Lasso regression for Mathematics

schoolMS	.
sexM	0.817565729
age	-0.132818044
addressU	0.304060079
famsizeLE3	0.409395784
PstatusT	-0.114561905
Medu	0.253366503
Fedu	.
Mjobhealth	1.201629335
Mjobother	.
Mjobservices	0.853817716
Mjobteacher	.
Fjobhealth	0.090248757
Fjobother	-0.008503719
Fjobservices	.
Fjobteacher	0.716317798
reasonhome	.
reasonother	0.385420721
reasonreputation	0.305236995
guardianmother	.
guardianother	.
traveltime	-0.130990747
studytime	0.257531647
failures	-1.661880666
schoolsupyes	-0.825500102
famsupyes	-0.460077155
paidyes	0.059052807
activitiesyes	.
nurseryyes	.
higheryes	1.022163827
internetyes	0.198002688
romanticyes	-0.716109590
famrel	0.046775346
freetime	0.077935168
goout	-0.361925108
Dalc	.
Walc	.
health	-0.083314647
absences	0.028320085

rmse / mean = 0.4099021 = 41%

Lasso regression for Portuguese

schoolMS	-1.169855763
sexM	-0.541573048
age	0.125430588
addressU	0.271464984
famsizeLE3	0.227962892
PstatusT	0.066550266
Medu	0.056463333
Fedu	0.135308342
Mjobhealth	0.647768858
Mjobother	-0.001435661
Mjobservices	0.241713393
Mjobteacher	0.333843234
Fjobhealth	-0.219519785
Fjobother	.
Fjobservices	-0.368704030
Fjobteacher	0.712918611
reasonhome	.
reasonother	-0.426584298
reasonreputation	0.185843092
guardianmother	-0.306121785
guardianother	.
traveltime	.
studytime	0.396403997
failures	-1.387835893
schoolsupyes	-1.224305008
famsupyes	.
paidyes	-0.276345736
activitiesyes	0.163176843
nurseryyes	-0.126035046
higheryes	1.701942399
internetyes	0.239014556
romanticyes	-0.367335308
famrel	0.120839565
freetime	-0.117894039
goout	-0.041381922
Dalc	-0.201984863
Walc	-0.083216544
health	-0.170237680
absences	-0.031423576

rmse / mean = 0.2309279 = 23%

Part 2

I performed Lasso regression. Displayed above are the coeffecients for each predictor.

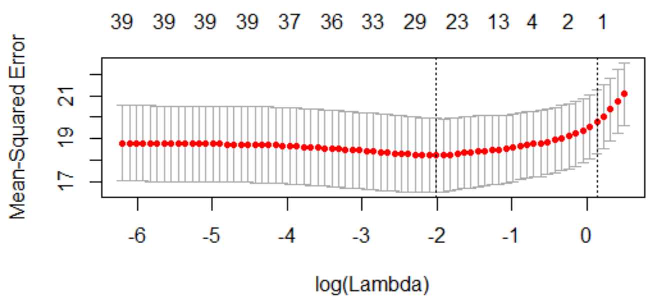
Prediction accuracy for "final grade"

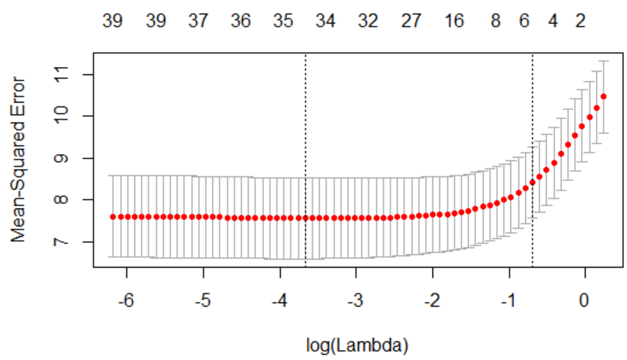
Mathematics: the model is quite poor, with a root-mean-squared-error of 41% of the sample mean of the final grade.

Portuguese: the model is better than the one for Mathematics, with a root-mean-squared-error of 23% of the sample mean of the final grade.

<u>Interpretation</u>

There are 4 nominal variables. “mother’s job” and “father’s job” have “at home” as the base case, so the relevant coefficients are relative to the “at home” case. E.g. a Maths student whose mother works in “health”, is predicted to score 1.2 more points on the Maths test, than if the student’s mother worked “at home”. “Reason” has course preference as the base case. “Guardian” has “father” as base case. Similarly, binary variables have “no” as the base case.





Above are the plots of cross-validation Mean-Squared Error vs log(Lambda) for Maths and Portuguese respectively. Each log(Lambda) value corresponds to a unique Lasso model. The left vertical dotted line indicates the log(Lambda) value of the point with least cross-validation error. The right dotted line indicates the smallest model (the one with least predictors) whose error is within one standard error from the model indicated by the left dotted line.

The wide range of values for log(Lambda) between the two dotted lines, shows that there are many possible models each with different number of predictors, that produce similarly accurate predictions. This means that the coefficients should not be interpreted too seriously, because the Lasso regression does not give us much confidence in the truthfulness of any chosen model corresponding to a particular number of predictors. E.g. for Mathematics, the model at the right dotted line produces a model with only one predictor (“failures”).

Logistic regression for bank data

age	-0.0001415071
jobblue-collar	-0.2652712116
jobentrepreneur	-0.0624738735
jobhousemaid	-0.1808366518
jobmanagement	0.0150510370
jobretired	0.6058998363
jobself-employed	.
jobservices	.
jobstudent	0.4458274979
jobtechnician	-0.0492008822
jobunemployed	-0.4024502777
jobunknown	0.3975623497
maritalmarried	-0.3396536185
maritalsingle	-0.1026129203
educationsecondary	.
educationtertiary	0.2318290735
educationunknown	-0.3581899935
defaultyes	0.4082144270
balance	.
housingyes	-0.2237008673
Loanyes	-0.5561009736
contacttelephone	-0.0167035130
contactunknown	-1.2850063330
day	0.0113335586
monthaug	-0.1744552316
monthdec	0.1282695945
monthfeb	0.1925716377
monthjan	-0.8497269008
monthjul	-0.5958861945
monthjun	0.4933176625
monthmar	1.5320362249
monthmay	-0.3976177022
monthnov	-0.6670507511
monthoct	1.4138041976
monthsep	0.6826810408
duration	0.0040917507
campaign	-0.0608291238
pdays	.
previous	.
poutcomeother	0.4502936237
poutcomesuccess	2.3777377588
poutcomeunknown	-0.0985305483

Confusion matrix

	bankPred	
	no	yes
no	3919	81
yes	344	177

• cross-validated Misclassification error = 0.09666003 = 9.7%

Part 3

I performed cross-validated logistic regression with regularisation.

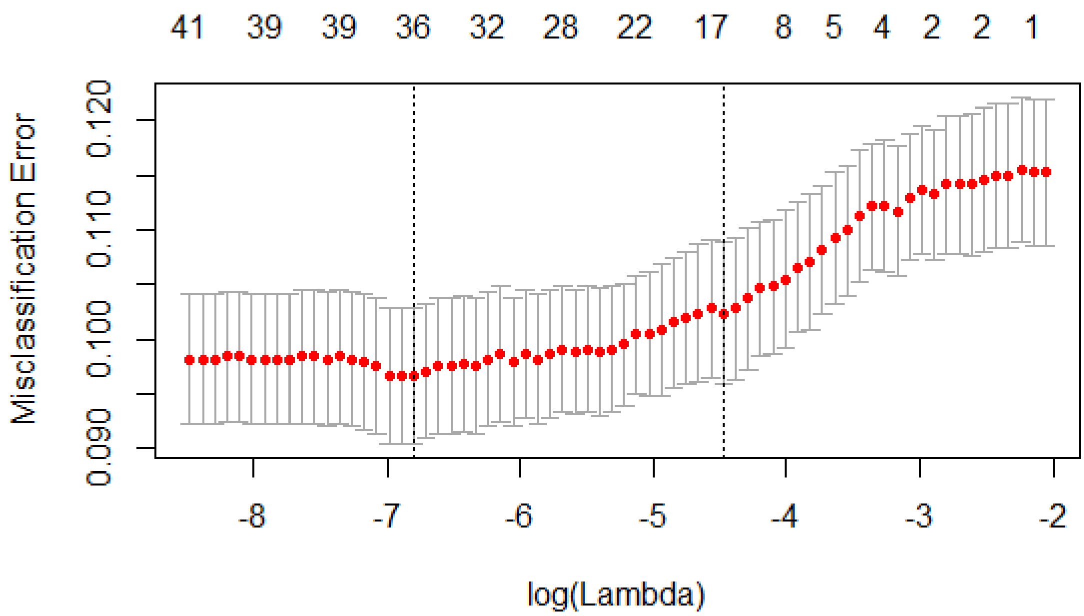
Prediction accuracy

The cross-validated misclassification error of 9.7% is quite low, which means that using the above model, about 9.7% of our predictions will be wrong.

• positive predictive value = 177 / (81+177) = 0.6860465 = 69%

• subscription rate estimate (from sample) = 521/4521 = 0.11524 = 11.5%

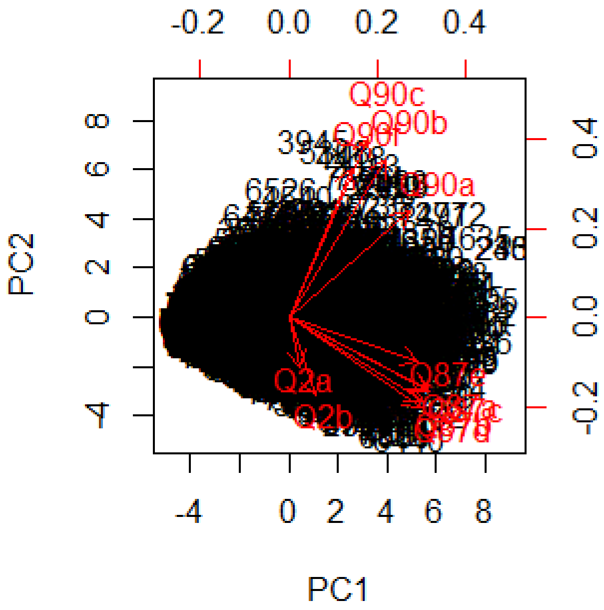
Using the confusion matrix, we can see that out of the people we predict to be subscribers, we expect 69% of them will truly subscribe. This is much better than the expected 11.5% (estimated from sample) subscription rate if we asked people at random. This means that using the predictive model, resources put towards getting new subscribers will be well spent indeed (at 69% success rate).



<u>Interpretation</u>

The variables with no coefficient ("."), are found to compromise the model's predictive accuracy, and hence are dropped from the logistic regression. The x-axis denotes a range of models with different regularisation values ("Lambda"). The top row of the diagram indicates the number of predictors used. The model at the left dotted line is the one we have chosen here, since it produces the lowest cross-validated misclassification error. As the graph indicates, there is a whole range of possible models ranging from one with 36 predictors (the one we've chosen), to one with as few as 17 predictors, that give similar predictive accuracies. So we should not make too much of any one partiular variable.

A negative coefficient means that higher values of the variable are associated with proportionately lower chances of subscrition. A positive coefficient means that higher values of the variable are associated with proportionately higher chances of subscrition. (Keeping the values of other variables constant.) E.g. given two otherwise identical people, the retiree is e^0.6 = 1.8 times more likely to subscribe than the non-retiree. E.g.(2) the person who has personal loan is 1/e^0.556 = 0.57 = 57% as likely to subscribe as the person who has no personal loan.



Part 1

The subquestions are all highly correlated within the main questions of Q87 and Q90 (vectors are long and in the same direction). As seen in the above principal components analysis biplot: the direction of the (principal component loading) vectors {Q87a,Q87b,…} are nearly identical. Same for {Q90a,Q90b,…}.

This is reasonable because Q87 concern positive traits about general life, while Q90 are about positive traits of the workplace. i.e. within the main question, the subquestions are very similar.

It is interesting to observe that general life (Q87) and workplace (Q90) are almost perfectly orthogonal (have no association with each other). This strongly suggests that people tend to compartmentalise (mentally separate) their assessment of general life and assesment of workplace.

We should not make too much of Q2a and Q2b, because of their short vector lengths from this biplot perspective. Though, we can try to note the weak positive association between the Q2 variables and Q87, i.e. being female / older is associated with better general life. Simarly, a weak negative associataion between Q2 and Q90, i.e. being female/older is associated with worse workplace. This does confirm some stereotypes about females / older people being less suited to the workplace, while tending to have a more positive attitude in general life.