**ECMT 670 EXAM 2** **(Instructions)**

The file Exam.R contains codes to load data from examdata.csv. You just need to put the code file and data file in the same folder. The sample is split into a training sample and a test sample. There is one model, linear model, already written in the codes. I would like you to come up with your own candidate models.

In each model, the estimation is performed using the training sample. If there are tuning parameters, you need to pick a criterion for selecting the tuning parameters. The test MSE is calculated using the test sample. The test MSE should be reported along with the candidate model’s name, the selected tuning parameters, and the corresponding criterions.

Each of you need to submit your own codes and a report of all the test MSEs via Canvas. (The report could be either a word or pdf file.)

Your score depends on the performance of your best model, the variety of your candidate models, as well as a short summary of your findings.

The deadline is Saturday, December 4th, at 11:59 pm.

**(Results)**

The model with the best test MSE was a boosted tree model which is presented last.

1. The linear model already written in the code produced a test MSE of 13.435.
2. From preforming a best subset selection, a model with five variables(x01-x05) has the lowest BIC and a model with 7 variables (x01-x05, x09, x15) has the lowest Cp.

After testing the two, the test MSE for the 5 variable model produced a better MSE of 12.792675859

1. After fitting a PCR and PLS model with M chosen by cross-validation, the test MSE was worse at 13.77096 for PCR and 13.435 for PLS.
2. Next, I ran a ridge regression with k-fold cross-validation to find the best lambda. The best lambda chosen was 1.242251 and the test MSE was 12.85553
3. After that I ran a lasso regression with k-fold cross validation to get the best lambda. The lambda chosen was 0.2402913 and the test MSE was 12.45728. The lasso regression used x01-x05, x09, x15, x20 as important variables and set the rest to 0.

22 x 1 sparse Matrix of class "dgCMatrix"

(Intercept) -1.11266702

(Intercept) .

x01 0.55543402

x02 0.95558730

x03 0.94523961

x04 0.95253985

x05 1.12146593

x09 -0.16927517

x15 -0.19585467

x20 0.02672137

1. Fitting a GAM model with just the first 5 variables, what is determined from the best subset selection, gives a worse MSE of 12.77296. All five variables are important variables in predicting y and all five are linear.

Anova for Parametric Effects

Df Sum Sq Mean Sq F value Pr(>F)

s(x01, df = 2) 1 177.44 177.44 12.658 0.0004728 \*\*\*

s(x02, df = 2) 1 495.35 495.35 35.337 1.314e-08 \*\*\*

s(x03, df = 2) 1 254.46 254.46 18.152 3.214e-05 \*\*\*

s(x04, df = 2) 1 288.41 288.41 20.574 1.017e-05 \*\*\*

s(x05, df = 2) 1 392.75 392.75 28.018 3.314e-07 \*\*\*

Residuals 189 2649.40 14.02

---

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

Anova for Nonparametric Effects

Npar Df Npar F Pr(F)

(Intercept)

s(x01, df = 2) 1 0.78163 0.3778

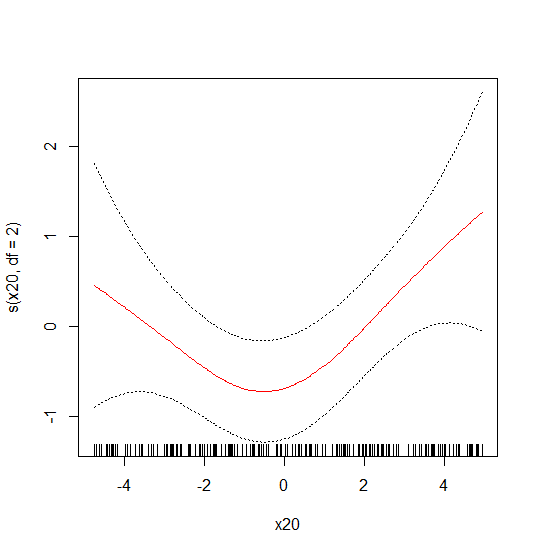
s(x02, df = 2) 1 0.82450 0.3650

s(x03, df = 2) 1 0.23483 0.6285

s(x04, df = 2) 1 0.82202 0.3657

s(x05, df = 2) 1 1.05766 0.3051

1. Fitting a GAM model with the same variables as the Lasso model shows that x20 is not linear. The variable x20 changes directions at what appears to be 0.



This produces a test MSE of 12.08347 which is lower than the lasso model. Although ANOVA shows that variables x09 and x15 are not as significantly important.

Anova for Parametric Effects

Df Sum Sq Mean Sq F value Pr(>F)

s(x01, df = 2) 1 178.56 178.56 13.5126 0.0003112 \*\*\*

s(x02, df = 2) 1 500.96 500.96 37.9112 4.576e-09 \*\*\*

s(x03, df = 2) 1 256.47 256.47 19.4086 1.793e-05 \*\*\*

s(x04, df = 2) 1 287.30 287.30 21.7417 5.999e-06 \*\*\*

s(x05, df = 2) 1 353.15 353.15 26.7249 6.109e-07 \*\*\*

s(x09, df = 2) 1 34.46 34.46 2.6077 0.1080692

s(x15, df = 2) 1 56.74 56.74 4.2941 0.0396472 \*

s(x20, df = 2) 1 16.29 16.29 1.2326 0.2683547

Residuals 183 2418.18 13.21

---

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

Anova for Nonparametric Effects

Npar Df Npar F Pr(F)

(Intercept)

s(x01, df = 2) 1 0.5318 0.466780

s(x02, df = 2) 1 0.8333 0.362526

s(x03, df = 2) 1 0.3392 0.561013

s(x04, df = 2) 1 0.7379 0.391452

s(x05, df = 2) 1 0.8716 0.351739

s(x09, df = 2) 1 2.3418 0.127671

s(x15, df = 2) 1 0.4239 0.515825

s(x20, df = 2) 1 8.2297 0.004606 \*\*

---

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

1. After doing this with removing either and both variables (x09 and x15), the best model is found with variables (x01-x05, x20). The test MSE is 11.75981.

Anova for Parametric Effects

Df Sum Sq Mean Sq F value Pr(>F)

s(x01, df = 2) 1 174.97 174.97 12.9303 0.0004136 \*\*\*

s(x02, df = 2) 1 502.28 502.28 37.1188 6.201e-09 \*\*\*

s(x03, df = 2) 1 260.32 260.32 19.2379 1.924e-05 \*\*\*

s(x04, df = 2) 1 291.73 291.73 21.5594 6.447e-06 \*\*\*

s(x05, df = 2) 1 349.40 349.40 25.8208 9.037e-07 \*\*\*

s(x20, df = 2) 1 24.96 24.96 1.8446 0.1760436

Residuals 187 2530.41 13.53

---

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

Anova for Nonparametric Effects

Npar Df Npar F Pr(F)

(Intercept)

s(x01, df = 2) 1 0.6679 0.414824

s(x02, df = 2) 1 1.0621 0.304075

s(x03, df = 2) 1 0.2982 0.585682

s(x04, df = 2) 1 0.6944 0.405738

s(x05, df = 2) 1 1.0846 0.299017

s(x20, df = 2) 1 7.0340 0.008686 \*\*

---

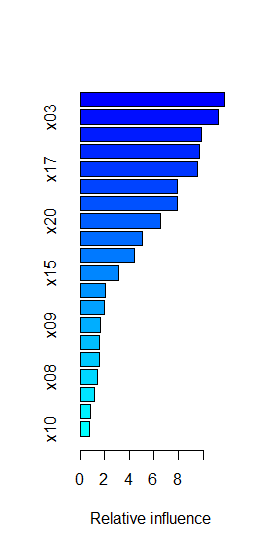
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

1. Finally, the last method was to use boosting to predict y. Boosting with 1000 trees for a range of lambda values to find the lambda (0.0199526) with the best test MSE of 10.40531. The relative influence is shown below.

var rel.inf

x05 x05 11.7467656

x03 x03 11.2632521

x02 x02 9.8867497

x04 x04 9.7030163

x17 x17 9.5206071

x18 x18 7.9384648

x19 x19 7.8856786

x20 x20 6.5143870

x01 x01 5.0510431

x16 x16 4.4538587

x15 x15 3.1133462

x07 x07 2.0307552

x14 x14 1.9463625

x09 x09 1.6321654

x06 x06 1.5948902

x12 x12 1.5542220

x08 x08 1.3917768

x13 x13 1.1386073

x11 x11 0.8711736

x10 x10 0.7628777