Stat 112 – Final Project

Tier-2 Group 12

Group member:

YuLin Zhang	905118491
Xiangyu Zhou	504817670
Kaixin Huo	905188667
Simeng Shen	404826665
Qiyu Dai	804805350

1）Abstract: A summary of up to 100 words stating the problem to be solved within context, list of outcomes and predictors, statistical methods used to answer the question to be answered, statement of conclusions within context, limitations, and next steps. 5 points

The purpose of this project is to use the perception of confidence in academic performance, perception of sense of exclusion, international students(or not), as well as transfer students(or not) to predict the perception related to sense of belonging and perception of stress at UCLA. The predictors are Academic, Exclusion, International, and Transfer. The outcomes are Belonging and Stress. We used multivariate multiple linear regression to tackle this problem, and reached the conclusion that the models are able to explain some, but not all, of the variance in the outcome. There might be other important variables we didn't include and our sample dataset is heavily biased towards to statistics undergraduate community at UCLA. The next step would be feature selection on a broader level to find other variables related to Belonging and/or Stress.

2) Statement of the Problem: Description of the question to be answered within context. 5 points

How are multiple variables including UCLA students' perception of confidence in academic performance, perception for feeling of exclusion, whether they are transfer students, whether they are international students and the interaction effect of transfer and international students related to students' sense of belonging to UCLA and their stress levels?

Outcome variable:

- Belonging – indicator for perception related to sense of belonging to UCLA
- Stress – indicator for perception of stress

Predictor variable:

- Academic – indicator for perception of confidence in academic performance
- Exclusion - indicator for perception for feeling of exclusion
- International - whether students are international students or not
- Transfer - whether students are transfer students or not

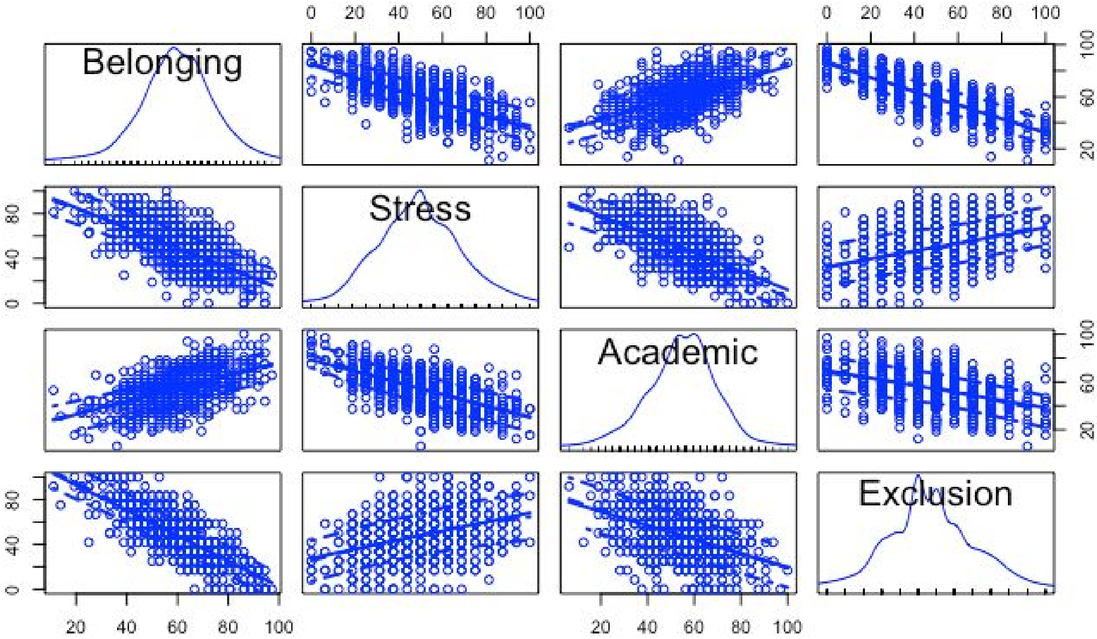
3) Exploratory data analysis: frequency tables, scatterplots, histograms, table of means, sample size and standard deviation for numerical predictors, correlation matrix of numerical variables) Do not include more than one plot or table per page. If there is enough space, write the description next to the plot or table, if not do it in the next slide. 15 points

Frequency table for the variables Transfer and International:

	Domestic	International	Total
Non-Transfer	399	125	524
Transfer	35	45	80
Total	434	170	604

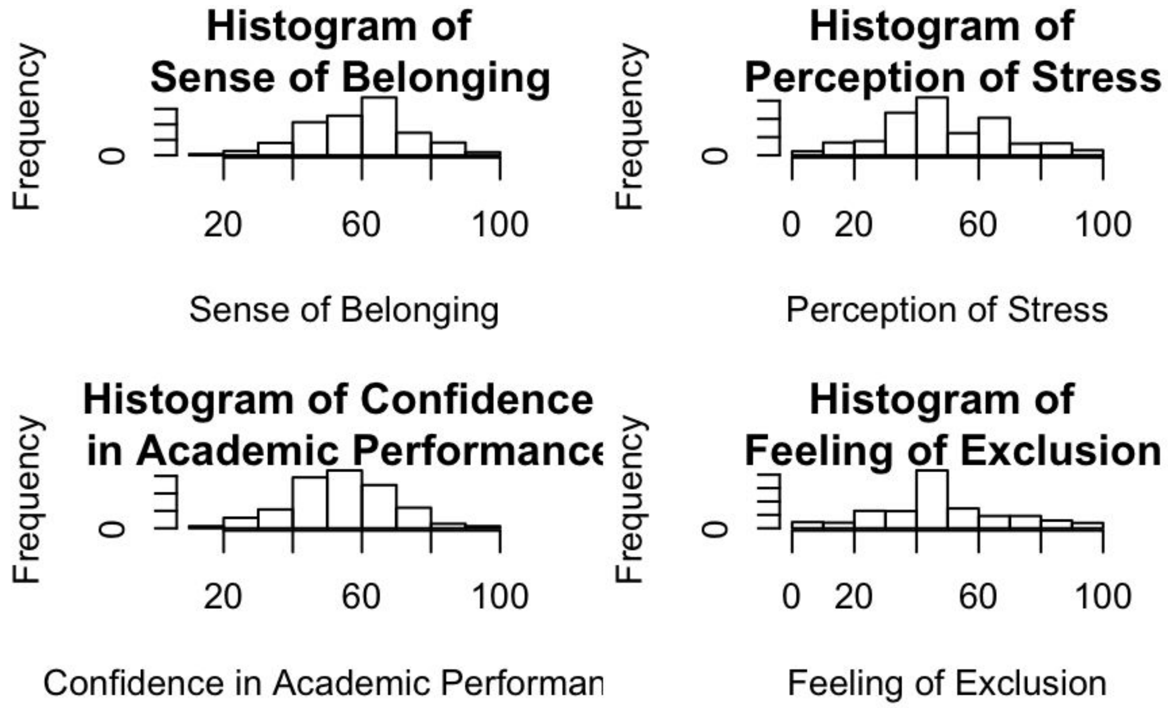
According to the frequency table, we can see that for domestic students, around 8% are transfer students, whereas for international students, around 26% are transfer students, which indicates that for international students there exists a high proportion of transfer students. In addition, for non-transfer students, around 24% are international students, whereas for transfer students, around 56% are international students, indicating a high proportion of international students in the transfer student community. Therefore, we expect to see an interaction between these two variables.

Scatterplots for the numerical variables Belonging, Stress, Academic, and Exclusion:



The scatterplots indicate pretty strong linear relationships between the two outcome variables and the two numerical predictors. The plots in the middle also show that both the outcome variables and the predictors are roughly normally distributed.

Histograms for the numerical variables Belonging, Stress, Academic, and Exclusion:



The histograms show that both the outcome variables and the numerical predictors are roughly normally distributed.

Table of means and standard deviations for the numerical variables:

	Mean	Standard Deviation
Outcome Variables:		
Belonging	59.62599	14.56283
Stress	49.52632	19.58026
Predictors:		
Academic	54.41291	14.47612
Exclusion	48.30397	20.73543

Sample size: 604.

Correlation matrix for numerical variables:

	Belonging	Stress	Academic	Exclusion
Belonging	1.00	-0.69	0.56	-0.77
Stress	-0.69	1.00	-0.65	0.41
Academic	0.56	-0.65	1.00	-0.47
Exclusion	-0.77	0.41	-0.47	1.00

A further looking into the correlation matrix shows the bivariate correlation between the two numerical outcomes (i.e., Belonging and Stress) and the two numerical predictors (i.e., Academic and Exclusion) to be high.

4) A table listing the outcome variables, predictors, and describing how they were measured, for categorical variables list the levels. If you changed a numerical variable to a factor, explain how you did it, If you pooled some categories explain how you did it, for instance if you combined strongly agree and disagree, make it clear, if need be change the order of a factor so it makes sense. (10 points)

	Type	Range/Level
Outcome Variables:		
Belonging	Numerical; Continuous (Indicator for perception related to sense of belonging to UCLA)	0-100
Stress	Numerical; Continuous (Indicator for perception of stress at UCLA)	0-100
Predictors:		
Academic	Numerical; Continuous (Indicator for perception of confidence in academic performance at UCLA)	0-100
Exclusion	Numerical; Continuous (Indicator for perception for the feeling of exclusion at UCLA)	0-100
Transfer	Categorical; Discrete (Indicator for whether the student is a transfer student or not)	2 levels: “No” (0) / “Yes” (1)
International	Categorical; Discrete (Indicator for whether the student is an international student or not)	2 levels: “No” (0) / “Yes” (1)

All the data entries with NA’s in one of the columns are removed from the original 866-entry dataset, resulting in 604 entries in our final dataset. Furthermore, the two levels of the categorical variables, i.e., “No” and “Yes”, are re-coded to be “0” and “1” for clarity purposes.

5) Summarizing the results of the relevant model in the form of a clear table. Do not cut and paste R outputs, Report your results in the form of clear tables. 10 points

Table for predicting **Belonging** from **Academic, Exclusion, Transfer or not, International or not and interaction of Transfer and International.**

	Coefficient	Standard Error	t	p
Intercept	67.57403	2.10791	32.057	<2e-16
Academic	0.27225	0.02771	9.827	<2e-16
Exclusion	-0.45643	0.01907	-23.941	<2e-16
Transfer-Yes	-0.35474	1.51158	-0.235	0.8145
International-Yes	-2.07871	0.89219	-2.330	0.0201
Transfer-Yes: International-Yes	-1.11145	2.12083	-0.524	0.6004

Residual standard error: 8.556 on 598 degrees of freedom
Multiple R-squared: 0.6577,   Adjusted R-squared: 0.6548
F-statistic: 229.8 on 5 and 598 DF, p-value: < 2.2e-16

Model: Predicted_Belonging = 67.57403 + 0.27225*Academic - 0.45643*Exclusion - 0.35474*Transfer(Yes) - 2.07871*International(Yes) - 1.11145*Transfer(Yes):International(Yes)

Table for predicting **Stress** from **Academic, Exclusion, Transfer or not, International or not and interaction of Transfer and International.**

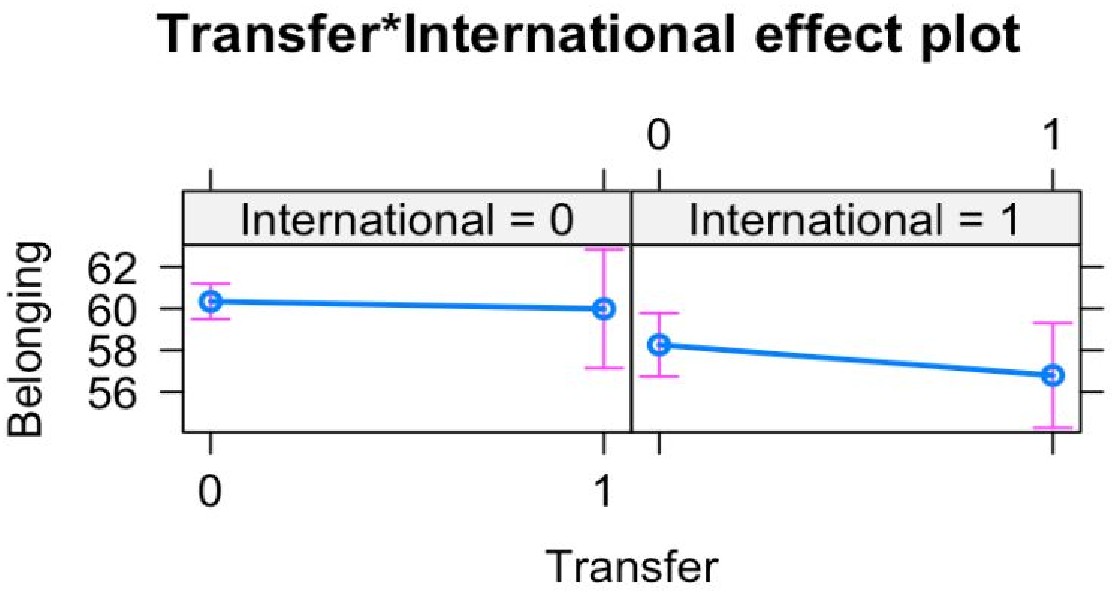
	Coefficient	Standard Error	t	p
Intercept	86.58059	3.60653	24.007	<2e-16
Academic	-0.78352	0.04740	-16.529	<2e-16
Exclusion	0.13238	0.03262	4.058	5.6e-05
Transfer-Yes	-3.54971	2.58623	-1.373	0.1704
International-Yes	-3.05915	1.52649	-2.004	0.0455
Transfer-Yes: International-Yes	6.92293	3.62863	1.908	0.0569

Residual standard error: 14.64 on 598 degrees of freedom
Multiple R-squared: 0.4457, Adjusted R-squared: 0.441
F-statistic: 96.15 on 5 and 598 DF, p-value: < 2.2e-16

Model: Predicted_Stress = 86.58059 - 0.78352*Academic + 0.13238*Exclusion - 3.54971*Transfer(Yes) - 3.05915*International(Yes) + 6.92293*Transfer(Yes):International(Yes)

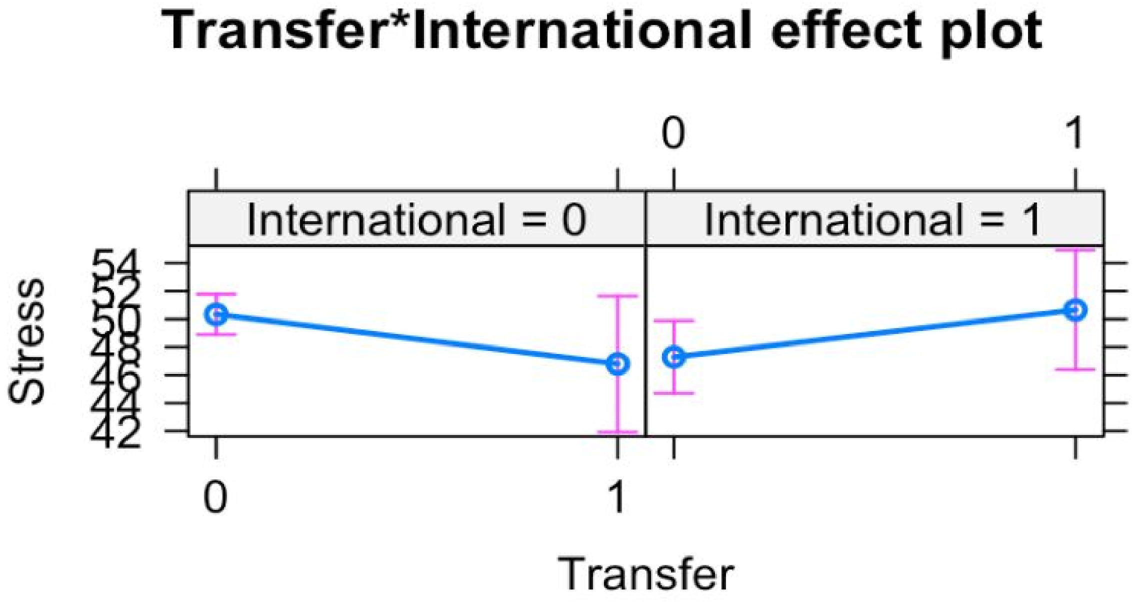
6) Plot of interaction effect (5 points)

Plot of interaction effect for predicting **Belonging** from **Academic, Exclusion, Transfer or not, International or not and interaction of Transfer and International.**



According to the interaction effect plot, since the lines are close to paralleling each other, we could say that the average sense of belonging to UCLA is similar for students who are international students and those who are not, regardless of whether they are transfer students or not.

Plot of interaction effect for predicting **Stress** from **Academic, Exclusion, Transfer or not, International or not and interaction of Transfer and International.**



According to the interaction effect plot, we could say that the average sense of stress at UCLA is slightly different for students who are transfer students and those who are not, and it depends on whether they are international students or not. For domestic students, if they are also transfer students, then their average sense of stress is lower than if they are non-transfer students. For international students, if they are also transfer students, then their average sense of stress is higher than if they are non-transfer students.The interaction effect is not obvious as it is consistent with the result from R that the p-value of interaction is 0.057, which is very close to the threshold which indicates statistical significance.

7) Interpretation of partial slopes, interaction effect, and adjust R-squared within context (20 points)

Interpretation for predicting **Belonging** from **Academic, Exclusion, Transfer or not, International or not and interaction of Transfer and International.**

1. Interpretation of partial slopes for Academic, Exclusion, Transfer, International, respectively:

   Keeping all other variables constant,

   For one unit of increase in the perception of confidence in academic performance at UCLA, on average the perception of sense of belonging will increase by 0.272 units. This is statistically significant.

   For one unit of increase in the perception of feelings of exclusion at UCLA, on average the perception of sense of belonging will decrease by 0.456 units. This is statistically significant.

   On average the mean of perception of sense of belonging is 0.355 units lower for transfer students than a non-transfer students. This is not statistically significant.

   On average the mean of perception of sense of belonging is 2.079 units lower for international students than domestic students. This is statistically significant.

2. Interpretation of the interaction effect:

   Interaction effect is not statistically significant indicating that that the effect of transfer from community college on the perception of sense of belonging is similar to students who are international students and those who are not.

3. Interpretation of the adjusted R-squared:

   65.5% of variance in the perception of sense of belonging is explained by the perception of confidence in academic performance at UCLA, the perception of feelings of exclusion at UCLA and whether students are international or not.

4. Conclusion to non-statistics audience

UCLA students with higher perception of academic confidence have more sense of belonging to UCLA.

UCLA students with higher perception of feeling of exclusion have less sense of belonging to UCLA.

International students have less perception of sense of belonging to UCLA than domestic students have.

Interpretation for predicting **Stress** from **Academic, Exclusion, Transfer or not, International or not and interaction of Transfer and International.**

1. Interpretation of partial slopes for Academic, Exclusion, Transfer, International, respectively:

   Keeping all other variables constant,

   For one unit of increase in perception of confidence in academic performance at UCLA, on average, perception of stress will decrease by 0.784 units. This is statistically significant.

   For one unit of increase in perception of feelings of exclusion at UCLA, on average, perception of stress will increase by 0.132 units. This is statistically significant.

   On average the mean of perception of stress is 3.550 units lower for transfer students than for non-transfer students. This is not statistically significant.

   On average the mean of perception of stress is 3.059 units lower for international students than for domestic students. This is statistically significant.

2. Interpretation of the interaction effect:

   The interaction effect is not statistically significant, indicating that the effect of transfer from community college on the perception of stress is similar to students who are international students and those who are not.

3. Interpretation of the adjusted R-squared:

44.1% of variance in the perception of stress is explained by the perception of confidence in academic performance at UCLA, the perception of feelings of exclusion at UCLA and whether students are international or not.

4. Conclusion to non-statistics audience

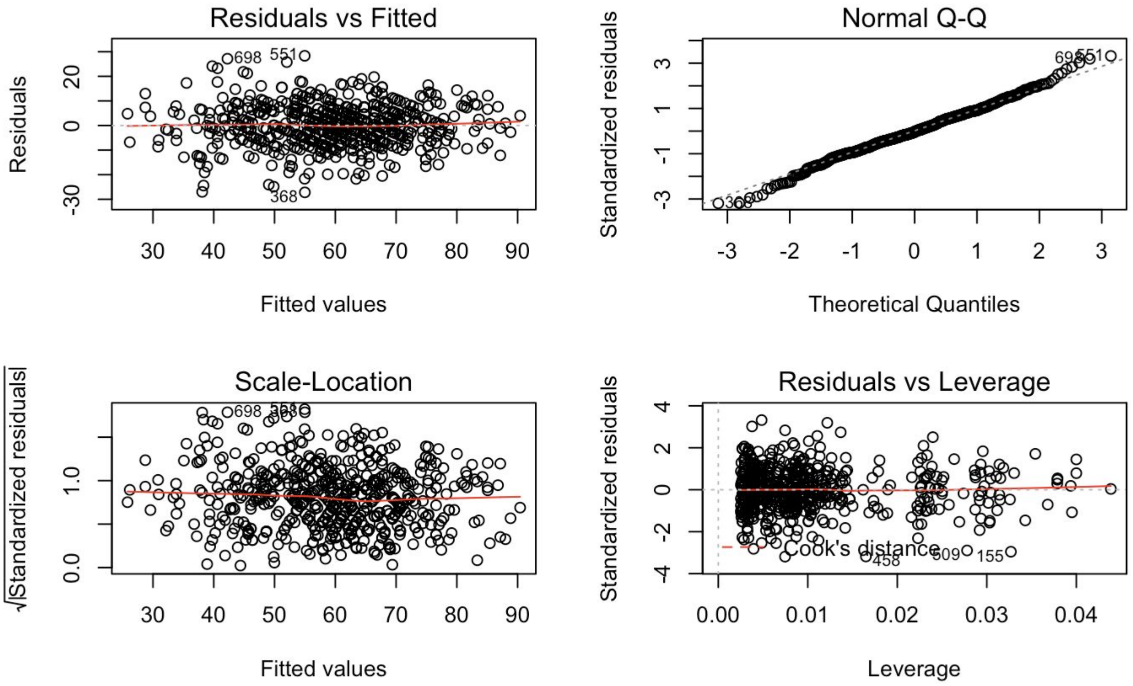
UCLA students with higher perception of academic confidence have less stress at UCLA.

UCLA students with higher perception of feeling of exclusion have higher sense of stress at UCLA.

International students feel less stressed at UCLA than domestic students do.

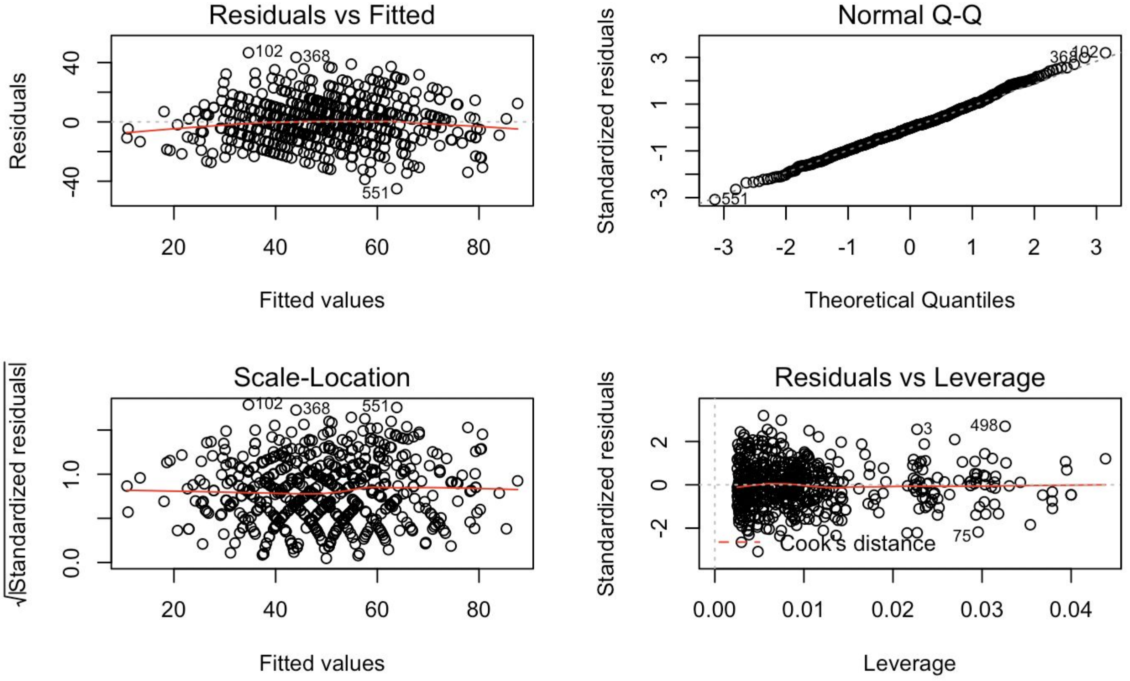
8) Checking the relevant assumptions and indicating whether they are met (15 points)

Assumptions for predicting **Belonging** from Academic, Exclusion, Transfer or not, International or not and interaction of Transfer and International:



All basic model assumptions are met. The residuals are randomly distributed, the error variance is constant, and Belonging is roughly normally distributed. However, there are some bad leverage points.

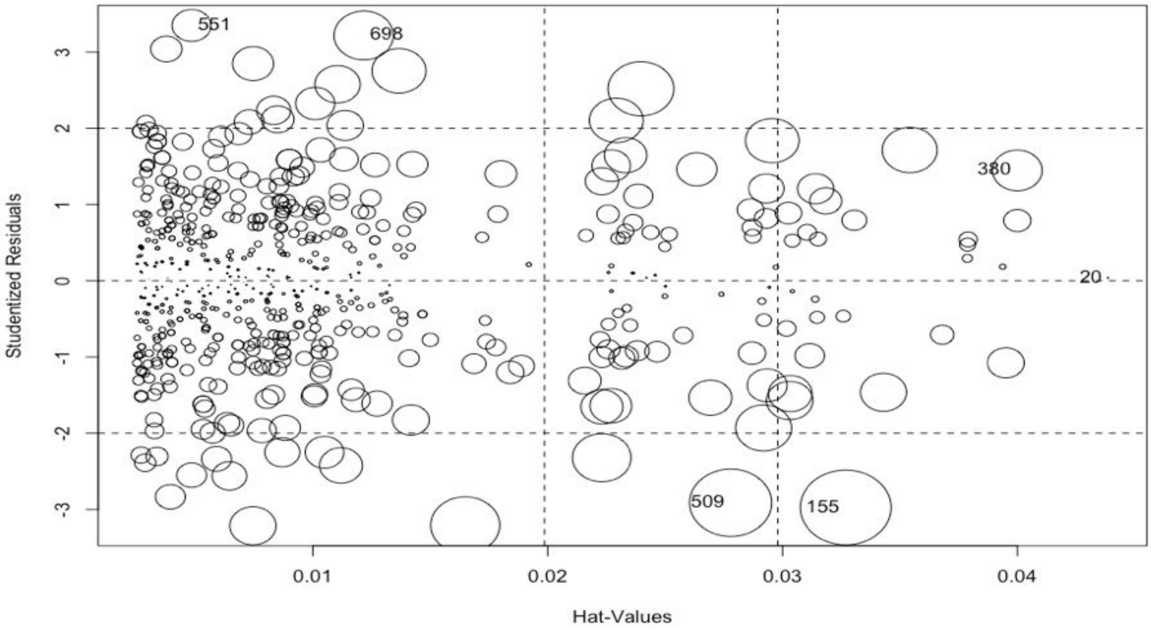
Assumptions for predicting **Stress** from Academic, Exclusion, Transfer or not, International or not and interaction of Transfer and International:



All basic model assumptions are met. The residuals are randomly distributed, the error variance is constant, and Stress is roughly normally distributed. However, there are some bad leverage points.

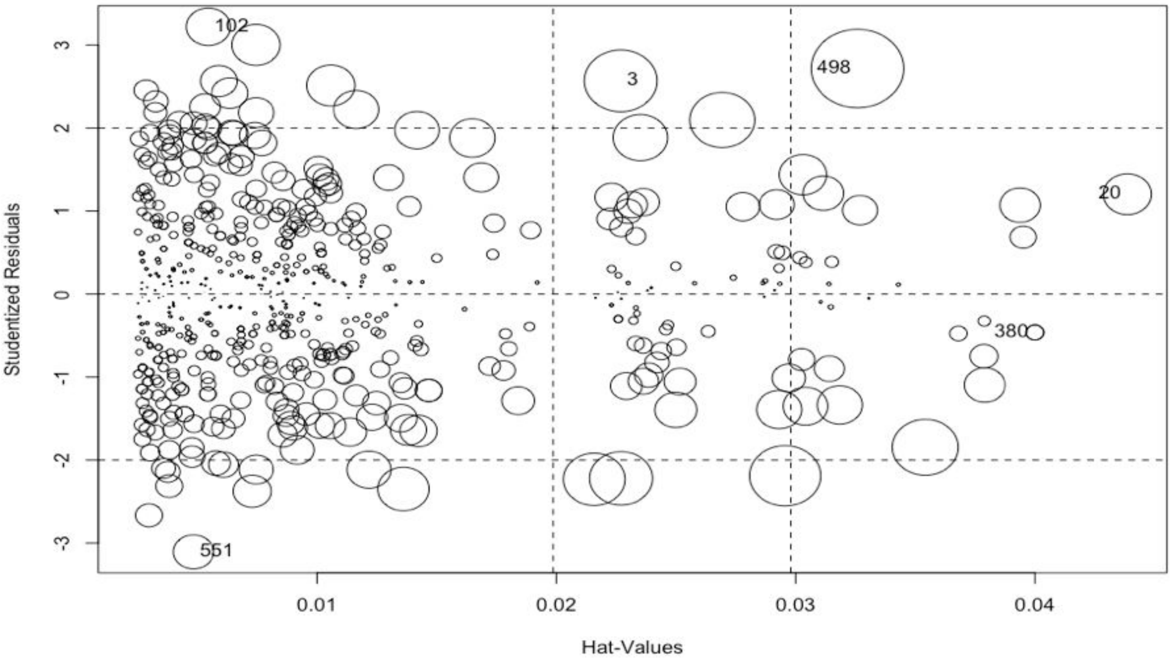
9) Checking for bad leverage - plots are needed (5 points)

Influence plot for predicting **Belonging** from Academic, Exclusion, Transfer or not, International or not and interaction of Transfer and International:



Since there are 604 observations in the model, the leverage threshold for bad leverage is 0.006622517. After calculating the hat leverage values, the points 1, 53, 121, 172, 181, 285, 368, 373, 393, 494, 509, 518, 560, 612, 698 have leverages are bigger than our leverage threshold and have standardized residuals which are larger than +/- 2. Therefore, there are a total of 15 bad leverage points.

Influence plot for predicting **Stress** from Academic, Exclusion, Transfer or not, International or not and interaction of Transfer and International:



Since there are 604 observations in the model, the leverage threshold for bad leverage is 0.006622517. After calculating the hat leverage values, the points 1, 3, 75, 368, 405, 498, 519, 511, 537, 560, 698, 752 have leverages are bigger than our leverage threshold and have standardized residuals which are larger than +/- 2. Therefore, there are a total of 12 bad leverage points.

10) Deleting bad leverage (if there are any) and running the new model (5 points).

Predicting **Belonging** from Academic, Exclusion, Transfer or not, International or not and interaction of Transfer and International after deleting bad leverage points:

	Coefficient	Standard Error	t	p
Intercept	67.66092	2.10549	32.135	<2e-16
Academic	0.27571	0.02771	9.950	<2e-16
Exclusion	-0.46295	0.01915	-24.177	<2e-16
Transfer-Yes	-0.26637	1.49842	-0.178	0.859
International-Yes	-2.15896	0.89382	-2.415	0.016
Transfer-Yes: International-Yes	-1.20702	2.11257	-0.571	0.568

```
Residual standard error: 8.47 on 584 degrees of freedom
Multiple R-squared:  0.6656,	Adjusted R-squared:  0.6627
F-statistic:  232.4 on 5 and 584 DF,  p-value: < 2.2e-16
```

According to the summary of the new model, we can see that Academic, Exclusion, and International are still significant. However, the adjusted R-squared has improved from 65.5% to 66.3%, which is not quite much.

Predicting **Stress** from Academic, Exclusion, Transfer or not, International or not and interaction of Transfer and International after deleting bad leverage points:

	Coefficient	Standard Error	t	p
Intercept	86.15950	3.60930	23.872	<2e-16
Academic	-0.78014	0.04746	-16.438	<2e-16
Exclusion	0.13723	0.03261	4.208	2.98e-05
Transfer-Yes	-3.56398	2.57495	-1.384	0.1669
International-Yes	-3.13988	1.53262	-2.049	0.0409
Transfer-Yes: International-Yes	6.17351	3.63069	1.700	0.0896

Residual standard error: 14.56 on 588 degrees of freedom
Multiple R-squared: 0.4515, Adjusted R-squared: 0.4468
F-statistic: 96.8 on 5 and 588 DF, p-value: < 2.2e-16

According to the summary of the new model, we can see that Academic, Exclusion, and International are still significant. However, the adjusted R-squared has improved from 44.1% to 44.7%, which is not quite much.

11) Limitations and recommendations (5 points)

We are only using four predictors here, so there might be other important factors that we didn’t include, which could in turn limit us from better results. In addition, our model did not have high external validity. The data was collected from students taking undergraduate statistics courses at UCLA, which could be highly biased. Most of the students who participated in generating our dataset are students who major or minor in statistics, which gives a high bias towards the statistics undergraduate community at UCLA, making our results hard to generalize to other populations. For example, our model might not give useful insight or accurate predictions about the psychology undergraduate community at UCLA. The recommendation for the next step is to first conduct feature selection. Using forward stepwise selection could be a good start. When an ideal number of predictors are included, we can use either the linear regression model or other non-linear approaches, such as random forest or gradient boosting, to get some insight into the relationship between the variables.