Machine Learning in Economics
Assiat Zhassaganbergen
Home assignment I

Question 1.
Part a)
The minimized expression in case of the LASSO regression is:

$$\Sigma(y_i - \beta_i)^2 + \lambda\sum|\beta_i|$$

The minimized expression in case of the ridge regression is:

$$\Sigma(y_i - \beta_i)^2 + \lambda\sum|\beta_i{}^2|$$

Part b)
The minimization problem in case of the ridge regression:

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\Sigma(y_i - \beta_i)^2 + \lambda\sum|\beta_i| = -2\Sigma(y_i - \beta_i) + \lambda\sum sign(\beta_i) = 0$$

Here, the sign(β_i) is the sign function of β_i. After solving for β_i, we get the following:

$$\beta = sign(y) * max(|y| - \frac{\lambda}{2}, 0)$$

So, in LASSO, the expression above indicates how estimated $\beta - s$ depend on λ. As increases, some estimated $\beta - s$ can shrunk to zero and some can be exactly equal to zero.

The minimization problem in case of the ridge regression:

$$\frac{d}{d\beta}\Sigma(y_i - \beta_i)^2 + \lambda\sum|\beta_i{}^2| = -2\Sigma(y_i - \beta_i) + 2\lambda\sum\beta_i = 0$$

After solving for β_i, we get the following:

$$\beta = (\frac{1}{1+\lambda})\, y$$

Here, as λ increases, the estimated $\beta - s$ decrease and get closer to zero.

The solutions to minimization problems show the main difference between LASSO and ridge regression. Even though both are regularization tools, the penalty term in LASSO results in some parameter estimates shrinking exactly to zero and others closer to zero. Meanwhile, the ridge regression shrinks all parameter estimates closer to zero, but without setting any of them to zero.

Question 2.

To calculate the probability that a sample element does not appear in a bootstrap sample for a sample size of *n*, some parameters must be determined first. Let X represent the number of times a sample element appears in a bootstrap sample of size *n*. Since each bootstrap sample is formed by independently sampling with replacement from the original sample, it follows that X follows a binomial distribution with the parameters of *n* and probabilities of *p=1/n* and *q= 1-1/n*.

If the probability that a sample element does not appear in a bootstrap is p(X=0), then we have:

$$P(X=0) = \binom{n}{0} p^0 q^n$$

This results in

$$P(X=0) = (1 - \frac{1}{n})^n$$

Finding the limit of this expression as $n \to \infty$ gives,

$$\lim_{n\to\infty} \left(1 - \frac{1}{n}\right)^n = e^{-1}$$

As can be noted, calculate the probability that a sample element does not appear in a bootstrap sample for a sample size of *n is* e^{-1} or approximately 0.368.

Question 3.

Part a)

First regression model:

$$wage = \beta_0 + \beta_1 BA + \beta_2 female + \beta_3 black + \beta_4 hispanic + u$$

The second regression model:

$$wage = \beta_0 + \beta_1 BA + \beta_2 female + \beta_3 black + \beta_4 hispanic + \beta_5 exper + \beta_6 exper^2 + u$$

The third regression model where observations from small city are used as a base group:

$$wage = \beta_0 + \beta_1 BA + \beta_2 female + \beta_3 black + \beta_4 hispanic + \beta_5 BA\_female + \beta_6 BA\_black + \beta_7 BA\_hispanic + \beta_8 exper + \beta_9 exper^2 + \beta_{10}\text{medcity} + \beta_{11}\text{submed} + \beta_{12}\text{lgcity} + \beta_{13}\text{sublg} + \beta_{14}\text{vlgcity} + \beta_{15}\text{subvlg} + u$$

The results are following:

	Model 1	Model 2	Model 3
intercept	11.747***	7.277***	6.770*
BA	2.304***	2.354***	2.242***
female	-3.144***	-2.667***	-2.610***
black	-1.357***	-1.264***	-1.434***
hispanic	-0.865**	-0.782**	-1.049***
exper		0.048***	0.049***
$exper^2$		-0.0001*	-0.0001*

BA_female			-0.149
BA_black			0.356
BA_hispanic			1.290
medcity			0.765***
submed			0.496
lgcity			0.416
sublg			1.347***
vlgcity			1.734***
subvlg			1.896***

According to the first model, all included variables concerning educational attainment and demographic background are very important and significantly correlated with one's hourly wage. As expected, having bachelor's degree is positively associated with wage since it increases one's hourly wage by 2.3 units. In the meantime, being female and of black and hispanic origin are negatively associated with the hourly wage.

From the second model, it appears that experience is also significant and positively associated with hourly wage. However, the model shows that there is an inverted-U shaped relationship between experience and hourly wage.

The third model shows that type of city is also an important determinant of hourly wage with the base group consisting of observations from small city. As coefficients for medium, suburb large, very large, and suburb very large city show, bigger the city, higher is the hourly wage. According to the coefficient estimates of interaction variables, they are not significant.

The evaluation of performance of these three regression models is summarized in the table below.

models	AIC	BIC	RMSE	R-squared
Model 1 (training set)	24,499.59	24,487.44	4.91	0.15
Model 2 (training set)	24,330.68	24,381.15	4.84	0.17
Model 3 (training set)	24,275.32	24,382.56	4.80	0.19
Model 1 (test set)	8,290.00	8,321.26	5.16	0.12
Model 2 (test set)	8,249.87	8,291.55	5.07	0.15
Model 3 (test set)	8,216,77	8,305.34	4.98	0.18

The evaluation both on the training and test sets show that the best model is the third regression model that has the lowest AIC and BIC estimates and the lowest RMSE.

Part b)

First regression model:

$$log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + u$$

The second regression model:

$$log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 female + \beta_4 jc\_female + \beta_5 univ\_female + u$$

The third regression model:

$$log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 female + \beta_4 jc\_female + \beta_5 univ\_female + \beta_6 exper + \beta_7 exper^2 + + u$$

The results are following:

	Model 1	Model 2	Model 3
intercept	2.086***	2.293***	1.794***
jc	0.074***	0.051***	0.062***
univ	0.071***	0.050***	0.064***
female		-0.371***	-0.262***
jc_female		0.033*	0.014
univ_female		0.032***	0.017***
exper			0.003***
$exper^2$			0.0001

According to the first model, total 2-year and 4-year credits are also significantly and positively associated with dependent variable, the logarithm of an hourly wage. A unit increases in total 2-year and 4-year credits result in 7.4% and 7.1% higher hourly wage.

The second model that incorporates gender variable and its interactions. It appears that the effect of being female on hourly wage decreases as she gets more of total 2-year and 4-year credits.

The third model shows experience is also significantly and positively correlated with hourly wage, a 1-year increase in wage is associated with 3% increase in hourly wage. The expected inverse U-shaped relationship between wage and experience is not confirmed since coefficient estimate for quadratic term is not statistically significant.

The evaluation of performance of these three regression models is summarized in the table below.

models	AIC	BIC	RMSE	R-squared
Model 1 (training set)	5,250.30	5,275.53	0.46	0.11
Model 2 (training set)	4,782.90	4,4827.06	0.44	0.21
Model 3 (training set)	4,467.25	4,524.02	0.42	0.27
Model 1 (test set)	1,779.75	1,800.59	0.47	0.11
Model 2 (test set)	1, 654.63	1,691.10	0.44	0.20
Model 3 (test set)	1,526.74	1,573.63	0.42	0.27

The evaluation both on the training and test sets show that the best model is the third regression model that has the lowest AIC and BIC estimates and the lowest RMSE.

Part c)

The best model from part a) is

$$wage = \beta_0 + \beta_1 BA + \beta_2 female + \beta_3 black + \beta_4 hispanic + \beta_5 BA\_female + \beta_6 BA\_black + \beta_7 BA\_hispanic + \beta_8 exper + \beta_9 exper^2 + \beta_{10}\text{medcity} + \beta_{11}\text{submed} + \beta_{12}\text{lgcity} + \beta_{13}\text{sublg} + \beta_{14}\text{vlgcity} + \beta_{15}\text{subvlg} + \text{u}$$

The best model from part b) is

$$wage = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 female + \beta_4 jc\_female + \beta_5 univ\_female + \beta_6 exper + \beta_7 exper^2 + + u$$

The evaluation of these two best models for the level of the wage on the test set is summarized in the table below.

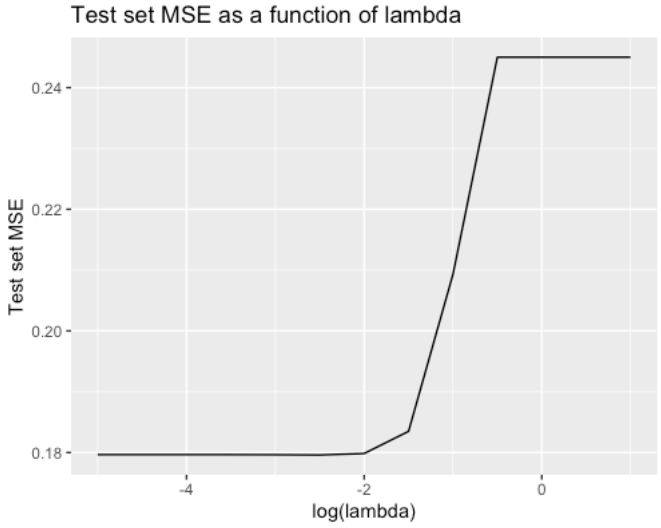
models	AIC	BIC	RMSE	R-squared
Model 1 (test set)	8,216.77	8,305.34	4.98	0.18
Model 2 (test set)	8,150.40	8,197.29	4.89	0.21

As the table shows, the best model from part b) has lower AIC and BIC estimated and lower RMSE, thus appears to be better in predicting for the level of the wage.

Part d) and e)

The best model from part c) was used on the training set to predict log(wage) on the test set using various λ values. The predicted log(wage) has a minimum of 1.546 and a maximum of 2.791 with the mean of 2.245

Below, there is a graph of the test set MSE as a function of different λ values.

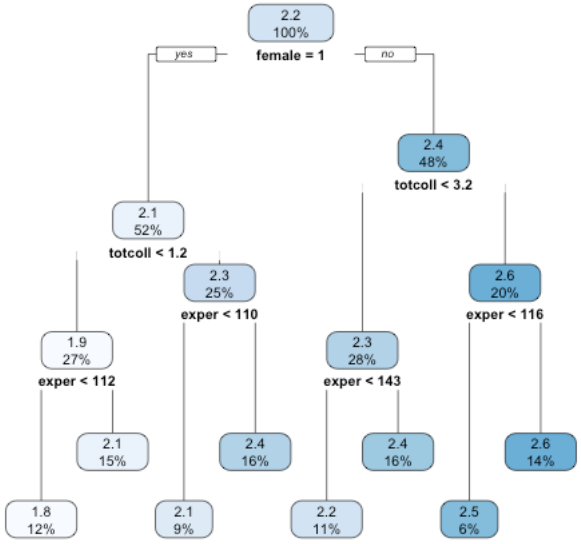


As can be seen, the graph of the test set MSE as a function of λ is upward sloping. It shows that the test set MSE is minimized when the log(λ) is up until -2. After -2, it starts to increase and rises sharply immediately.

By cross-validation, the best LASSO model is the best model from part c) because LASSO model shrinks all coefficient estimates towards zero compared to other models.

Part f)

The figure below demonstrates a pruned regression tree, the result of binary decisions, to the training and test set combined.



The very first root node represents the entire sample population. Then, the first decision rule follows regarding the female variable being set to 1 or not and this results in two smaller samples, each having 48% and 52% of the initial sample size. As female variable appears at the top, it seems that it has the largest impact on the dependent variable. The set of next decision rules concern the total college years that divides the sample into four smaller samples. After that, the decision rule regarding experience comes into effect. At the end, there are overall of eight smaller samples.