Individual Outline

Coco Kusiak

11/30/2017

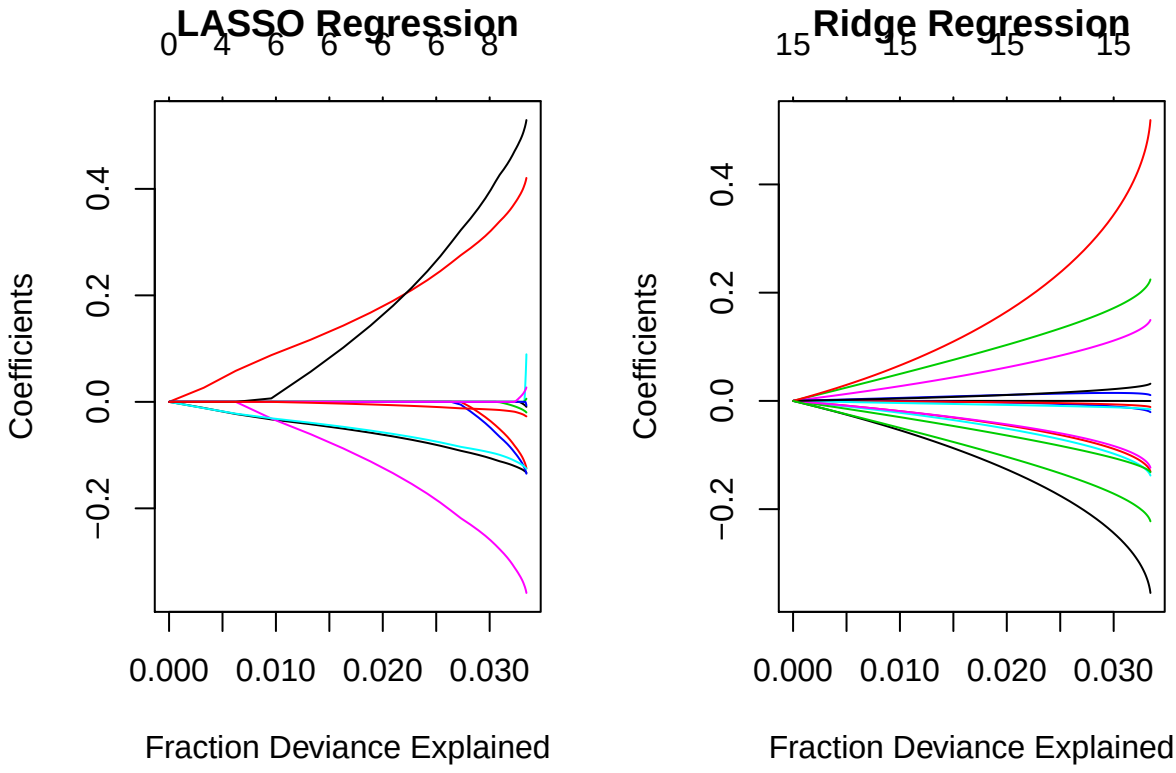
Overview

The goal of my project is to test the performance of the different model selection algorithms lasso, ridge and stepwise regressions. I will begin by using each of these models to find the best model for our data. From there I will choose which model has the best fit based on RMSE and AUC. I will use this model to simulate our outcome variable `mental` 1,000 times. Next, I will run the model selections algorithms again on each simulated set of the outcome. Finally I will compare the AUCs for each of these models again to determine overall which has the best performance.

Step 1 Running the LASSO, ridge, and stepAIC algorithms

The first two methods we will use for model specification is Least Absolute Shrinkage Selection Operator (LASSO) regression and ridge regression. In addition to minimizing the residual some of squares, these methods penalize covariates in our model based on some constraints.

For ridge regression this can be written as: $\sum_{j=1}^{p} \beta_j^2 < c$ with p = # of predictors and c a constant. For LASSO, this can be written as $\sum_{j=1}^{p} |\beta_j| < c$. Both methods were run on our data.



As shown in the plots above, the LASSO method shrinks some of the variables' coefficients down to zero, removing them from the model. The ridge method shrinks coefficeints as well, but only to be **close** to 0.

The final selection algorithm we will use is stepwise regression. This method begins with a set of candidate predictors and adds and removes them based on improving the model's AIC. This is implemented on our data in both the forward and backward directions.

	LASSO	Ridge	AIC
(Intercept)	0.77	0.73	0.65
state	0.00	-0.00	-0.00
sleep	-0.12	-0.09	-0.14
sex	0.37	0.17	0.46
employ	-0.01	-0.01	-0.02
sexorient	-0.09	-0.09	0.00
trans	0.00	0.11	0.00
actlimit	-0.31	-0.23	-0.37
metro	0.00	-0.01	-0.01
exer30	-0.08	-0.10	-0.13
race	0.00	0.01	0.01
bmi	0.00	-0.01	0.00
income	-0.11	-0.08	-0.13
smoker	0.00	0.02	0.03
binge	0.46	0.33	0.54
male	-0.02	-0.17	0.00
Number of Predictors	11.00	15.00	11.00

Table 1: Selection Method Coefficients

The selection algorithms seem to yield fairly similar results. For example, `trans` is weighted very little across the three models and `binge` has a lot of weight across the three. The stepwise and LASSO yield the same bumber of predictors.

Now, to test the performance of each model, we will calculate the Root Mean Squared Error and Area Under the Curve (from ROC) for each.

	Algorithm	RMSE	AUC
1	LASSO	0.55	0.52
2	Ridge Regression	0.56	0.50
3	Step AIC	0.55	0.52

As can be seen in the table above, the stepAIC method has the highest area under the curve and the lowest root mean squared error. This is fairly surprising because stepwise regressions have been proven problematic in many ways suach as having highly biased parameter estimates and R^2 values.

Evenso, we will continue and this model will be used to simulate additional mental health outcomes.

Step 2 Simulation Study

I will add some additional variation to the `sleep`, `employ`, and `income` variables by adding random noise to each observation.

For each simulation:

$$sleep_{i,n} = sleep_n + norm_{i,n}$$

$$employ_{i,n} = employ_n + unif_{i,n}$$

$$income_{i,n} = income_n + unif_{i,n}$$

with $norm_{i,n} \sim N(0,1)$ and $unif_{i,n} \sim Unif(-3,3)$ with $i = 1, ..., I = \#$ of simulations and with $n = 1, ..., N = \#$ of observations.

New outcomes for out response variable `mental` will be calculated using the model selected by the stepwise procedure explained in step 1. This speciication is written below.

The stepwise AIC model:
$mental = 0.651 - 0.0002 * state - 0.136 * sleep + 0.461 * sex - 0.023 * employ - 0.369 * actlimit - 0.012 * metro - 0.122 * exer30 + 0.010 * race - 0.130 * income + 0.033 * smoker + 0.545 * binge$

I'll then run the three model selection algorithms on these new simulated data and compare the AUCs for each set of model specifications. I have run this so far with 4 simulations but am having some trouble adding enough variation in the data to see substantial changes in the AUCs between the true and simulated data.

```
simulated <- subset(ment, select = -c(y.lasso, y.ridge, y.aic))
auc_lasso <- c()
auc_ridge <- c()
auc_aic <- c()

for (i in 1:2){
  trial <- simulated
  randos <- rnorm(n = nrow(trial), mean = 0, sd = 5)
  noise_employ <- sample(x = -3:3, size = nrow(trial), replace = TRUE)
  noise_income <- sample(x = -3:3, size = nrow(trial), replace = TRUE)
  trial <- mutate(trial, sleep = sleep + randos,
                  employ =employ + noise_employ,
                  income  = income + noise_income,
                  binge = 1 - binge)
  trial <- mutate(trial, sleep = ifelse(sleep < 0, yes = 0, no = sleep),
                  income = ifelse(income < 0, yes = 0, no = sleep))
  trial <- mutate(trial, mental = as.factor(ifelse(predict(model.AIC,
                                                   newx = covars,
                                                   type = "response") < .5,
                                           yes = 0, no = 1)))
  covars <- subset(x = trial, select = -mental)
  covars <-as.matrix(covars)
  ys <-as.matrix(trial$mental)
  fit.lasso <- cv.glmnet(covars, ys, family = "binomial", alpha = 1)
  fit.ridge <- cv.glmnet(covars, ys, family = "binomial",  alpha = 0)
  binom <- glm(mental ~ ., data = trial, family = "binomial")
  fit.aic <- stepAIC(binom, direction= "both", trace = FALSE)$formula
  trial <- mutate(trial,
                  y.lasso = as.numeric(predict(fit.lasso, newx = x, type = "class")),
                  y.ridge = as.numeric(predict(fit.ridge, newx = x, type = "class")),
                  y.aic = ifelse(predict(model.AIC, newx = x, type = "response") < .5,
                                           yes = 0, no = 1))
  auc_lasso[i] <- auc(roc(predictor = trial$y.lasso, response = trial$mental))
  auc_ridge[i] <- auc(roc(predictor = trial$y.ridge, response = trial$mental))
  auc_aic[i] <- auc(roc(predictor = trial$y.aic, response = trial$mental))
}
comparisons <- data.frame(`Mean AUC` = c(mean(auc_lasso), mean(auc_ridge), mean(auc_aic)),
                          `AUC Variance` = c(sd(auc_lasso), sd(auc_ridge), sd(auc_aic)))
rownames(comparisons) <- c("LASSO", "Ridge", "Stepwise AIC")
xtable(comparisons)
```

	Mean.AUC	AUC.Variance
LASSO	0.66	0.00
Ridge	0.67	0.00
Stepwise AIC	1.00	0.00

Conclusions

None too surprisingly, the average AUC for the stepwise regression is almost perfect because the model chosen by this algorithm before was used to simulate the additional data. The ridge regression method has the lowest AUC across the 3 models.

Note I plan on adding more to this conclusion once I am able to complete this for more iterations!