College students GPA data analysis with SAS

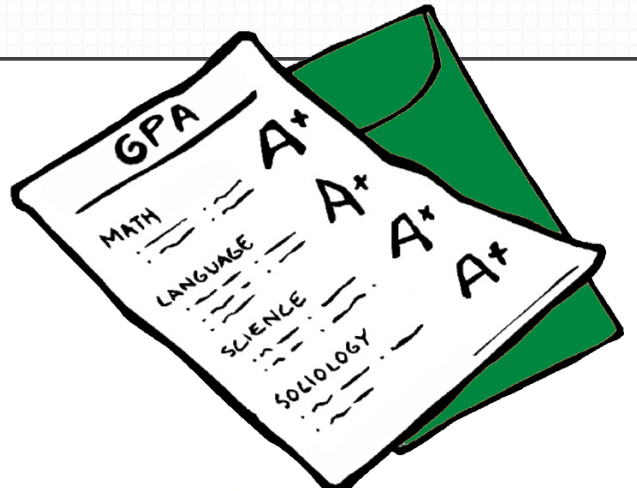
Speakers: 11711331刘润祺
11712621 张子怡
11712804 郭明磊

Contents

1 Introduction

Background

- GPA stands for Grade Point Average. It is a standard way of measuring academic achievement in the U.S and many other countries.



- Today, PC plays a significant role in universities and undergraduate's study. Students are able to look up information, complete their work on the computer and entertain themselves with computers.

Objectives

- **GPA in college**

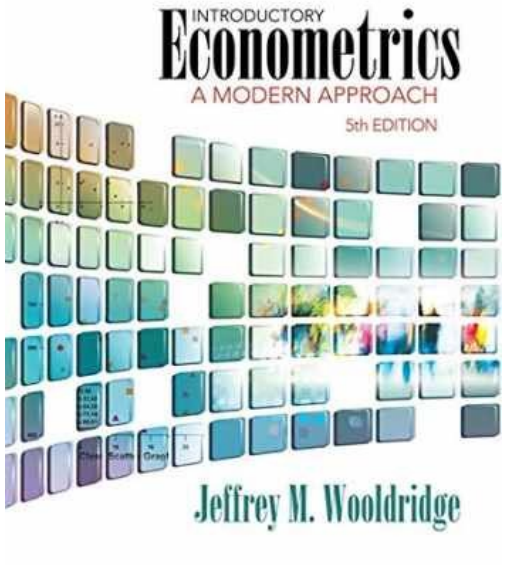
1. Determine what might affect students' GPA in college.
2. Among several factors that might have impacts on GPA, find out the ones that have deeper influence.
3. Construct a model to describe the relationship between the GPA and influencing factors.

- **PC**

1. Recognize the type of students that are more likely having a PC.
2. Build a regression model to predict whether a student has a PC based on certain variables.

Data

- The data we used in this study are extracted from Introductory Econometrics: A Modern Approach (5th edition) by Jeffrey M. Wooldridge.
- There are 29 variables in data GPA1.RAW, including the college grade point average (colGPA), high school GPA (hsGPA), possession of PC (PC) and several other information of a sample of 141 students from a large university.
- Both college and high school GPAs are on a four-point scale.



Data

Obs: 141

1. age	in years
2. soph	=1 if sophomore
3. junior	=1 if junior
4. senior	=1 if senior
5. senior5	=1 if fifth year senior
6. male	=1 if male
7. campus	=1 if live on campus
8. business	=1 if business major
9. engineer	=1 if engineering major
10. colGPA	MSU GPA
11. hsGPA	high school GPA
12. ACT	'achievement' score
13. job19	=1 if job <= 19 hours
14. job20	=1 if job >= 20 hours
15. drive	=1 if drive to campus
16. bike	=1 if bicycle to campus
17. walk	=1 if walk to campus
18. voluntr	=1 if do volunteer work
19. PC	=1 of pers computer at sch
20. greek	=1 if fraternity or sorority
21. car	=1 if own car
22. siblings	=1 if have siblings
23. bgfriend	=1 if boy- or girlfriend
24. clubs	=1 if belong to MSU club
25. skipped	avg lectures missed per week
26. alcohol	avg # days per week drink alcohol
27. gradMI	=1 if Michigan high school
28. fathcoll	=1 if father college grad
29. mothcoll	=1 if mother college grad

2 Exploratory data analysis

Data processing

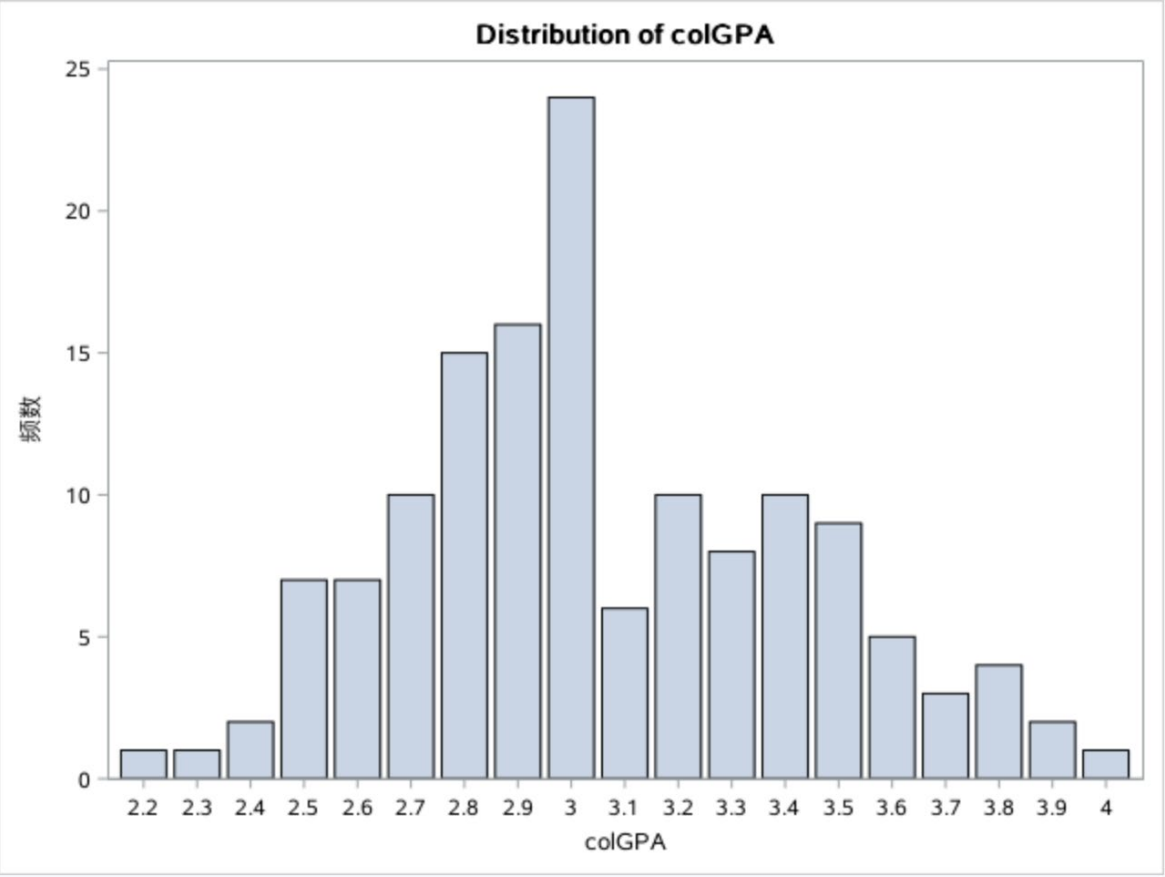
- No missing values; 141 observations in total; 56 out of 141 have their own PC.
- To simplify the data, we combined related binary variables into one variable with several levels and discretize age.

1. Grade: “soph or junior”, “senior” , “senior5”
2. Transport: “drive”, “bike”, “walk”
3. Job: “no job”, “<=19h”, “>=20”
4. Major: “business”, “engineer”, “other”
5. Parcoll: 0 (if neither of parents are college graduated),1(if one of parents is college graduated), and 2 (if both of parents are college graduated).
6. Age: “19”, “20”, “21”, “22” ,“23+”.

Correlation and multicollinearity

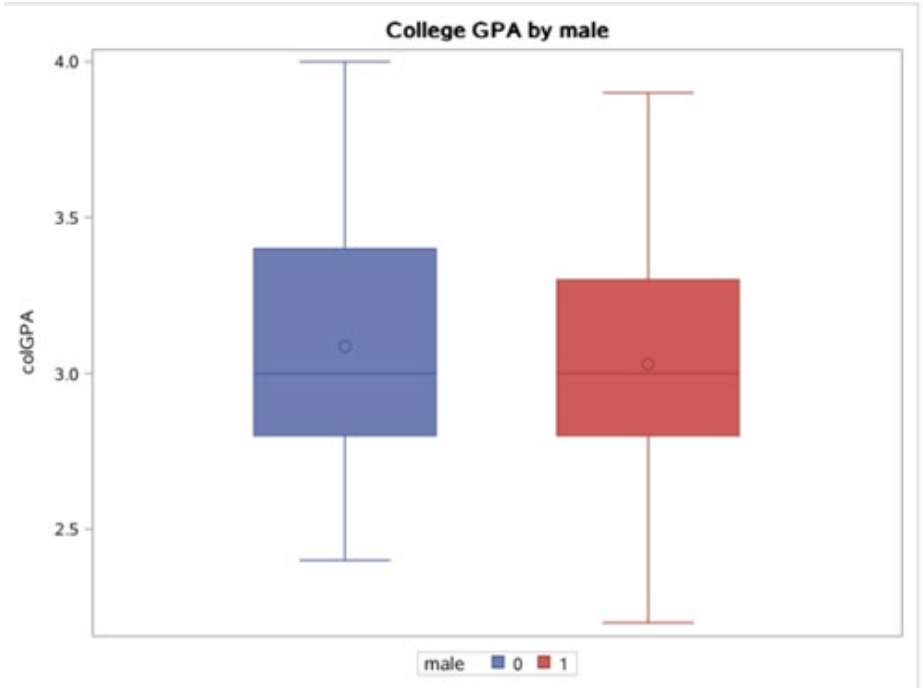
- The correlation matrix of raw data: No high correlation is revealed; some variables may influence colGPA, such as senior (r=-0.10), hsGPA (r=0.41), ACT(r=0.21), drive (r=-0.11), car (r=-0.12), clubs (r=0.16), skipped (r=-0.26) and gradMI (r=0.17)
- Looked for the variables with relatively strong relationship with PC: The correlation matrix suggests that colGPA (r=0.22), skipped (r=-0.207) and parcoll (r=0.20) are likely to determine whether a student has a PC
- Multicollinearity does not exist.

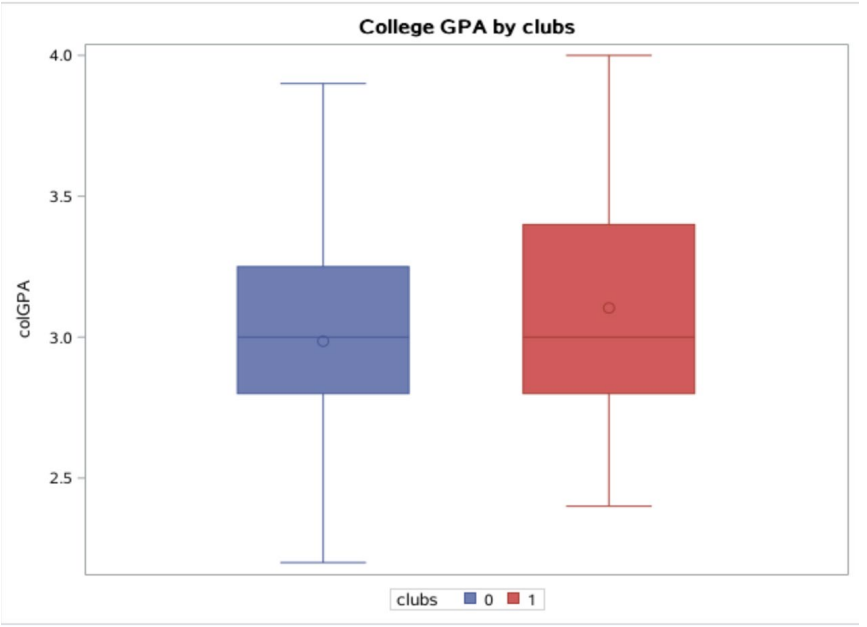
Understanding colGPA



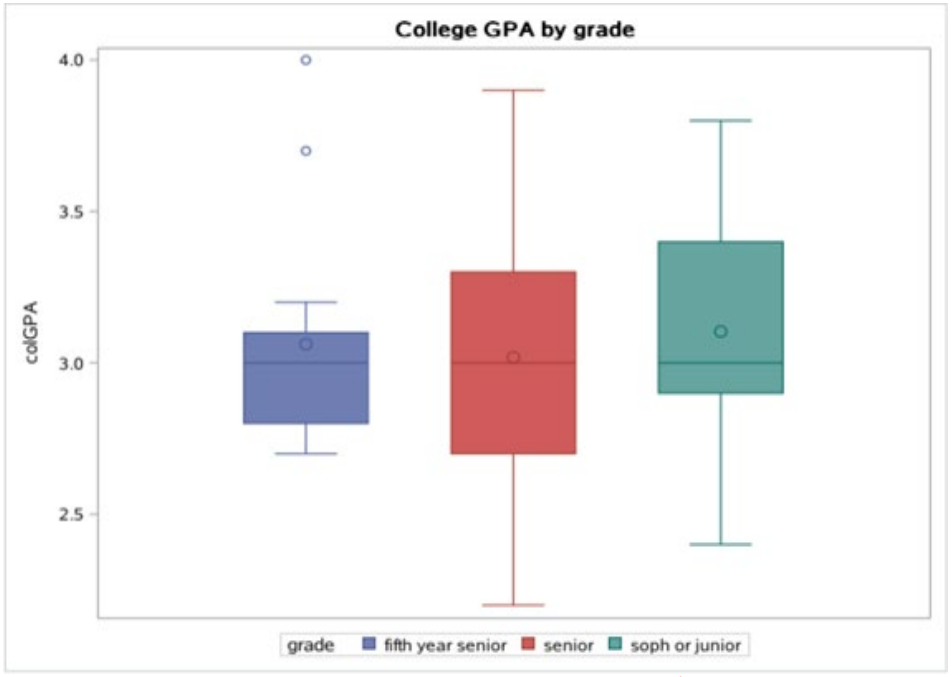
Despite the odd in colGPA=3.1, the colGPA approximately follows Normal distribution.

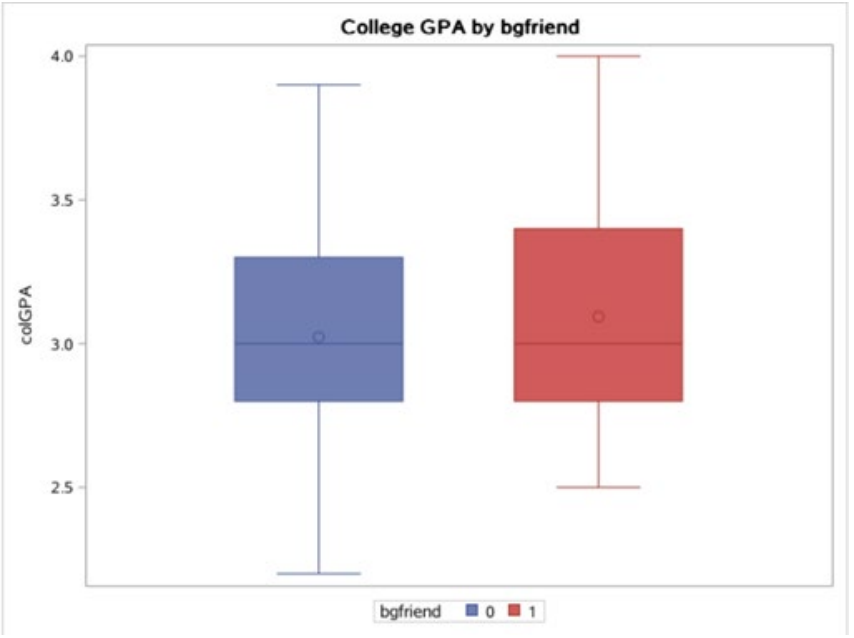
Understanding colGPA



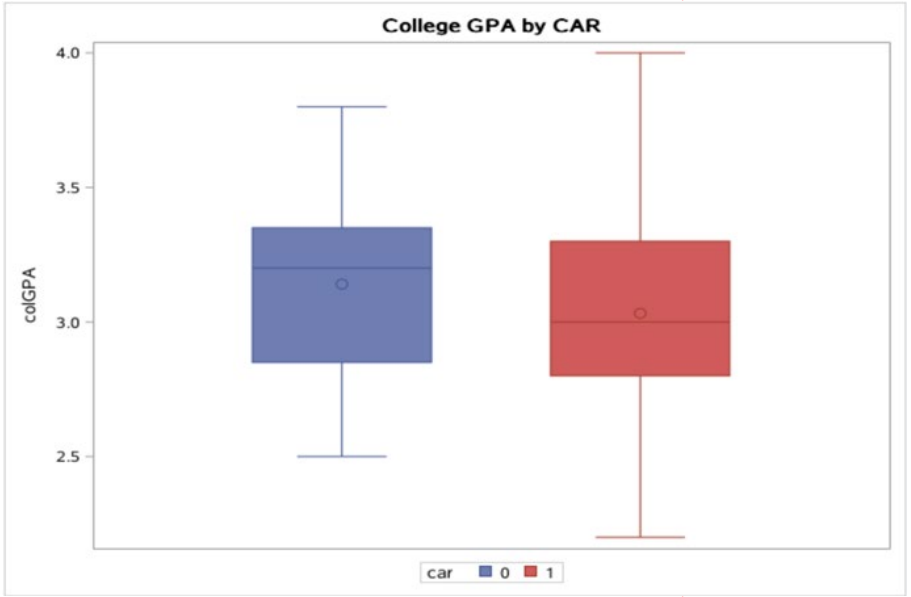


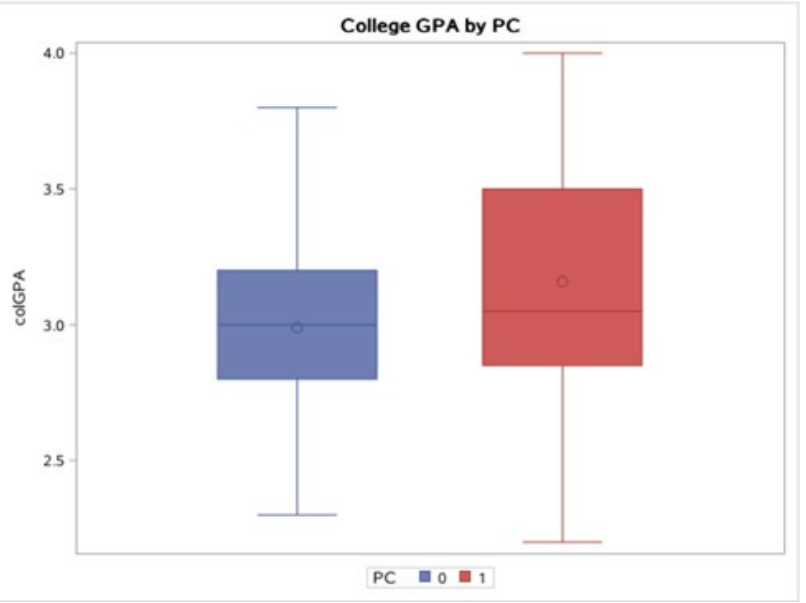
Understanding colGPA



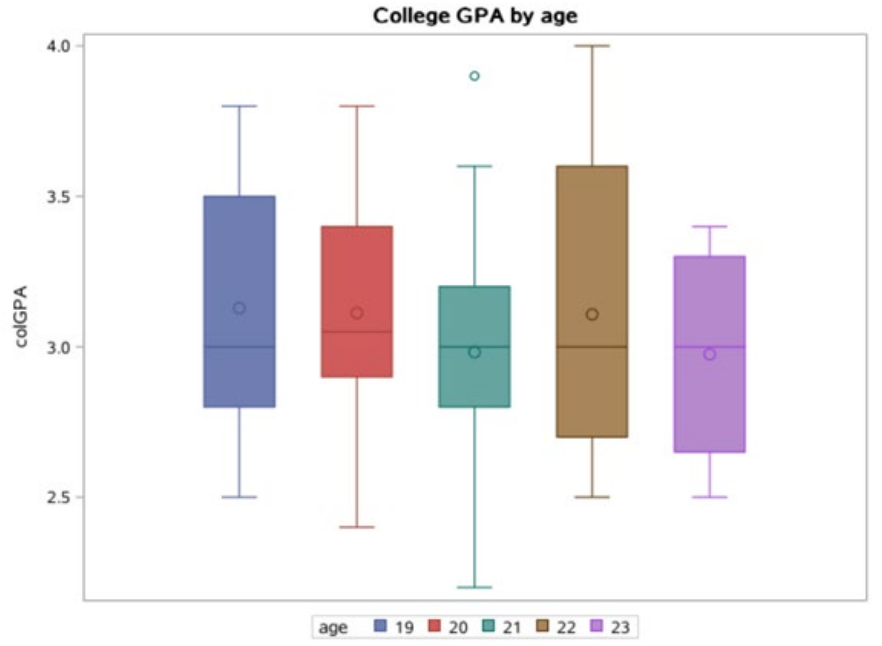


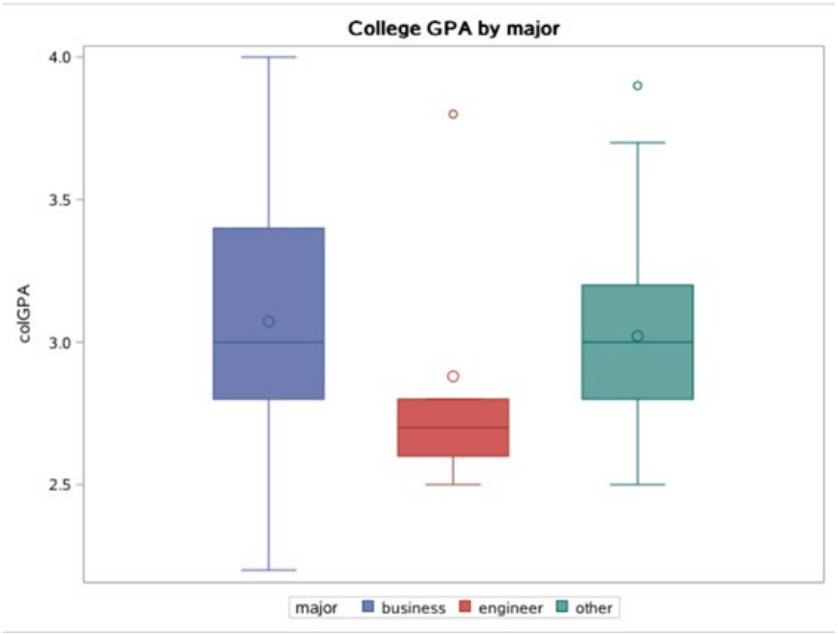
Understanding colGPA



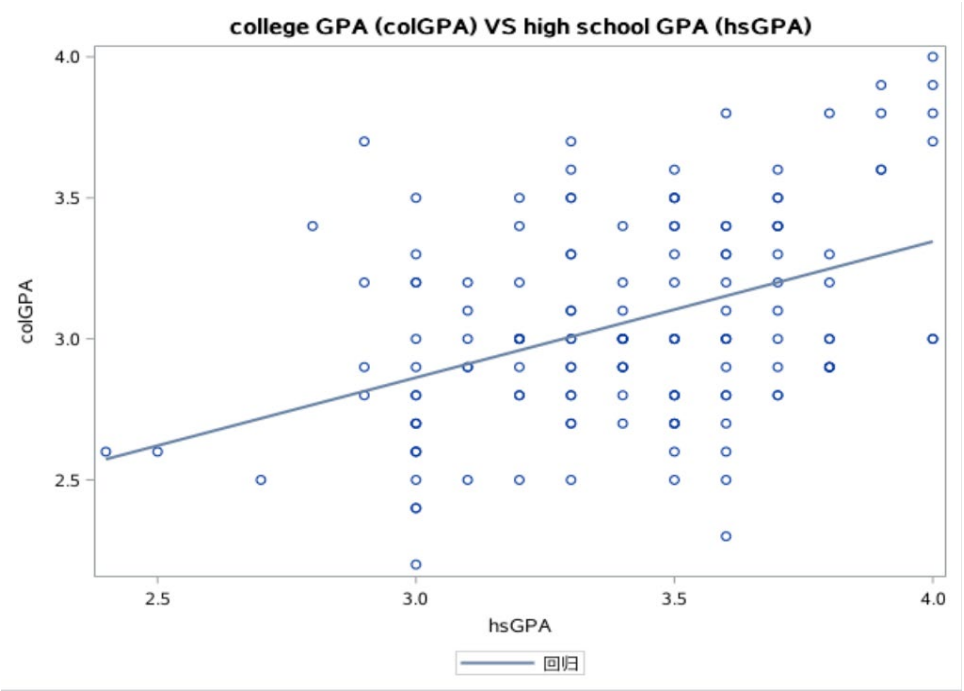


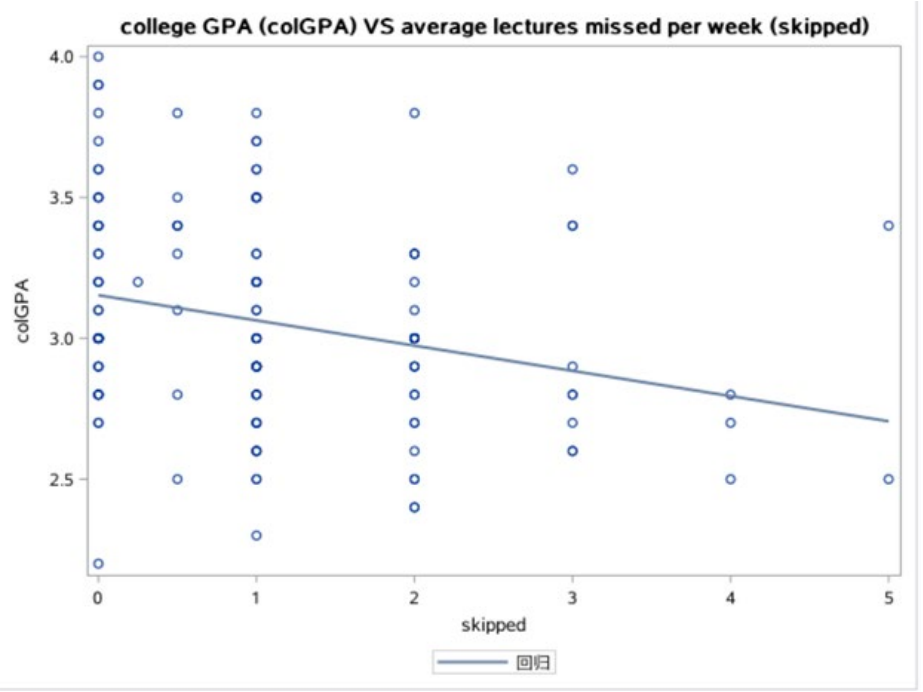
Understanding colGPA



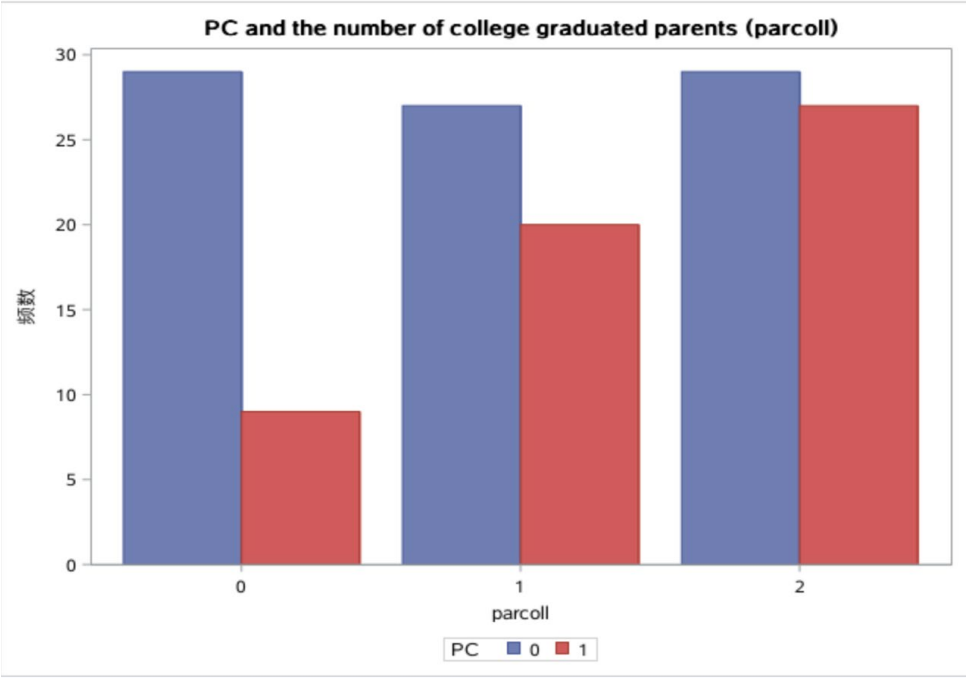


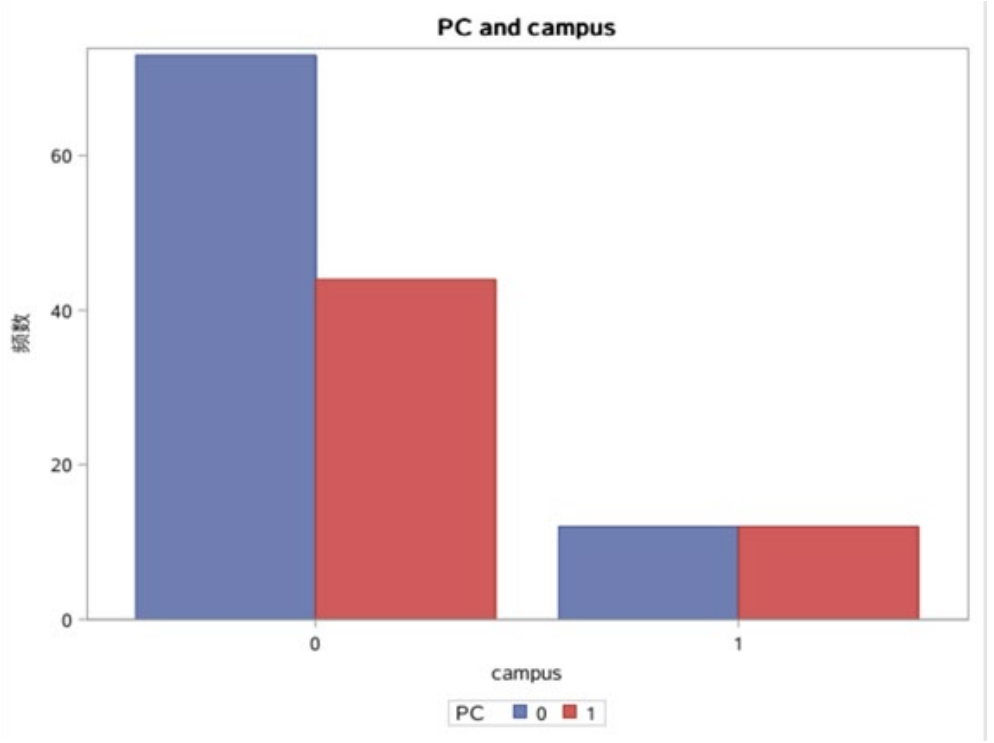
Understanding colGPA



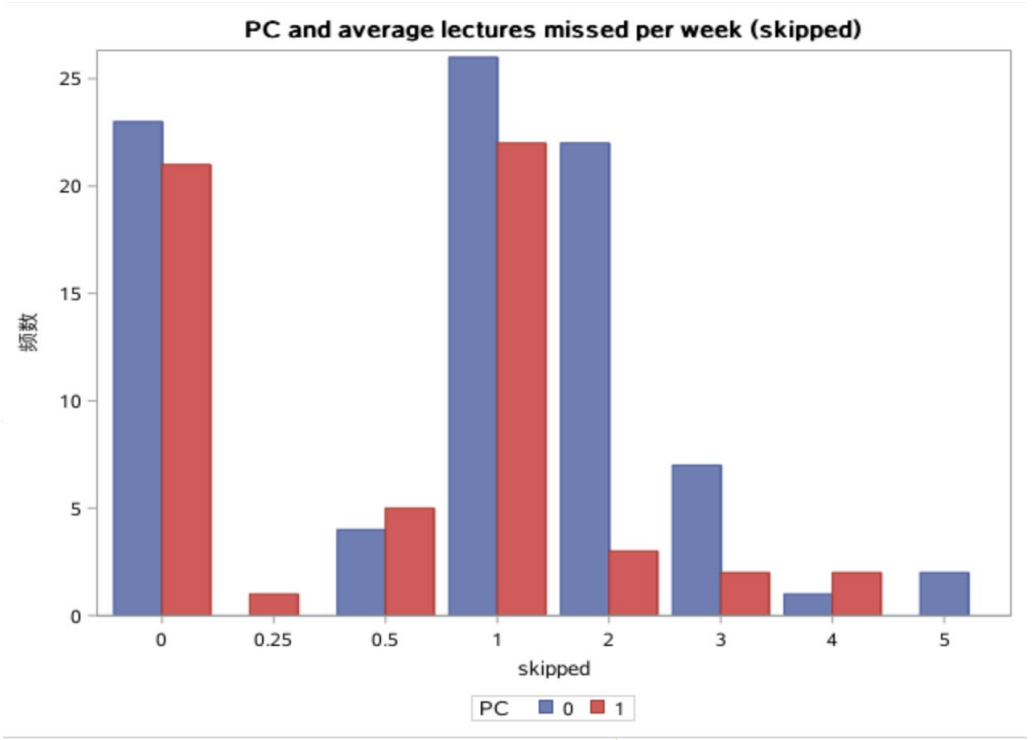


Understanding PC





Understanding PC



3 Modeling

- For colGPA: Linear Model
- For PC: Logistic Regression Model

For colGPA

LM

LM: Train and Test Set

```
/* Randomly split the data into train (70%) and test (30%) */
DATA GPA_LM_Train GPA_LM_Test;
        SET GPA_LM;
        CALL STREAMINIT(520);
        IF RAND("Uniform") <= 0.3 THEN OUTPUT GPA_LM_Test;
        ELSE OUTPUT GPA_LM_Train;
RUN;
```

LM: All Variables, Remove High VIF

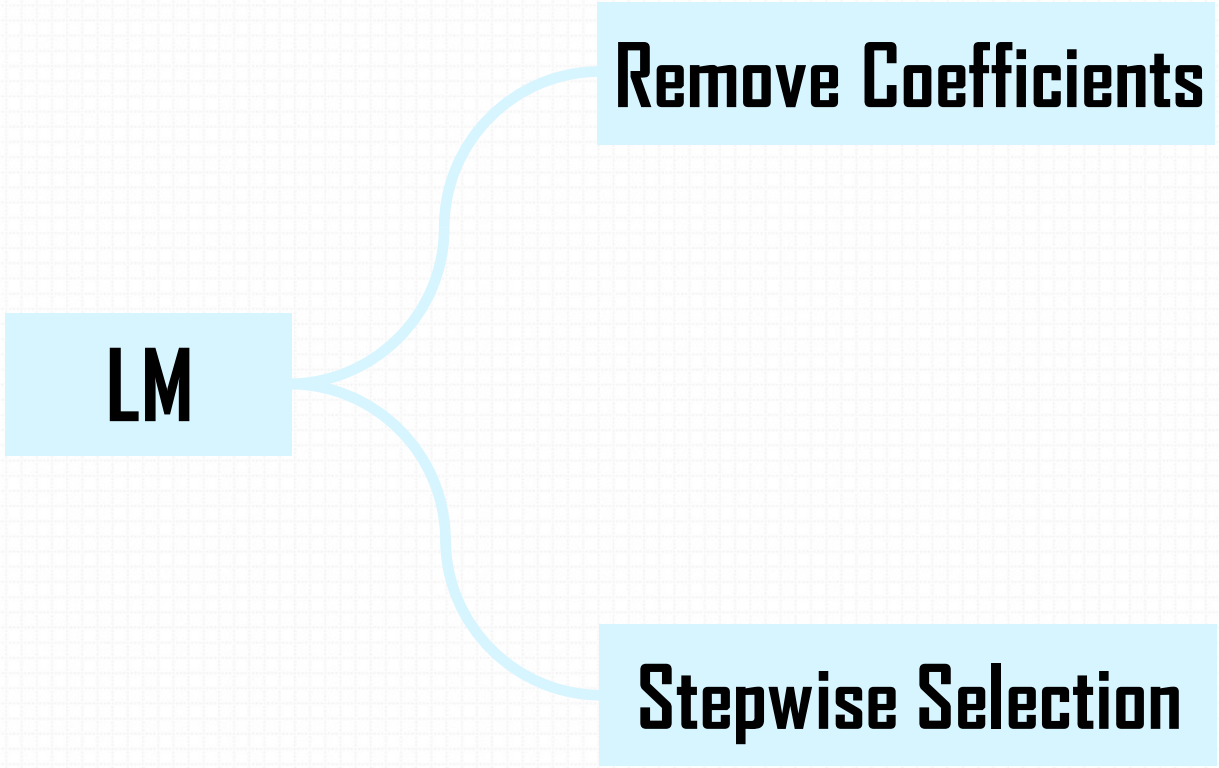
参数估计

变量	自由度	参数估计	标准误差	t 值	Pr > |t|	方差膨胀
Intercept	B	1.46055	0.88213	1.66	0.1017	0
ACT	1	-0.00053107	0.01312	-0.04	0.9678	1.65314
age	1	0.04770	0.03325	1.43	0.1553	2.19125
alcohol	1	0.02871	0.02911	0.99	0.3269	1.92811
bgfriend	1	0.12870	0.06664	1.93	0.0570	1.19625
bike	B	-0.00184	0.08283	-0.02	0.9824	1.67222
business	1	0.08541	0.09101	0.94	0.3508	1.50610
campus	1	-0.05855	0.09080	-0.64	0.5209	1.29750
car	1	-0.08950	0.08657	-1.03	0.3043	1.31895
clubs	1	0.13111	0.06892	1.90	0.0607	1.23032
drive	B	-0.06302	0.10332	-0.61	0.5436	2.05980
engineer	1	0.16201	0.22936	0.71	0.4820	1.54474
fathcoll	1	0.10924	0.07898	1.38	0.1705	1.57357
gradMI	1	0.14026	0.10837	1.29	0.1993	1.35060
greek	1	0.11870	0.07533	1.58	0.1190	1.36510
hsGPA	1	0.33464	0.12146	2.76	0.0073	1.78408
job19	1	-0.01128	0.07825	-0.14	0.8857	1.55960
job20	1	-0.02718	0.09753	-0.28	0.7812	1.43431
junior	B	-0.57900	0.25840	-2.24	0.0278	16.06254
male	1	-0.01111	0.08041	-0.14	0.8904	1.74161
mothcoll	1	-0.13406	0.07761	-1.73	0.0880	1.59055
PC	1	0.14670	0.07001	2.10	0.0393	1.25928
senior	B	-0.71878	0.26710	-2.69	0.0087	19.13565
senior5	B	-0.65447	0.28807	-2.27	0.0258	8.90340
siblings	1	-0.13495	0.12414	-1.09	0.2802	1.27915
skipped	1	-0.05840	0.03642	-1.60	0.1127	1.67857
soph	0	0	.	.	.	.
voluntr	1	-0.18138	0.08763	-2.07	0.0417	1.43956
walk	0	0	.	.	.	.

- walk, senior

参数估计

变量	自由度	参数估计	标准误差	t 值	Pr > |t|	方差膨胀
Intercept	1	0.74177	0.87885	0.84	0.4012	0
ACT	1	-0.00053107	0.01312	-0.04	0.9678	1.65314
age	1	0.04770	0.03325	1.43	0.1553	2.19125
alcohol	1	0.02871	0.02911	0.99	0.3269	1.92811
bgfriend	1	0.12870	0.06664	1.93	0.0570	1.19625
bike	1	-0.00184	0.08283	-0.02	0.9824	1.67222
business	1	0.08541	0.09101	0.94	0.3508	1.50610
campus	1	-0.05855	0.09080	-0.64	0.5209	1.29750
car	1	-0.08950	0.08657	-1.03	0.3043	1.31895
clubs	1	0.13111	0.06892	1.90	0.0607	1.23032
drive	1	-0.06302	0.10332	-0.61	0.5436	2.05980
engineer	1	0.16201	0.22936	0.71	0.4820	1.54474
fathcoll	1	0.10924	0.07898	1.38	0.1705	1.57357
gradMI	1	0.14026	0.10837	1.29	0.1993	1.35060
greek	1	0.11870	0.07533	1.58	0.1190	1.36510
hsGPA	1	0.33464	0.12146	2.76	0.0073	1.78408
job19	1	-0.01128	0.07825	-0.14	0.8857	1.55960
job20	1	-0.02718	0.09753	-0.28	0.7812	1.43431
junior	1	0.13978	0.08041	1.74	0.0860	1.55555
male	1	-0.01111	0.08041	-0.14	0.8904	1.74161
mothcoll	1	-0.13406	0.07761	-1.73	0.0880	1.59055
PC	1	0.14670	0.07001	2.10	0.0393	1.25928
senior5	1	0.06431	0.11213	0.57	0.5679	1.34892
siblings	1	-0.13495	0.12414	-1.09	0.2802	1.27915
skipped	1	-0.05840	0.03642	-1.60	0.1127	1.67857
soph	1	0.71878	0.26710	2.69	0.0087	1.41000
voluntr	1	-0.18138	0.08763	-2.07	0.0417	1.43956

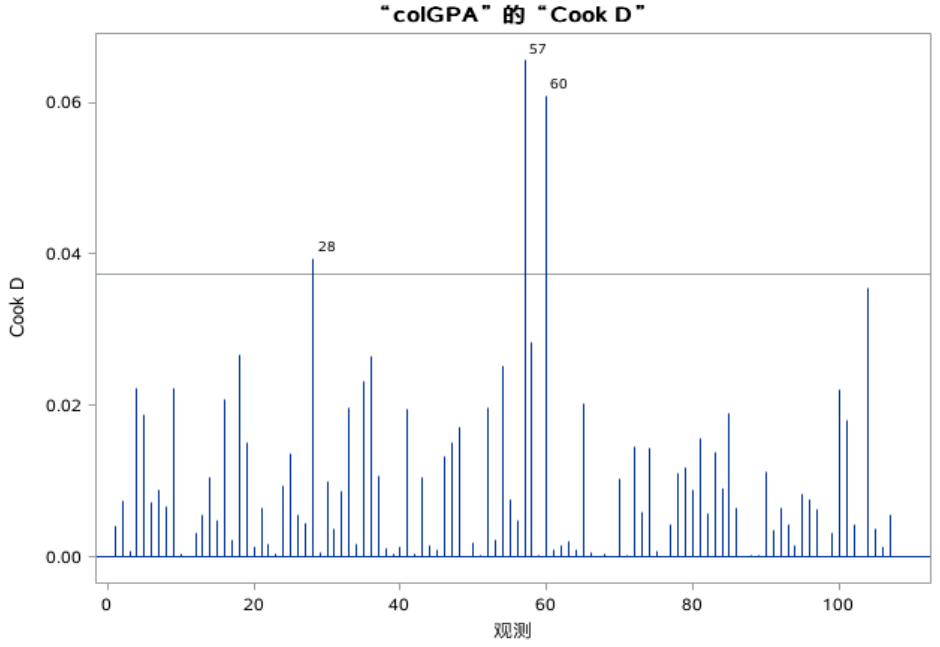

LM
Remove Coefficients
Stepwise Selection

LM1: Remove Coefficients

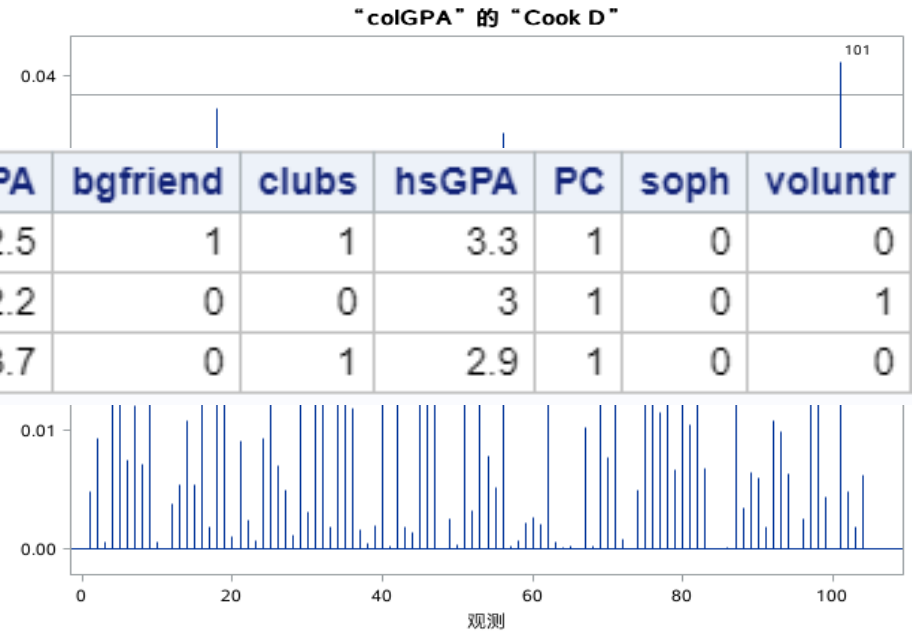
- Remove 4 times
- 1st p-value threshold is set to 0.15
- Next 3 times are all 0.10

参数估计						
变量	自由度	参数估计	标准误差	t 值	Pr > |t|	方差膨胀
Intercept	1	1.55279	0.31877	4.87	<.0001	0
bgfriend	1	0.11405	0.06311	1.81	0.0738	1.04759
clubs	1	0.11250	0.06586	1.71	0.0907	1.09709
hsGPA	1	0.40101	0.09568	4.19	<.0001	1.08082
PC	1	0.15334	0.06347	2.42	0.0175	1.01077
soph	1	0.46051	0.23349	1.97	0.0513	1.05205
voluntr	1	-0.12406	0.07646	-1.62	0.1078	1.07027

LM1: Influential Observations

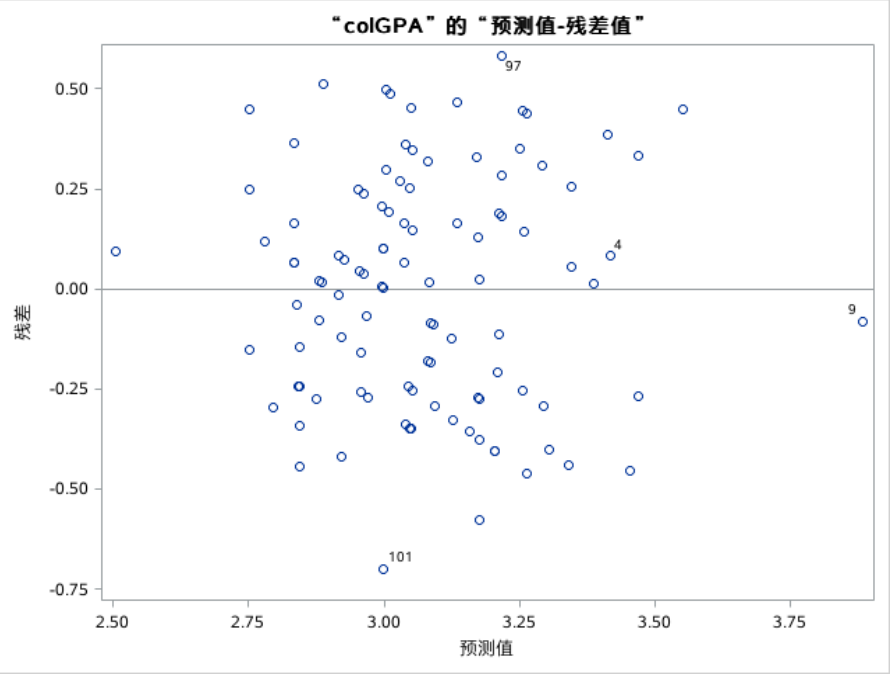


colGPA	bgfriend	clubs	hsGPA	PC	soph	voluntr
2.5	1	1	3.3	1	0	0
2.2	0	0	3	1	0	1
3.7	0	1	2.9	1	0	0



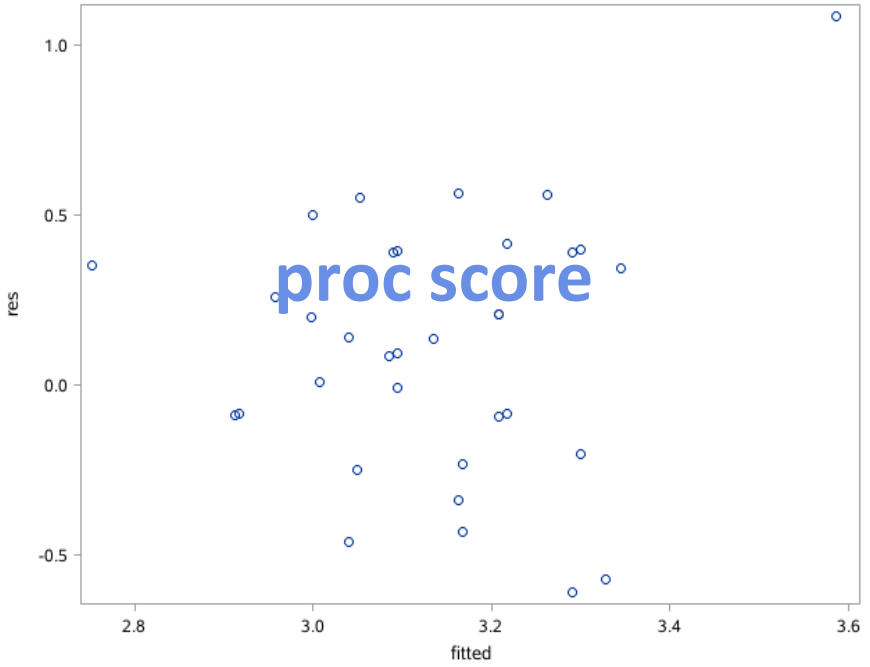
LM1: Model Diagnosis

Train Set



Shapiro-Wilk test: p-value=0.0661 > 0.05

Test Set



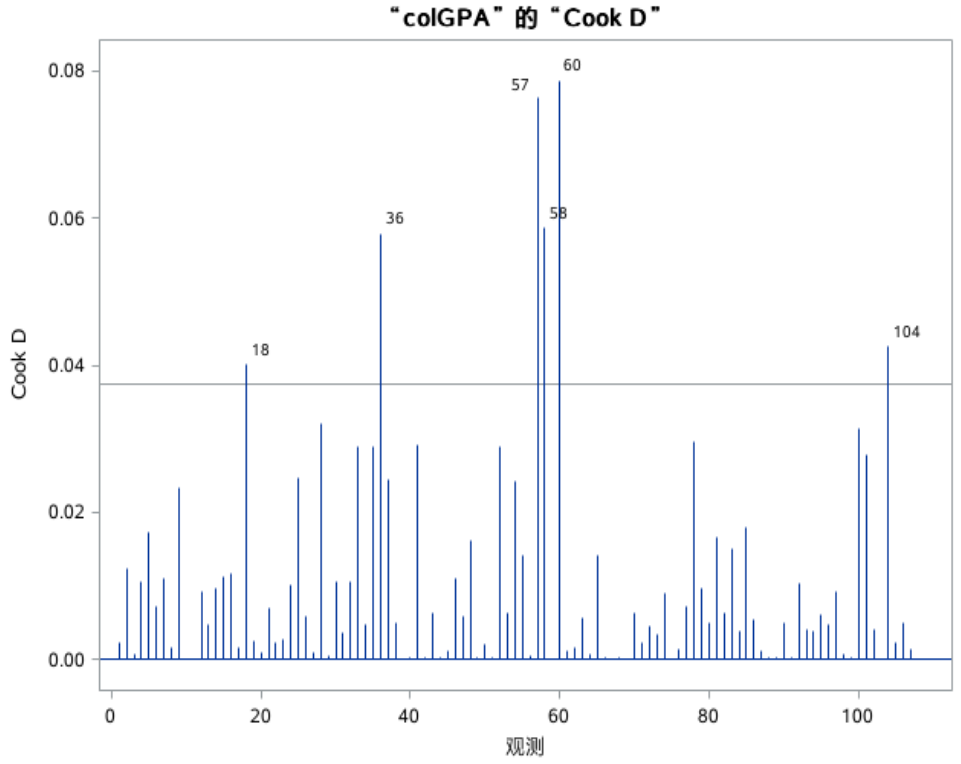
proc score

LM2: Stepwise Selection

- After removing the high VIF variable
- SBC as criteria

参数估计				
参数	自由度	估计	标准误差	t 值
Intercept	1	1.488028	0.321283	4.63
hsGPA	1	0.449694	0.094976	4.73
PC	1	0.167530	0.065156	2.57

LM2: Influential Observations

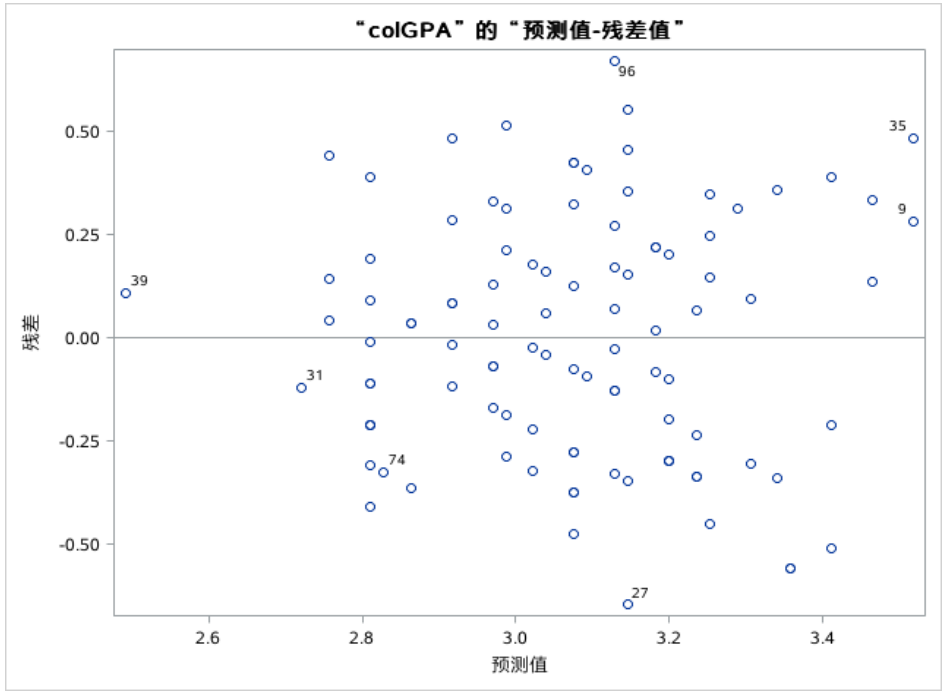


colGPA	hsGPA	PC
3	4	1
4	4	1
2.2	3	1
3.4	2.8	0
3.7	2.9	1
2.3	3.6	0

- 18, 57, 58, 60, 104

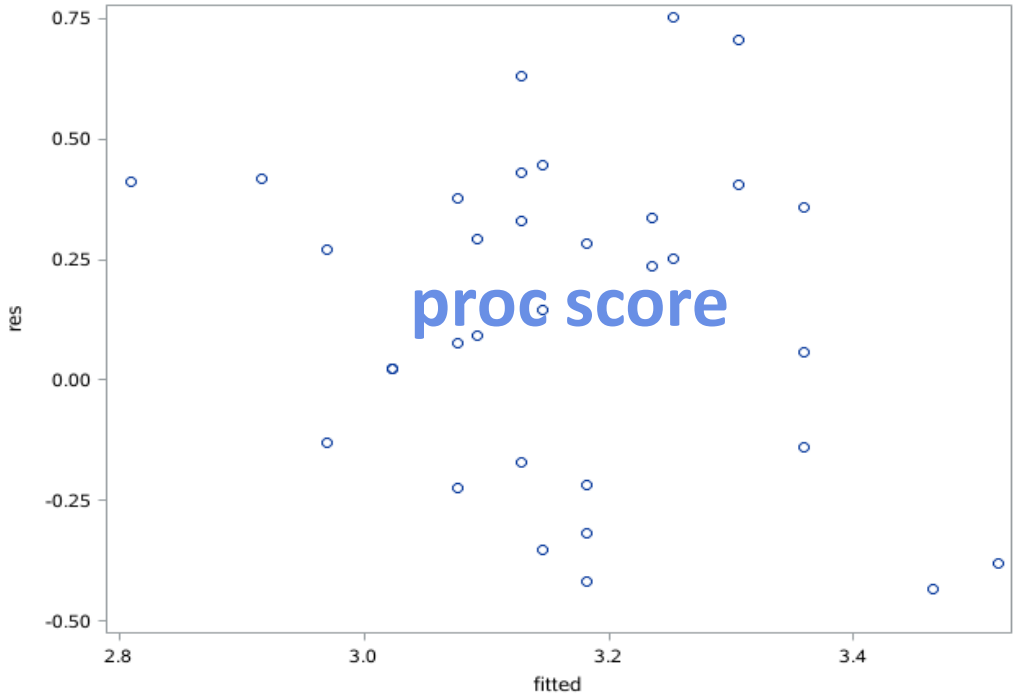
LM2: Influential Observations

Train Set



Shapiro-Wilk test: p-value=0.2262 > 0.05

Test Set



proc score

LM1 & LM2: A Notation

LM1

方差分析					
源	自由度	平方和	均方	F 值	Pr > F
模型	6	4.29563	0.71594	7.98	<.0001
误差	97	8.70427	0.08973		
校正合计	103	12.99990			

均方根误差	0.29956	R 方	0.3304
因变量均值	3.07404	调整 R 方	0.2890
变异系数	9.74476		

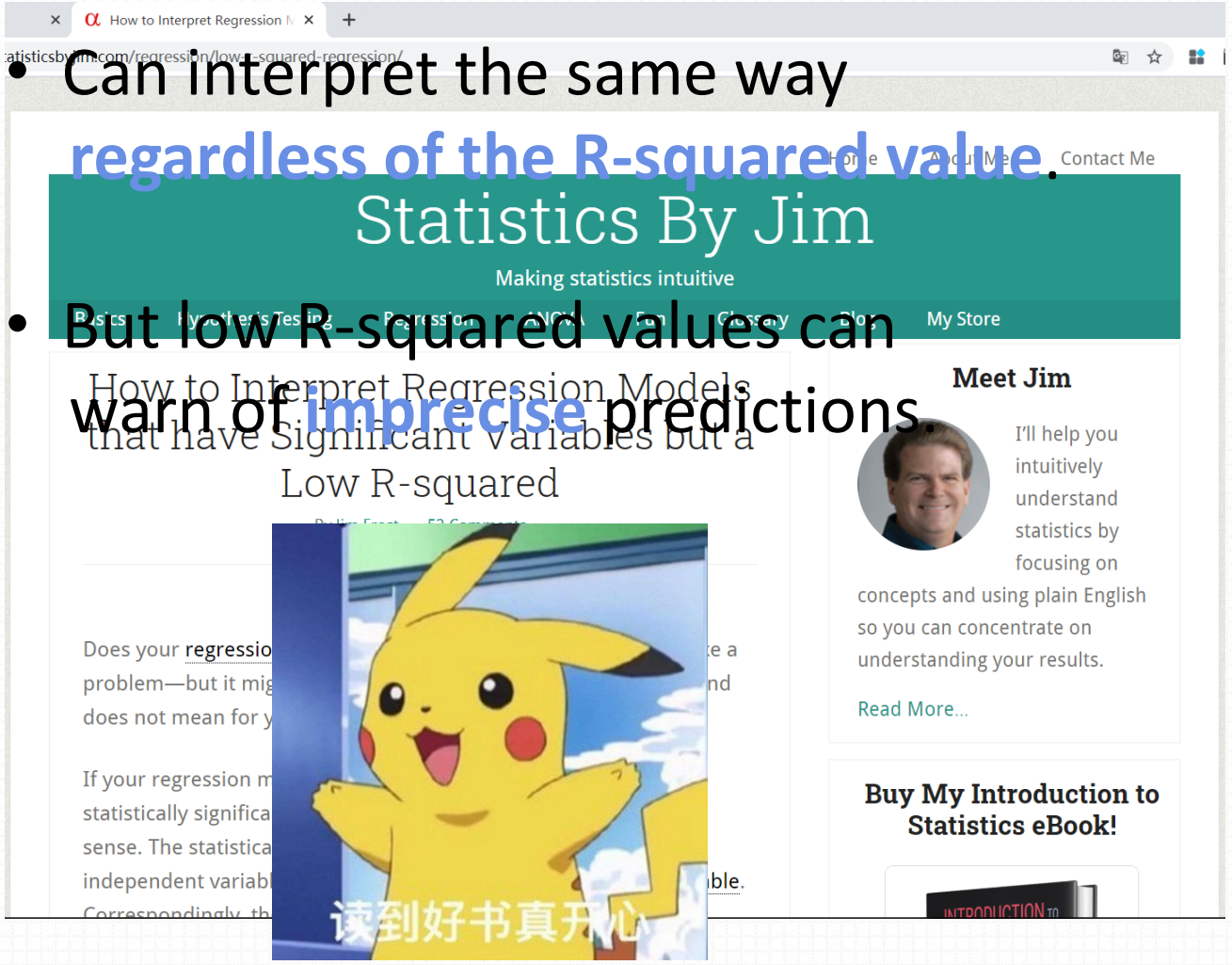
LM2

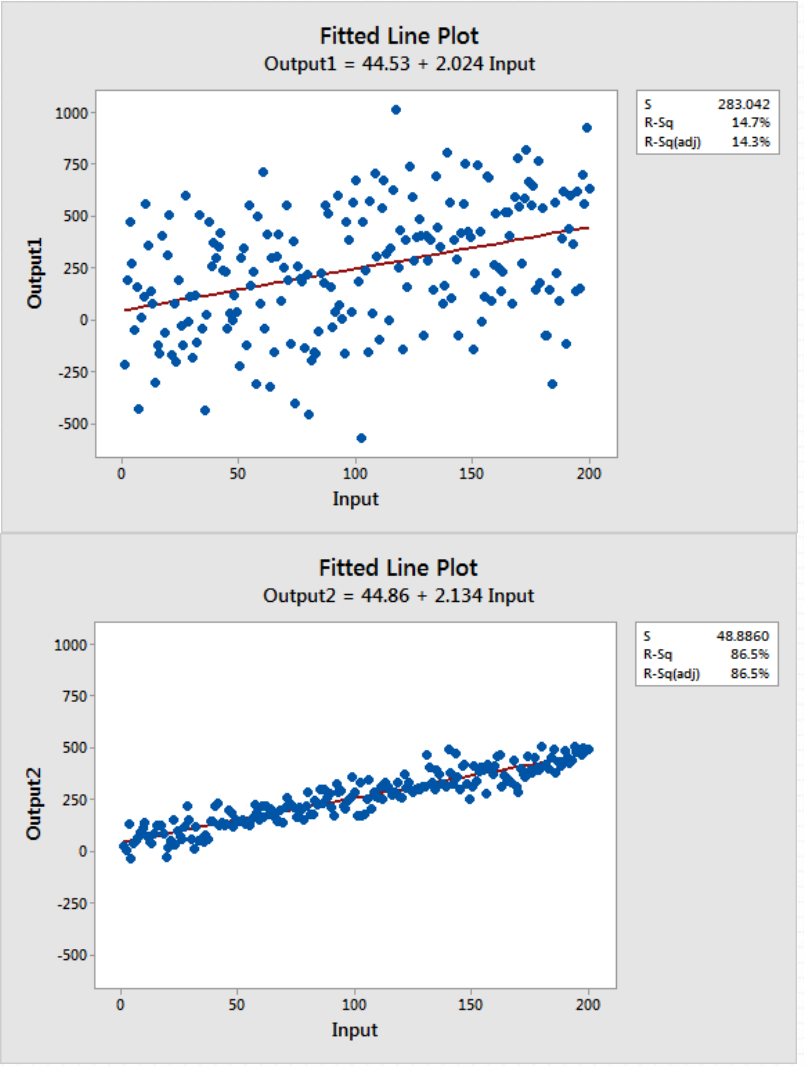
方差分析					
源	自由度	平方和	均方	F 值	Pr > F
模型	2	3.92333	1.96166	22.33	<.0001
误差	99	8.69520	0.08783		
校正合计	101	12.61853			

均方根误差	0.29636	R 方	0.3109
因变量均值	3.07353	调整 R 方	0.2970
变异系数	9.64239		

LM1 & LM2: A Notation

- Can interpret the same way **regardless of the R-squared value**.
- But low R-squared values can warn of **imprecise** predictions.

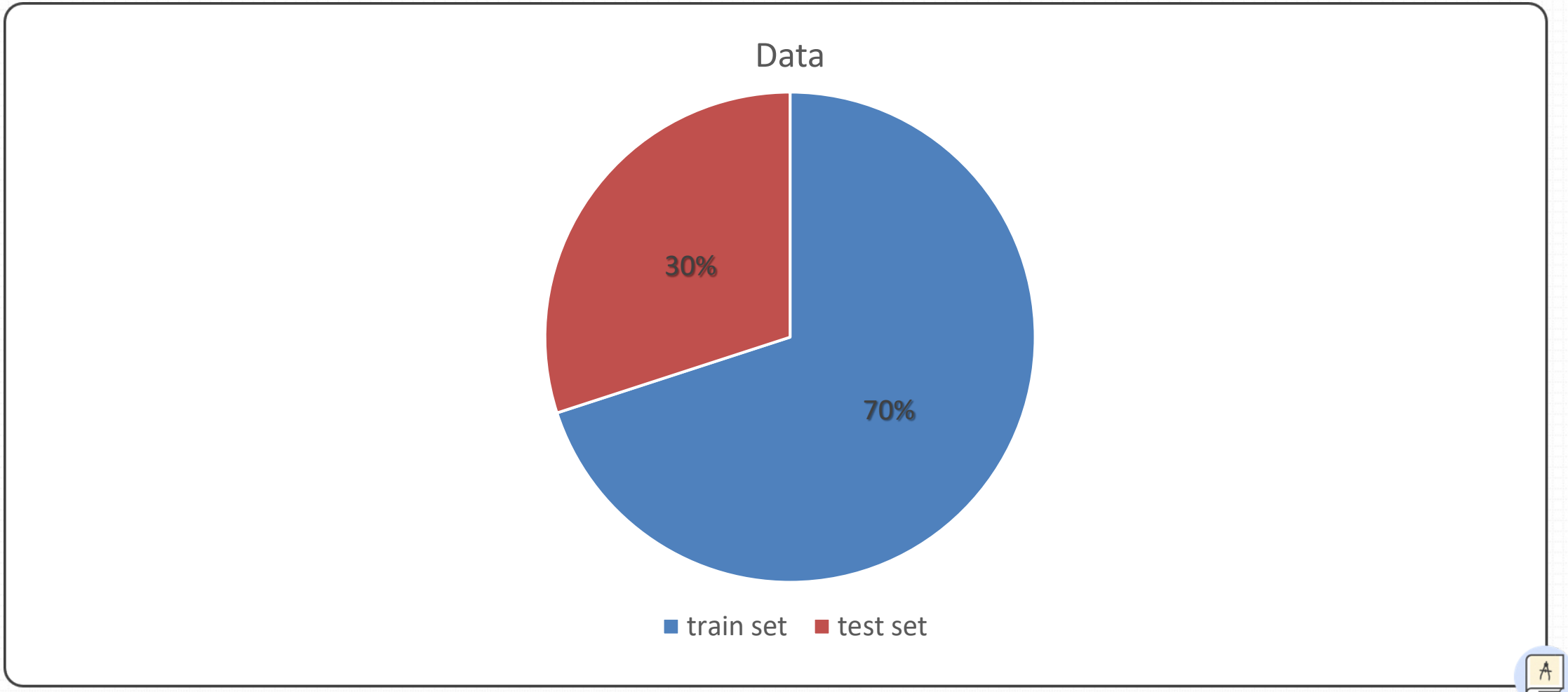


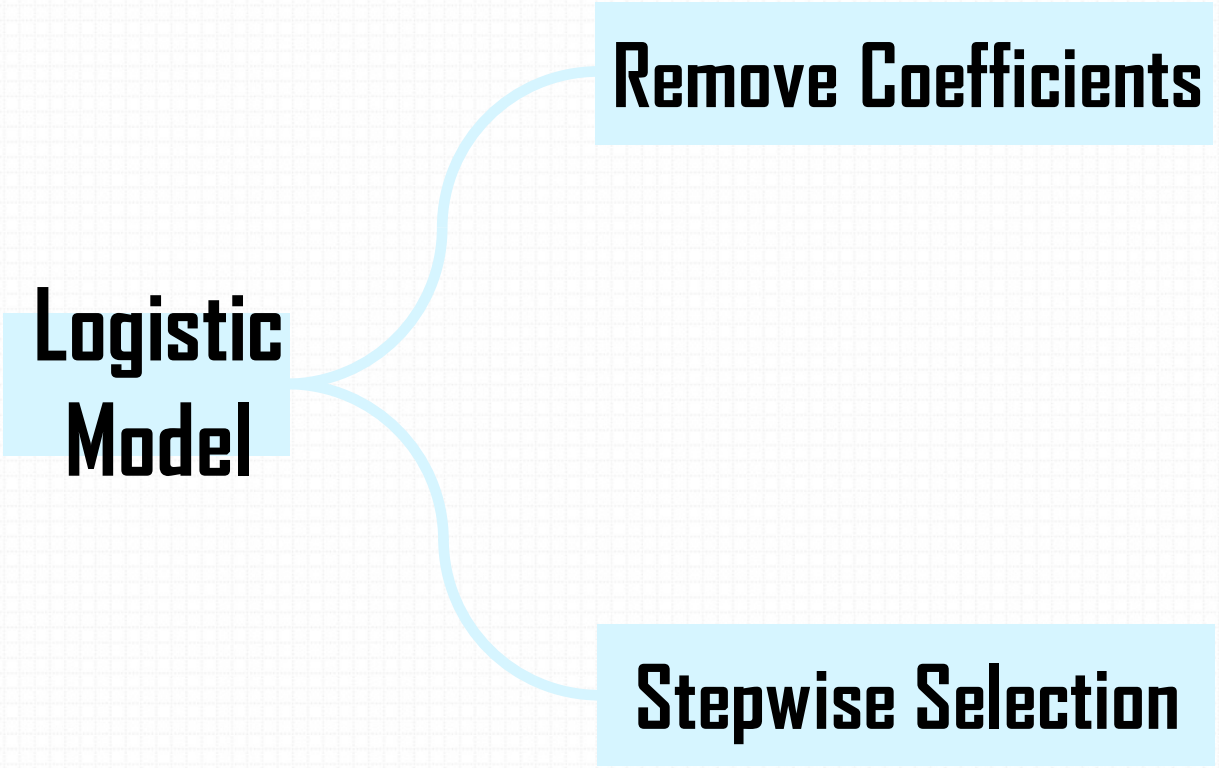


For PC possession

Logistic Model

Logistic Model: Train and Test Set




Logistic Model
Remove Coefficients
Stepwise Selection

Logistic Model 1: Remove Coefficients

- Remove 4 times
- **Unsure** of last removing step of deleting variable **skipped**

最大似然估计分析						
参数		自由度	估计	标准误差	Wald 卡方	Pr > 卡方
Intercept		1	-4.6205	1.9688	5.5078	0.0189
colGPA		1	1.1448	0.6033	3.6002	0.0578
parcoll	1	1	1.0074	0.6250	2.5983	0.1070
parcoll	2	1	1.3181	0.5917	4.9632	0.0259
skipped		1	-0.2863	0.2183	1.7196	0.1897

最大似然估计分析						
参数		自由度	估计	标准误差	Wald 卡方	Pr > 卡方
Intercept		1	-5.3988	1.8890	8.1681	0.0043
colGPA		1	1.3026	0.5903	4.8703	0.0273
parcoll	1	1	0.9864	0.6197	2.5335	0.1115
parcoll	2	1	1.3448	0.5871	5.2467	0.0220

Logistic Model 1: Remove Coefficients

Before deleting variable **skipped**

预测概率和观测响应的关联			
一致部分所占百分比	64.3	Somers D	0.293
不一致部分所占百分比	35.0	Gamma	0.295
结值百分比	0.7	Tau-a	0.146
对	280	c	0.646

After deleting variable **skipped**

预测概率和观测响应的关联			
一致部分所占百分比	58.9	Somers D	0.211
不一致部分所占百分比	37.9	Gamma	0.218
结值百分比	3.2	Tau-a	0.105
对	280	c	0.605

AUC decreased a lot after deleting variable **skipped**!!!

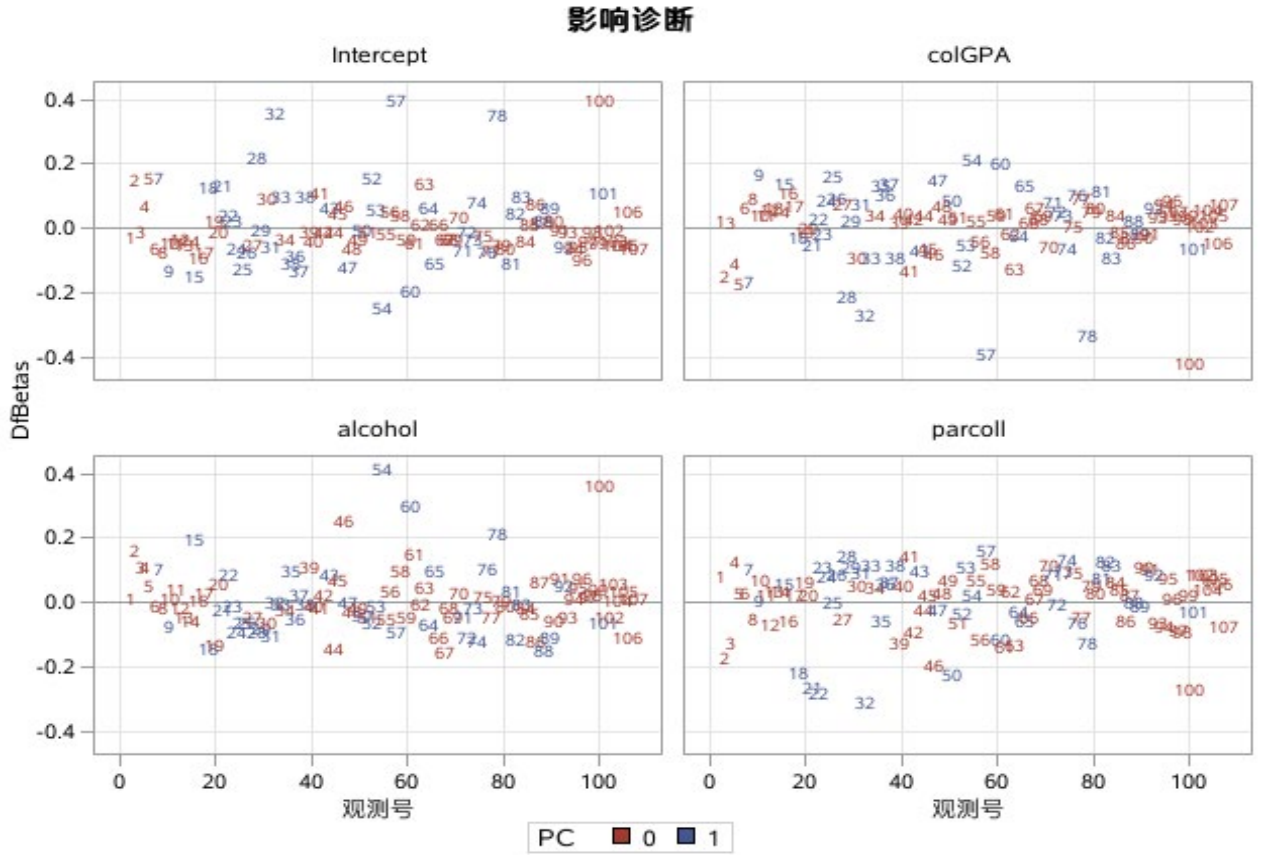
Talk later…

Logistic Model 2: Stepwise Selection

- After removing the high VIF variable
- SBC as criteria

最大似然估计分析						
参数		自由度	估计	标准误差	Wald 卡方	Pr > 卡方
Intercept		1	-5.3505	1.9309	7.6785	0.0056
alcohol		1	-0.2325	0.1539	2.2817	0.1309
colGPA		1	1.3795	0.6074	5.1582	0.0231
parcoll	1	1	1.1658	0.6366	3.3531	0.0671
parcoll	2	1	1.5699	0.6140	6.5365	0.0106

Logistic Model 2: Influential Observations



PC	colGPA	alcohol	parcoll
1	3.8	7	2
1	2.2	1	2
0	3.8	0	2

- 54, 57, 100

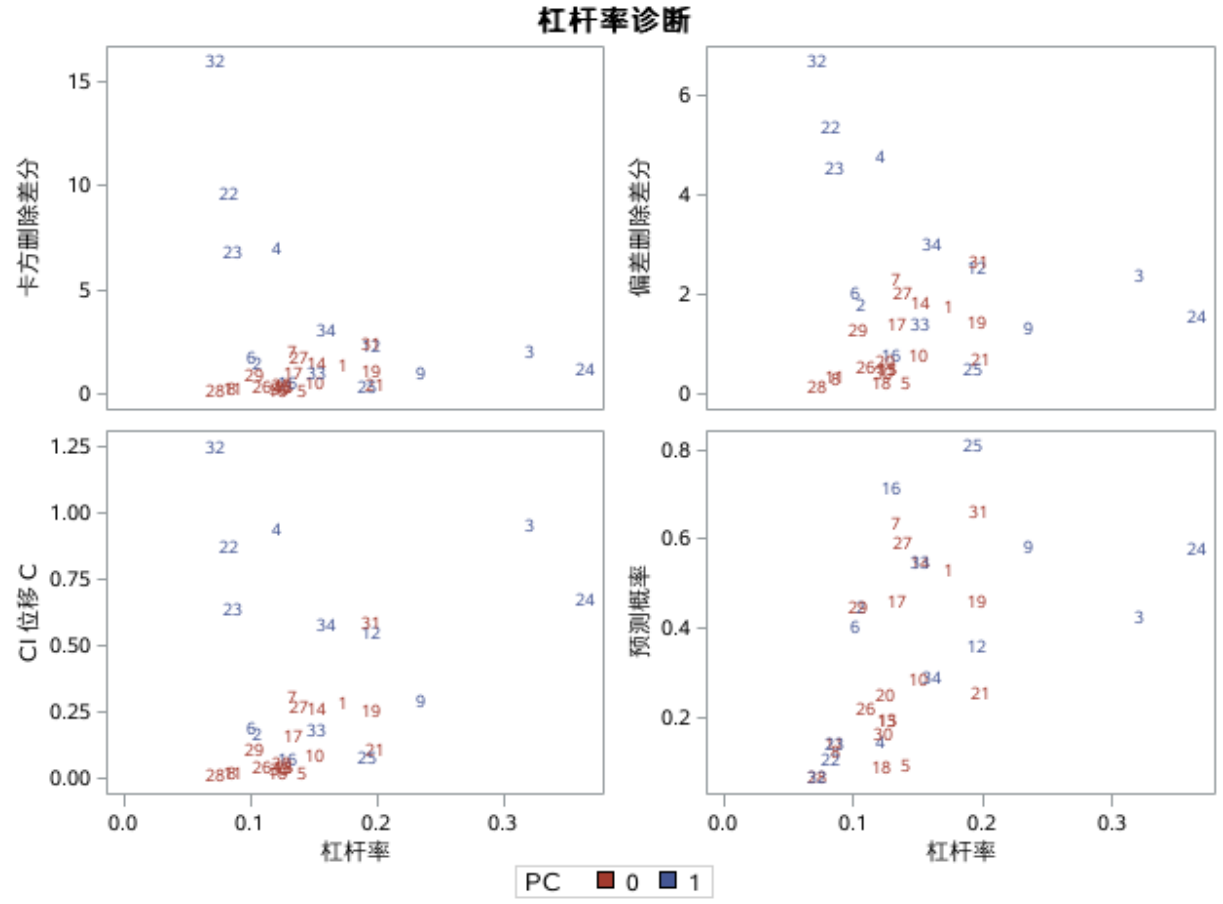
Apply Logistic Model 2 on test set

检验全局原假设: BETA=0			
检验	卡方	自由度	Pr > 卡方
似然比	-3.3318	4	.
评分	2.5445	4	0.6367
Wald	5.4625	4	0.2430

预测概率和观测响应的关联			
一致部分所占百分比	57.9	Somers D	0.171
不一致部分所占百分比	40.7	Gamma	0.174
结值百分比	1.4	Tau-a	0.086
对	280	c	0.586

Terrible!!

Logistic Model 2: Influential Observations in test set



Remove 3, 4, 22, 32

Apply Logistic Model 2 on test set again

检验全局原假设: BETA=0			
检验	卡方	自由度	Pr > 卡方
似然比	4.5156	4	0.3407
评分	6.2772	4	0.1794
Wald	4.6473	4	0.3254

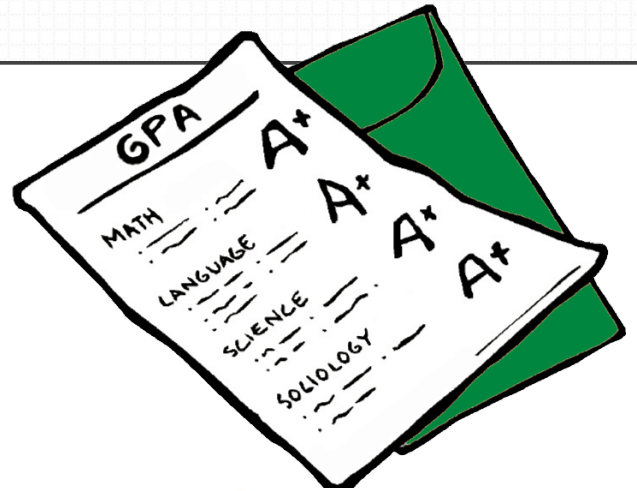
预测概率和观测响应的关联			
一致部分所占百分比	71.0	Somers D	0.435
不一致部分所占百分比	27.5	Gamma	0.442
结值百分比	1.5	Tau-a	0.200
对	200	c	0.718

Improve a lot!!!

4 Conclusion

Conclusion

- GPA has a lot to do with high school grades and whether or not you have a computer.



- Predicting whether a college student owns a computer through a model can be difficult. However, in general, the more educated the parents (i.e., college graduates) and the higher the student's GPA, the more likely the student is to own a computer.

Thank you for listening!