The Insight of Job Offer for Chinese International Students with a Major in Statistics

(Team member: Danjing Lin, Guimin Dong, Jian Sun, Xiaoqing Wu)

Abstract

This project focuses on investigating the contributing factors to the full-time job and internship offers of the students in the Department of Statistics at the George Washington University. To analyze the parameters concerning the results of job hunting, based on our data set, this paper performed linear regression, logistic regression, cross validation, principal component analysis, and linear discriminant analysis to detect the constructive foundation of getting a job.

Background

There are currently more than three hundred thousand students from China in colleges in the United States - a third of all international students in the country - which is almost a fivefold increase since 2000. Among these Chinese international students, the percentage of Statistics majors has increased rapidly in recent years. As graduate students who major in Statistics in GWU, we are quite concerned about our job-hunting situation. What aspects affect our job-hunting results the most?

There is no such related dataset available on the Internet, so we decided to collect the dataset on our own. Our sample study was the GWU Statistics major students. The dataset was obtained by processing the questionnaires we distributed to the graduate statistics students at GWU, and 66 samples were collected from 75 questionnaires distributed. After cleaning the data set, 53 samples were analyzed. 17 parameters in our data collection includes height, weight, age, gender TOEFL score, time spent in English speaking country, driver's license, programming skill, courses taken outside the statistics department, undergraduate GPA, undergraduate major, undergraduate school, graduate GPA, working experience, and the number of interview and job offers received. The selection of these parameters was based on our presumptions and expectations about the factors that would influence the results of job hunting. However, because of the design of the questionnaire, some students who received job offers may have been more willing to share the information, and others who had not received offers may not have been as willing to provide such information. As a result, the data set may be biased. The main research interest is to discover which and how the parameters influence job placement after graduation of statistics graduate students in our department.

Data Analysis and Results

Linear Regression

Firstly we construct the full linear model by setting the number of job offer as response variable and other parameters as explanatory variables:

$$O = H + W + A + T.ESC + fg + fdl + T + P + fc + U.gpa + fuschool + fm + G.gpa + X + I$$

Here H is height, W is weight, A is age, T.ESC is time in English speaking country, fg is gender as factor, fdl is driver license as factor, T is toelf score, P is programming skill measured by the

number of programming language, courses taken outside statistics department, U.gpa and G.gpa is undergraduate and graduate gpa, fuschool is undergraduate school as factor, fm is undergraduate major as factor, X is experience, I and O is number of interview and job offer separately.

By using "lm" function in R, we can detect that the p-values of T.ESC, fmE(undergraduate major in economics), X and I are less than 0.05, indicating that the coefficients of these covariates are significantly different from 0. What's more the p-value of A is 0.08 which is greater than 0.05 a little bit, and can still be considered in our remodelization. To verify the assumption of linear regression, the diagnosis plot is shown below, where an obvious outlier can be detected. After deleting the outlier and remodelization, the summary result from R indicating that the p-values of T.ESC, X, and I is less than 0.05. Then we construct our reduced linear model by considering these three covariates. Furthermore, by checking the diagnosis plot for the full and reduced model, after the deleting the outlier, the assumption of linear regression can be supported to approve that our reduced model is reasonable.

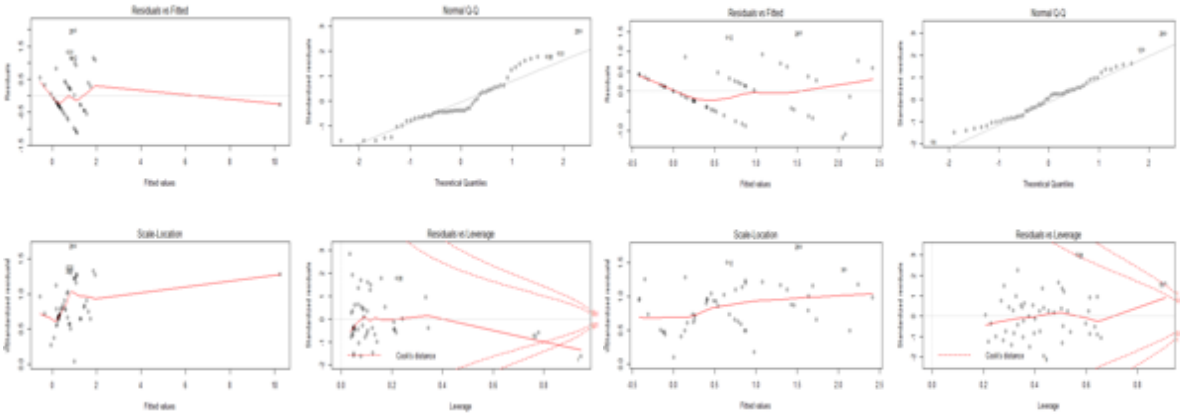

Figure 1. Diagnosis plot for full model & Diagnosis plot for reduced model

We use the ANOVA analysis to investigate whether is simplified linear model is enough to explain our response variable. The p-value of F test is 0.6463 implying that the simpler linear is not significantly different from the full model. Thus the reduced model will serve purpose very well.

Linear Regression Extension and Logistic Regression

Furthermore, we used the step-wise selection to investigate the linear relationship among the variables. By using the forward and backward selection, the step-selected linear model is $O = W + G.gpa + X + I + T.ESC$. Similarly, by performing ANOVA analysis, this model is significantly different from the reduced model we discover in the last section, with p-value 0.027. Then with the inference of cross-validation errors of these two model, the errors of the first reduced model and second reduced model are 0.523 and 0.464 respectively. Thus the second reduced model can perform better than the first reduced model with smaller error, and with 65.54% explanation of the whole data set. Thus our simplified linear model is

$$O = -4.383 + 0.019W + 0.924G.gpa + 0.036X + 0.300I - 0.067T.ESC$$

Previously, we deemed the number of offers as response. Now, we just regard the observation whether get offers or not, and deemed the response as binomial distribution. So we used the logistic regression method to get the equation:

$$\text{Offer} = -15.85 + 0.141\text{T} + 0.114\text{X} + 0.763\text{I}$$

And this logistic linear model implies that TOEFL score, experience, and interview will provide positive contribution to get a job offer.

Principal component analysis

We do linear analysis to get multiple linear relationship among the predict variables and response variable. And we do not want to stop here, since we guess there will be more information behind the dataset, so we do the principal component analysis to do more exploration.

After linear regression analysis, we leave 5 variables: Age; The time of staying in English speaking country; work experience; the number of interview; the number of offer.

Then by using the first four variables, Age, The time of staying in English speaking country, Work experience and The number of interview, we get a variance-covariance matrix S:

	FP.G.gpa	FP.T.ESC	FP.X	FP.I
FP.G.gpa	0.0351	-0.000844	0.1968	0.003897
FP.T.ESC	-0.000844	5.1693	2.3008	-0.68197
FP.X	0.19683	2.3008	26.387	3.26597
FP.I	0.003897	-0.68197	3.266	47.09144

Table 1. The variance-covariance matrix S

By the help of R, we draw the eigenvalues and eigenvectors of the S easily.

Eigenvalues	λ_1	λ_2	λ_3	λ_4
	47.5969	26.1546	4.8975	0.03355
Eigenvectors	e_1	e_2	e_3	e_4
1	-0.0007078	0.00738	0.00461	0.99996
2	0.007679	0.11256	-0.9936	0.00375
3	-0.1514	0.9823	0.11008	-0.00786
4	-0.98845	-0.14957	-0.02458	0.00052

Table 2. Eigenvalues and eigenvectors of variance-covariance matrix S

The next, we set the Cumulative Proportion as 99%, so we decided to choose the first three principal component from the summary(prcomp(PCA)).

Importance of components:	PC1	PC2	PC3	PC4
Standard deviation	6.8991	5.1142	2.21303	0.18316

Proportion of Variance	0.6049	0.3324	0.06224	0.00043
Cumulative Proportion	0.6049	0.9373	0.99957	1

Table 3. Information of principle components

According to Table 3, 60.49% proportion of variance is attributed to λ_1, 33.24% proportion of variance is attributed to λ_2, 6.22% proportion of variance is attributed to λ_3.

To check our result, we made a scree plot, the following is the picture. And we noticed that there is a clear bent in PC3, so we tested that our former result is right, we kept the first three principal components.

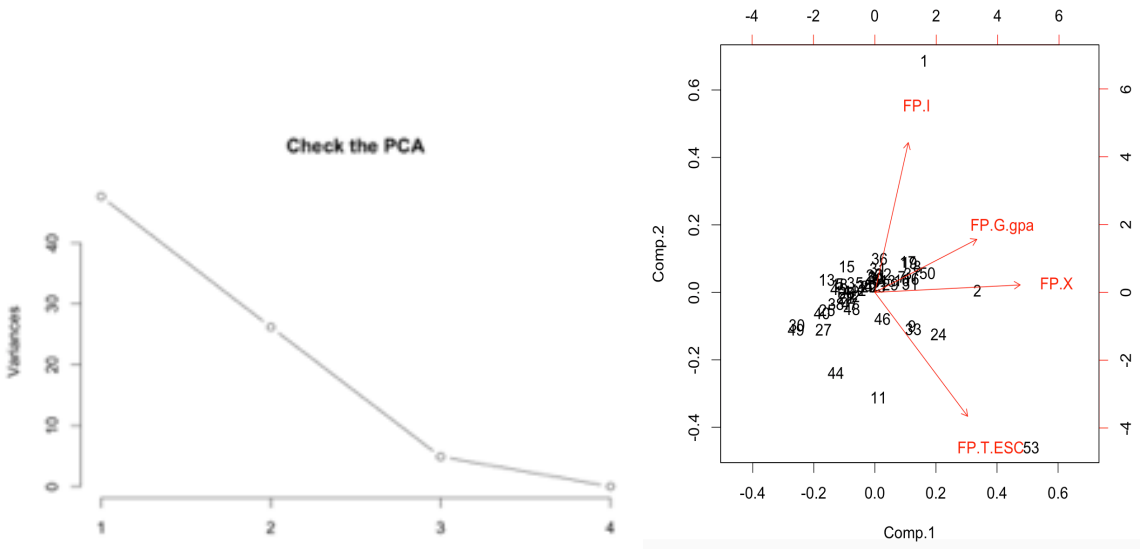

Figure 2. Scree plot & biplot

The left picture of Figure 2, is a scree plot. We use it to assure the number of PC we should keep. And the right picture, is a biplot for principal component. It is a scatter plot for PC to help us know the distribution of dataset in different direction.

The result of 9th ,33th, 53th sample stays so many years in English speaking country. The result of 2th sample has more working experience than other samples. The result of 17th,19th 50th sample are influenced by working experience, years and age. The results of 36th is influenced by the numbers of interview.

By the above analyzing, we wrote down the three PC:

PC1: Y1=-0.00071*G_GPA + 0.0077*TESC - 0.15*X - 0.988*I

PC2: Y2=0.0074*G_GPA + 0.113*TESC + 0.982*X - 0.1496*I

PC3: Y3=0.00461*G_GPA - 0.9936*TESC + 0.1101*X - 0.02458*I

In the next, we tried to explain our PC.

PC1: the degree of melting. The more years you stay in English speaking country, the more you know about western world. It has positive influence on your job offers. PC2: comprehensive ability. A person who is awesome at study, English and working experience will have more

chance to get an offer. PC3: diligence makes perfect. A person who studies hard and works hard is luckier in job hunting.

Finally, we deduct our dataset into 3 dimensions by using principal component analysis. They are PC1 represents the degree of melting; PC2 represents comprehensive ability; PC3 represents diligence making perfect.

PC1: Y1 = -0.00071*G_GPA + 0.0077*TESC - 0.15*X - 0.988*I

PC2: Y2 = 0.0074*G_GPA + 0.113*TESC + 0.982*X - 0.1496*I

PC3: Y3 = 0.00461*G_GPA - 0.9936*TESC + 0.1101*X - 0.02458*I

During the research, we also found there are some disadvantages about principal component analysis. 1.Principal component analysis is an unsupervised learning technique(do not use class information). 2.In details, when we analyzed the data, we did not separate the data according to their degree or the courses they chose.

These two variables will lead to a different result if we separate the data following them. So to do a more precise analysis, we did classification in the next part. And we also visualize the difference between PCA and LDA by ggplot.

Linear Discriminant Analysis

LDA, closely related to PCA, looks for linear combinations of variables, which best explain the data. And based on the R result. We found that the first two linear discriminants, listed below:

LD1: Z1 = 1. 56*G.gpa-0.12*T.ESC+0.09*X+0.58*I

LD2: Z2 = 1.61*G.gpa+0.08*T.ESC-0.20*X+0.37*I

explain the between-class variance very well because of the high proportion of trace, which is 99.89%, similar result to the principle component analysis.

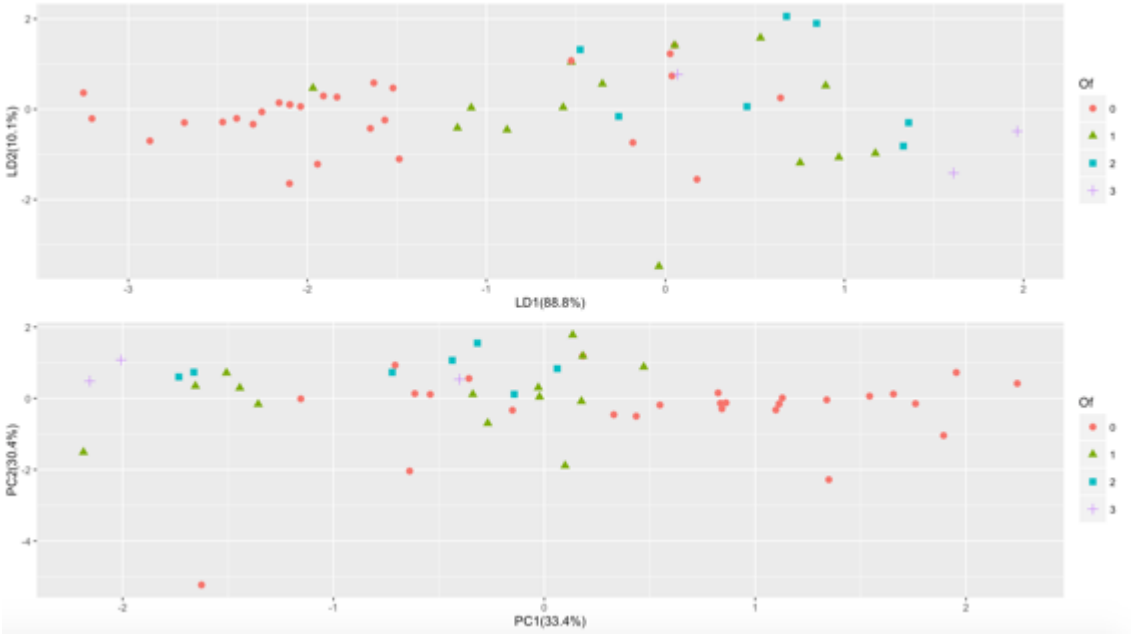

Figure 3. Visualization the difference between LDA and PCA

The difference between the LDA result and PCA can be visualized as Figure 3. It seems that LDA run better result than PCA because the same classes show a little bit more concentrate in LDA. However, neither of them performed good job here. And the error rate from hold out table turned out to be quite high, 41.18%.

To reduce the error rate and perform better result, if the job offer received or not became the new y parameter instead, we redid the LDA and got new result that,

LD1: Z1 = 2.2533*G.gpa-0.1069*T.ESC+0.0794*X+0.5471*I

would be the new linear discriminant with error rate 21.6%.

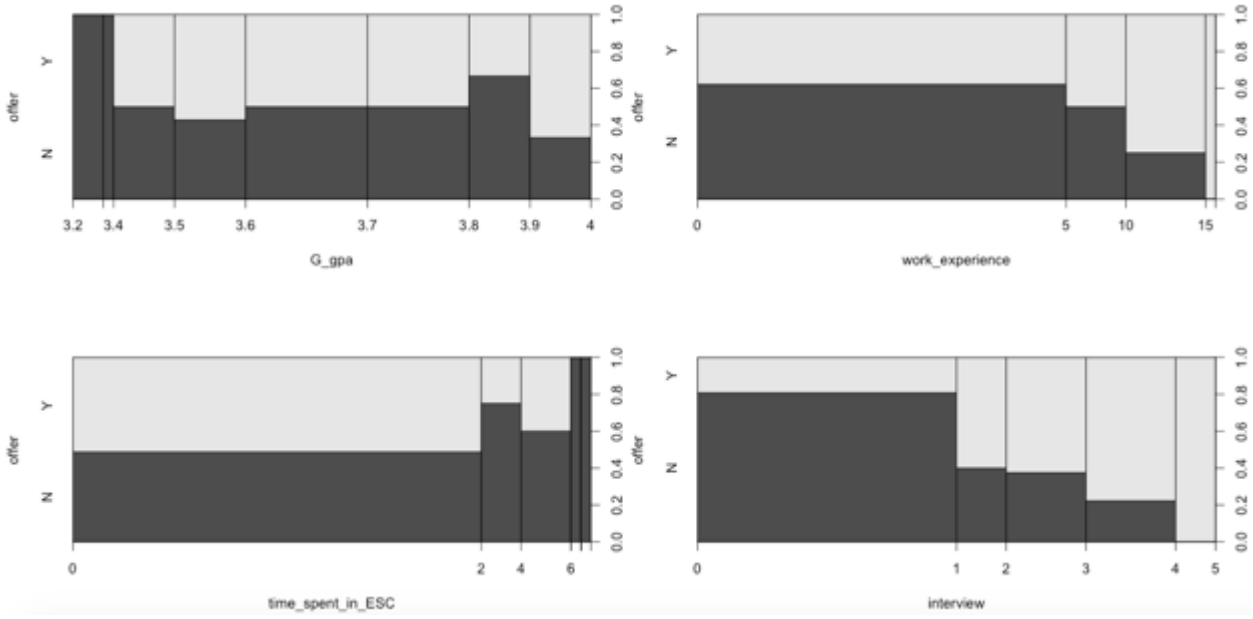

Figure 4. Histogram of job offer gets or not vs. graduate GPA, work experience, time spent in an English-speaking country and number of interviews separately.

From the histogram above, we can easily find that there is none of them who get any job offer has a graduate GPA below 3.4. And obviously, job offer has positive linear relation with either work experience or number of interviews.

Mean test

In the end, the mean test with $H_0 : \mu_1 - \mu_2 = 0$ was done to check the mean difference between group with job offers received and group without a job offer. And the null hypothesis was rejected because $T^2 = 38.06492 > c^2 = 30.92589$. From the former analysis, we considered that there is difference between the mean vectors with two groups. The null hypothesis as following,

$H_0 = \mu_1 - \mu_2 = \delta = (0,\ 0,\ 3.5,\ 4)'$, $T^2 = 5.722345 < c^2 = 10.9321$. H_0 should not be rejected and this verified that students who received a job offer probably have more work experience and more interviews.

Conclusions and Discussions

After this analysis, three points of advice are provided for Chinese international students with a major in Statistics. First, GPA is important; it does not need to be extremely high, but you should keep it over 3.4. Second, practice your English. The time you spend in an English-speaking country provides an advantage for obtaining job offers, but it does not play an essential role in the final offer decisions like we thought it would in the beginning. The last and also the most significant suggestion is to be active. Chance favors only the prepared mind. There will be a higher chance of you getting a job in the United States if you are more active by interviewing more and getting experience.

In addition, some improvements for future studies could be useful. 53 samples are too small compared to the population of 300,000 Chinese students in the United States. If big data is used, the results will be much more significant. In this case, we did not take networking into account, even though 70% of work positions are taken by networking in America. The number of connections on LinkedIn is probably significant for the job hunt.

Reference

Richard A., Dean W. (2007). *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ: Pearson.

Roles

Linear regression and background did by Guimin Dong. Linear Regression Extension and Logistic Regression did by Danjing Lin. PCA part did by Jian Sun. DCA, mean test and conclusion did by Xiaoqing Wu. Dataset collection did by Guimin Dong, Jian Sun and Xiaoqing Wu.

Appendix

R code:

```
data<-read.xlsx("C:\\Users\\Chauncey\\Desktop\\data_set 6215.xlsx",1)
fdl<-as.factor(data$DL)
fc<-as.factor(data$C)
 fm<-as.factor(data$M)
 fuschool<-as.factor(data$U.school)
 fg<-as.factor(data$G)
 fit0<-lm(O~H+W+A+T.ESC+fg+fdl+T+P+fc+U.gpa+fuschool+fm+G.gpa+X+I,data=data)
summary(fit0)
par(mfrow=c(2,2))
plot(fit0)
data1<-data[2:53,]
 fit1<--lm(I~I+X+ T.ESC,data=data1)
plot(fit1)
summary(fit1)
anova(fit0,fit1)
step(fit0,direction="both")
fit1<-glm(O~W+G.gpa+X+I+T.ESC,data=data1)
 fit<-glm(O~T.ESC+X+I,data=data1)
```

```
 cv.err1<-cv.glm(data1,fit1,K=10)$delta
 cv.err<-cv.glm(data1,fit,K=10)$deltaf<-lm(O~T.ESC+I+X,data=data1)
 f1<-lm(O~W+G.gpa+X+I+T.ESC,data=data1)
 anova(f,f1)
for (i in 1:52){
 if (o[i]>0){
 o[i]=1}
 else {
o[i]=0
}
o[i]
}
data2<-cbind(o,data1[,1],data1[,2],data1[,4],data1[,5],data1[,6],data1[,8],data1[,10],data1[,13],data1[,14],data1[,15])
colnames(data2)<-c("Offer","Height","Weight","Age","T","T.ESC","P","U.gpa","G.gpa","X","I")
data2<-data.frame(data2[1:52,])
logit_f<-glm(Offer~.,data=data2,family=binomial())
summary(logit_f)
logit_f2<-glm(Offer~T+X+I,data=data2,family=binomial())
summary(logit_f2)
FP=read.csv("/Users/LoveChina/Documents/6215/Final Project/3n.csv",header = TRUE)#input the sample data
NWD=data.frame(FP$G.gpa,FP$T.ESC,FP$X,FP$I,FP$O)#choose the data that you want to analysis
CND=cov(NWD[,-5])#calculate the parameters' variance-covariance matrix S
Xbar=colMeans(NWD[,-5])
PCA=prcomp(NWD[,-5],retx = TRUE, scale = FALSE)
summary(PCA) #check the cumulative proportion
#2 get some plot
PCA2=princomp(NWD[,-5],cor = TRUE)
pre=predict(PCA2)
head(pre)
screeplot(PCA,type="lines",main = "Check the PCA")#help to choose the number of principal component
biplot(PCA2)#check the principal component
pca <- prcomp(ND1[,-5],center = TRUE,scale = TRUE)
proppca = pca$sdev^2/sum(pca$sdev^2)
lda <- lda(ND1$O~.,ND1,prior=c(0.25,0.25,0.25,0.25))
proplda = lda$svd^2/sum(lda$svd^2)
plda <- predict(object = lda,newdata = ND1)
dataset = data.frame(Of = ND1[,"O"],pca = pca$x, lda = plda$x)
str(dataset)
dataset[,1]=factor(dataset[,1])
p1 <- ggplot(dataset) + geom_point(aes(K1,K2, colour = Of, shape = Of), size = 2.5) +
  labs(x = paste("LD1(", percent(proplda[1]), ")", sep=""),
       y = paste("LD2(", percent(proplda[2]), ")", sep=""))
p2 <- ggplot(dataset) + geom_point(aes(L1,L2, colour = Of, shape = Of), size = 2.5) +
  labs(x = paste("PC1(", percent(proppca[1]), ")", sep=""),
       y = paste("PC2(", percent(proppca[2]), ")", sep=""))
grid.arrange(p1, p2)
library(MASS)
lda.fit<-lda(O~.,data=na.omit(data_for_classification))
lda.fit
lda.fit$class<-predict(lda.fit)$class
table(na.omit(data_for_classification)$O,lda.fit$class)
```