1. LASSO and Ridge Regression

a.The minimized expressions for ridge regression and LASSO are:
Ridge regression:
LASSO:

$$\min_{\beta} \Sigma(y_i - \beta_i)^2 \quad \text{s.t.} \sum|\beta_i| \leq s$$

function of λ: $F(\beta_i, \lambda) = \sum(y_i - \beta_i)^2 + \lambda(|\beta_i| - s)$

Ridge:

$$\min_{\beta} \Sigma(y_i - \beta_i)^2 \quad \text{s.t.} \sum\beta_i{}^2 \leq s$$

function of λ: $F(\beta_i, \lambda) = \sum(y_i - \beta_i)^2 + \lambda(\beta_i{}^2 - s^2)$

b. The minimization problem should be solved by taking derivative.
LASSO:

$$\frac{\partial(\sum(y_i-\beta_i)^2+\lambda\sum|\beta_i|)}{\partial\beta_{i-s}} = -2(y_i - \beta_i) + (-)\,\lambda$$

1) If $\beta_i > 0$,
$-2(y_i - \beta_i) + \lambda = 0$
$\beta_i = y_i - \lambda/2$
2) If $\beta_i < 0$,
$-2(y_i - \beta_i) - \lambda = 0$
$\beta_i = y_i + \lambda/2$

When $\lambda \to 0$, $\beta_i = y_i$

When $\lambda \to \infty$, performs variable selection by setting some varibables to zero as λ increases.

Ridge :

$$\frac{\partial(\sum(y_i-\beta_i)^2+\lambda\sum(\beta_i)^2)}{\partial\beta_{i-s}} = -2(y_i - \beta_i) + 2\,\lambda\beta_i$$

$$\beta_i = \frac{y_i}{1+\lambda}$$

When $\lambda \to 0$, $\beta_i = y_i$

When $\lambda \to \infty$, $\beta_i = 0$

In LASSO, the penalty term sets some coefficients exactly to zero and delete the variables; In Ridge, the penalty term sets coefficients close to zero. And for LASSO , it performs variable selection by setting some varibables to zero as λ increases.

2. bootstrap sample

If we select simple random sampling with alternatives to have bootstrap samples, then the probability of selecting a specific element in any one draw is 1/n, so the probability that it is not selected in any of the n draws is (1-1/n)^n.To calculate the probability that a sample element does not appear in a bootstrap sample, we can should raise this probability to the power of n (the sample size), since each sample has the same probability of not appearing in a bootstrap sample. So the probability that a sample is not selected in a bootstrap sample is:

[(1-1/n)^n]^n = (1-1/n)^(n^2)

To find the limit of this expression as $n \to \infty$, we can take the limit of the natural logarithm of the expression:

lim($n\to\infty$) ln[(1-1/n)^(n^2)]
= lim($n\to\infty$) ((-1/(n-1)) / (-2/n^3))
= 0

Therefore, as $n \to \infty$, the probability that a sample does not appear in a bootstrap sample move towards 0.

3.Twoyear Dataset Modeling

PART(a)(b)(c) Wage Or Lwage？

Table 1 Models 1, 2,3 with wage as dependent variable- training set

Table X

	Dependent variable:		
	wage		
	(1)	(2)	(3)
female	-3.574***	-2.699***	-2.759***
	(0.155)	(0.182)	(0.158)
phsrank	0.034***	0.017***	0.017***
	(0.003)	(0.003)	(0.003)
BA	1.987***	0.387*	-0.253
	(0.175)	(0.220)	(0.269)
AA	0.536	-0.315	-0.369
	(0.368)	(0.416)	(0.415)
black	-0.893***	-0.806***	-0.806***
	(0.260)	(0.251)	(0.251)
hispanic	-0.129	0.090	0.100
	(0.356)	(0.344)	(0.344)
exper		0.029***	0.029***
		(0.002)	(0.002)
jc		0.507***	0.478***
		(0.113)	(0.113)
univ		0.588***	1.204***
		(0.047)	(0.157)
smcity		-0.137	-0.205
		(0.237)	(0.167)
medcity		0.379	0.390*
		(0.232)	(0.231)
female:smcity		-0.160	
		(0.326)	
univsq			-0.098***
			(0.024)
Constant	10.087***	6.208***	6.209***
	(0.203)	(0.364)	(0.360)
Observations	4,058	4,058	4,058
R2	0.179	0.234	0.238
Adjusted R2	0.177	0.232	0.235
Residual Std. Error	4.791 (df = 4051)	4.629 (df = 4045)	4.619 (df = 4045)
F Statistic	146.741*** (df = 6; 4051)	103.188*** (df = 12; 4045)	105.014*** (df = 12; 4045)
Note:			*p<0.1; **p<0.05; ***p<0.01

(Source: R)

In question (a)， we use different models in training set by following " domin language" : model 1 is a multivariate model, model 2 is also a multivariate model plus interactive variable(female and small city), model 3 is a multivariate model plus one quadratic variable of univ. The regression results in shown in Table 1. Most of the variables that are selected are significant under 95% confidence level. When explaining the factors affecting wage, I concluded the following opinions: being as a woman can cause wage decrease; the better rank the high school has , the wage would be better ; having a BA degree can lead to higher wage ; associate degree doesn't cause significant effect on wage ; being a African-American would cause wage decreasing, whilst being Hispanic doesn't cause significant effect on wage; every one year of the work experience increase would cause wage to increase 0.029; total 2 year credits and total four year credits also have positive effect on wage. The size of city overall doesn't have significant effect on wage, being a woman doesn't strengthen or weaken the effect that city size has for wage. The polynomial variable indicates that with the total 4-year credits increase, the positive effect would be less and less, this

might cause by that someone focus too much on schoolwork and neglect the experiment chance in work place such as internship opportunities. From the intercept, I concluded that even though someone doesn't have had education experience or work experience, he or she still has the basic salary that doesn't start from 0. The average wage is 10.63452, the people who have 0 secondary education and 0 work experience are expected to have wage under average.

Table 2 Models 4,5,6with lwage as dependent variable – training set

Table X

	Dependent variable:		
	lwage		
	(1)	(2)	(3)
female	-0.349***	-0.240***	-0.248***
	(0.014)	(0.016)	(0.014)
phsrank	0.003***	0.002***	0.002***
	(0.0003)	(0.0003)	(0.0003)
BA	0.193***	0.043**	-0.020
	(0.016)	(0.019)	(0.024)
AA	0.083**	-0.011	-0.016
	(0.033)	(0.037)	(0.037)
black	-0.098***	-0.088***	-0.088***
	(0.024)	(0.022)	(0.022)
hispanic	-0.016	0.005	0.006
	(0.032)	(0.031)	(0.030)
exper		0.004***	0.004***
		(0.0002)	(0.0002)
jc		0.052***	0.049***
		(0.010)	(0.010)
univ		0.057***	0.118***
		(0.004)	(0.014)
smcity		-0.015	-0.027*
		(0.021)	(0.015)
medcity		0.022	0.023
		(0.021)	(0.020)
female:smcity		-0.026	
		(0.029)	
univsq			-0.010***
			(0.002)
Constant	2.195***	1.684***	1.685***
	(0.018)	(0.032)	(0.032)
Observations	4,058	4,058	4,058
R2	0.201	0.291	0.295
Adjusted R2	0.200	0.289	0.293
Residual Std. Error	0.435 (df = 4051)	0.410 (df = 4045)	0.409 (df = 4045)
F Statistic	170.269*** (df = 6; 4051)	138.618*** (df = 12; 4045)	141.017*** (df = 12; 4045)
Note:			*p<0.1; **p<0.05; ***p<0.01

In question (b), we use similar models as Table 1, The only difference outcome variable is log(wage), it is shown in Table 2. Most of the variables that are selected are significant under 95% confidence level. When explaining how the factors effect wage, the explanation will be based on how the factors affect the percentage change of wage.

Table 3 Information criteria and prediction accuracy

No.	models	AIC	BIC	RMSE	RSquared	k
1	reg1	24435.6	24486.06	4.903173	0.170772572	6
2	reg2	24148.6	24236.89	4.725811	0.229678715	12
3	reg3	24137.4	24225.76	4.719335	0.231788394	12
4	test1	8133.99	8175.667	4.860087	0.181099344	6
5	test2	8010.31	8083.252	4.622416	0.259233699	12
6	test3	8010.43	8083.369	4.622617	0.259169213	12
7	reg4	4900.88	4951.344	0.44173	0.185619061	6
8	reg5	4466.35	4554.664	0.418083	0.27047687	12
9	reg6	4452.81	4541.123	0.417386	0.27290712	12
10	test4	1655.11	1696.788	0.443433	0.202911807	6
11	test5	1456.77	1529.713	0.410271	0.31767429	12
12	test6	1455.41	1528.356	0.410065	0.318358337	12

(Source: R, colorized by Excel)

In the table 3 blue session, they are the three models with wage as outcome. I generalize the AIC , BIC , RMSE , Rsquare and K for Model 1,2, 3 in training set and I use same models in the test set (the results is named as test 1 , test 2 and 3); In the yellow session, they are models with lwage as outcome. I generalize the information criteria for Model 4, 5, 6 and use them in the test set (name as test 4,5,6) .

Among the models123 with wage as outcome, model 3 has the best performance, for the reason of smallest AIC , smallest BIC , smallest RMSE and best Rsquare. Comparing the results between training set and test set , the AIC and BIC is very different; but the prediction accuracy, presented by RMSE, are very similar between training set and test set.

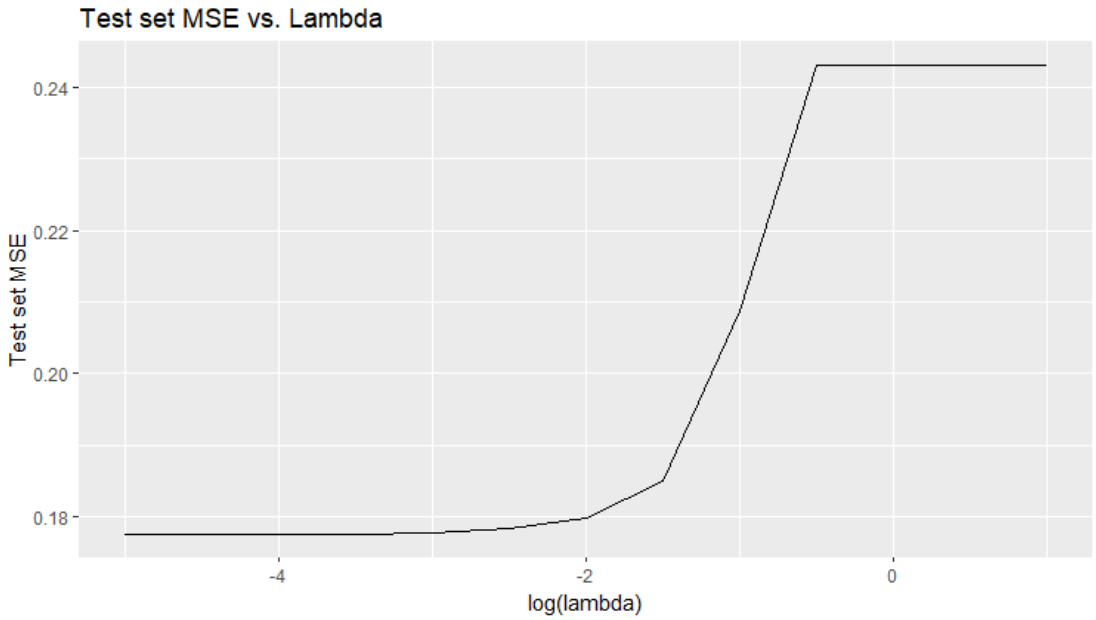
Among the models456 with lwage as outcome, model 6 has smallest AIC, smallest BIC, smallest RMSE and best Rsquare. Comparing the results between training set and test set , the AIC and BIC is very different; but the prediction accuracy, presented by RMSE, are very similar between training set and test set.

Based on a rule of thumb, it can be said that RMSE values between 0.2 and 0.5 shows that the model can relatively predict the data accurately. In all , the prediction accuracy is satisfying in these models.

Compare the best model in part a) and b), which is model 3 and model 6 , the RMSE is smaller in model 6, this can verifies two things: 1. model 6 are better in terms of prediction accuracy, 2.log(wage) is a more optimal outcome variable than wage

Part(d) MSE and λ

Graph 1



In this graph, as lambda increases, the MSE generally increases, indicating that the model is becoming more biased and less complex. However, at the point where log(lamda) is around -2, the increase in bias outweighs the reduction in variance, and the MSE starts to increase rapidly.

Regrading using various lambda values in LASSO to estimated the lwage, the effect of lambda on the prediction of the lambda in LASSO depends on the balance between bias and variance. If the initial model is overfit and has high variance, increasing lambda can improve the model's performance by reducing overfitting. However, if the initial model is underfit and has high bias, increasing lambda can lead to an increase in the model's error. The optimal value of lambda for the test set , in our case, is 0.0002358847, which is the smallest lambda.

Part (e)

Table 4

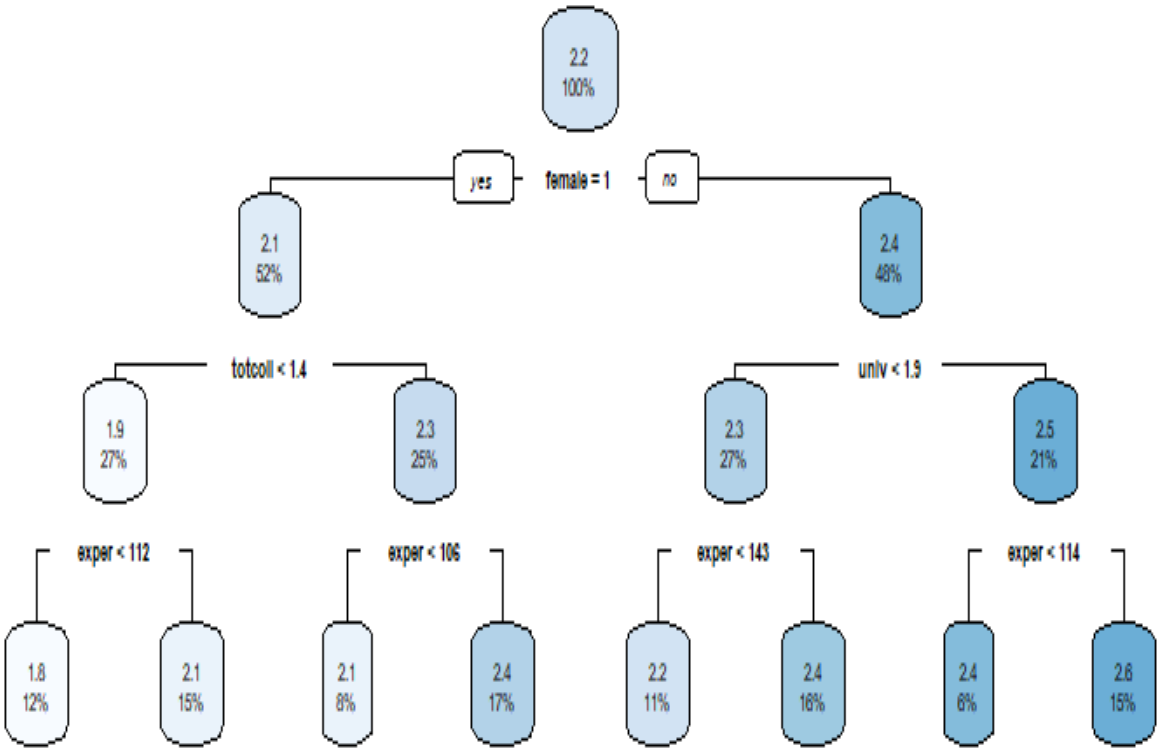
```
13 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept)  1.706116455
female      -0.237934842
phsrank      0.001296532
BA           .
AA           0.031610371
black       -0.100491915
hispanic    -0.036646738
exper        0.003752094
jc           0.057237366
univ         0.092095768
smcity      -0.024759190
medcity      0.009064102
univsq      -0.005030839
```

By applying cross-validation with 5 folds, I first find the best lamda, 0.0003162278, then I use the best lambda value to find the best LASSO model. The explanatory variables and coefficients are as table 4, the outcome variables is lwage. The reason why it didn't delete any variables is that, there is not many cubic or fourth time variables, so it still keep as the same as OLS.

Part(f)Regression trees

Min split is set to 1000 to ensure that each terminal node has at least 1000 observations, And maxdepth is set to 10 to limit the depth of the tree, and xval is set to 5 to perform 5-fold cross-validation. The xerror column represents the cross-validation error. The row with the lowest cross-validation error represents the optimal value for the cost complexity parameter.Cost complexity parameter controls the trade-off between the complexity of the tree and its fit to the training data. It this case, cost complexity= 0.0047506. When controlling the tree branches, I control that each terminal node has at least 1000 observations.

Graph 2 Regression Trees



By this tree, the following expectations can be illustrated:
The first split in the CART finds that gender would divide the observations into two groups with largest difference in their average lwage. If the observation is a woman, then her average lwage is 2.1 , the percentage of women in the workout set is 52%; If the observation is a man, then his average lwage is 2.1 , the percentage of man in the workout set is 48%. For women with total credits higher than 1.4, their wage would be 2.3 in average, even reach to be averagely 2.4 if the work experiences is above 106 ; if women with total credits lower than14, then their lwage would 1.9 in average. For men with the university credits higher than 1.9, the average lwage for them can be as high as 2.5; if not then the average lwage is 2.3. Among the man with university credits higher than 1.9, if they has work experience larger than 114, the averaged lwage is expected to be 2.6 ; otherwise 2.4. Among the man with university credits lower than 1.9, if they has work experience larger than 143, the averaged lwage is expected to be 2.4 ; otherwise 2.2. In all , gender , education level, work experience are three main factors on deciding lwage.

Part(g) random forest

Random Forest is an ensemble algorithm that builds a large number of decision trees and combines their predictions to make a final prediction. It contains bootstrap and

aggregating (bagging steps). When it comes to the tuning parameters,
I did five random forests with different tuning parameters to the training set, controlled by mtry or the number of the trees. For rf4 and rf5 , the number of trees is 25,larger number of trees can improve the accuracy of the model, such as rf1 , rf2 and rf 3, but larger tree number also increases the computational cost and may lead to overfitting. Meanwhile, the no. of variables helps to reduce the correlation between trees and improve the diversity of the ensemble. One of the advantages of random forest is that it can provide measures of variable importance, which can help with feature selection and interpretation of the model.
The result is shown in Table 5:

Table 5

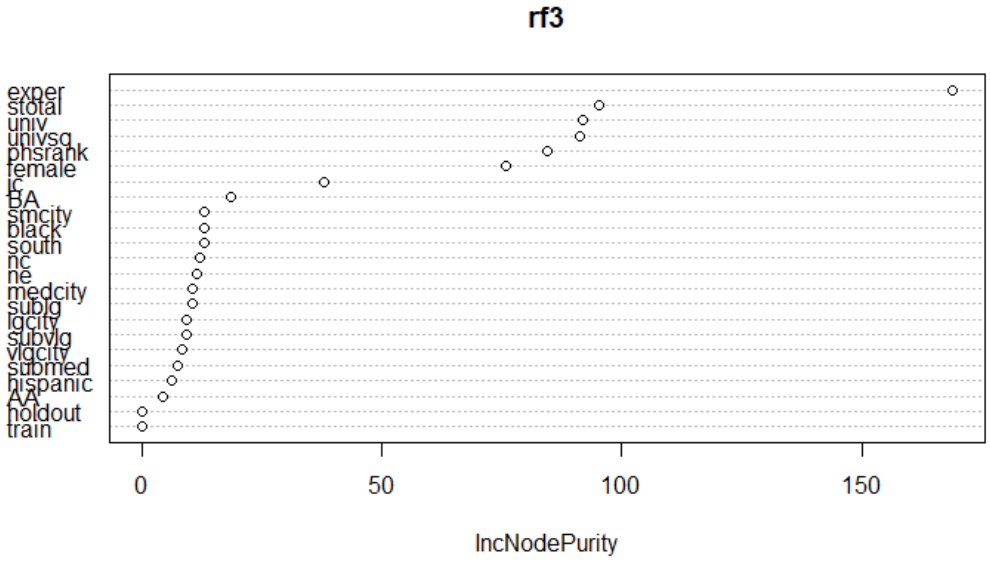
```
> rf1

Call:
 randomForest(formula = lwage ~ . - wage - totcoll - id, data = twoyear_train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 7

          Mean of squared residuals: 0.1741088
                    % Var explained: 27.21
> rf2

Call:
 randomForest(formula = lwage ~ . - wage - totcoll - id, data = twoyear_train,      mtry = ncol(twoyear_train))
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 23

          Mean of squared residuals: 0.1796276
                    % Var explained: 24.9
> rf3

Call:
 randomForest(formula = lwage ~ . - wage - totcoll - id, data = twoyear_train,      mtry = sqrt(ncol(twoyear_train))
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 5

          Mean of squared residuals: 0.1722796
                    % Var explained: 27.97
> rf4

Call:
 randomForest(formula = lwage ~ . - wage - totcoll - id, data = twoyear_train,      ntree = 25)
               Type of random forest: regression
                     Number of trees: 25
No. of variables tried at each split: 7

          Mean of squared residuals: 0.1907086
                    % Var explained: 20.27
> rf5

Call:
 randomForest(formula = lwage ~ . - wage - totcoll - id, data = twoyear_train,      ntree = 500)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 7

          Mean of squared residuals: 0.1740057
                    % Var explained: 27.25
```

After running random forests on training set, I made predictions in test set, and calculate the MSE, so as to compare their performance and choose the best model :

Table 6 Random Forest and MSE

	Training set MSE	Test set MSE
Rf1	0.1741088	0.1690667
Rf2	0.1796276	0.1749268
Rf3	0.1722796	0.1658911
Rf4	0.1907086	0.1703309
Rf5	0.1740057	0.1687

The Rf 3 is the best model among the 5 models.
I also plot the importance of variables using Rf3:
Graph 3 Variable Importance



(Source: R)
Through this plot, the exper , stotar, univ , and high school ranking , gender , BA degree are the most important factors.

Part(h) Boosted trees

The key idea behind boosted trees is to focus on the observations that are most difficult to classify correctly. The algorithm assigns weights to each observation, with more weight given to the observations that are misclassified. This ensures that the subsequent trees focus on the areas where the model is weak. In this question, I use different tree numbers and fit in 5 different boosted trees and use them on test set.

Table 7Boosted Trees and MSE

	Test set MSE
Gbm1	0.164582
Gbm 2	0.1672431
Gbm 3	0.1991614
Gbm 4	0.2165675
Gbm 5	0.1632268

(Source:R)
Gbm 5 with random forest at T = 500 trees is the best model among 5 models.

Part (i) Extrapolate best model to holdout data

Table 8 MSE for all the best models

Best model	Name in R	Holdout MSE
Part b	Model6	0.1786598
Part e	LASSO.model	0.1760435
Part f	Tree	0.2174534
Part g	Rf3	0.1657825
Part h	Gmb5	0.1543546

(Source: R)

A frequently used statistic for assessing the effectiveness of a regression model is mean squared error (MSE). It calculates the average squared difference between the response variable's expected and actual values. It is a crucial indicator for assessing the efficacy and effectiveness of regression models. It offers a straightforward and understandable indicator of how well the model predicts the response variable, and it may be used to contrast many models or tweak the settings of a single model to enhance performance.

In all , the model in Part h , denoted by “gmb5 “ for the excellent performance in MSE.

Gmb5 used the random tree approach, boosted trees are a powerful ensemble learning technique that combines multiple weak decision trees to create a strong model. Boosted trees are known for their high accuracy in prediction tasks, particularly in structured data problems such as tabular data. They are particularly effective in handling non-linear relationships between input features and output variables. Boosted trees can automatically perform feature selection by assigning higher weights to more informative features. This can improve model performance and reduce overfitting.

Reference

Békés, G., & Kézdi, G. (2021). Data analysis for business, economics, and policy. Cambridge University Press.

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). An introduction to statistical learning: with applications in R. Spinger.