Report 3

Vanessa Machuca and Luis Espino

May 1, 2018

1 Task 1 - Sparse and Smooth Linear Models

Introduction

The data we are working with looks at student achievement in Portuguese course at two secondary education Portuguese schools. We found the data through the UCI Machine Learning Repository. The observational units for the dataset are students. The dataset includes demographic variables like studentâĂŹs school name (binary), sex (binary), age (numeric), and whether the student lives in an urban or rural area.

It also includes variables that relate to family, including family size, parents cohabitation status (living apart or together), mother and father's education (ordinal), mother and fatherâĂŹs job ('teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other'), studentâĂŹs guardian (mother, father or other), and quality of family relations (1 being very bad to 5 being excellent).

School-related variables include home to school travel time (ordinal), weekly study time (ordinal), and number of past class failures (1-3 if $n < 3$, 4 otherwise). For each of two courses, students were given a graded after the first and second period as well as a final grade. Academic routine, habits and performance were also measured. These variables included travel time from home to school (nominal), weekly study time (nominal), number of past class failures (nominal), reason to choose school (categorical), and number of absences (numeric). Additionally, educational resources were assessed through binary measurements of extra educational support, family educational support, extra paid classes within the course subject, extracurricular activities, attended nursery school, college aspiration, and home internet access.

Finally, nominal variables that measured more personal aspect of life included frequency of outings with friends, workday alcohol consumption, weekend alcohol consumption, current health status, and being in a romantic relationship (binary).

RR and LASSO

We will now run ridge regression (RR) and LASSO on the full varaible set, using cross validation to find lambda, and compare them. First up, RR.

```
[1] 0.0225702
```

We omit categorical variables, as RR does not seem to like them very much. The lambda value which yields the lowest SSE is 0.0225702. Onward to LASSO.

```
[1] 0.03125716
```

The lambda value which yields the lowest SSE is 0.03125716. Now to compare these two models by looking at their coefficients. Previously, we looked at plots of cefficients.

MLR

```
> fwdselec <- lm(G3 ~ failures + school + sex + higher + health + Mjob, data=student_por_n)
> summary(fwdselec)

Call:
lm(formula = G3 ~ failures + school + sex + higher + health +
```

```
    Mjob, data = student_por_n)

Residuals:
     Min       1Q   Median       3Q      Max
-12.1608  -1.5552  -0.0308   1.6777   6.9644

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.72992    0.49645  23.628  < 2e-16 ***
failures     -1.55182    0.19375  -8.009 5.47e-15 ***
schoolMS     -1.47695    0.23563  -6.268 6.72e-10 ***
sexM         -0.78110    0.22634  -3.451 0.000595 ***
higheryes     1.95805    0.37591   5.209 2.57e-07 ***
health       -0.21861    0.07581  -2.884 0.004062 **
Mjobhealth    1.18435    0.47486   2.494 0.012878 *
Mjobother     0.17151    0.29723   0.577 0.564130
Mjobservices  0.50851    0.34646   1.468 0.142676
Mjobteacher   1.04287    0.42005   2.483 0.013294 *
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 2.75 on 639 degrees of freedom
Multiple R-squared:  0.2855,        Adjusted R-squared:  0.2755
F-statistic: 28.37 on 9 and 639 DF,  p-value: < 2.2e-16
```

Pairs plot comparing LASSO and RR.

```
> coefols <- as.numeric(coef(fwdselec)[c(-1)])
> coef.ridge <- as.numeric(coef(pr.ridge.cv, s="lambda.min")[c(-1)])
> coef.lasso <- as.numeric(coef(pr.lasso.cv, s="lambda.min")[c(-1)])
> compare <- cbind(coefols,coef.ridge,coef.lasso)
> pairs(compare)
```

The pairs plot shows that the linear model coefficients donâĂŹt have high correlation with either the ridge regression or lasso coefficients. Most nonzero constants for OLS are very close to zero for ridge regression, and the biggest ridge regression coefficients are not the biggest OLS coefficients. The coefficient plot between lasso and OLS is fairly similar to the coefficient plot between ridge regression and OLS, except the coefficients hug around the line x=0 tighter. Between ridge regression and lasso, most of the coefficients that hug around 0 (on both sides of 0) for ridge are positive and even closer to 0 for lasso. For both plots, only three coefficients seem to be visibly larger relative to all the other ones.

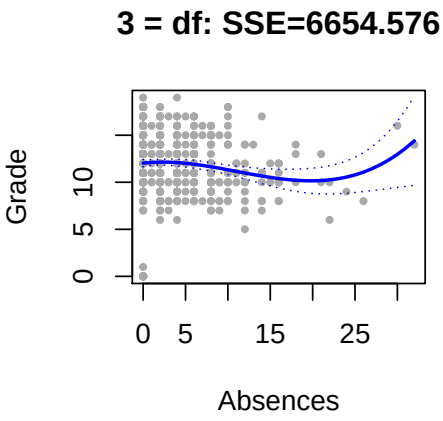
Below is a plot comparing responses predicted using MLR, RR, and LASSO.

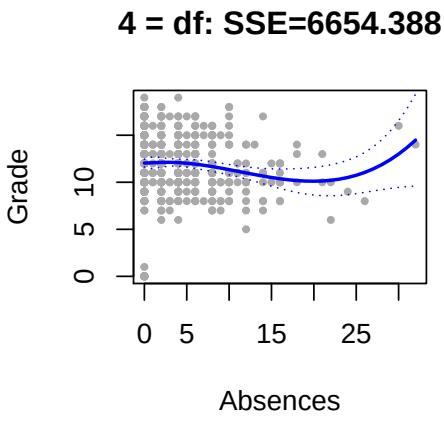
```
[1] 4832.25
```

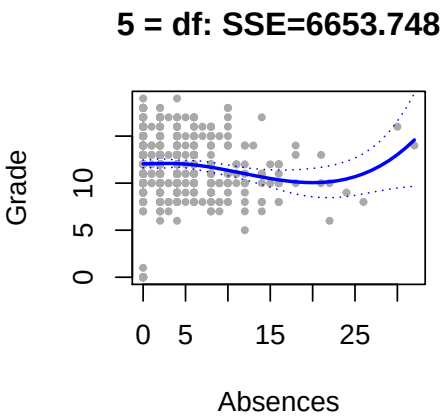
```
[1] 992.5012
```

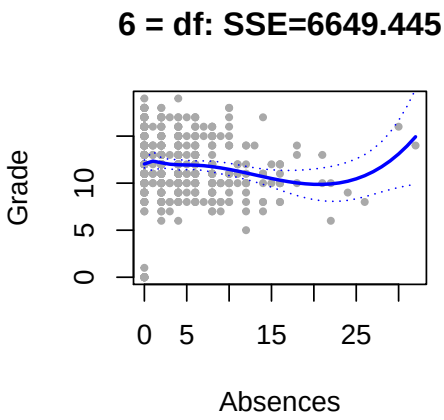
```
[1] 1001.052
```

Now to choose one variable and run smoothing spline and kernel smoother models on it. We change the parameters so that we have four different models for each method. We chose "absences", as this was found to be the most significant variable in report 1. Let us first use regression splines for four different degrees of freedom: 3,4,5, and 6

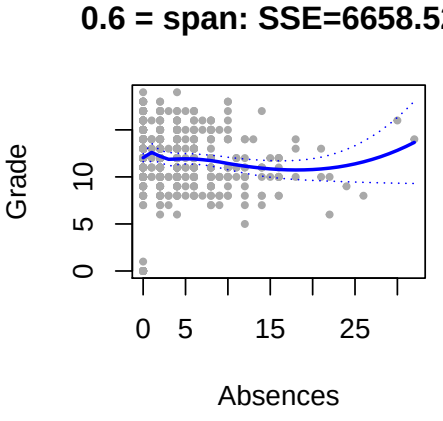


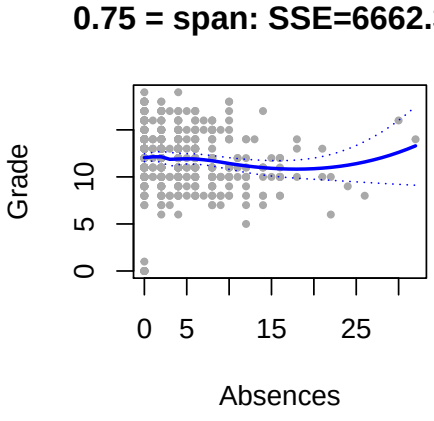


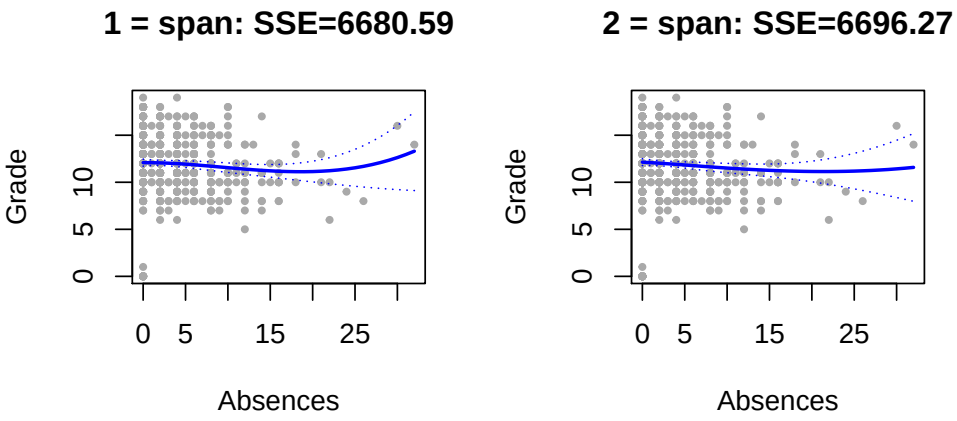




And now onward to LOESS.







Without cross validating, we would choose the best model. Here's why. Now to summarize our results for this section. In it, we will comment on anything of interest that occurred. Did anything surprise us?

2 Task 2 - Something New

3 Task 3 - Summary