OLLSCOIL NA hEIREANN, CORCAIGH
THE NATIONAL UNIVERSITY OF IRELAND, CORK

COLAISTE NA hOLLSCOILE, CORCAIGH
UNIVERSITY COLLEGE, CORK

ST4060 - ST6015 - ST6040

Continuous assessment 2 - 2020-21

Eric Wolsztynski
`eric.w@ucc.ie`

List of (possibly) useful R functions:

```
apply()
c()
cbind()
coef()
colnames()
cut()
cv.glmnet()
data.frame()
fitted()
glmnet()
library()
lm()
mean()
model.matrix()
na.omit()
nrow()
numeric()
plot()
points()
predict()
rbind()
round()
sample()
set.seed()
summary()
which()
```

Question 1

Note: if you do not manage to answer a question item, provide the R code you would have used, or a comment on the answer you would expect for that question, as relevant.

Consider R's dataset `airquality`. For this question, remove any observation with missing values by running the R instruction:

```
dat = na.omit(airquality)
```

and use `dat` as the dataset in what follows. Run `set.seed(4060)` before executing your R code for this question.

(a) Fit a model of the form

$$Y = a + bX_1 + cX_1^2 + \epsilon$$

to the whole dataset, where X_1 and Y are respectively variables `Wind` and `Ozone`.

(i) Quote the R expression you have used to fit this model (only one line of code is required).

(ii) Quote the coefficient estimates for this model.

(iii) Quote the Mean Squared Error (MSE) corresponding to the model fit.

(b) Fit a model of the form

$$Y = a + bX_1 + cX_2 + \epsilon$$

to the whole dataset, using both predictors $X_1 =$ `Wind` and $X_2 =$ `Temp` to explain observations $Y =$ `Ozone`.

(i) Quote the R expression you have used to fit this model (only one line of code is required).

(ii) Quote the coefficient estimates for this model.

(iii) Quote the Mean Squared Error (MSE) corresponding to the model fit.

(c) Fit the lasso model (using `library(glmnet)`) to the whole dataset, using both predictors $X_1 =$ `Wind` and $X_2 =$ `Temp` to explain observations $Y =$ `Ozone`.

(i) Quote the value of regularization parameter you used in fitting the model, and show the R code you used to obtain that value.

(ii) Quote the coefficient estimates for this model.

(iii) Quote the Mean Squared Error (MSE) corresponding to the model fit.

(d) Fit a ridge regression model with regularization parameter $\lambda = 0.21$ to the whole dataset, using both predictors $X_1 =$ `Wind` and $X_2 =$ `Temp` to explain observations $Y =$ `Ozone`.

(i) Quote the coefficient estimates for this model.

(ii) Quote the Mean Squared Error (MSE) corresponding to the model fit.

(e) Create a scatterplot to illustrate this analysis, showing the (X_1, Y) data points as black dots (we omit X_2 here in order to create a simple 2D plot), and the fitted values obtained in (b) and (c) respectively in blue and red.

(f) Compare the three model fits obtained in (b), (c) and (d) above, in terms of MSE or any other aspect. Does one of these models outperform the others? Why? What does this indicate about the original multilinear model of (b)?

Solution:

(a) MSE should be 542.64

(b) MSE should be 459.36

(c) MSE should be 460.09 (and $\lambda = 0.7403$ in (i))

(d) MSE should be 459.37

(e) plot as per R code below

(f) Here thew simple linear model is appropriate to describe the pattern in the data. Regularisation is not yielding any improvement (this is not very surprising given the small number of covariates).

R code:

```r
library(glmnet)
set.seed(4060)

dat = na.omit(airquality)
y = dat$Ozone
x1 = dat$Wind
x2 = dat$Temp

# (a)
mod.a = lm(y~x1+I(x1^2))
coef(mod.a)
(mse.a = mean( (mod.a$residuals)^2 ))
mean( (fitted(mod.a)-y)^2 ) # either works

# (b)
mod.b = lm(y~x1+x2)
coef(mod.b)
(mse.b = mean( (mod.b$residuals)^2 ))
mean( (fitted(mod.b)-y)^2 ) # either works

# (c)
cv.c = cv.glmnet(cbind(x1,x2), y, alpha=1)
(l.min = cv.c$lambda.min)
mod.c = glmnet(cbind(x1,x2), y, alpha=1, lambda=l.min)
coef(mod.c)
yhat.c = predict(mod.c, newx=cbind(x1,x2))
(mse.c = mean( (yhat.c-y)^2 ))

# (d)
mod.d = glmnet(cbind(x1,x2), y, alpha=0, lambda=0.21)
coef(mod.d)
yhat.d = predict(mod.d, newx=cbind(x1,x2))
(mse.d = mean( (yhat.d-y)^2 ))

# (e)
plot(x1, y, pch=20)
points(x1, fitted(mod.b), col=4, pch=15)
points(x1, yhat.c, col=2, pch=20)
## NB: checking 3d hyperplane:
# library(scatterplot3d) # loads in current R environment
# out = scatterplot3d(cbind(x1, x2, y), pch=20)
# out$plane3d(mod.b, draw_polygon=TRUE, draw_lines=FALSE)

# (f)
c(mse.b, mse.c, mse.d)
```

Question 2

Table 1 and Figure 1 capture the output of 10-fold cross-validation of two distinct multilinear models applied to a dataset with $N = 263$ observations of major baseball league players on $P = 19$ variables.

(a) Indicate what errors **A**, **B**, **C** and **D** are likely to quantify in this comparative analysis, and why. Include an indication of which model each of these four quantities could relate to and why.

(b) Provide a possible explanation for the greater variances observed for distributions **C** and **D** compared to those of **A** and **B**.

(c) Name two multilinear models that could yield these results, and why.

	A	**B**	**C**	**D**
K=1	0.34	0.36	0.57	0.50
K=2	0.33	0.35	0.58	0.53
K=3	0.36	0.38	0.26	0.24
K=4	0.36	0.38	0.28	0.30
K=5	0.33	0.34	0.55	0.60
K=6	0.35	0.37	0.40	0.39
K=7	0.36	0.38	0.27	0.26
K=8	0.34	0.35	0.50	0.53
K=9	0.34	0.36	0.42	0.41
K=10	0.35	0.37	0.39	0.38
Mean	**0.34**	**0.36**	**0.42**	**0.41**

Table 1: Mean squared errors obtained from K-fold cross-validation of the two distinct multilinear models.

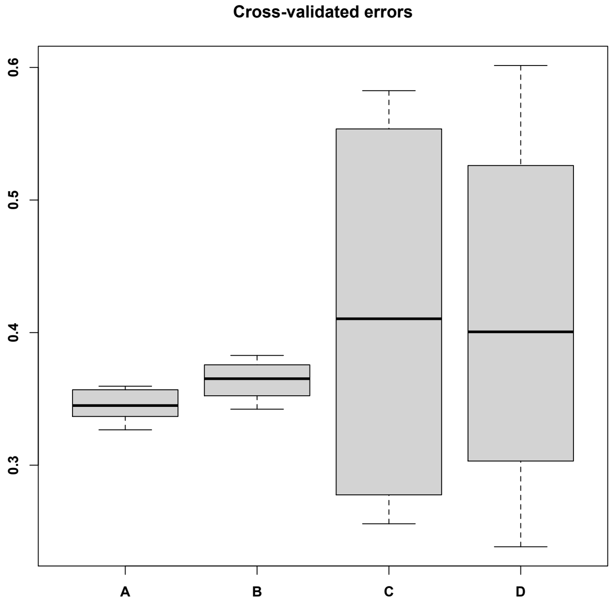

Figure 1: Distributions of cross-validated mean squared errors shown in Table 1.

Solution:

Here, we know that there are 2 models, that models these are *linear* and that they each use $P = 19$ variables X to describe $N = 263$ observations Y. We also know that they are trained and test via 10-fold cross-validation.

(a) Error sets A and B clearly have lower variance and lower mean and median that error sets C and D (error means/medians being between about 14% and 24% higher in C and D than in A and B.) This is typical of a difference between training error and test error in cross-validation, in particular where overfitting is present. Since 2 models are considered, we are therefore looking at train and test set errors for each model. So a reasonable proposal is as follows:

A: training set error for model 1.

B: training set error for model 2.

C: test set error for model 1.

D: test set error for model 2.

Note you could use "validation" instead of 'test' interchangeably, as this convention is rather arbitrary.

(b) Larger variance corresponds to smaller sample size used in the test set samples (the validation folds).

(c) Based on the above, and observing that $mean_A < mean_B$, and that $(mean_C - mean_A) > (mean_B - mean_D)$, we could reasonably consider that overfitting has been reduced with model 2 compared to model 1; hence some reasonable proposals are as follows:

- Model 1 is an ordinary linear regression model, and model 2 a regularized linear model, such as lasso, ridge regression, or elastic net for example (since the purpose of regularization is to reduce overfitting).
- Models 1 and 2 are two different regularization models, one being more effective than the other (if one considers that the different between train and test errors for either model is acceptable, and does not indicate significant overfitting – which is arguable).
- Models 1 and 2 are distinct evaluations of a same regularized model using different regularization parameter values (with a more effective calibration of the regularization parameter in the second model).

Amy of the above proposals constitute a correct answer. Reasonable alternative proposals can be accepted as correct answers so long as they are properly justified.