5054 Midterm Project

ZHANG Juntao
20908272

October 19, 2022

1 Problem 1: Speed and Stopping Distances of Cars

1.1 Question 1

Building models on distance and speed according to different degrees of the polynomial. Use leave-one-out cross-validation, and get the CV errors in Table 1. Plot CV errors versus the degree of the polynomial, as shown in Figure 1.

It's clear to see that the CV error is smallest when the degree of polynomial is 2. And the CV error gets larger with the degree of polynomial increase, except when the degree is 9.

So it can be concluded that fitting the model with quadratic polynomial works best, which shows that it is most reasonable to describe the distance taken to stop in terms of a quadratic polynomial of speed.

Table 1: LOOCV

Degree	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CV Error	246.4054	243.0292	246.8288	250.0914	279.6864	327.5014	408.9479	476.4366	318.1917	696.9015

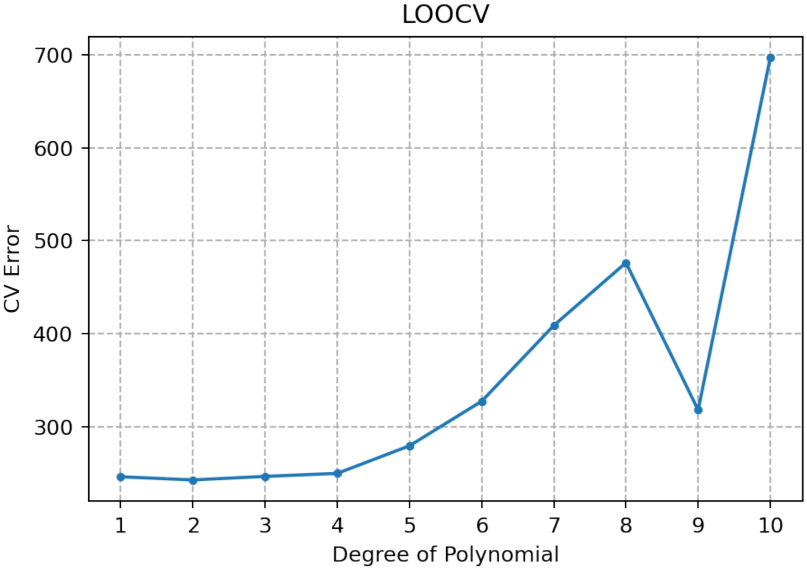


Figure 1: LOOCV

1.2 Question 2

Similarly, building models on distance and speed according to different degrees of the polynomial. Use 5-Fold cross-validation, and get the CV errors in Table 2. Plot CV errors versus the degree of the polynomial, as shown in Figure 2.

In 5-Fold cross-validation, there are many different scenarios for splitting the data set into five groups, and we get different CV errors in different splitting methods. In general, the CV error is small with degree 1-4.

Here we choose one situation to present, where the data set is split into the following five groups:

[1 10 13 15 20 22 27 30 36 42], [2 9 17 18 29 32 34 35 41 46], [5 6 12 16 21 24 31 33 38 45],

[0 7 8 11 26 37 40 43 48 49], [3 4 14 19 23 25 28 39 44 47],

where 0-49 represents the label of observations.

Roughly speaking, the CV error gets larger with the degree of polynomial increase. And the CV error is smallest(237.2272) when the degree of polynomial is 3. But the CV errors of degree 2 and degree 4 are 237.5886 and 236.8141, which are both very close to the smallest CV error.

We can consider the CV errors of degree 2, degree 3, and degree 4 are not significantly different and are all the smallest. So from ”Occam’s Razor principle”, choose the simplest one, which means that using a quadratic polynomial to fit the model is best, i.e. it is most reasonable to describe the distance taken to stop in terms of a quadratic polynomial of speed.

Table 2: 5-Fold CV

Degree	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CV Error	242.9784	237.5886	237.2272	236.8141	264.4524	286.3088	411.7961	402.0928	556.2113	2704.7814

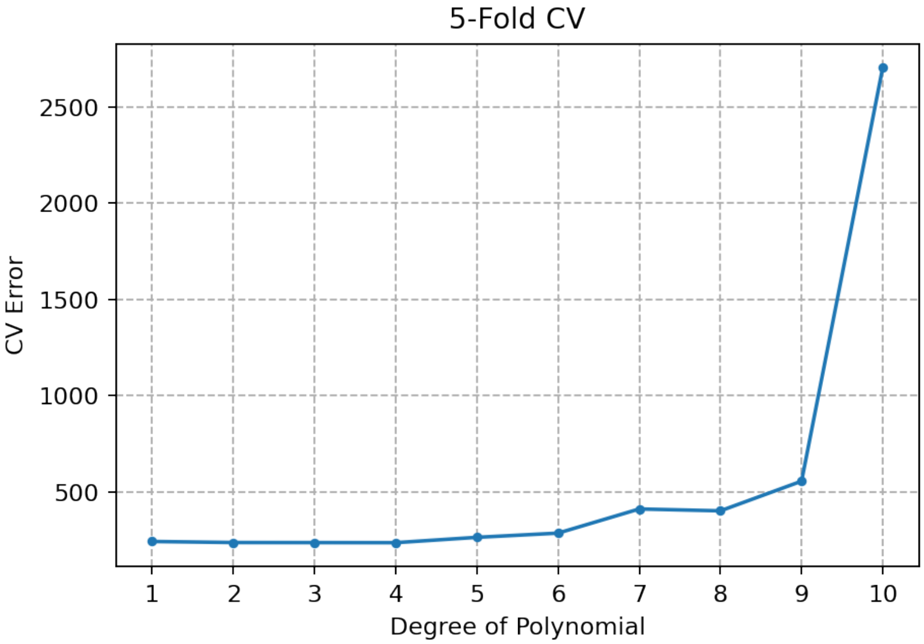


Figure 2: 5-Fold CV

1.3 Question 3

Firstly, we got the same conclusion according to these two approaches that it's reasonable to describe the distance taken to stop in terms of a quadratic polynomial of speed. Then compare LOOCV and 5-Fold CV, get the following findings:

1. LOOCV takes longer time of computation than 5-Fold CV since the model needs to be fit 50 times in LOOCV but only 5 times in 5-Fold CV. So LOOCV is computationally expensive.
2. The CV error in LOOCV is smaller than in 5-Fold CV since the training data size is close to the entire data size in LOOCV.
3. The CV error is fixed in LOOCV, but varies in 5-Fold CV according to the different splitting methods. Especially for small data sets like in this problem, the variation of CV error is large among different grouping methods.

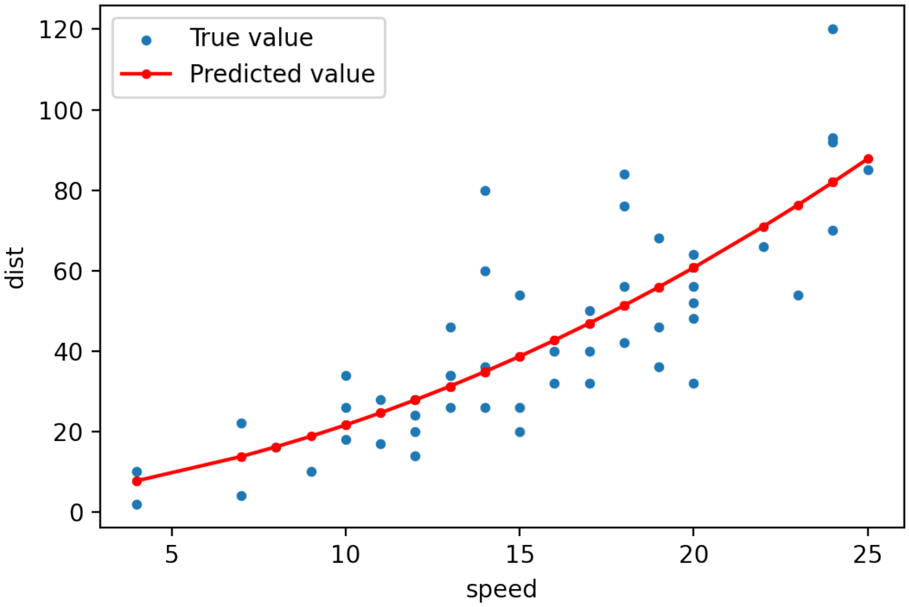


Figure 3: Prediction of best model(quadratic polynomial)

2 Problem 2: Titanic – Survival or Not

2.1 Question 1

Treat survival as the response variable, fit a logistic regression model using predictors pclass, sex, age, sibsp and fare.

2.1.1 Deal with missing value

Firstly check if there exist missing values in these variables, and find that there are 177 missing values in age. Then consider filling in missing values in the age variable. As Figure 4 shows, the age of passengers in different Pclass and sex has different distributions. So grouping passengers through Pclass and sex, and filling in the missing value of age with the mean age of its group.

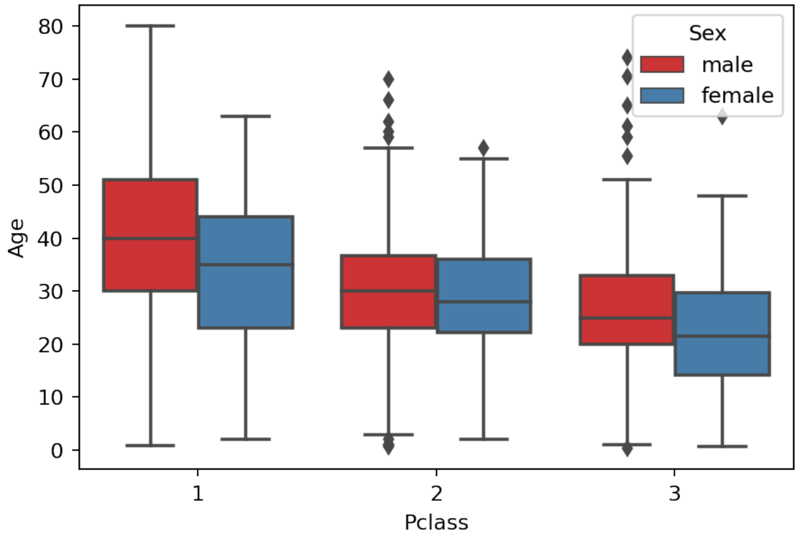


Figure 4: Age on different sex and Pclass

2.1.2 Logistic regression

For the variable with n category, we can set either n or n-1 dummy variables for it. When setting n dummy variables, we can see the relationships between the response variable and all categories. And when setting n-1 dummy variables, the relationship between the remaining category and the response variable can't be seen directly, but is mixed into the coefficients of other variables.

So for a better view of how each pclass and each sex influence the survival status, setting 3 dummy variables for Pclass($Pclass_1, Pclass_2, Pclass_3$) and 2 dummy variables for Sex (Sex_{female}, Sex_{male}). Then fit a logistic regression model and get the result in Figure 5.

The coefficient of Sex_{male} is -0.9180, and the 95% confidence interval of it is [-1.143, -0.693].

The coefficient of $Pclass_{3rd}$ is -0.9239, and the 95% confidence interval of it is [-1.173, -0.675].

So both Sex_{male} and $Pclass_{3rd}$ are significant with 95% confidence.

2.2 Question 2

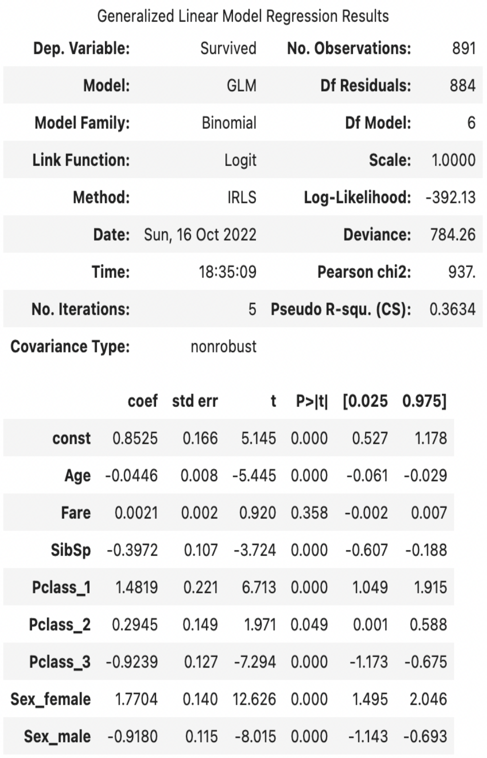
Apply Bootstrap with 1000 repetitions, then a logistic regression model again, and get the result in Figure 6.

The coefficient of Sex_{male} is -0.7095, and the 95% confidence interval of it is [-0.909, -0.509].

The coefficient of $Pclass_{3rd}$ is -0.8468, and the 95% confidence interval of it is [-1.063, -0.631].

Both Sex_{male} and $Pclass_{3rd}$ are significant with 95% confidence.

Compare with the reported confidence intervals in 2.1.2, we can find that each confidence interval(with Bootstrap) becomes smaller than the original confidence interval. It actually shows our estimated coefficients become more accurate with the Bootstrap method.

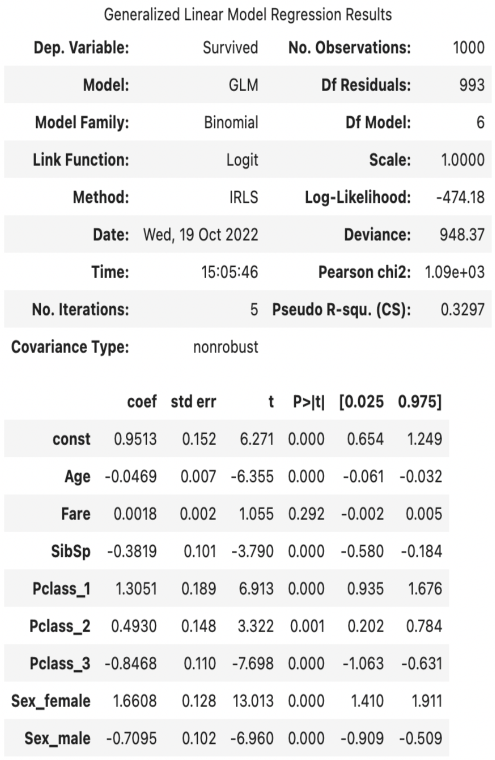


Generalized Linear Model Regression Results

Dep. Variable:	Survived	**No. Observations:**	891
Model:	GLM	**Df Residuals:**	884
Model Family:	Binomial	**Df Model:**	6
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-392.13
Date:	Sun, 16 Oct 2022	**Deviance:**	784.26
Time:	18:35:09	**Pearson chi2:**	937.
No. Iterations:	5	**Pseudo R-squ. (CS):**	0.3634
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.8525	0.166	5.145	0.000	0.527	1.178
Age	-0.0446	0.008	-5.445	0.000	-0.061	-0.029
Fare	0.0021	0.002	0.920	0.358	-0.002	0.007
SibSp	-0.3972	0.107	-3.724	0.000	-0.607	-0.188
Pclass_1	1.4819	0.221	6.713	0.000	1.049	1.915
Pclass_2	0.2945	0.149	1.971	0.049	0.001	0.588
Pclass_3	-0.9239	0.127	-7.294	0.000	-1.173	-0.675
Sex_female	1.7704	0.140	12.626	0.000	1.495	2.046
Sex_male	-0.9180	0.115	-8.015	0.000	-1.143	-0.693

Figure 5: Summary of LR model



Generalized Linear Model Regression Results

Dep. Variable:	Survived	**No. Observations:**	1000
Model:	GLM	**Df Residuals:**	993
Model Family:	Binomial	**Df Model:**	6
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-474.18
Date:	Wed, 19 Oct 2022	**Deviance:**	948.37
Time:	15:05:46	**Pearson chi2:**	1.09e+03
No. Iterations:	5	**Pseudo R-squ. (CS):**	0.3297
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.9513	0.152	6.271	0.000	0.654	1.249
Age	-0.0469	0.007	-6.355	0.000	-0.061	-0.032
Fare	0.0018	0.002	1.055	0.292	-0.002	0.005
SibSp	-0.3819	0.101	-3.790	0.000	-0.580	-0.184
Pclass_1	1.3051	0.189	6.913	0.000	0.935	1.676
Pclass_2	0.4930	0.148	3.322	0.001	0.202	0.784
Pclass_3	-0.8468	0.110	-7.698	0.000	-1.063	-0.631
Sex_female	1.6608	0.128	13.013	0.000	1.410	1.911
Sex_male	-0.7095	0.102	-6.960	0.000	-0.909	-0.509

Figure 6: Summary of LR model with Bootstrap

2.3 Question 3

Explore the original dataset.

From the coefficients of logistic regression above, we can see that female has a larger probability of survival than male, and passengers of $Pclass_1$ have larger probability of survival than passengers of $Pclass_2$, passengers of $Pclass_2$ has a larger probability of survival than passengers of $Pclass_3$. So we explore the survival status from two aspects: Pclass and Sex. The visualization shows in Figure 7 and Figure 8.

And get the following findings:

1. 342 passengers survived and 549 passengers did not survive, only 38.4% of passengers survived. In survived passengers, 233 are female and 109 are male. But there are 577 males and 314 females in total, which means that 74.2% of females survived and 18.9% of males survived.

2. In 342 survived passengers: 136 with $Pclass_1$, 87 with $Pclass_2$ and 119 with $Pclass_3$.

   In all passengers, there are 216 with $Pclass_1$, 184 with $Pclass_2$ and 491 with $Pclass_3$.

   It shows that 57.6% of passengers with $Pclass_1$ survived, 47.3% of passengers with $Pclass_2$ survived, and only 24.2% of passengers with $Pclass_3$ survived.

3. In 136 survived passengers with $Pclass_1$, 91 are female and 45 are male;

   In 87 survived passengers with $Pclass_2$, 70 are female and 17 are male;

   In 119 survived passengers with $Pclass_3$, 72 are female and 47 are male.

4. To summarize, females are more likely to survive on the titanic, and passengers with greater ticket class are more likely to survive on the titanic.

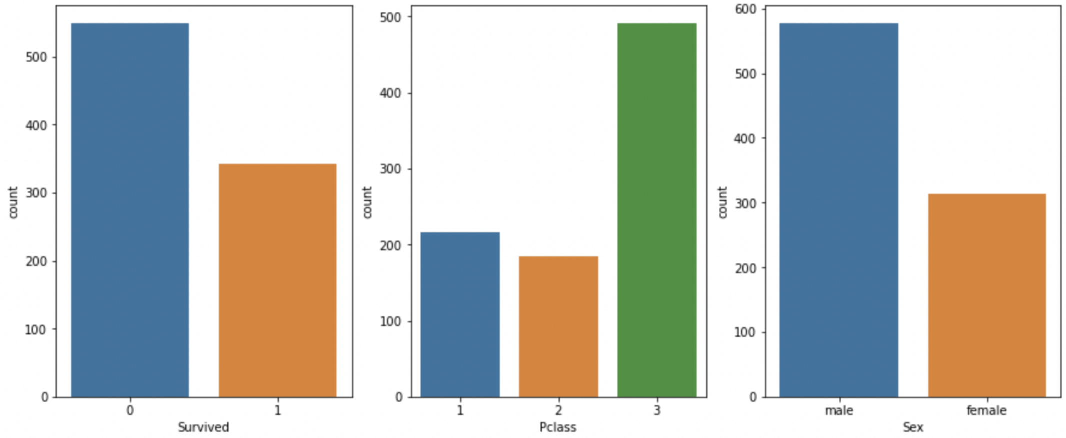


Figure 7: EDA-1

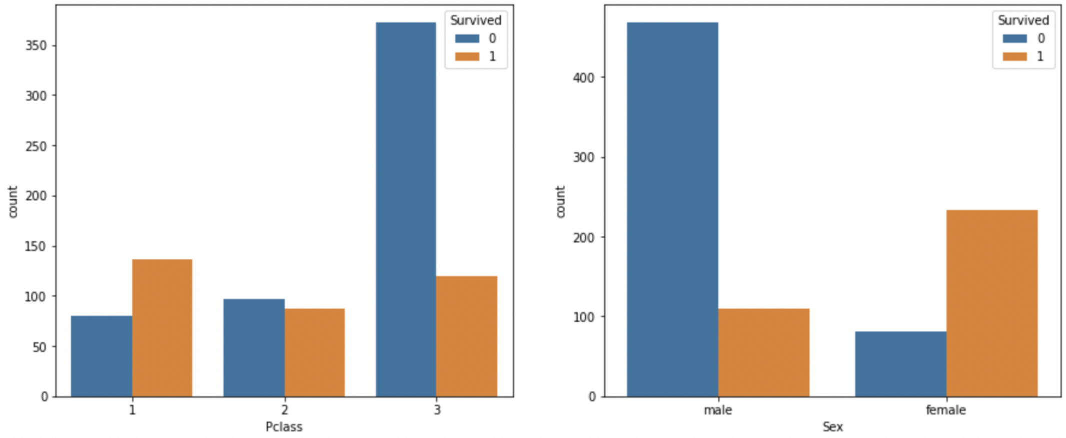


Figure 8: EDA-2

3 Problem 3: Predicting First-Year College Students' GPA

3.1 Question 1

Use all variables to predict students' first-year GPA, by best subset selection up to the size of 8. Report the 8 best linear models in Table 3, and plot the R-square versus model size in Figure 9.

Show the adjusted R-square of these eight best models from best subset selection in Table 4.

It is clear to see that the 6th model has the largest adjusted R-square. So from the adjusted R-square aspect, the 6th model is the best model.

Table 3: Best subset selection

Model size	HSGPA	SATV	SATM	Male	HU	SS	FirstGen	White	CollegeBound
1 (1)	*								
2 (1)	*				*				
3 (1)	*				*			*	
4 (1)	*	*			*			*	
5 (1)	*	*			*	*		*	
6 (1)	*	*		*	*	*		*	
7 (1)	*	*		*	*	*	*	*	
8 (1)	*	*		*	*	*	*	*	*

Table 4: Adjusted R-square of models from best subset selection

Model size	1	2	3	4	5	6	7	8
Adj. R-square	0.1960203	0.2629543	0.3136440	0.3251153	0.3283423	0.3285134	0.3278000	0.3247430

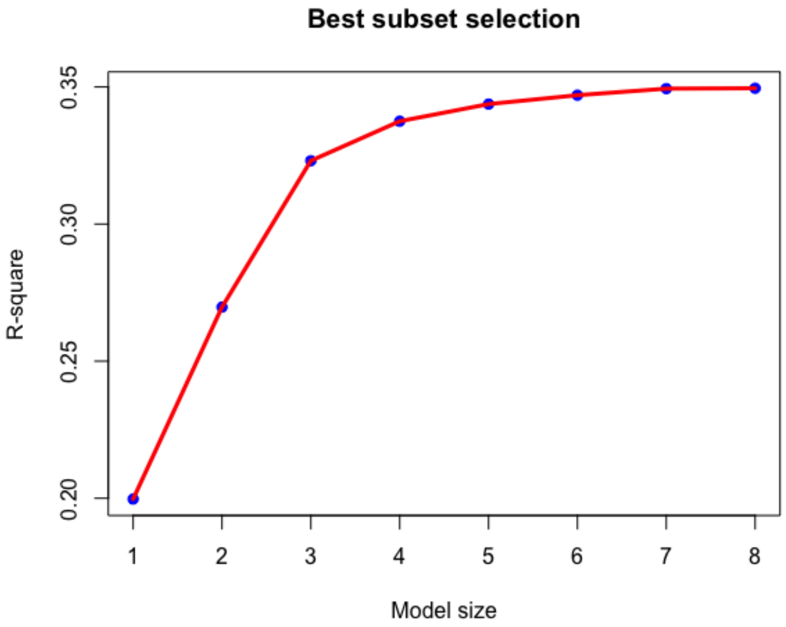


Figure 9: R-square versus model size

3.2 Question 2

Choose the best model from above eight models by 5-Fold CV, and get the CV errors in Figure 10. Obviously, the 4th model has the smallest CV error, so the 4th model is the best model according to CV error.

From Table 3 we can get that the 4th model's predictors are 'HSGPA', 'SATV', 'HU', and 'White'. Report the summary of this model in Figure 11.

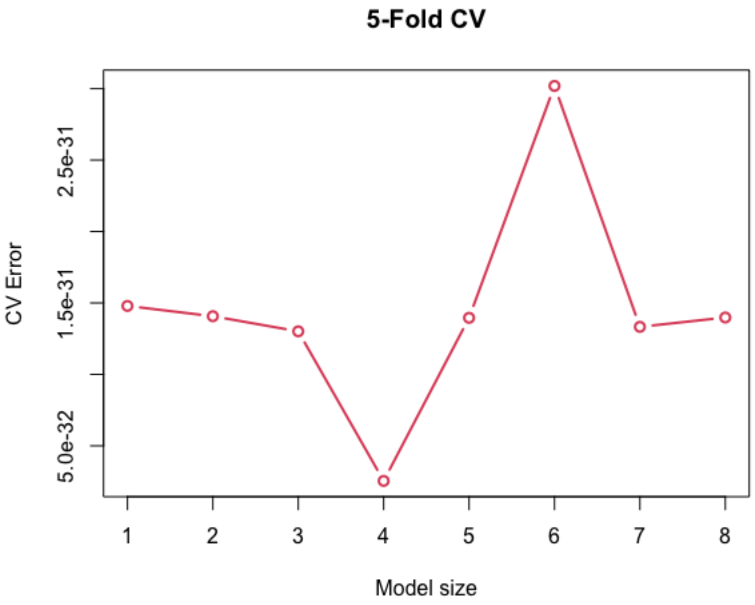
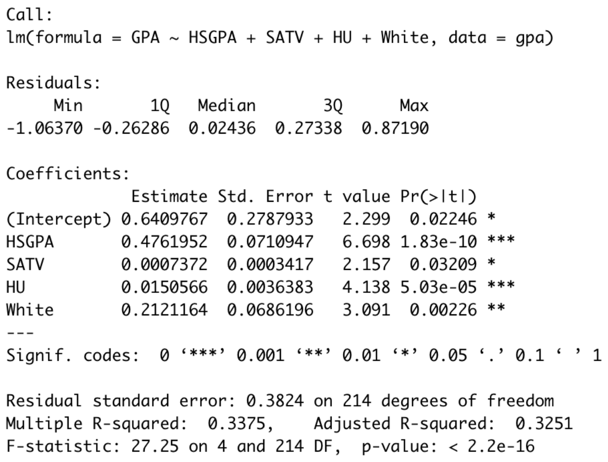


Figure 10: CV error versus model size



```
Call:
lm(formula = GPA ~ HSGPA + SATV + HU + White, data = gpa)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.06370 -0.26286  0.02436  0.27338  0.87190 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.6409767  0.2787933   2.299  0.02246 *  
HSGPA       0.4761952  0.0710947   6.698 1.83e-10 ***
SATV        0.0007372  0.0003417   2.157  0.03209 *  
HU          0.0150566  0.0036383   4.138 5.03e-05 ***
White       0.2121164  0.0686196   3.091  0.00226 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3824 on 214 degrees of freedom
Multiple R-squared:  0.3375,	Adjusted R-squared:  0.3251 
F-statistic: 27.25 on 4 and 214 DF,  p-value: < 2.2e-16
```

Figure 11: Summary of best model

3.3 Question 3

Use all variables to predict students' first-year GPA, by forward stepwise selection up to the size of 8. Report the 8 best linear models in Table 5, and plot the adjusted R-square versus model size in Figure 12.

Show the BIC of these eight best models from forward stepwise selection in Table 6.

It is clear to see that the 3rd model has the smallest BIC. So from the BIC aspect, the 3rd model is the best model.

Table 5: Forward stepwise selection

Model size	**HSGPA**	**SATV**	**SATM**	**Male**	**HU**	**SS**	**FirstGen**	**White**	**CollegeBound**
1 (1)	*								
2 (1)	*				*				
3 (1)	*				*			*	
4 (1)	*	*			*			*	
5 (1)	*	*			*	*		*	
6 (1)	*	*		*	*	*		*	
7 (1)	*	*		*	*	*	*	*	
8 (1)	*	*		*	*	*	*	*	*

Table 6: BIC of models from forward stepwise selection

Model size	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
BIC	-38.01046	-52.66931	-63.90100	-63.22406	-59.91041	-55.60773	-51.02158	-45.67917

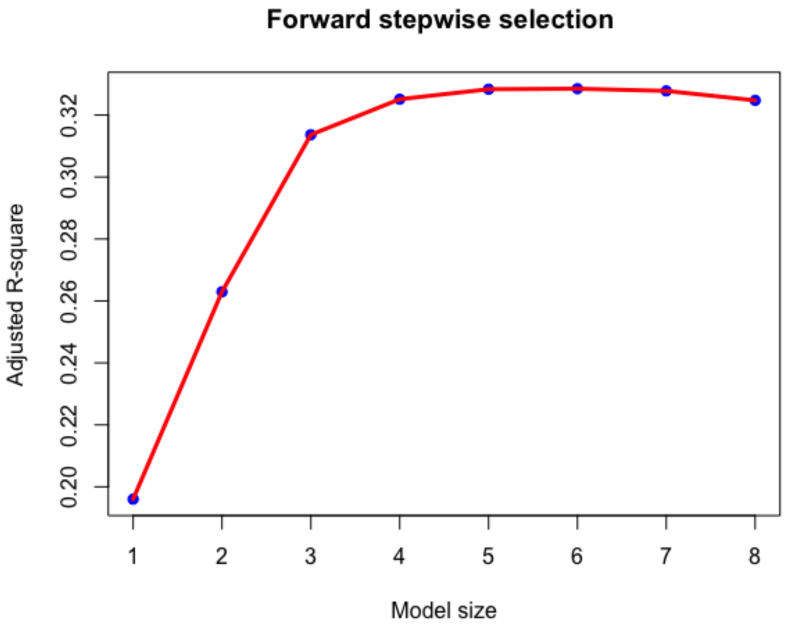


Figure 12: Adj. R-square versus model size

3.4 Question 4

Choose the best model from above eight models by 5-Fold CV, and get the CV errors in Figure 13. Obviously, the 3rd model has the smallest CV error, so the 3rd model is the best model according to CV error.

From Table 5 we can get that the 3rd model's predictors are 'HSGPA', 'HU', and 'White'. Report the summary of this model in Figure 14.

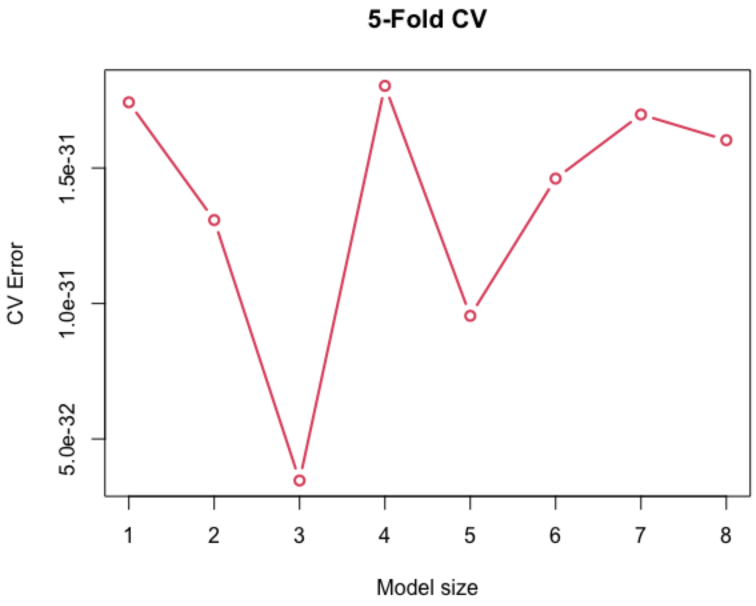
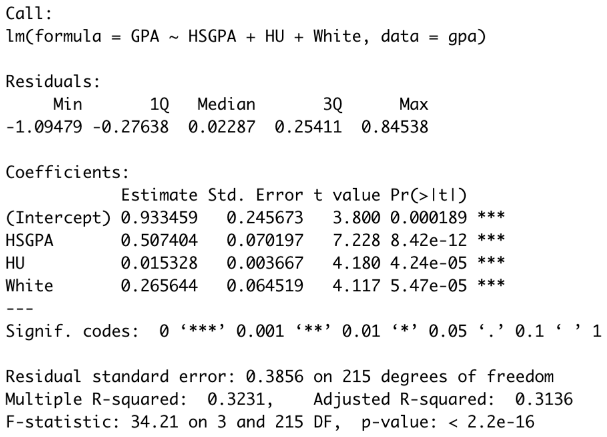


Figure 13: CV error versus model size



```
Call:
lm(formula = GPA ~ HSGPA + HU + White, data = gpa)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.09479 -0.27638  0.02287  0.25411  0.84538 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.933459   0.245673   3.800 0.000189 ***
HSGPA       0.507404   0.070197   7.228 8.42e-12 ***
HU          0.015328   0.003667   4.180 4.24e-05 ***
White       0.265644   0.064519   4.117 5.47e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3856 on 215 degrees of freedom
Multiple R-squared:  0.3231,	Adjusted R-squared:  0.3136 
F-statistic: 34.21 on 3 and 215 DF,  p-value: < 2.2e-16
```

Figure 14: Summary of the best model

4 Problem 4: Prediction of the Progression of Diabetes

4.1 Question 1

In the dataset, 'age', 'sex', 'bmi' are all standardized, and the value of other predictors are all belong to [-1,1], so the dimensions(scale) of the predictors in the dataset are the same. Thus, we don not need to perform standardization on inputs before trying LASSO regression.

Use the train dataset to fit LASSO estimators with regularization parameter λ chosen from the grid $10^{seq(4,-2,length=100)}$. Plot the coefficients versus the l_1 norm in Figure 15.

We can find that as l_1 norm increases, the absolute value of some coefficients increases first, while the absolute value of some coefficients increases later. And the absolute value of some coefficients decrease latter with l_1 norm increase. In the end, most of the coefficients tend to be stable and change little with l_1 norm increase, while there are still a few coefficients continue to increase with l_1 norm increase.

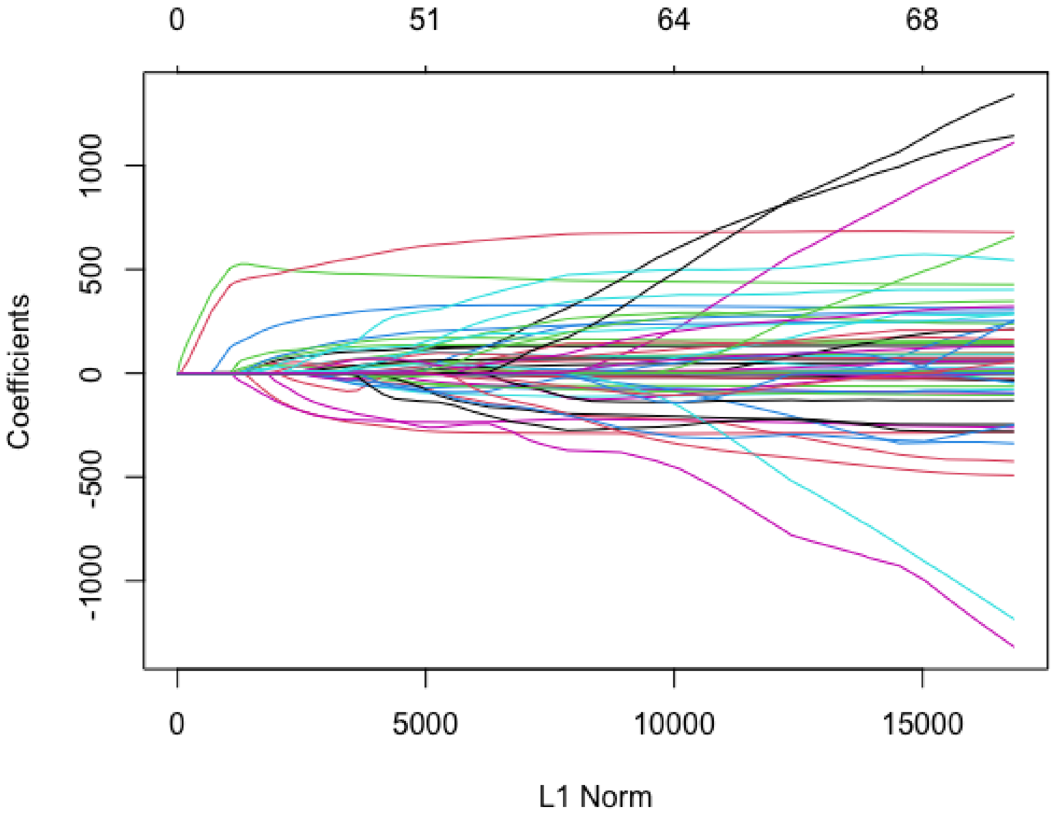


Figure 15: Coefficients versus l_1 norm

4.2 Question 2

Use train dataset to fit LASSO and apply 10-Fold cross validation. Plot the CV error versus the values of λ in Figure 16.

According CV error, the best λ value is 2.838049, then report the linear model using the best λ:

1. All coefficients are shown in Figure 17, and we only care about non-zero coefficients.
2. There are 25 variables actually included in the model, they are: sex, bmi, map, hdl, ltg, glu, sex.1, bmi.1, map.1, hdl.1, ltg.1, glu.1, age.2, bmi.2, ltg.2, glu.2, age.sex, age.ltg, sex.bmi, bmi.map, bmi.ltg, tc.hdl, tc.tch, ldl.ltg, tch.glu

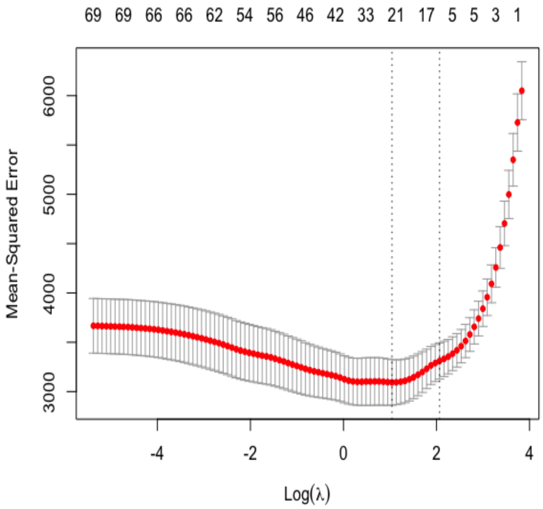
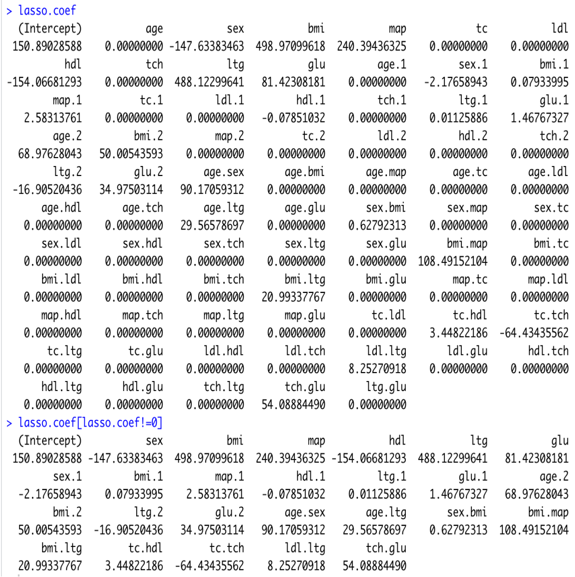


Figure 16: CV error versus λ

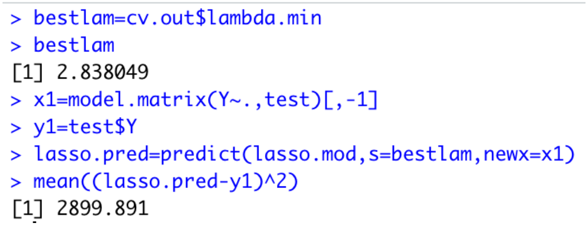


```
> lasso.coef
   (Intercept)           age           sex           bmi           map            tc           ldl
  150.89028588    0.00000000 -147.63383463  498.97099618  240.39436325    0.00000000    0.00000000
           hdl           tch           ltg           glu         age.1         sex.1         bmi.1
 -154.06681293    0.00000000  488.12299641   81.42308181    0.00000000   -2.17658943    0.07933995
         map.1          tc.1         ldl.1         hdl.1         tch.1         ltg.1         glu.1
    2.58313761    0.00000000    0.00000000   -0.07851032    0.00000000    0.01125886    1.46767327
         age.2         bmi.2         map.2          tc.2         ldl.2         hdl.2         tch.2
   68.97628043   50.00543593    0.00000000    0.00000000    0.00000000    0.00000000    0.00000000
         ltg.2         glu.2       age.sex       age.bmi       age.map        age.tc       age.ldl
  -16.90520436   34.97503114   90.17059312    0.00000000    0.00000000    0.00000000    0.00000000
       age.hdl       age.tch       age.ltg       age.glu       sex.bmi       sex.map        sex.tc
    0.00000000    0.00000000   29.56578697    0.00000000    0.62792313    0.00000000    0.00000000
       sex.ldl       sex.hdl       sex.tch       sex.ltg       sex.glu       bmi.map        bmi.tc
    0.00000000    0.00000000    0.00000000    0.00000000    0.00000000  108.49152104    0.00000000
       bmi.ldl       bmi.hdl       bmi.tch       bmi.ltg       bmi.glu        map.tc       map.ldl
    0.00000000    0.00000000    0.00000000   20.99337767    0.00000000    0.00000000    0.00000000
       map.hdl       map.tch       map.ltg       map.glu        tc.ldl        tc.hdl        tc.tch
    0.00000000    0.00000000    0.00000000    0.00000000    0.00000000    3.44822186  -64.43435562
        tc.ltg        tc.glu       ldl.hdl       ldl.tch       ldl.ltg       ldl.glu       hdl.tch
    0.00000000    0.00000000    0.00000000    0.00000000    8.25270918    0.00000000    0.00000000
       hdl.ltg       hdl.glu       tch.ltg       tch.glu       ltg.glu
    0.00000000    0.00000000    0.00000000   54.08884490    0.00000000
> lasso.coef[lasso.coef!=0]
   (Intercept)           sex           bmi           map           hdl           ltg           glu
  150.89028588 -147.63383463  498.97099618  240.39436325 -154.06681293  488.12299641   81.42308181
         sex.1         bmi.1         map.1         hdl.1         ltg.1         glu.1         age.2
   -2.17658943    0.07933995    2.58313761   -0.07851032    0.01125886    1.46767327   68.97628043
         bmi.2         ltg.2         glu.2       age.sex       age.ltg       sex.bmi       bmi.map
   50.00543593  -16.90520436   34.97503114   90.17059312   29.56578697    0.62792313  108.49152104
       bmi.ltg        tc.hdl        tc.tch       ldl.ltg       tch.glu
   20.99337767    3.44822186  -64.43435562    8.25270918   54.08884490
```

Figure 17: Coefficients of LASSO regression

4.3 Question 3

Then use the best model in Question 2 to predict the progression of diabetes on the test dataset, and get the mean test error is 2899.891 as Figure 18 shows.



```
> bestlam=cv.out$lambda.min
> bestlam
[1] 2.838049
> x1=model.matrix(Y~.,test)[,-1]
> y1=test$Y
> lasso.pred=predict(lasso.mod,s=bestlam,newx=x1)
> mean((lasso.pred-y1)^2)
[1] 2899.891
```

Figure 18: Mean test error

5 Appendix

Use python in Problem 1 and Problem 2.

Use R in Problem 3 and Problem 4.

All codes can be seen in the following link:

https://github.com/KlausZhangjt/HKUST_DDM/tree/main/5054/Mid_Project