南方科技大学

SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

College students’ GPA analysis with SAS

Author: Minglei GUO, Runqi LIU, Ziyi ZHANG

Course: MA409 Statistical Data Analysis (SAS)

Professor: Cong XU

Date: June 12, 2020

Content

Abstract

GPA represents one of the most important measures of students' academic performance in school. In this paper, we use SAS analysis software to develop a linear model for predicting college students' GPAs. We find that GPA is closely related to high school GPA, the presence of a boyfriend or girlfriend, and the presence of a computer. Besides, because computers have become indispensable to college students, we also established a logistic model to determine whether students have a computer. The model is not very robust and can only determine the average case. For some more extreme student situations, the model is a poor predictor.

Keywords: GPA, PC, linear model, logistic model, SAS

1 Introduction

GPA stands for Grade Point Average. It is a standard way of measuring academic achievement in the U.S and many other countries. As a direct index reflecting the academic achievement and the competitiveness of students, GPA is very important for undergraduate students. GPA is also a decisive factor that influences students' careers and further study.

Being aware of the importance of GPA, our team wonder what factors might affect students' GPA in college. Among several factors that might have an impact on GPA, we want to determine which of them would have a deeper influence. For instance, Does the educational background of parents affect children's GPA? Does the high school GPA have lasting effects on college GPA? Does being in a relationship or going to clubs affect GPA, and which of them has greater impacts?

Also, since the first IBM personal computer (PC) was introduced in 1981, PC has entered every corner of people's daily life. Today, the PC plays a significant role in universities and undergraduate studies. Students can look up information, complete their work on the computer, and entertain themselves with computers. We wonder what type of students are more likely to have a PC. Therefore, we investigated the relationship between the possession of PC and other variables and would like to build a regression model predicting whether a student has a PC.

The data we used in this study are extracted from Introductory Econometrics: A Modern Approach (fifth edition) by Jeffrey M. Wooldridge. The variables in data GPA1.RAW includes 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.

2 Data processing and analysis

The raw data we used in this study consist of 141 observations and 29 variables. These variables provide information about student's gender, age, grade, residence, major, processions, modes of transport, academic performance, after-school life such as part-time job and clubs, family background, etc. 23 out of 29 variables are binary and the rest are numeric. The response variable we are interested in is the GPA in MSU, colGPA, and the possession of PC, PC.

2.1 Data import and missing value detection

As the first step, we imported the raw data into SAS. Then, we checked whether there exist missing values in raw data. The results suggest that no missing values are detected. Therefore, there are 141 observations in total and 56 of 141 have their PC.

2.2 Data processing

To simplify the data, we combined related binary variables into one variable with several levels. For instance, the raw data uses variable "soph", "junior", "senior" and "senior5" to specify the grade of observation. We combined them into a new variable "grade" with 3 levels "soph or junior", "senior" and "senior5". Note that there are no first-year students in data and we collapsed the group "soph" and "junior" into one group "soph or junior" because the sample size of group "soph" is too small (n=3).

Similarly, we combined binary variables “drive”, “bike” and “walk” into a new categorical variable “transport” with levels “drive”, “bike” and “walk”; we combined variables “job19” and “job20” into categorical variable “job” with levels “<=19h”, “>=20” and “no job”; we combined variables “business” and “engineer” into “major” with levels “business”, “engineer” and “other”; we combined variables “fathcoll” and “mothcoll” into new categorical variable “parcoll” with level 0 (if neither of parents is college graduated),1(if one of the parents is college graduated), and 2 (if both of parents are college graduated). Moreover, for the numeric variable “age”, we discretized it into a categorical variable with 5 levels “19”, “20”, “21”, “22” and “23+”.

3 Exploratory data analysis

3.1 Correlation and multicollinearity

Firstly, we computed the correlation matrix of raw data with SAS. No high correlation is revealed in the matrix. Also, from the matrix, we found some variables that 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) (The absolute values of correlation coefficients between these variables and colGPA are greater than 0.1).

Secondly, we computed the correlation matrix of simplified data with SAS and looked for the variables with a 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 (The absolute values of correlation coefficients between these variables and PC are greater than 0.1).

As no collinearity is detected in correlation matrix, we then check the multicollinearity of data with the simplified data. Use tolerance less than or equal 0.1 as the criterion of multicollinearity, and the results suggest that multicollinearity does not exist.

3.2 Exploratory data analysis through plots

3.2.1 colGPA

Figure 1 shows the distribution of colGPA. Despite the odd in colGPA=3.1, the GPA in college approximately follows Normal distribution. Graphical methods were used to illustrate the relationships between the colGPA and other variables. The Figure 2 (a)-(k) are boxplots of several variables. On average, the students with no cars, students going to clubs, students graduating from Michigan high school, female students, students with PC, students that have a boyfriend or girlfriend tend to have a higher GPA. Also, on average, the sophomore or junior students have the highest GPA, while the senior students have the lowest GPA. The finding is consistent with the average GPA in different age groups: The groups 19-year-old and 20-year-old have the highest GPA, while the groups 21-year-old and above-23-year-old have the lowest GPA. The GPA of students spending different time on jobs does not differ much. On the other hand, the GPA of students in different majors varies much. Note that there are outliers in engineer and other groups, so we compare the median of three groups: the engineer students have the lowest GPA, the median is 2.7, and the business students have the highest GPA, the median is 3.0. Besides, the students with both parents college graduated have a higher college GPA than others on average, while the difference of GPA between students with no college graduated parents or with only one college graduated parents is slight.

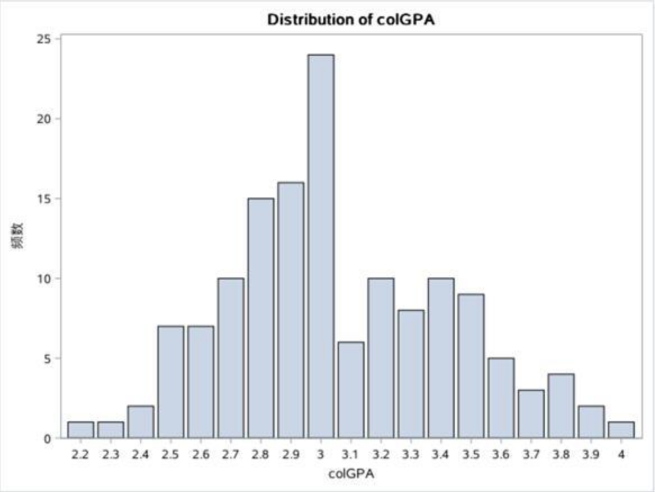
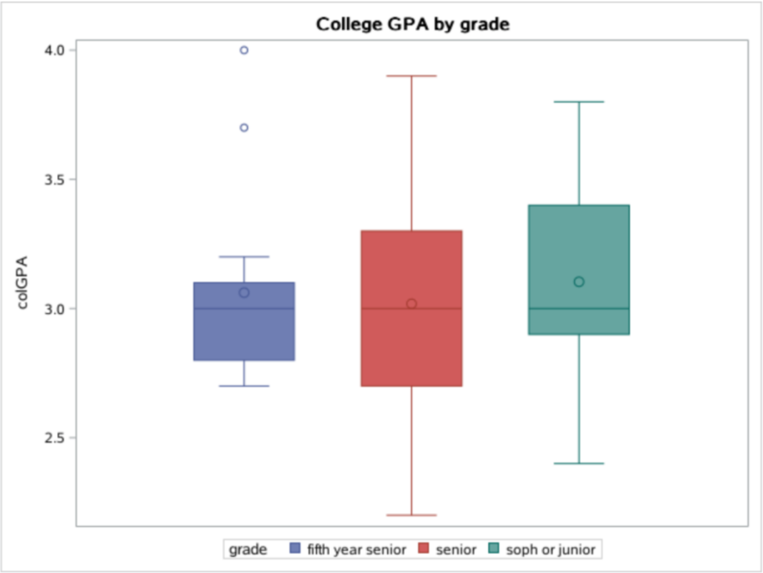
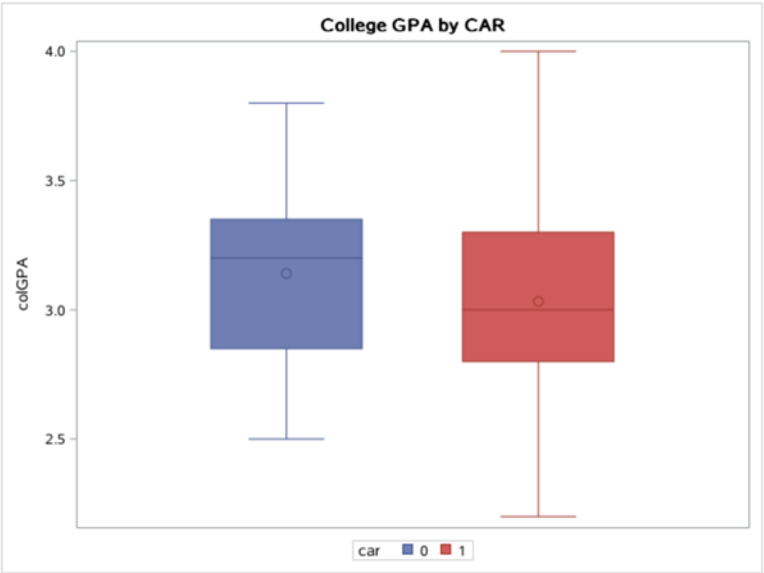


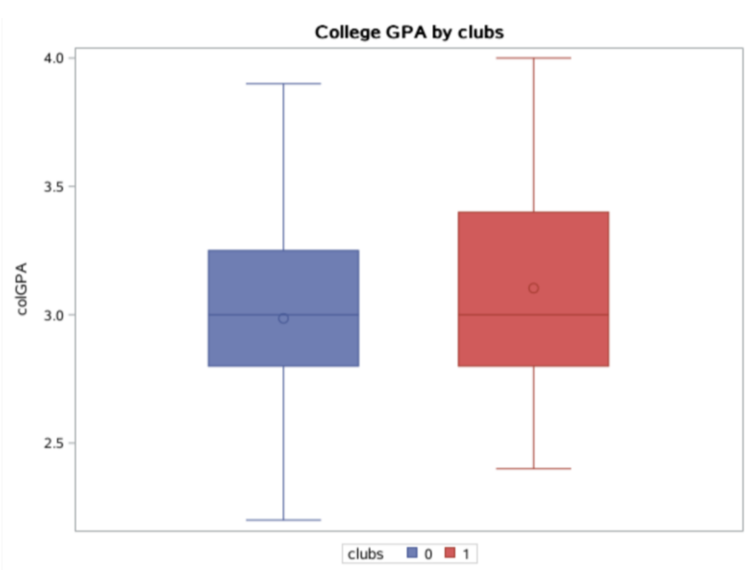
Figure 1. distribution of GPA in college



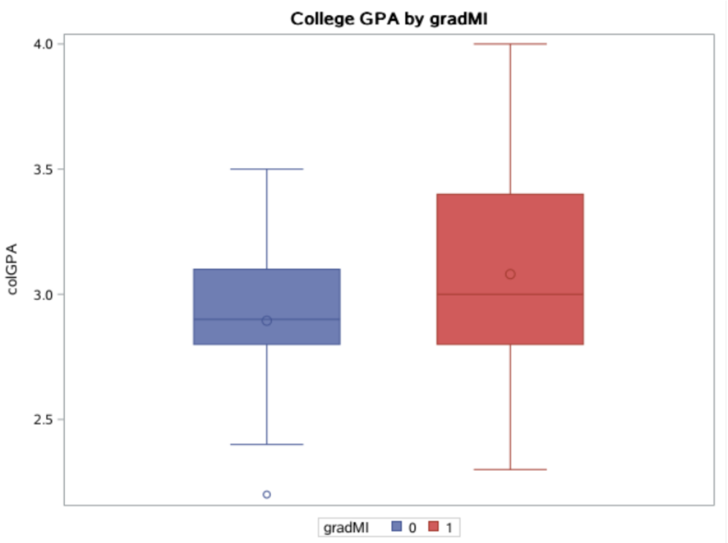
(a) College GPA by grade



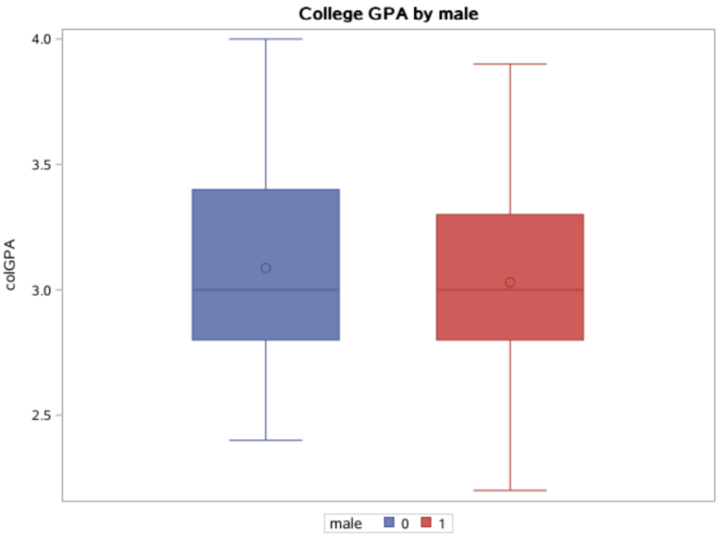
(b) College GPA by possession of car



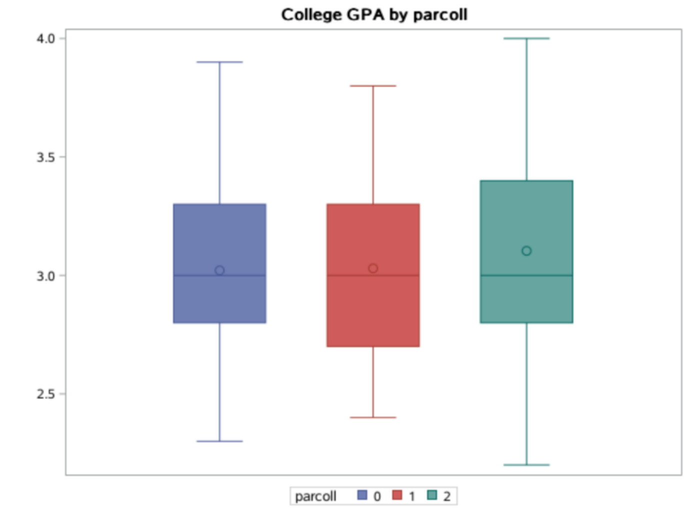
(c) College GPA by clubs



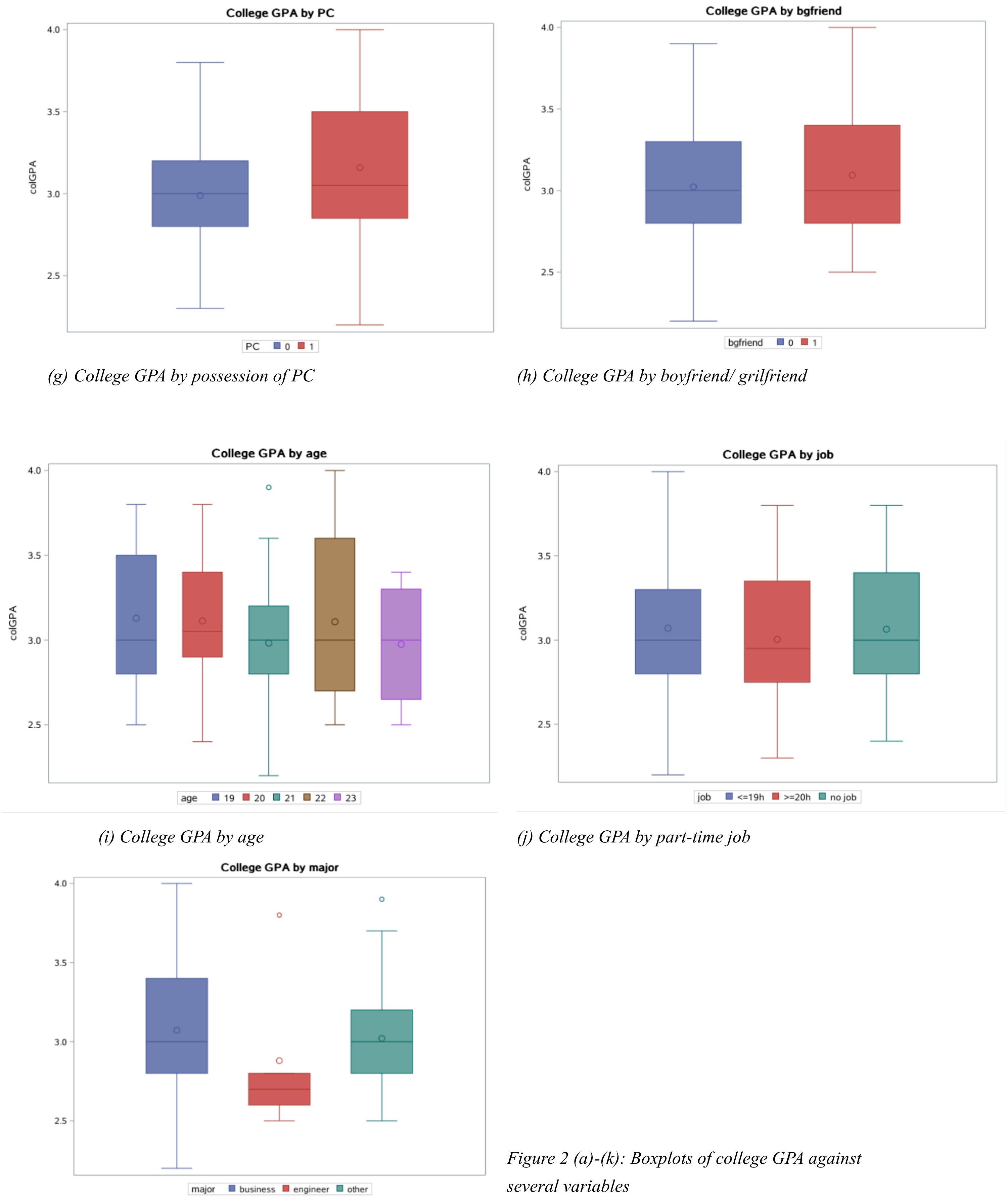
(d) College GPA by if graduated from Michigan high school



(e) College GPA by gender



(f) College GPA by parents' education background



(g) College GPA by possession of PC

(h) College GPA by boyfriend/ grilfriend

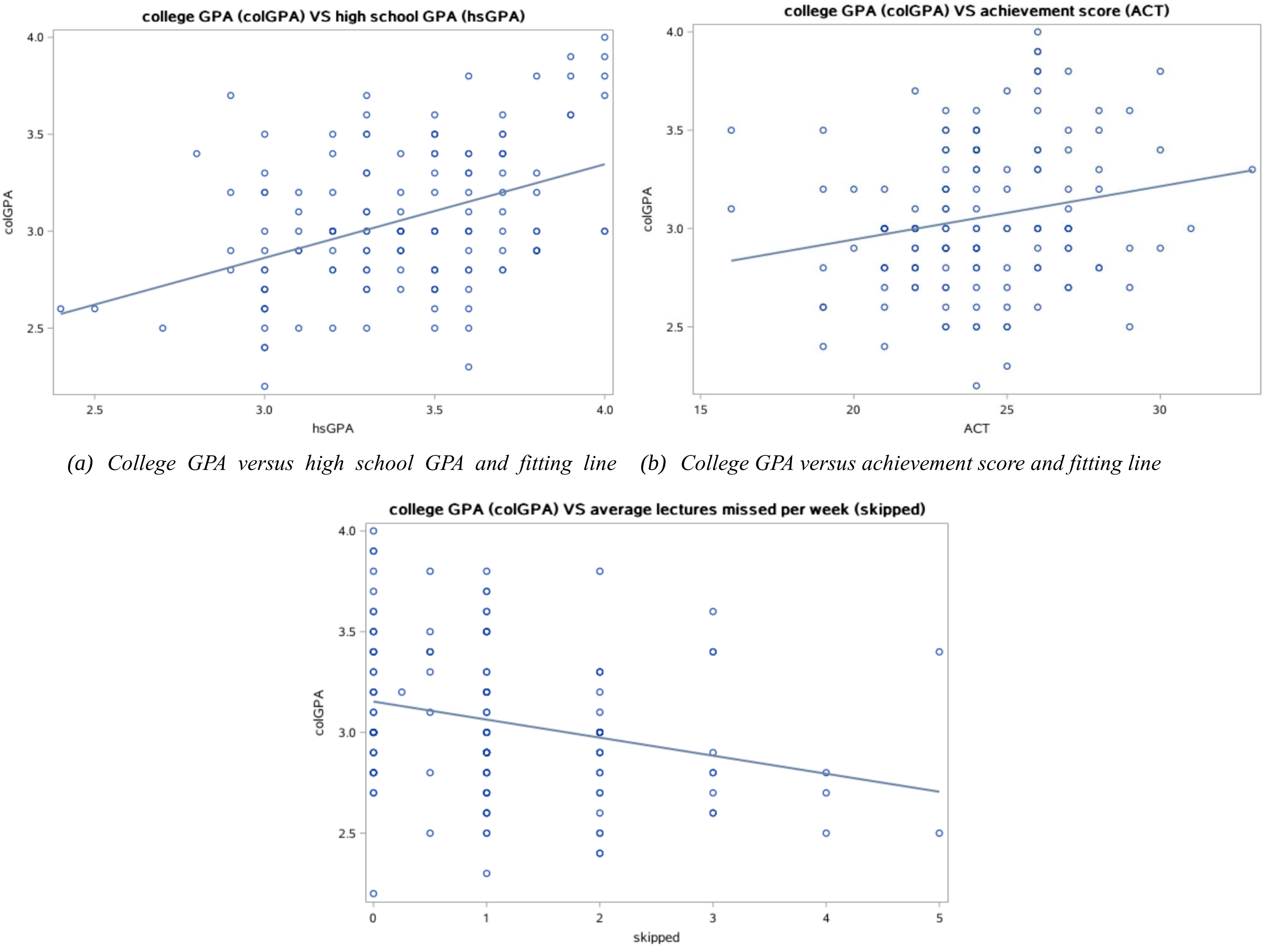
(i) College GPA by age

(j) College GPA by part-time job

(k) College GPA by age

Figure 2 (a)-(k): Boxplots of college GPA against several variables

Furthermore, we illustrated the relationship between the colGPA and some numeric variables using scatter plots with fitting lines. The Figures 3(a)-(c) corresponds to the variables hsGPA, ACT, and skipped. From these figures, we can see that the hsGPA and ACT positively related to colGPA. That is, the students with a higher GPA in high school and higher achievement scores tend to have a higher GPA in college. In contrast, the students missing more lectures tend to have a lower college GPA.



(a) College GPA versus high school GPA and fitting line *(b) College GPA versus achievement score and fitting line*

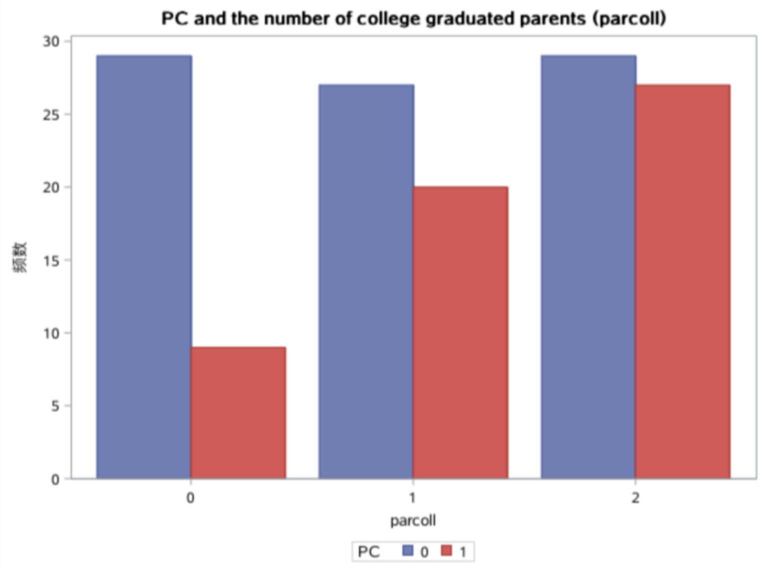
(c) College GPA versus number of lectures missed and fitting line

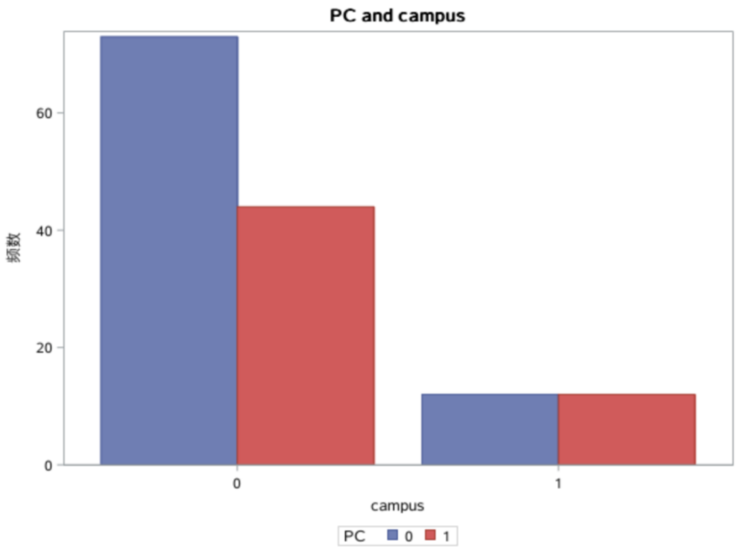
Figure 3 (a)-(c): Scatter plot of college GPA versus variables and fitting lines

3.2.1 PC

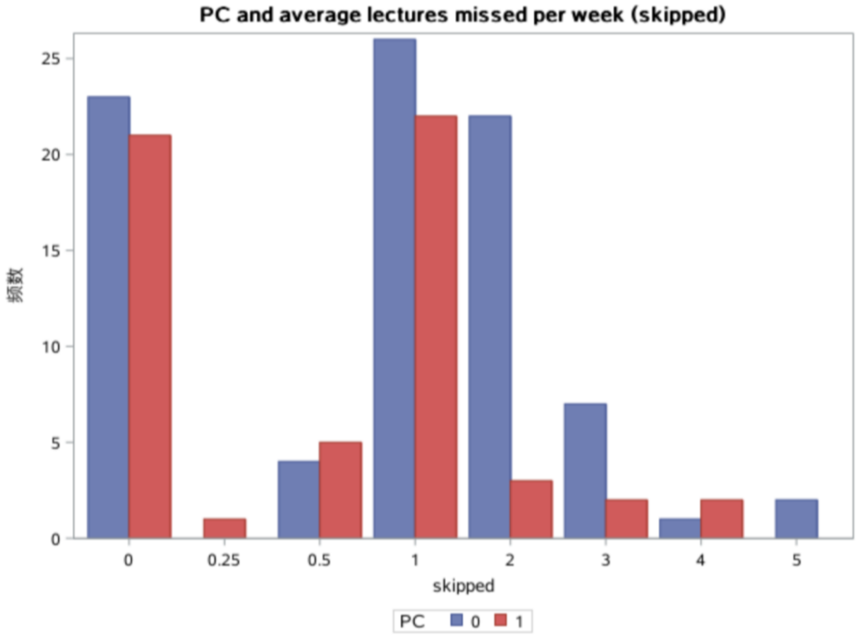
Graphical methods were used to illustrate the relationships between the PC and other variables. In general, there are more students without their PC (n=85) than students with their PC (n=56). We have found that students with PC are more likely to get a high college GPA in section 3.2.1. The figures 3(a)-(c) illustrate the distribution of variables parcoll, campus, and skipped in two groups of students, the students with PC and the students without PC. The students with college graduated parents are more likely to have their PC than the students from the non-college-graduated family. Also, the students with both parents college-graduated are more likely to own a PC than those students having only one college-graduated parent.

The proportion of PC possession in the group of students live on campus is larger than that in the group of students lie outsides campus. Moreover, the students who miss equal or less than one lecture and who miss four lectures on average per week are more likely to have their PC than the others.





(a) PC and education background of parents *(b) PC and if live on campus*



(c) PC and the number of lectures missed

Figure 3(a)-(c): Bar charts of variables parcoll, campus, and skipped by possession of PC

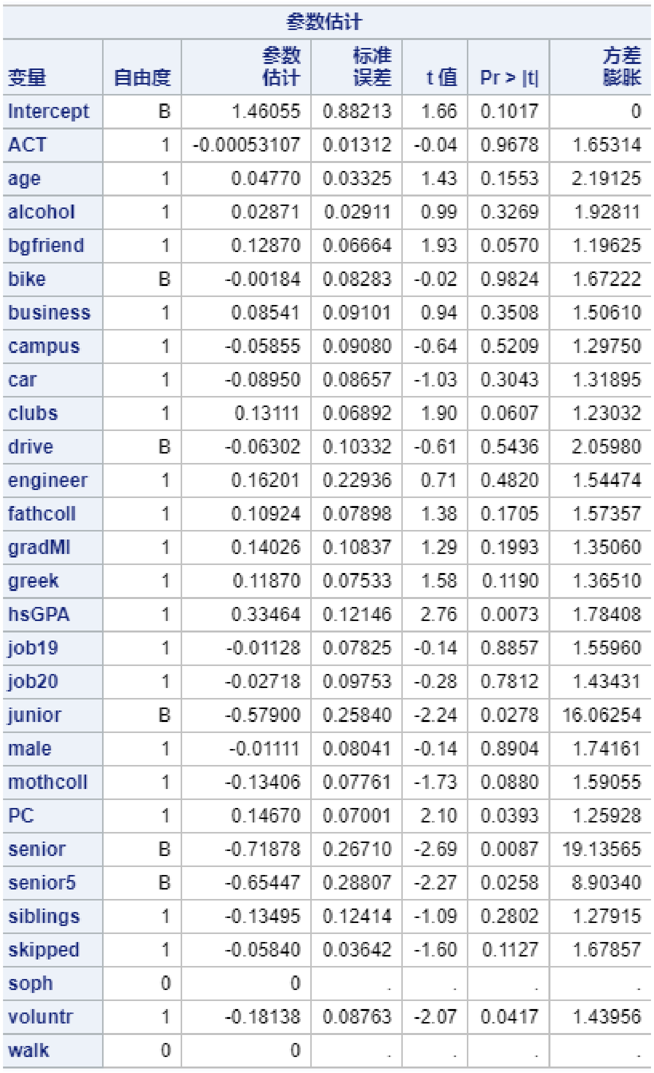
4 Modeling

4.1 Linear Model

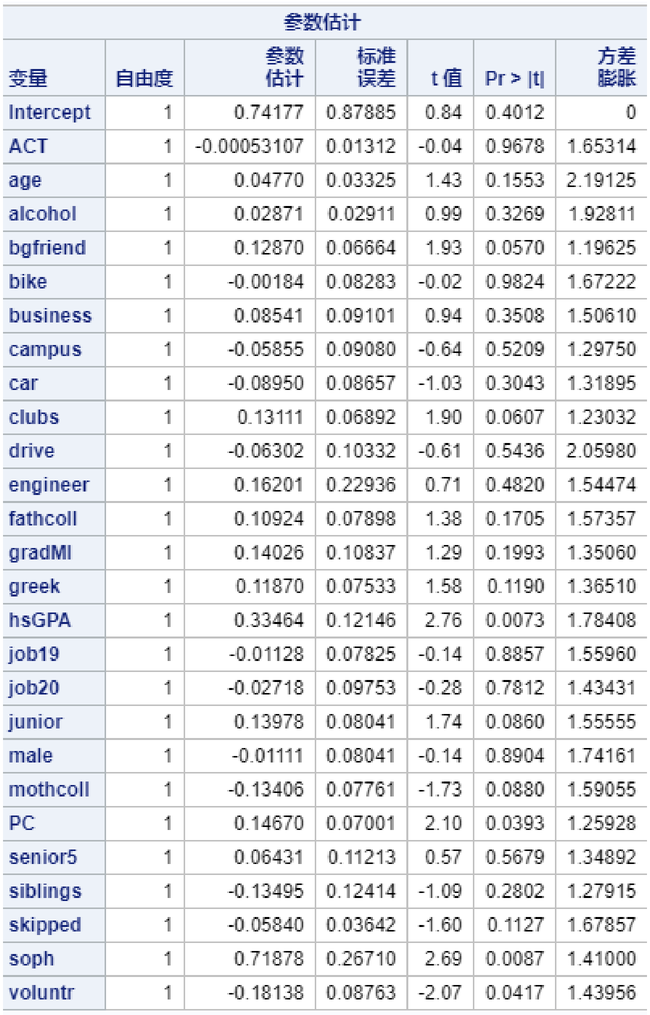
4.1.1 linear model-1 for colGPA

The next step is to build a model to predict GPA in college. Before doing so, we first divide the dataset into a train set and test set. We randomly select 70% of the data to form the train set and the rest as the test set. We then start with the simplest linear model which we denote as linear model-1.

We first performed linear regression with all variables and used the VIF index to detect collinearity. From the Table 1, we can easily find that the senior and junior variables have high VIF values, as well as the walk variable. We delete them. After we do this, the results get better. There is no collinearity now.



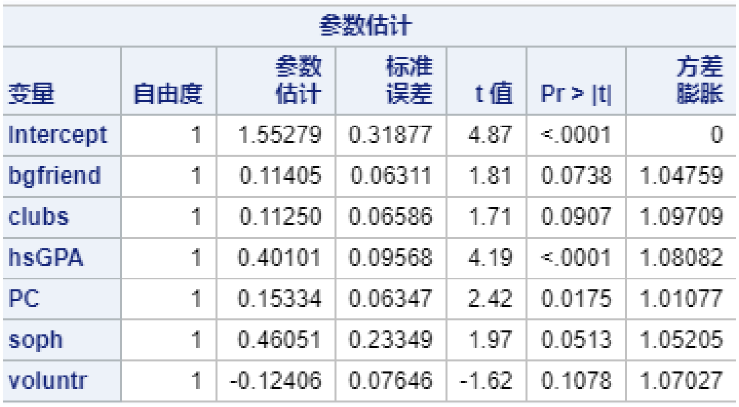
参数估计						
变量	自由度	参数估计	标准误差	t 值	Pr > |t|	方差膨胀
Intercept	B	1.46055	0.88213	1.66	0.1017	0
ACT	1	-0.00053107	0.01312	-0.04	0.9678	1.65314
age	1	0.04770	0.03325	1.43	0.1553	2.19125
alcohol	1	0.02871	0.02911	0.99	0.3269	1.92811
bgfriend	1	0.12870	0.06664	1.93	0.0570	1.19625
bike	B	-0.00184	0.08283	-0.02	0.9824	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	.	.	.	.



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

Table 1 Estimate Results before and after deleting high VIF variables

Then we try to remove all the insignificant variables in which p-values are larger comparatively. We performed a total of four operations to remove insignificant variables. The first time the p-value threshold was set to 0.15, and the next three times it was set to 0.10. Finally, our model only retains 6 variables which are bgfriend, clubs, hsGPA, PC, soph, voluntr.



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

Table 2 Final estimate results after removing coefficients

As we can see in Table 2, most coefficients' p-values are smaller than 0.10. Besides, the number of influential observations is small (only 3). Let’s look at these three students. For observation 28 and 60, firstly they both have a large difference in colGPA and hsGPA, besides, the four explanatory variables are different from most students (means: too many 1), these information shows that the two students may pay much attention to other things that are not directly linked to study, and they have a great change in GPA between college and high school time. While for observation 57, it is another story, this student does not have a boyfriend/girlfriend or take in any clubs, but his/her grade still decreases a lot compared to high school time, he/she may meet some study problem. In summary, it seems like our linear model does not predict well for these students, we decide to remove them all.

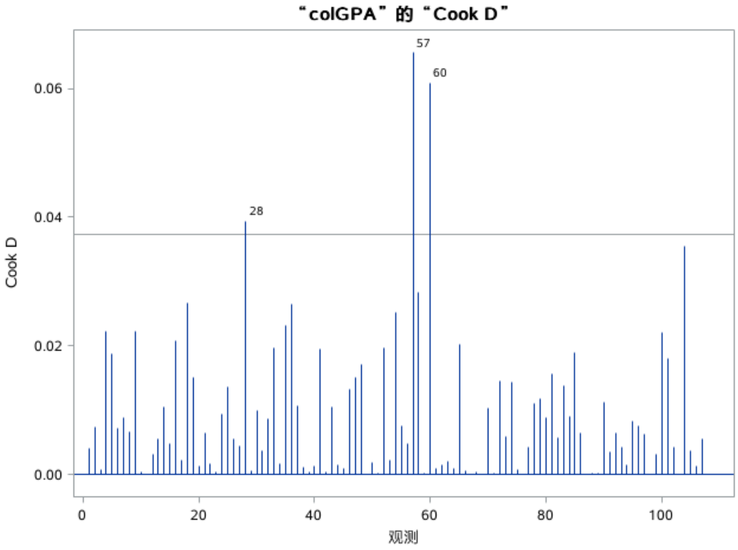


Figure 4 cook distance of model-1 on train set

After we remove these three observations, the cook's distance plot is much better.

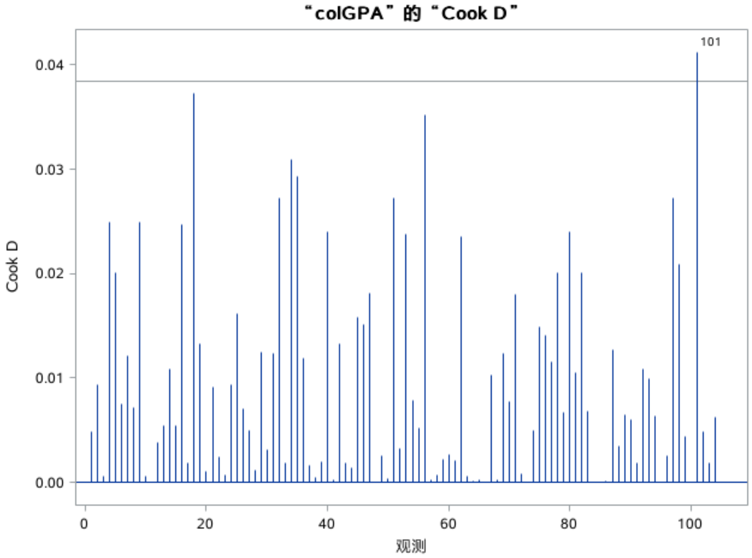


Figure 5 cook distance of model-1 on train set after removing influential points

Finally, we perform model diagnostics. First is the residual-fitted value plot, the mean is close to 0, thus satisfies the assumption of linearity. Also, homoscedasticity seems not any problem. Besides, we test whether the residual normality assumption is satisfied. Since the *Shapiro-Wilk* test shows that *p*-value equals 0.0661 which is bigger than 0.05, the residuals follow the normality assumption.

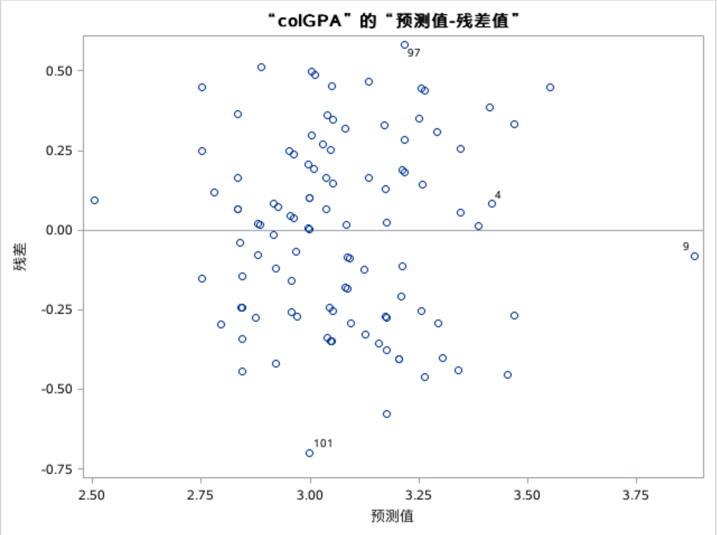


Figure 6 residual-fitted value plot on train set

Then we apply the model in the test data set and get the residual-fitted values plot. As we can see here, the mean of residual is about zero which means model-1 is good.

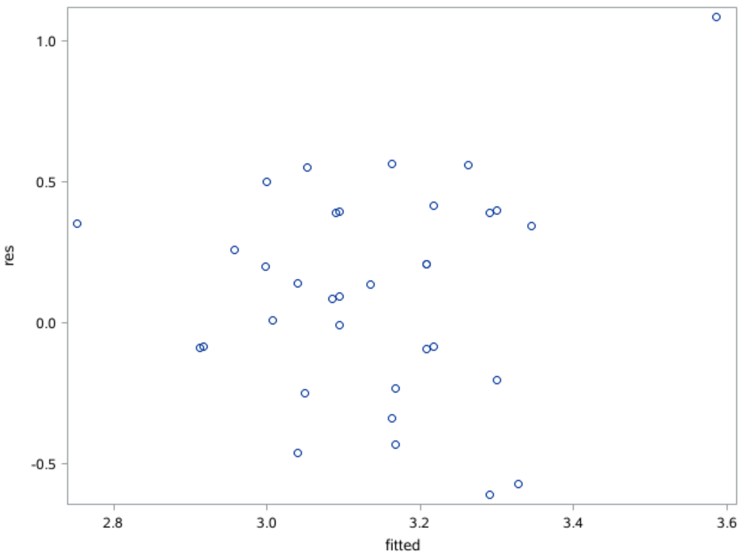
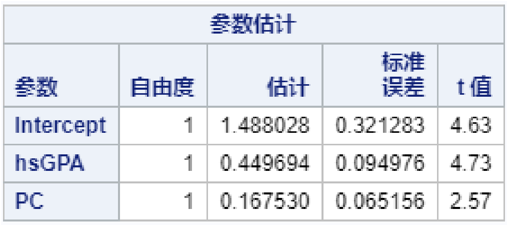


Figure 7 residual-fitted value plot on test set

4.1.2 linear model-2 for colGPA

In model two, we use the stepwise method with SBC to select variables. The remaining variables are only hsGPA and PC.



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

Table 3 estimate results of stepwise selection method

Still, then we do the model diagnosis. For influential observations, most of them have the common feature that the difference between colGPA and hsGPA is large. It seems like that our linear model does not predict well for these students, thus, we decide to remove observations 18, 57, 58, 60, 104.

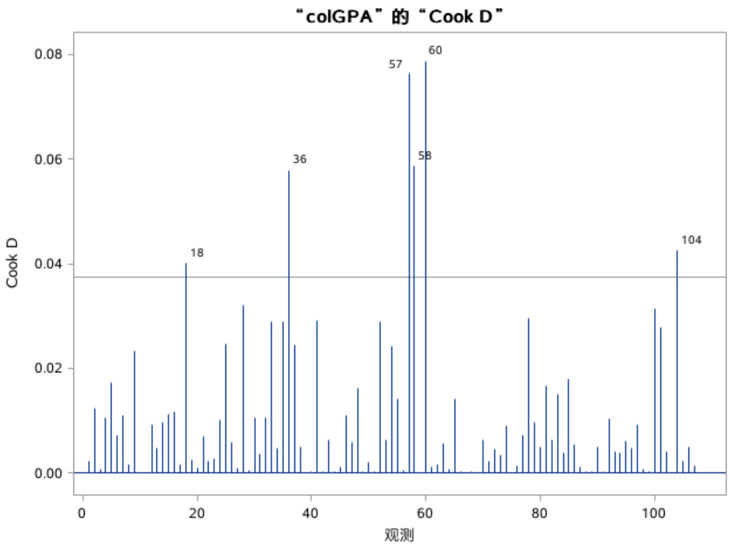


Figure 8 cook distance of model-2 on train set

Similarly, we test whether the residual normality assumption is satisfied. The residual plot reflects the linearity and homoscedasticity just like in model-1. And since the *Shapiro-Wilk* test shows that *p*-value equals 0.2262 which is bigger than 0.05, so the residuals follow the normal assumption.

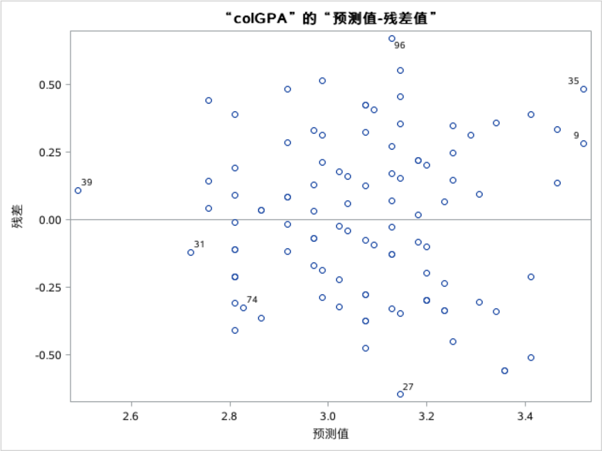


Figure 9 residual-fitted value plot on train set

Then we get the residual-fitted values plot in test data. Still, the mean of residual is about zero which means good model.

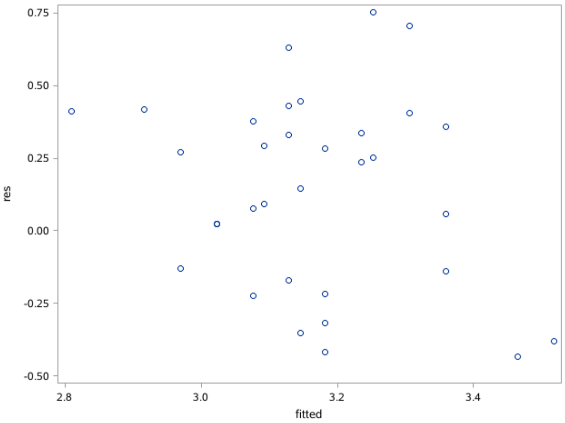
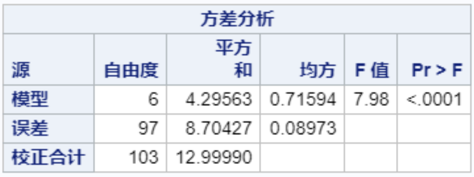


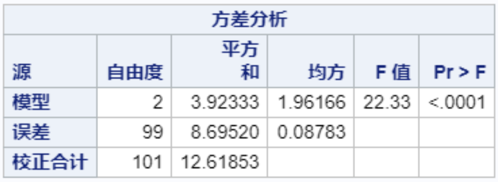
Figure 10 residual-fitted value plot on test set

4.1.3 A Notation

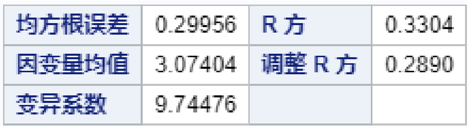
For the two linear models we obtained above, we found an interesting problem. Although both two models are significant (p-value of F test is small), the R^2 values of the models are small. As shown in the figure below.



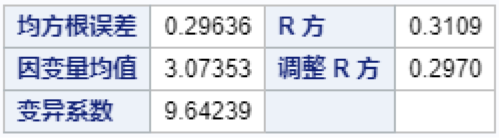
方差分析					
源	自由度	平方和	均方	F 值	Pr > F
模型	6	4.29563	0.71594	7.98	<.0001
误差	97	8.70427	0.08973		
校正合计	103	12.99990			



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	0.29956	R 方	0.3304
因变量均值	3.07404	调整 R 方	0.2890
变异系数	9.74476		



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		

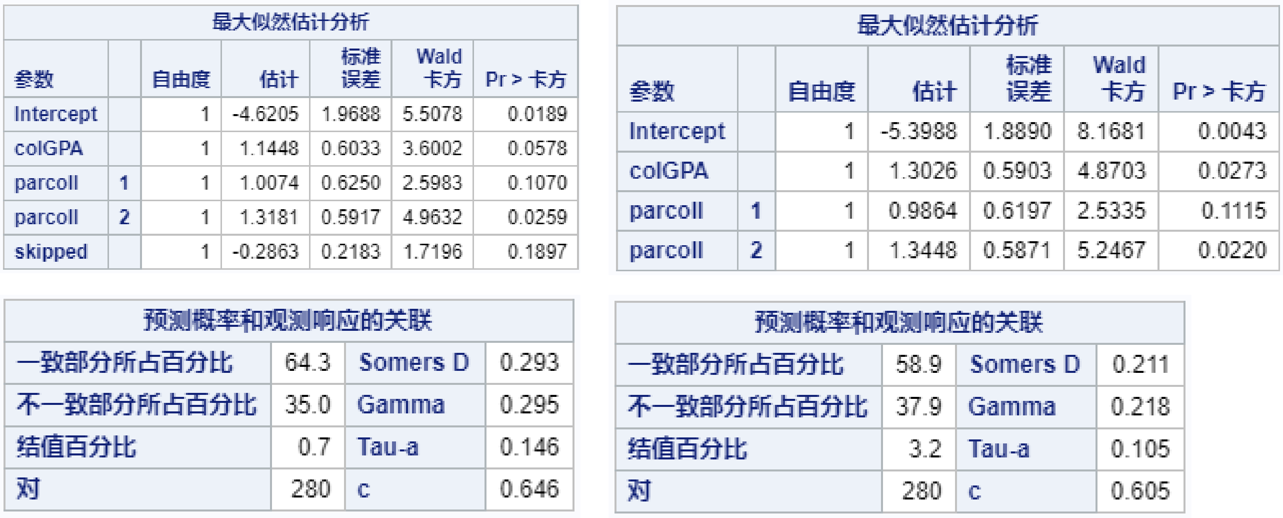
Since regression coefficients and fitted values represent means, R-squared and prediction intervals represent variability. We can interpret the coefficients for significant variables the same way regardless of the R-squared value. However, we should remember that low R-squared values are a warning of imprecise predictions. (Frost, 2018)

4.2 Logistic Model

Since the computer is now almost an essential learning device for college students, we decide to explore the relationship between the variable PC and other variables. We build a logistic model to predict whether a student has a PC. Similar to the linear model, we randomly select 70% of the data to form the train set and the rest as the test set.

4.2.1 Logistic model-1 for PC

Still, we try to remove all the insignificant variables in which p-values are large relatively. We performed a total of four operations to remove insignificant variables. However, we're hesitant about the last operation of the deleting variable **skipped**. For this step, the p-value of the variable skipped equals 0.1897 which is kind of big. The hypothesis test for BETA=0 is not to be rejected. But after we compared the performance of the model on the test set before and after removing this variable, we found that AUC decreased a lot which is from 0.646 to 0.605. We thought it is because there are too many influential observations in the test set, the specific proving process is in the following part.



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

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

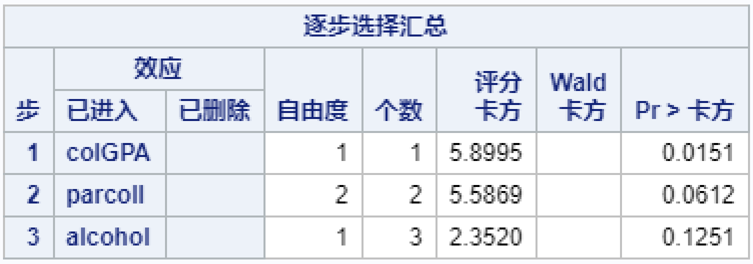
预测概率和观测响应的关联

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

Table 5 comparison before and after deleting variable skipped

4.2.2 Logistic model-2 for PC

In model two, we use the stepwise method with SBC to select variables. The remaining variables are only colGPA, parcoll, and alcohol.



逐步选择汇总

	效应						
步	已进入	已删除	自由度	个数	评分卡方	Wald卡方	Pr > 卡方
1	colGPA		1	1	5.8995		0.0151
2	parcoll		2	2	5.5869		0.0612
3	alcohol		1	3	2.3520		0.1251

Table 6 estimate results of stepwise selection

Then we detect influential observations. And we removed observations 54, 57, 100.

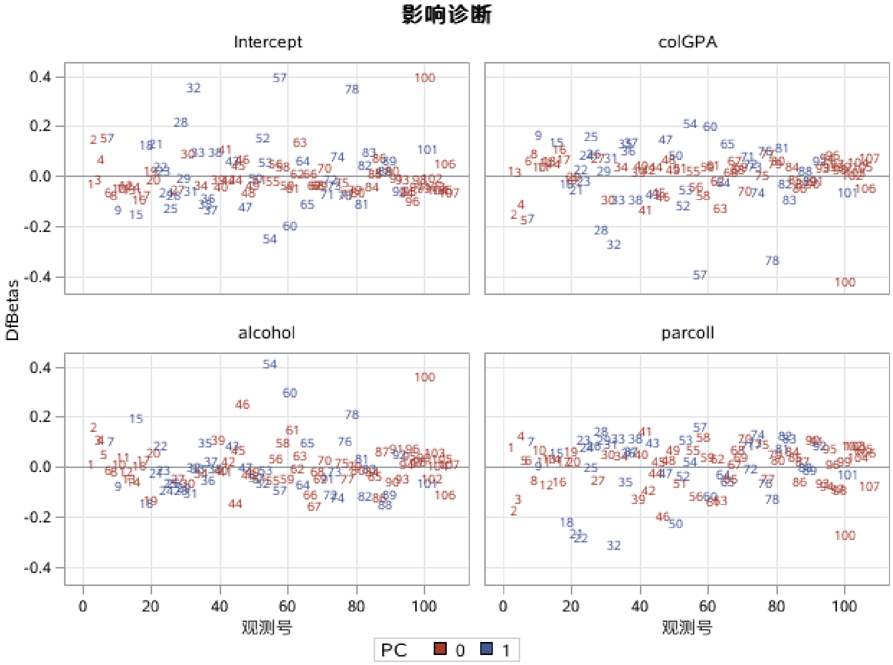
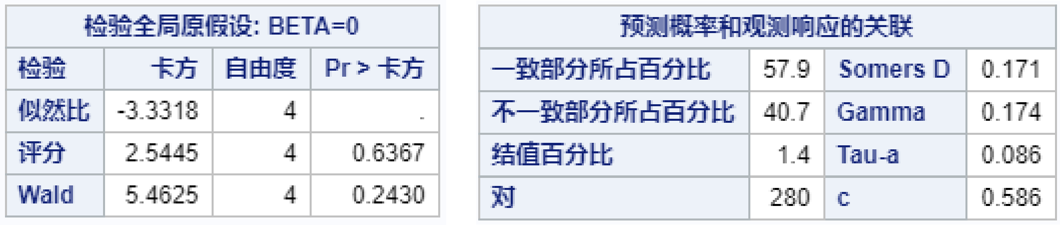


Figure 11 detect influential observations

Next, we apply model-2 on the test set. As we can see, the results (both hypothesis test for BETA=0 and AUC) in the test set are terrible.

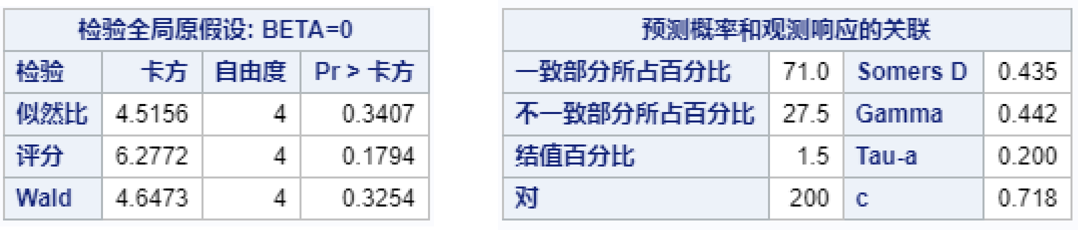


检验全局原假设: BETA=0			
检验	卡方	自由度	Pr > 卡方
似然比	-3.3318	4	.
评分	2.5445	4	0.6367
Wald	5.4625	4	0.2430

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

Table 7 model performance before removing influential observations

We try to detect influential observations in the test set and remove them. After removing 4 most influential observations which are 3, 4, 22, 32, our results (both hypothesis test for BETA=0 and AUC) in the test set become much nicer. This shows that PC is a complex variable that is affected by many reasons, and our logistic model can deal with the average cases, while does not predict well for an extreme case. We believe it is also the reason that model-1 cannot get good results in the test set. In summary, by the logistic model, we can only predict the probability of PC at a general level.



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

Table 8 model performance after removing influential observations

5 Conclusion

GPA has a lot to do with high school grades and whether or not you have a computer. That's why it's important to build a strong academic foundation in high school. Also, it is best to advise your parents to acquire a computer for you. This is a necessity for college studies.

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. The possible reason is that college-educated parents are more educated and enlightened. Also, a student with a higher GPA tends to be more self-disciplined. Parents will feel comfortable buying computers for them.

6 Reference

1. Frost, J. (2018). How to Interpret Regression Models that have Significant Variables but a Low R-squared - Statistics By Jim. Retrieved 10 June 2020, from https://statisticsbyjim.com/regression/low-r-squared-regression/
2. xu, w. (2012). 影响大学绩点的因素. Retrieved 10 June 2020, from http://f.dataguru.cn/thread-2043-1-1.html?tdsourcetag=s_pctim_aiomsg