STAA 577: HW5

Your Name

Problem 1

- **Model assessment**:
- Process of evaluating a model on unseen data to determine how well the model generalizes
- **Model selection**:
- Process of choosing the best model and model parameters from a set of candidate models based on comparative performance

Problem 2

- Direct estimates:
  - Adv: unbiased estimates of the true test and easy to explain
  - Disadv: Computationally intensive, require more data, and can have higher variance
- Indirect estimates:
  - Adv: Computationally simpler, as they do not require splitting or retraining
  - Disadv:biased towards training data, and adjustments do not account for all sources of error

Problem 3a

Problem 3b

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.76644 -0.35510 -0.00328  0.38087  1.55770
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.47839    0.07102  34.895  < 2e-16 ***
## lcavol       0.66515    0.10352   6.425 6.55e-09 ***
## lweight      0.26648    0.08607   3.096  0.00263 **
## age         -0.15820    0.08252  -1.917  0.05848 .
## lbph         0.14031    0.08402   1.670  0.09848 .
## svi          0.31533    0.09985   3.158  0.00218 **
## lcp         -0.14829    0.12566  -1.180  0.24115
## gleason      0.03555    0.11218   0.317  0.75207
## pgg45        0.12572    0.12312   1.021  0.31000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6995 on 88 degrees of freedom
## Multiple R-squared:  0.6634, Adjusted R-squared:  0.6328
## F-statistic: 21.68 on 8 and 88 DF,  p-value: < 2.2e-16
```

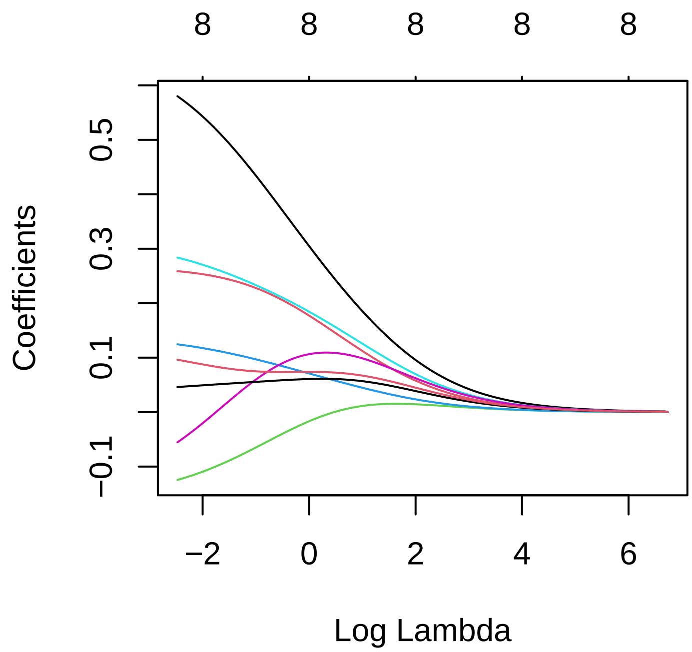
- It looks like lcavol lweight and svi are the most important variables. We have an ok rsquared at .66, and a signifficant P Value

Problem 3c

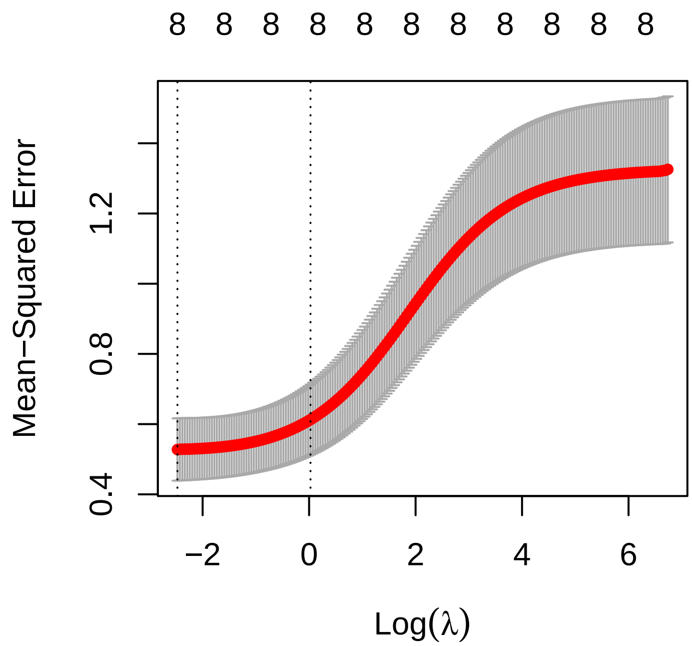
```
## [1] 0.5044407
```

CV MSE:0.5044407

Problem 3d



Problem 3e



Optimal $\lambda = 0.08434274$

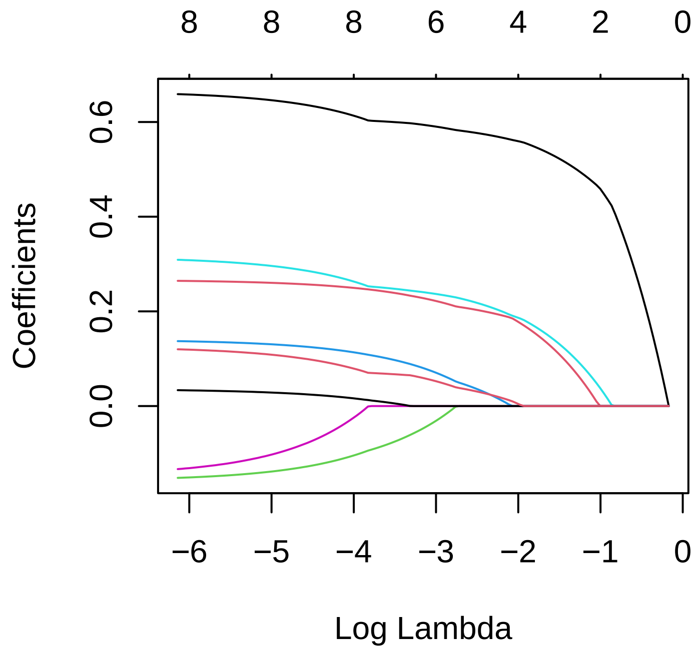
Problem 3f

```
## [1] 0.08434274
```

```
## [1] 0.5276459
```

CV MSE: .5276459

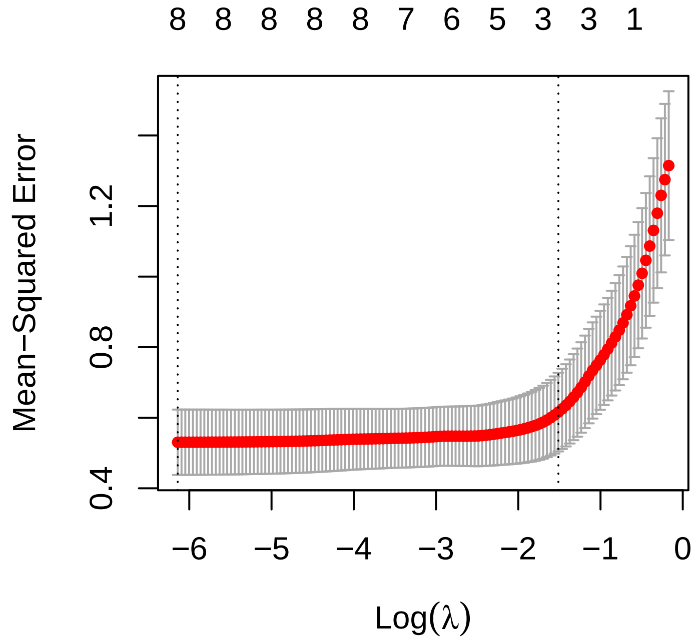
Problem 3g



Problem 3h

– Both plots show how the magnitude of the estimated coefficients changes as labmda increases. – In ridge regression, coefficients shrink continuously toward zero but none become exactly zero whereas in lasso, some coefficients are driven exactly to zero

Problem 3i




```
## [1] 0.002153193
```

```
## [1] 0.530523
```

- $\lambda = 0.002153193$
- CV MSE = 0.530523

Problem 3j

```
##                     s0    coef_mlr
## (Intercept) 2.47838688  2.47838688
## lcavol      0.57075673  0.66514667
## lweight     0.19578624  0.26648026
## age         0.00000000 -0.15819522
## lbph        0.02085106  0.14031117
## svi         0.20673566  0.31532888
## lcp         0.00000000 -0.14828568
## gleason     0.00000000  0.03554917
## pgg45       0.02219560  0.12571982
```

- Lasso selects all but age, lcp, and gleason features. it also pulls many of the coefficents for the features closer to zero like lbph and gleason.

Problem 4a & 4B

```
## [1] "Minimum CV MSE for Ridge: 0.940366233539713"
```

```
## [1] "AVG CV MSE for Ridge: 57.560673724125"
```

```
## [1] "AVG CV MSE for LASSO: 46.3994570187772"
```

- We can use the lasso, ridge, and traditional models in a cv=5 loop to predict the crime in boston
- Lasso Model has the best avg MSE accross folds of cross validation so I think we should go with that. It also provides a simpler model as it pulls some features to zero.

Problem 4c

- No. because I have chosen Lasso Regularization, the coefficients of some of our features are exactly zero.

Appendix

```
library(knitr)
# install the tidyverse library (do this once) install.packages('tidyverse')
library(tidyverse)
library(patchwork)
# set chunk and figure default options
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE, fig.width = 4,
    fig.height = 4, tidy = TRUE)
# problem 3
prostate <- read.csv("prostate.csv")
prostate <- prostate[, -ncol(prostate)]
prostate[, 1:8] <- scale(prostate[, 1:8])

# problem 3b

lmOut <- lm(lpsa ~ ., data = prostate)
```

```r
summary(lmOut)

# problem 3c
library(caret)
library(glmnet)
set.seed(577)
train_control <- trainControl(method = "cv", number = 10)
cv_model <- train(lpsa ~ ., data = prostate, method = "lm", trControl = train_control)

rmse <- cv_model$results$RMSE^2
print(rmse)
# problem 3d
lpsa <- prostate$lpsa
xmat <- data.matrix(prostate[, -which(names(prostate) == "lpsa")])
ridgeOut <- glmnet(x = xmat, y = lpsa, family = "gaussian", alpha = 0, nlambda = 200)
plot(ridgeOut, xvar = "lambda", label = TRUE)

# problem 3e
set.seed(577)
cv_ridge <- cv.glmnet(x = xmat, y = lpsa, family = "gaussian", alpha = 0, nlambda = 200)
plot(cv_ridge)

# Problem 3f
optimal_lambda_ridge <- cv_ridge$lambda.min
cv_mse_ridge <- min(cv_ridge$cvm)
optimal_lambda_ridge
cv_mse_ridge
# problem 3g
lassoOut <- glmnet(x = xmat, y = lpsa, family = "gaussian", alpha = 1, nlambda = 200)
plot(lassoOut, xvar = "lambda", label = TRUE)

set.seed(577)
cv_lasso <- cv.glmnet(x = xmat, y = lpsa, family = "gaussian", alpha = 1, nlambda = 200)
plot(cv_lasso)
optimal_lambda_lasso <- cv_lasso$lambda.min
cv_mse_lasso <- min(cv_lasso$cvm)
optimal_lambda_lasso
cv_mse_lasso

# problem 3j
lasso_fixed <- glmnet(x = xmat, y = lpsa, family = "gaussian", alpha = 1, lambda = 0.1)
coef_lasso <- coef(lasso_fixed)
coef_mlr <- coef(lmOut)
comparison <- cbind(as.matrix(coef_lasso), coef_mlr)
print(comparison)

# problem 4a
library(MASS)
data("Boston")

lm_model <- lm(crim ~ ., data = Boston)
lm_cv <- train(crim ~ ., data = Boston, method = "lm", trControl = trainControl(method = "cv",
    number = 5))
```

```r
mse <- mean(lm_cv$residuals^2)
print(paste("Minimum CV MSE for Ridge:", mean(cv_ridge$cvm)))

x <- model.matrix(crim ~ ., Boston)[, -1]  # remove intercept column
y <- Boston$crim

cv_ridge <- cv.glmnet(x, y, alpha = 0, nlambda = 100)
optimal_lambda_ridge <- cv_ridge$lambda.min
print(paste("AVG CV MSE for Ridge:", mean(cv_ridge$cvm)))


cv_lasso <- cv.glmnet(x, y, alpha = 1, nlambda = 100)
optimal_lambda_lasso <- cv_lasso$lambda.min
print(paste("AVG CV MSE for LASSO:", mean(cv_lasso$cvm)))
```