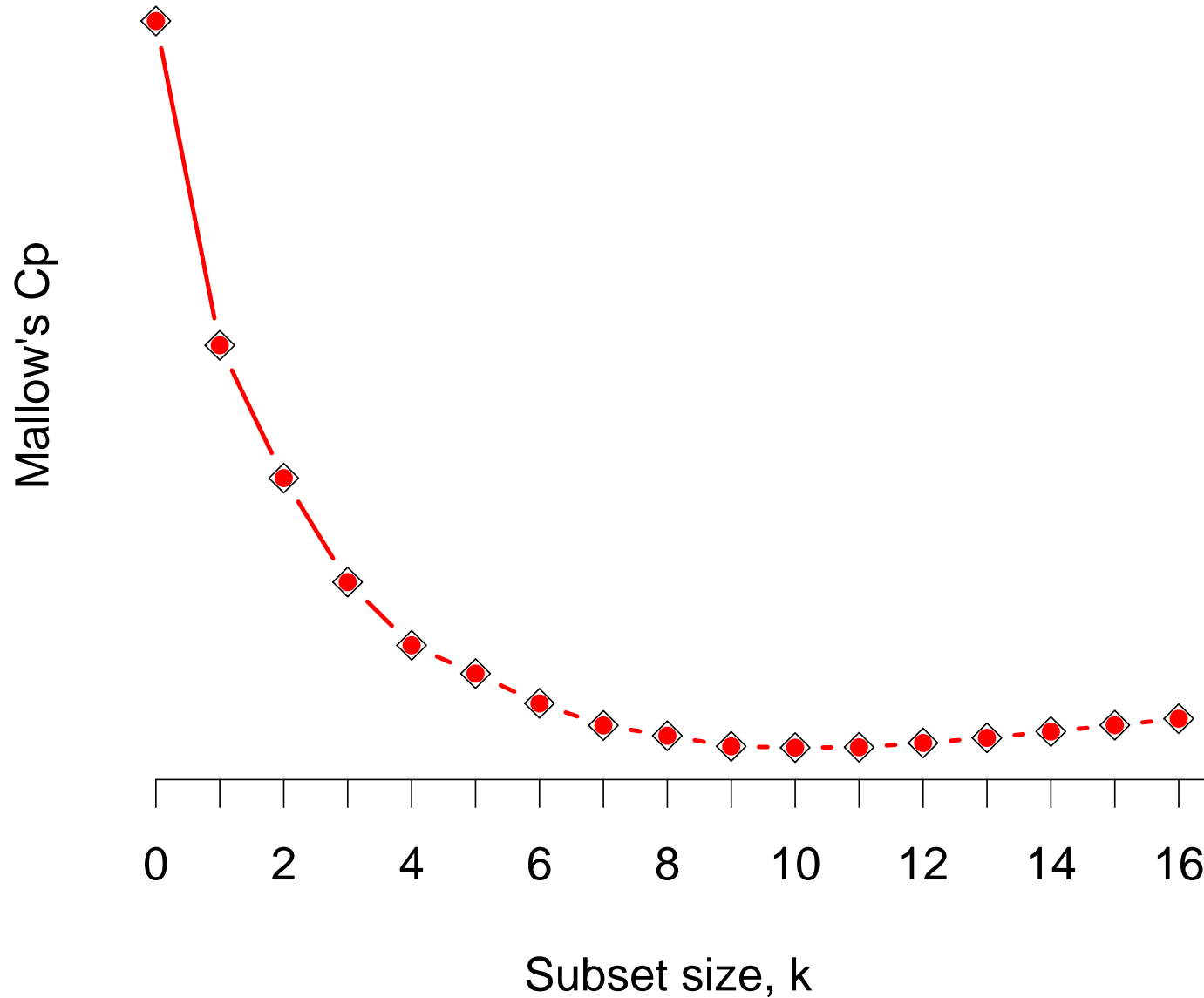
Summary

Best subsets regression in R

```
##
#####
#####      Now let's do best subsets with Cp as the criteria
#####
##
bestSubCp <- leaps(x=model.matrix(fitF),
                  y=weight$bwt, int=FALSE,
                  nbest=1, method="Cp")
## 'results' contains the subset size, k, and the Cp value
##
results <- NULL
results <- rbind(results, cbind(apply(bestSubCp$which, 1, sum)-1,
                                  bestSubCp$Cp))

##
minCp<- tapply(results[,2], results[,1], FUN=min)
```

Best subset selection for the King County 2001 birth weight data



Ex: King County Birth Weight Data

Best subsets regression

Ridge regression

Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

Ex: King County birth weight data

Best subsets regression in R

```
##
#####
#####          Compare which were selected in the k=10 models...
#####
##
> cbind( dimnames( model.matrix(fitF) )[[2]],
         bestSubCp$which[11,],
         bestSubRSS$which[11,] )

  [,1]      [,2]      [,3]
1 "(Intercept)" "TRUE"  "TRUE"
2 "genderM"     "TRUE"  "TRUE"
3 "age"         "FALSE" "FALSE"
4 "raceblack"   "FALSE" "FALSE"
5 "racehispanic" "TRUE"  "TRUE"
6 "raceother"   "FALSE" "TRUE"
7 "racewhite"   "TRUE"  "TRUE"
8 "parity"      "TRUE"  "TRUE"
9 "married"     "TRUE"  "TRUE"
A "smokeN"      "TRUE"  "TRUE"
B "drinkN"      "FALSE" "FALSE"
C "firststep"   "FALSE" "FALSE"
D "welfare"     "TRUE"  "TRUE"
E "smokerY"     "TRUE"  "TRUE"
F "drinkerY"    "FALSE" "FALSE"
G "wpre"        "TRUE"  "TRUE"
H "education"   "TRUE"  "TRUE"
```

Ex: King County Birth Weight Data

Best subsets regression

Ridge regression

Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

Ex: King County birth weight data

Stepwise procedures

- ▶ Instead of performing an exhaustive enumeration for each value for k , we can search for a ‘good path’
- ▶ Forward selection
 - ▶ start with an ‘intercept-only’ model and build up the model
- ▶ Backward selection
 - ▶ start with a ‘full’ model and reduce the model
- ▶ In R, see `stepAIC` in the MASS library

Ex: King County Birth Weight Data

Best subsets regression

Ridge regression

Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

Ex: King County birth weight data

Stepwise AIC in R

```
library(MASS)
##
#####
#####      Stepwise selection using AIC
#####
##
fitStepAIC <- stepAIC( fit0, scope=maxModel, direction="forward" )
.
.
.
>Step:  AIC=31425
bwt ~ wpre + smoker + gender + married + race + parity + welfare +
      education + smokeN

           Df Sum of Sq      RSS      AIC
<none>            7.12e+08 3.14e+04
+ drinker      1  1.96e+05 7.12e+08 3.14e+04
+ drinkN       1  4.37e+04 7.12e+08 3.14e+04
+ firststep    1  2.41e+04 7.12e+08 3.14e+04
+ age          1  1.88e+04 7.12e+08 3.14e+04
```

Ex: King County Birth Weight Data

Best subsets regression

Ridge regression

Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

Ex: King County birth weight data

Shrinkage methods

- ▶ Rather retaining some variables and discarding the rest, an alternative is to keep all the variables but impose restrictions on the size of the coefficients
 - ▶ the point estimates for β are subject to bias
 - ▶ results often don't suffer as much in terms of variability
- ▶ Ridge regression imposes an L_2 -type penalty
 - ▶ the solution is given by

$$\hat{\beta}^{\text{ridge}} = \operatorname{argmin}_{\beta} \operatorname{RSS}(\beta)$$

subject to the constraint:

$$\sum_{j=1}^p \beta_j^2 \leq s$$

- ▶ value of s influences how large the components of β can get

Ex: King County Birth Weight Data

Best subsets regression

Ridge regression

Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

Ex: King County birth weight data

Shrinkage methods

- ▶ An alternative way of writing the problem is

$$\hat{\beta}^{\text{ridge}} = \operatorname{argmin}_{\beta} \left\{ \operatorname{RSS}(\beta) + \lambda \sum_{j=1}^p \beta_j^2 \right\}$$

- ▶ Here, $\lambda \geq 0$ controls the amount of shrinkage
 - ▶ when $\lambda = 0$, we are performing ordinary least squares estimation
 - ▶ for large λ , minimizing the penalized RSS requires the components of β to be small
 - ▶ there is a one-to-one relationship between λ and s
- ▶ Minimization yields the solution

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

- ▶ Could allow λ to be a vector, and ensure no shrinkage among certain coefficients

Ex: King County Birth Weight Data

Best subsets regression

Ridge regression

Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

Ex: King County birth weight data

Shrinkage methods

- ▶ The solution is linear, and we can therefore obtain the effective degrees of freedom as

$$\text{df}(\lambda) = \text{tr}\{\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\}$$

- ▶ depends on the complexity/smoothing parameter λ
- ▶ Ridge regression for the linear model is implemented in \mathbb{R}

```
library(MASS)

ridgeFit <- lm.ridge(maxModel, data=weight,
                    lambda=c(0, 100, 1000, 10000))
```

Ex: King County Birth Weight Data

Best subsets regression

Ridge regression

Simulation study (AIC and BIC)

Estimation of the extra-sample error

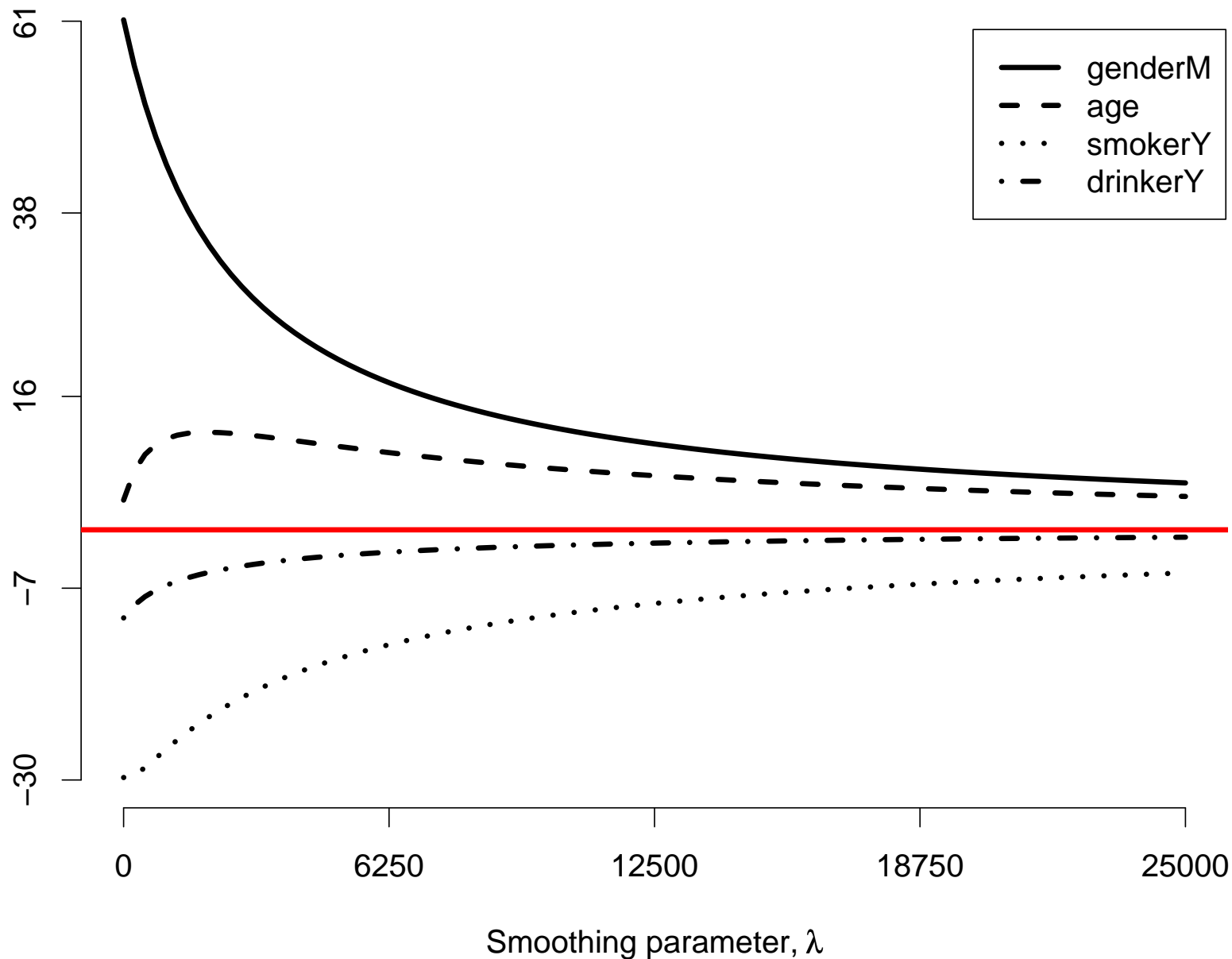
Cross-validation

Bootstrap methods

Summary

Ridge regression for the King County Birth data

Ridge regression coefficient estimates



Ex: King County Birth Weight Data

Best subsets regression

Ridge regression

Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

Ex: King County birth weight data

Ex: King County Birth Weight Data

Best subsets regression

Ridge regression

Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

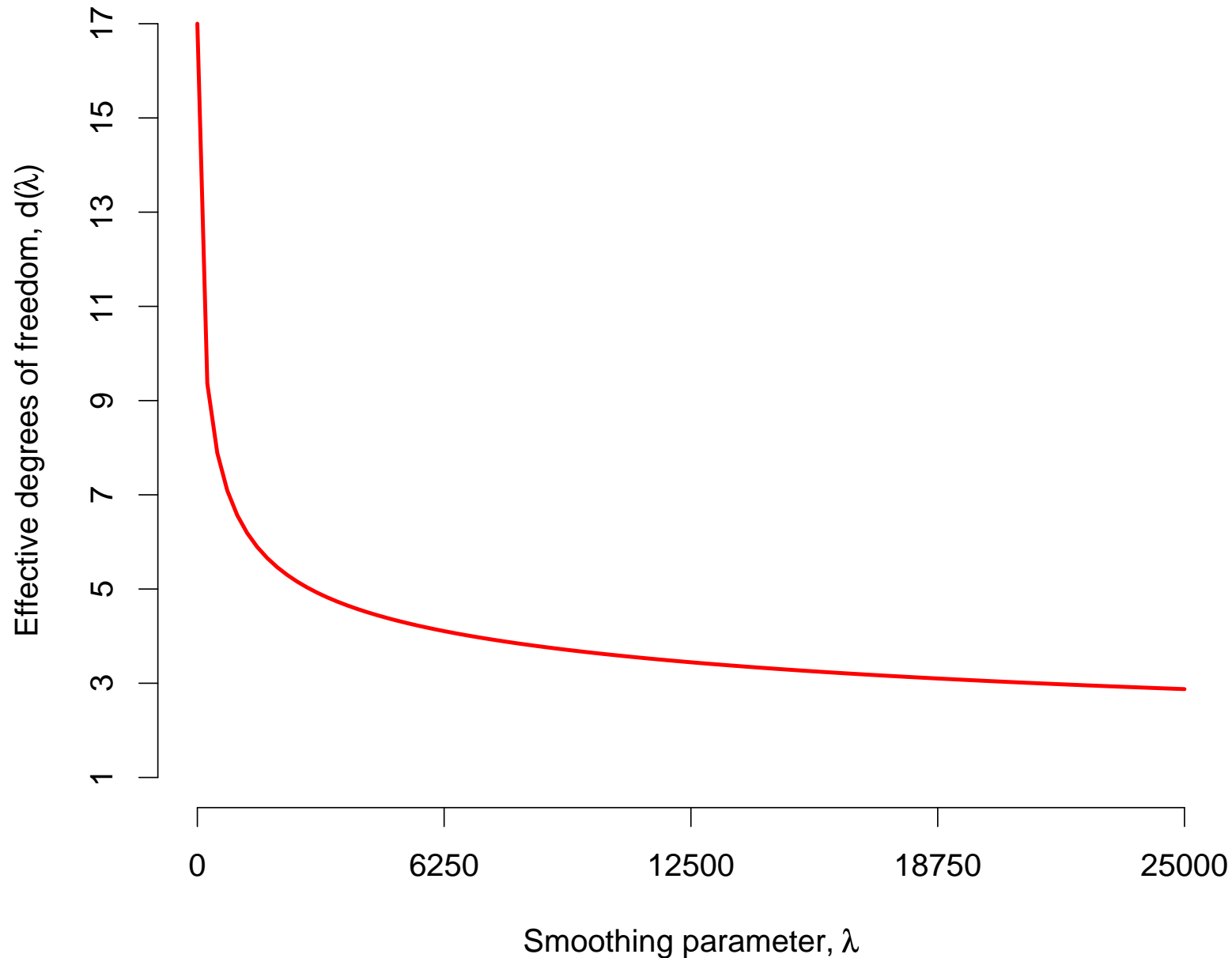
Summary

Ridge regression results for King County data

- ▶ Relationship between λ and $df(\lambda)$ is non-linear
- ▶ For these data, there is a dramatic reduction in 'complexity' of the model up to about $\lambda = 1,000$

Ridge regression for the King County Birth data

Smoothing parameter vs. effective degrees of freedom



Ex: King County Birth
Weight Data

Best subsets regression

Ridge regression

Simulation study (AIC and
BIC)

Estimation of the
extra-sample error

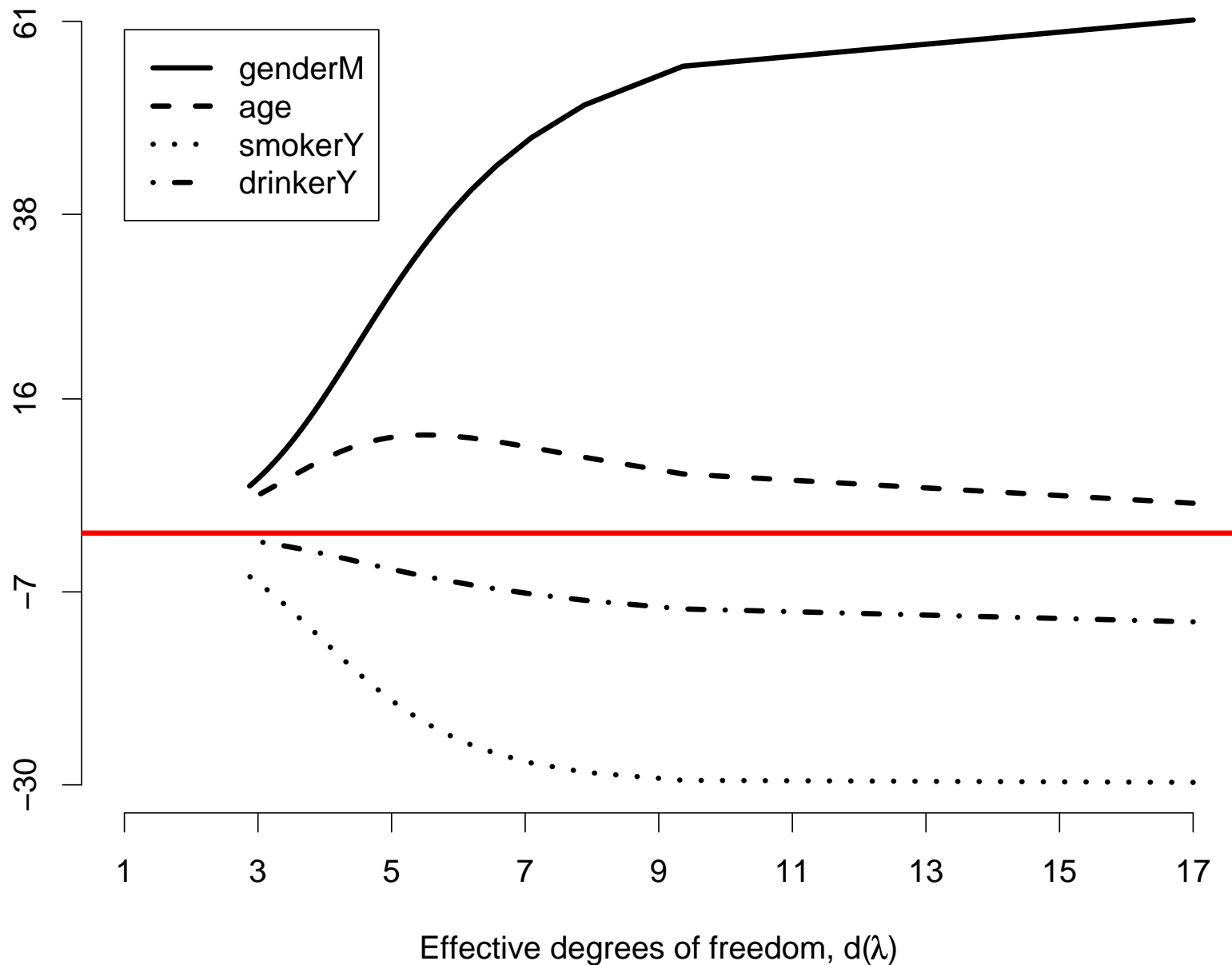
Cross-validation

Bootstrap methods

Summary

Ridge regression for the King County Birth data

Ridge regression coefficient estimates



Ex: King County Birth Weight Data

Best subsets regression

Ridge regression

Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

Model selection for 'complexity'

AIC vs. BIC

- ▶ Up to a constant of proportionality, AIC and BIC differ in terms of the penalty imposed on increasing complexity

$$\text{AIC} \Rightarrow 2p$$

$$\text{BIC} \Rightarrow (\log n)p$$

- ▶ for reasonable sample sizes, BIC imposes a heavier penalty
- ▶ Unfortunately, in practice, there isn't a clear choice between the two
- ▶ We can investigate their relative merits using the King county birth weight data
 - ▶ consider determining the value of λ in a ridge regression analysis which includes all 13 predictors

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression

Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation
Bootstrap methods

Summary

AIC and BIC for a ridge regression analysis of the King county birth weight data

Ex: King County Birth Weight Data

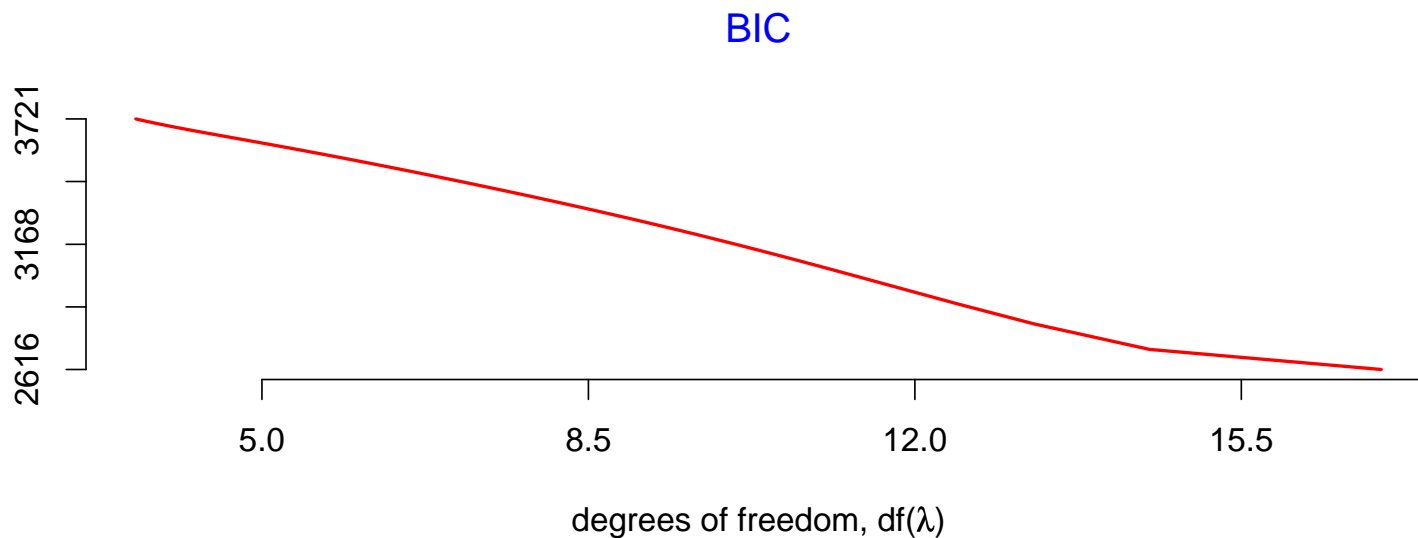
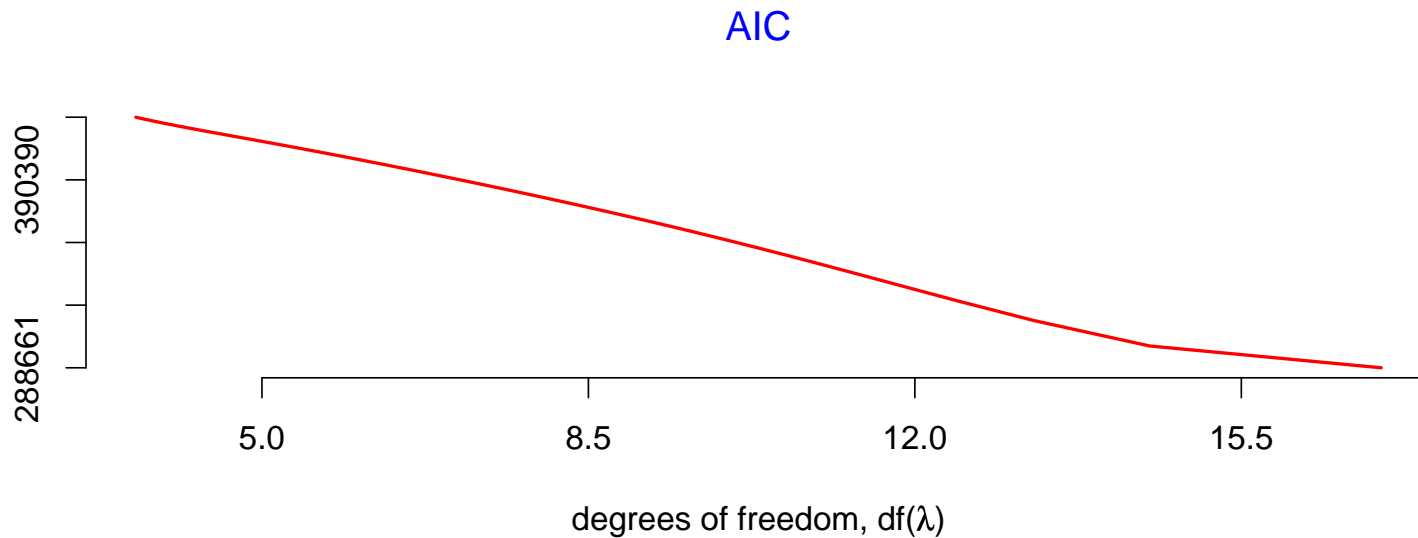
Best subsets regression
Ridge regression

Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation
Bootstrap methods

Summary



Model selection for 'complexity'

AIC vs. BIC

- ▶ It seems that both AIC and BIC choose the optimal value of λ to be zero
 - ▶ degrees of freedom = 17
- ▶ They both favor the most complex models
 - ▶ neither penalty seems to offset the reduction in RSS by increasing the complexity of the model
- ▶ Even though we are estimating 17 parameters with 2500 observations, seems that there should still be room for improvement in the model...

Ex: King County Birth Weight Data

Best subsets regression

Ridge regression

Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

Estimation of the extra-sample error

Extra-sample error

- ▶ While AIC and BIC permit an analytic treatment of assessing the predictive ability of a given model, their focus on the in-sample error, Err , is somewhat of a drawback
- ▶ Here we return to estimation of the extra-sample error,

$$EPE = E_{X,Y} \left[L(Y, \hat{f}(X)) \right],$$

interpreted as the generalization error when the prediction rule $\hat{f}(\cdot)$ is applied to an independent test sample, from the joint distribution of X and Y

- ▶ Both approaches we consider here involve the clever use and re-use of the training data

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation
Bootstrap methods

Summary

Cross-validation

- ▶ One possibility for choosing λ could be to attempt to minimize the observed mean squared error:

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

- ▶ However, this is typically a poor estimate of mean squared prediction error (or out-of-sample prediction error)
- ▶ One aspect of the problem is that the estimate $\hat{y}_i = \hat{f}(x_i)$ uses the observed outcome y_i , as well as the others, to predict y_i
- ▶ One solution to this would be to predict y_i using all the observations *except* the i^{th} case

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

Cross-validation

- ▶ If we denote the resulting prediction as $\hat{y}_{(i)}$, then the corresponding sum of squared residuals is referred to as the *predicted residual sum of squares*

$$\text{PRESS} = \sum_{i=1}^n (y_i - \hat{y}_{(i)})^2$$

- ▶ PRESS is also referred to as the *cross-validation* statistic
 - ▶ *leave-one-out cross-validation*
 - ▶ denote with CV
- ▶ In general situation the computational burden can be substantial
 - ▶ requires n fits of the model

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

Cross-validation

- ▶ However, calculation of the CV statistic is fairly straightforward for linear models
 - ▶ leave-one-out, or deleted, residuals are obtained from the residuals of the model based on all the data as well as the hat matrix, \mathbf{H}

$$y_i - \hat{y}_{(i)} = \frac{y_i - \hat{y}_i}{1 - H_{ii}}$$

where H_{ii} denote the i^{th} diagonal element of \mathbf{H}

- ▶ We therefore have

$$\text{CV} = \frac{1}{n} \sum_{i=1}^n \left[\frac{y_i - \hat{y}_i}{1 - H_{ii}} \right]^2$$

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

Generalized cross-validation

- ▶ The generalized cross-validation statistic arises when we approximate the H_{ii} by their average

$$\text{GCV} = \frac{1}{n} \sum_{i=1}^n \left[\frac{y_i - \hat{y}_i}{1 - \text{trace}(\mathbf{H})/n} \right]^2$$

- ▶ For the case of penalized regression, we replace $\text{trace}(\mathbf{H})$ with the effective degrees of freedom

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

K-fold cross-validation

- ▶ Leave-one-out cross-validation involves splitting the data into n parts
- ▶ The approach can be generalized somewhat by splitting the data into $K < n$ parts as follows

(1) Split the data into K roughly equal parts, and denote the collection of indexes for the k^{th} part as C_k , $k = 1, \dots, K$

(2) For each part, fit a model using all the remaining data,

$$\mathbf{y}^{(k)} = \{y_i \mid i \notin C_k\},$$

and denote the fitted model as $\hat{f}^k(x)$

(3) For all i such that $i \in C_k$, obtain a prediction via

$$\hat{y}_i = \hat{f}^k(x_i)$$

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

K-fold cross-validation

- ▶ Let $k(i)$ denote the part in which y_i resides
- ▶ The K -fold cross validation statistic, for a general loss function, is given by

$$CV_K = \frac{1}{n} \sum_{i=1}^n L(y_i, \hat{f}^{-k(i)}(x_i))$$

where L denotes a *loss function*. (We have been considering squared error loss so that

$$L(y_i, \hat{f}^{-k(i)}(x_i)) = (y_i - \hat{f}^{-k(i)}(x_i))^2$$

- ▶ As we decrease K , however, the bias of CV_K as an estimate of MS[P]E increases
 - ▶ CV_K is biased upward
 - ▶ extent depends on the sample size

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

Cross-validation for ridge regression

- ▶ The `select()` in the `MASS` library minimizes the generalized cross validation statistic for ridge regression
- ▶ Let's compare the complexity of the model when GCV is used as opposed to AIC and BIC

```
##  
#####  
##### How does AIC/BIC compare with GCV???  
#####  
##  
maxLambda <- 25000  
lambdaVal <- seq(from=0, to=maxLambda, length=100)  
select(lm.ridge(maxModel, data=weight, lambda=lambdaVal))  
Xmat      <- model.matrix(lm(maxModel, data=weight))  
  
calcDF(Xmat, lambda=252.53)  
> [1] 9.3619
```

So, the effective degrees of freedom using cross validation are 9.36, as compared to 17 for AIC and BIC

Ex: King County Birth
Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and
BIC)

Estimation of the
extra-sample error

Cross-validation

Bootstrap methods

Summary

Bootstrap estimates of prediction error

- ▶ Let $\hat{f}^b(\cdot)$ denote the estimate of $f(\cdot)$ obtained from the b^{th} bootstrap replicate, $b = 1, \dots, B$
- ▶ For each fit, keep a track of how well it predicts the original training data
 - ▶ evaluate the training error for each fit
- ▶ We could average across the B replicates to get an estimate of EPE

$$\widehat{\text{EPE}}_{\text{boot}} = \frac{1}{B} \sum_{b=1}^B \left[\frac{1}{n} \sum_{i=1}^n L(y_i, \hat{f}^b(x_i)) \right]$$

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

Leave-one-out bootstrap

- ▶ Typically $\widehat{EPE}_{\text{boot}}$ is not a good estimate of EPE since there is too much overlap between the bootstrap samples (which act as training data) and the training data (which acts as the test data)
- ▶ Cross-validation worked by averaging across replications where the training (sub-)data and test (sub-)data were explicitly separated
- ▶ We could mimic this by only evaluating the predictions for the i^{th} observation from bootstrap datasets in which it was not sampled

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

Leave-one-out bootstrap

- ▶ The *leave-one-out bootstrap* is defined by

$$\widehat{\text{EPE}}^{(1)} = \frac{1}{n} \sum_{i=1}^n \left[\frac{1}{|C_i|} \sum_{b \in C_i} L(y_i, \hat{f}^b(x_i)) \right]$$

- ▶ the set C_i denotes the indices of the bootstrap samples b that do *not* contain observation i
- ▶ $|C_i|$ is the number of such samples

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

.632 bootstrap estimator

- ▶ While the leave-one-out bootstrap estimator resolves the overfitting associated with $\widehat{EPE}_{\text{boot}}$, it can suffer in terms of bias analogous to that suffered by K -fold cross-validation when $K > 1$
- ▶ The average number of distinct observations in each bootstrap sample is $0.632n$

$$\begin{aligned}\Pr(\text{observation } i \in \text{bootstrap sample } b) &= 1 - \left(1 - \frac{1}{n}\right)^n \\ &\approx 1 - e^{-1} \\ &= 0.632\end{aligned}$$

- ▶ so $\widehat{EPE}^{(1)}$ behaves roughly in the same way as two-fold cross-validation

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

.632 bootstrap estimator

- ▶ The '.632 estimator' is design to alleviate the 'training-set-size' bias, and is defined by

$$\widehat{\text{EPE}}^{(.632)} = 0.368\text{err} + 0.632\widehat{\text{EPE}}^{(1)}$$

- ▶ intuitively, the estimator pulls the leave-one-out bootstrap estimator down towards the training error rate, and hence reduces its upward bias

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

Ex: King County birth weight data

Computation of prediction criteria for ridge regression models

- ▶ The function `ridge.predcrit()` on the course webpage will compute all of our commonly used estimates of prediction error for ridge regression models...

```
> set.seed(12345)
> source( "http://www.ics.uci.edu/~dgillen/
          Stat211/Code/ridgePredCrit.q" )
> maxModel <- as.formula(bwt ~ gender + age + race + parity +
+                         married + smokeN + drinkN +
+                         firststep + welfare + smoker +
+                         drinker + wpre + education)

> ridgeFit <- lm.ridge(maxModel, data=weight, lambda=252.53)
> ridge.predcrit( ridgeFit, formula=maxModel, data=weight,
                  K=10, B=500, boot=TRUE, sigmaSq="calculate" )
      df      mse    aic    bic      cv bs.mse bs.lout bs.632
9.3619 284974 38514 38568 287171 286657 290600 289149
```

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation

Bootstrap methods

Summary

Obtaining prediction criteria for an OLS fit in R

Obtaining prediction criteria for an OLS fit in R

- ▶ Similarly, the function `lm.predcrit()` will compute all of our commonly used estimates of prediction error for a standard OLS regression model...

```
##  
#####  
#####      Fit a standard liner regression model adjusting for  
#####      wpre, age, gender, and smokeN  
#####  
##  
> fit.lm <- lm( bwt ~ wpre + age + gender + smokeN, data=weight )  
  
> lm.predcrit( fit.lm, data=weight, K=10, boot=TRUE, B=100 )  
df      mse      Cp      aic      bic      cv      cv.k  bs.mse  bs.lout  bs.632  
5 291265 292432 38560 38589 292487 292512 291927 293132 292689
```

Ex: King County Birth
Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and
BIC)

Estimation of the
extra-sample error

Cross-validation

Bootstrap methods

Summary

Criteria to assess predictive accuracy

- ▶ Decision theoretic approach
- ▶ We measure errors between Y and $\hat{f}(X)$ by specifying a loss function $L(Y, \hat{f}(X))$
- ▶ The *test* or *generalization* error is the expected prediction error over an *independent* test sample

$$\text{EPE} = E_{X,Y} [L(Y, \hat{f}(X))]$$

- ▶ the expectation is taken over the joint distribution of X and Y
- ▶ the average error, were the prediction model to be applied to an independent sample from the population

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation
Bootstrap methods

Summary

Possibilities for estimating EPE

- ▶ Might consider *training error*

$$\text{err} = \frac{1}{n} \sum_{i=1}^n L(y_i, \hat{f}(x_i))$$

- ▶ Negatively biased....Overly optimistic
- ▶ Analytically, focus on *in-sample* error

$$\text{Err} = \frac{1}{n} \sum_{i=1}^n E_y \left[E_{Y^{\text{new}}} \left[L(Y_i^{\text{new}}, \hat{f}(x_i)) \right] \right]$$

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation
Bootstrap methods

Summary

Possibilities for estimating ERR

▶ AIC

- ▶ Consider (-2 times) the log-likelihood to be a loss function

$$\text{AIC} = -\frac{2}{n} \text{loglike} + 2\frac{p}{n}$$

▶ BIC

- ▶ Motivated by the Bayes factor in model selection

$$\Pr(\text{Data} | \mathcal{M}_m) \approx \log \Pr(\text{Data} | \mathcal{M}_m, \hat{\theta}_m) - (\log n) \frac{p_m}{2}$$

- ▶ Computed in practice as

$$\text{BIC} = -2\text{loglike} + (\log n)p$$

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation
Bootstrap methods

Summary

Resampling estimates EPE

- ▶ Using resampling to change the support of the observed predictors...
- ▶ General strategies that can be applied to any estimation technique (some quicker than others!)
- ▶ Cross-validation
 - ▶ Focus on the *predicted residual sum of squares*

$$\text{PRESS} = \sum_{i=1}^n (y_i - \hat{y}_{(i)})^2$$

- ▶ Easily computed for OLS fits
- ▶ Can be computationally intensive for more complicated regression models
- ▶ In this case, could focus on K -fold cross-validation

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation
Bootstrap methods

Summary

Resampling estimates EPE

▶ Bootstrapping

- ▶ Basic bootstrap is biased downwards
- ▶ Leave-one-out bootstrap is generally biased upwards
- ▶ Compromise is the .632 bootstrap

$$\widehat{EPE}^{(.632)} = 0.368\text{err} + 0.632\widehat{EPE}^{(1)}$$

Ex: King County Birth Weight Data

Best subsets regression
Ridge regression
Simulation study (AIC and BIC)

Estimation of the extra-sample error

Cross-validation
Bootstrap methods

Summary