Due Wednesday, October 5, 11:59pm on Gradescope

36-309 / 36-749 Homework 4: Italian Linear Regression

Code ▾

Due Wednesday, October 5, 11:59pm on Gradescope

Sanaz Saadatifar

This homework is based on data I've analyzed in my own research. The data (a simplified version of the real data) is here:

Hide

```
italyData = read.csv("https://raw.githubusercontent.com/zjbranson/stat309fall2022/master/i
talyData.csv")
#make sure that grant is a categorical variable
italyData$grant = factor(italyData$grant)
```

The dataset consists of 299 first-year college students at the University of Pisa or the University of Florence from 2004-2006. In Italy, state universities offer financial grants to students whose families are deemed low-income. The grant is substantial: Students have their tuition waived, they get free meals and accommodations, and they receive a stipend of around 3,000 euros. However, to receive the grant, students must apply for it. The lower a student's income, the more likely they are to apply for the grant, because they are more likely to believe that they are "low income enough" to get the grant. (In reality, any student who applies and whose annual family income is below 15,000 euros will get the grant, but the students and their families don't know about this cut-off value.)

The dataset includes the following information on the first-year college students:

- `income` : A measure of the student's annual family income (in euros)
- `hsgrade` : Summary measure of high school grades (in %)
- `grant` : Whether or not they received the financial grant (yes or no).
- `dropout` : A quantitative measure (between 0 and 1) for a student's risk of dropping out of college after their first year. A higher measure denotes a higher risk of dropping out. For the purposes of this homework, you can assume that this measure has high construct validity (i.e., it really does measure a student's risk of dropping out).

The Italian education system was interested in assessing whether offering low-income students the financial grant lowers their risk of dropping out of college.

1. (10pts) For this part, answer the following two questions.

- (5pts) First, using only the outcome variable (`dropout`) and the treatment variable (`grant`), perform the appropriate formal statistical analysis for assessing whether receiving the grant raises or lowers students' risk of dropping out of college. After running your analysis, state your scientific conclusion from that analysis. In your answer, please report the p-value, treatment effect point estimate *for the dropout under `grant = "yes"` minus dropout under `grant = "no"`* , and confidence interval for this quantity. Also be sure to mention whether this suggests that the grant is helpful or harmful. (**Hint**: Remember that `dropout` is a quantitative outcome and `grant` is a two-level categorical variable.)

Hide

```
t.test(dropout~grant, data = italyData)
```

```
##
##  Welch Two Sample t-test
##
## data:  dropout by grant
## t = -5.4001, df = 296.4, p-value = 1.368e-07
## alternative hypothesis: true difference in means between group no and group yes is not
equal to 0
## 95 percent confidence interval:
##  -0.07817721 -0.03641543
## sample estimates:
##  mean in group no mean in group yes
##         0.3680429         0.4253392
```

[Null hypothesis: receiving grant does not raise or lower students’ risk of dropping out of college (H0: μ(yes) = μ2(no)) Alternative hypothesis: receiving grant raises or lowers students’ risk of dropping out of college (HA: μ(yes) ≠ μ2(no)) Since the P-value is 1.368e-07 (less than 0.05), we reject the null hypothesis. So, point estimate is `grant = "yes"` minus dropout under `grant = “no (0.4253392- 0.3680429=0.0572963) This suggests that receiving grant is harmful in terms of having higher risk of dropping out because we are 95% sure that the true mean difference of risk of dropout for people who did not receive a grant minus people who did receive the grant falls between (-0.07817721, -0.03641543) range. It shows that people who did receive a grant have higher risk of dropout.]

- (5pts) In this homework, we will be running regression models involving the outcome (dropout risk score), the treatment variable (whether or not a student receives the financial grant), and the other explanatory variables. When we interpret different variables’ coefficients for these models, will we be able to make any claims about causation, or will we only be able to make claims about association? Explain your answer in 1-2 sentences.

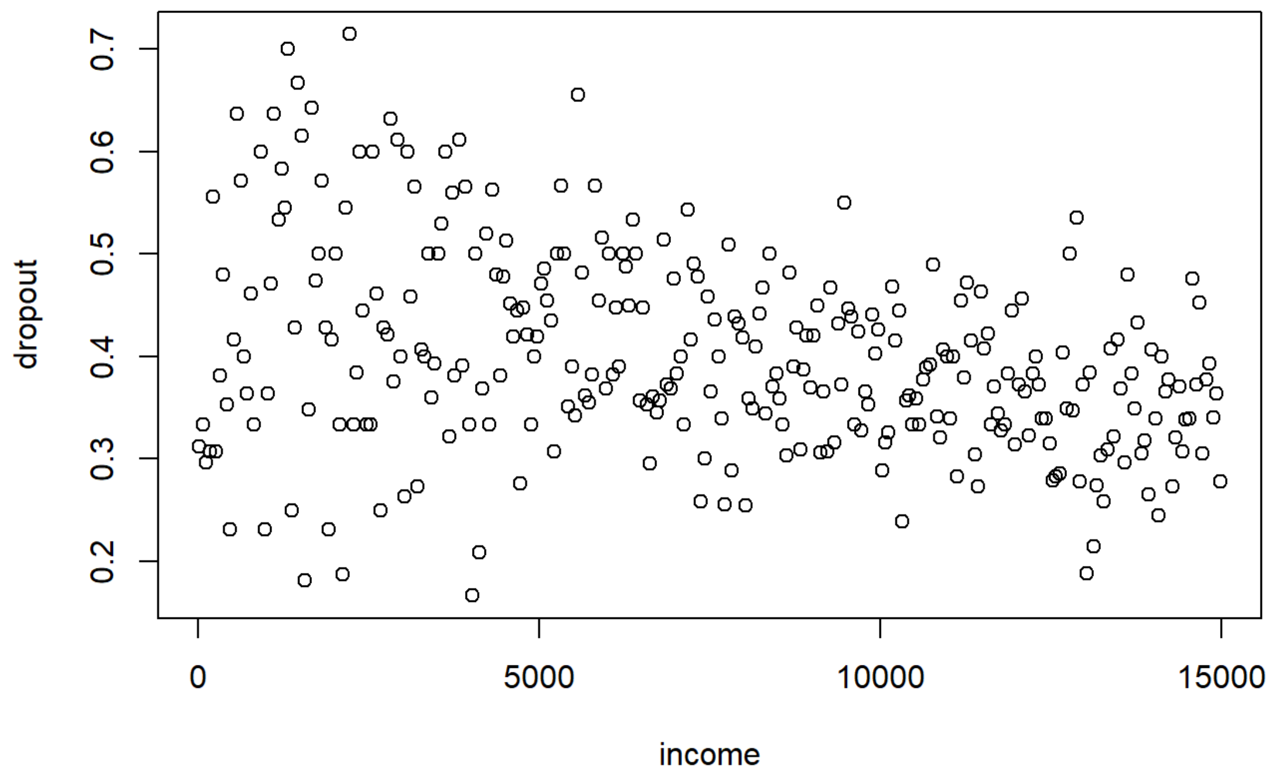
[we only be able to make claims about association because we do not have any information about internal validity of the 2 treatment groups. moreover, we only have limited explanatory variables here, but there might be other variables affecting the risk of the dropout which is nor included. Hence, it is association.]

2. (10pts) As some initial EDA, let’s make a few scatterplots for this dataset. For this part, answer the following two questions.

- (5pts) First, make a scatter plot with `income` on the x-axis and `dropout` on the y-axis. Does there appear to be a linear relationship between these two variables? Now make a scatter plot with `hsgrade` on the x-axis and `dropout` on the y-axis. Does there appear to be a linear relationship between these two variables? Your answer for this part should include two scatterplots and a discussion of each.

Hide

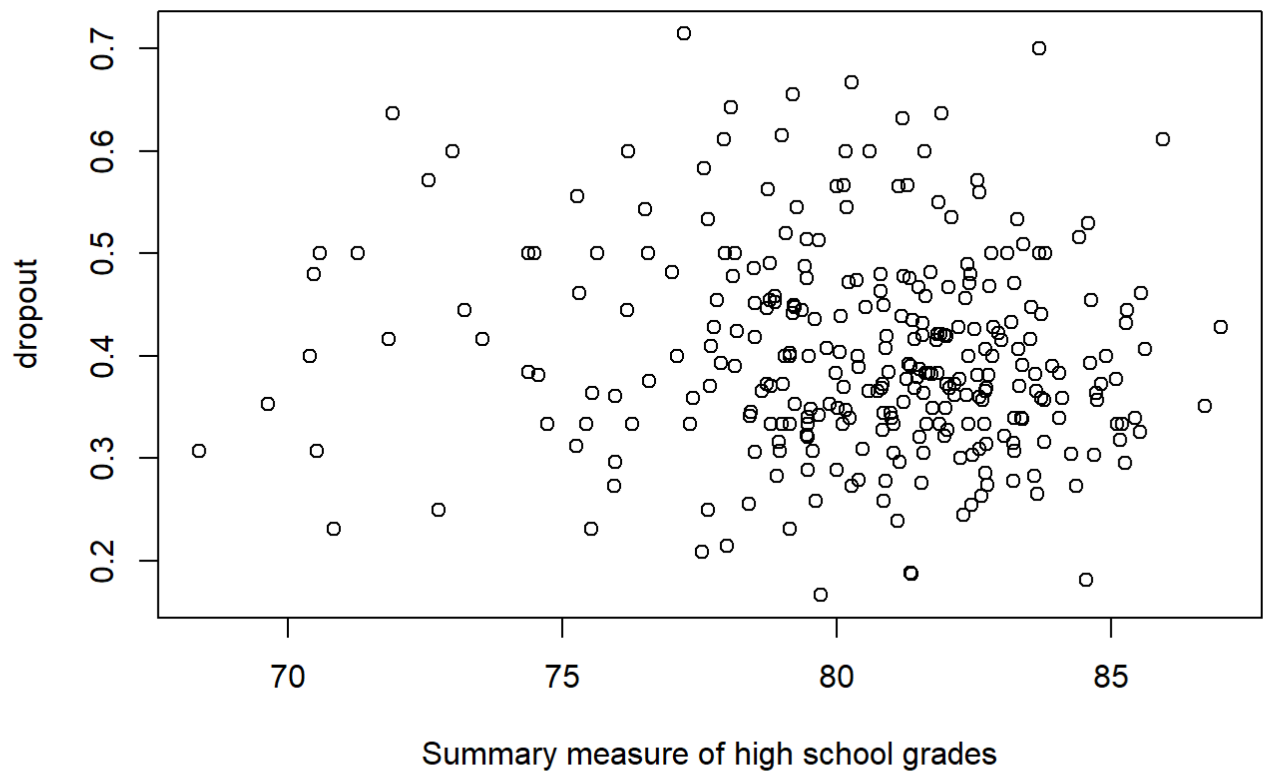
```
plot(italyData$income, italyData$dropout,
     xlab = "income", ylab = "dropout")
```



[There appears to be a linear relationship between dropout and income variables. As income increases the dropout risk decreases. Highest dropout risks are observed in lower incomes. However, there are also so many people with lower incomes but again lower risk of dropout.]

Hide

```
plot(italyData$hsgrade, italyData$dropout,
     xlab = "Summary measure of high school grades", ylab = "dropout")
```



[This scatter plot not necessarily shows a linear relationship between dropout and highs school grades variables. Because we can see different range of dropouts (both high or low range) in different grades such as the grades between 70 and 75 as well as the 85.]

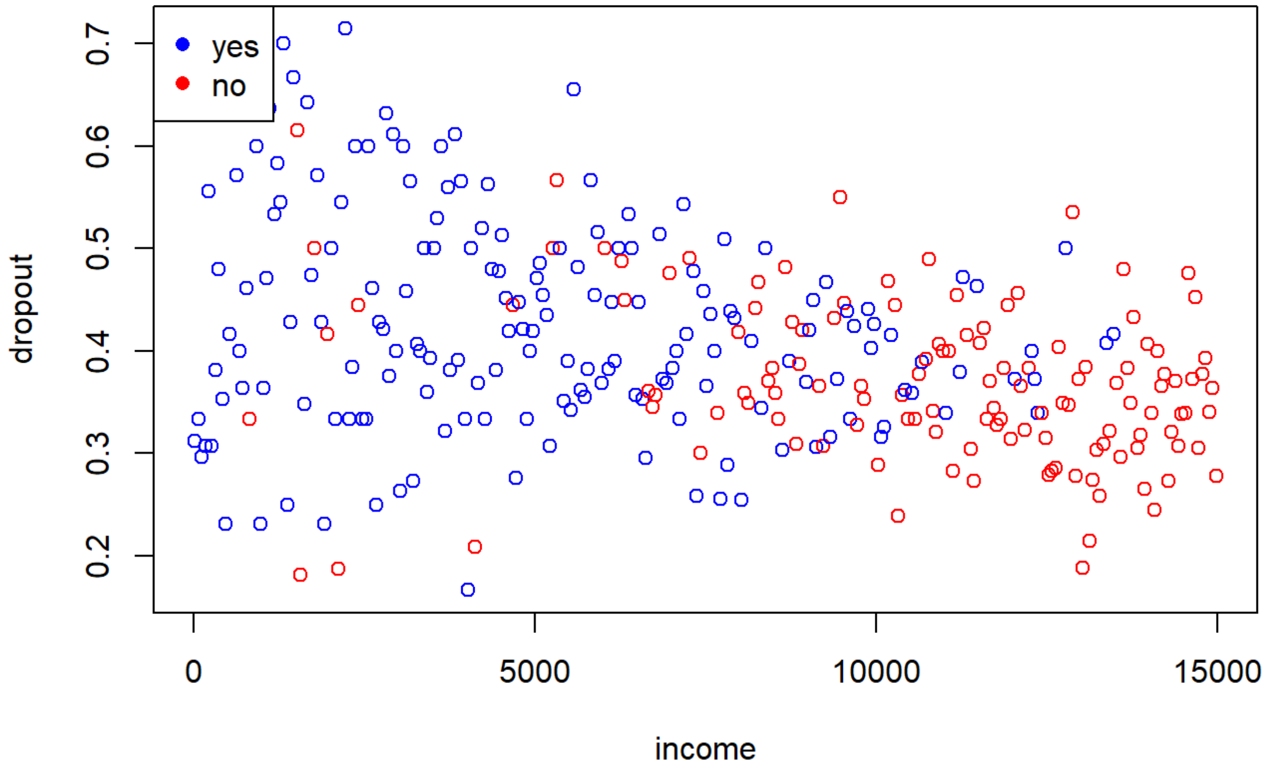
- (5pts) Now, similar to the previous part, make a scatter plot with `income` on the x-axis and `dropout` on the y-axis, but now color the points based on the `grant` variable. Then, do the same thing, but using `hsgrade` instead of `income` . After making your two scatterplots, write 1-2 sentences describing how these two scatter plots relate to **internal validity**. (**Hint**: First, make a `grantColor` variable using the `ifelse()` function, as in Lab5. Then, when using `plot()` , set `col = grantColor` . When defining `grantColor` , choose whatever two colors you want.)

Hide

```
grantColor = ifelse(italyData$grant == "yes", "blue","red")

plot(italyData$income, italyData$dropout,
     col = grantColor,
     xlab = "income", ylab = "dropout")

legend("topleft",
legend = c("yes", "no"),
col = c("blue", "red"),
pch = 16)
```



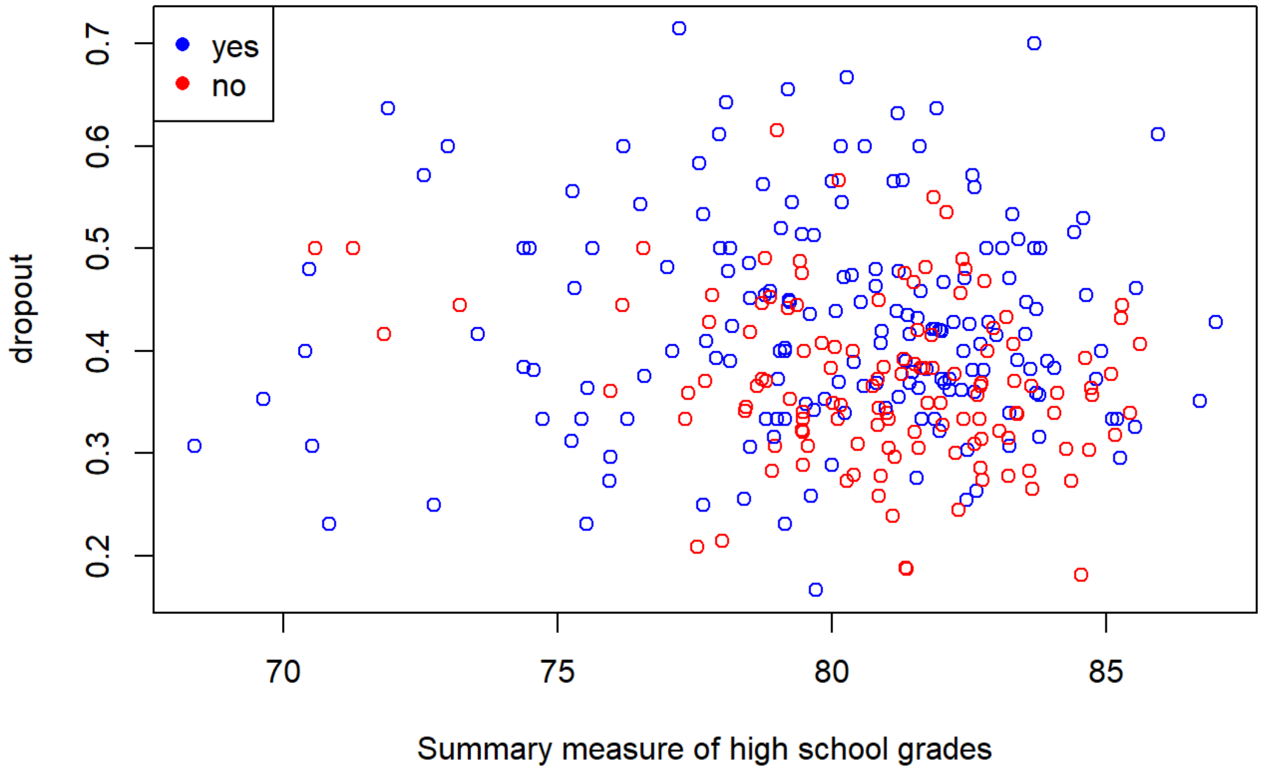
[This scatter plot shows a big difference between the yes and no groups in terms of income. The yes group are people with lower income and No group includes people with higher income, which implies low internal validity between two groups in terms of income.]

Hide

```
grantColor = ifelse(italyData$grant == "yes", "blue","red")

plot(italyData$hsgrade, italyData$dropout,
     col = grantColor,
     xlab = "Summary measure of high school grades", ylab = "dropout")

legend("topleft",
legend = c("yes", "no"),
col = c("blue", "red"),
pch = 16)
```



[This scatter plot does not show a big difference between the yes and no groups in terms of high school grades which implies high internal validity in terms of grades between two groups.]

3. (15pts) In this question, we'll consider three linear models, presented below. [I wrote the modeling equations in a particular way, such that they should display for folks with or without LaTeX on their computer.]

Model A: `dropout` is Normally distributed with mean equal to ß0 + ßi * income + ßg * Ig.

Model B: `dropout` is Normally distributed with mean equal to ß0 + ßh * hsgrade + ßg * Ig.

Model 1: `dropout` is Normally distributed with mean equal to ß0 + ßi * income + ßh * hsgrade + ßg * Ig.

Here, Ig denotes the indicator variable for grant, where Ig = 1 if someone received the grant and Ig = 0 if they did not receive the grant. Using these above model equations, answer the following questions.

- (5pts) For each model, write out your interpretation of the intercept parameter ß0.

[Model A: for the control group (group "NO"), the mean risk of dropout is estimated to be ß0 on average if the income is zero. Meanwhile , for the group "YES", the mean risk of dropout is estimated to be (ß0+ ßg) on average if the income is zero [ßg higher than control group]. Model B: for the control group (group "NO"), the mean risk of dropout is estimated to be ß0 on average if the hsgrade is zero. Meanwhile , for the group "YES", the mean risk of dropout is estimated to be (ß0+ ßg) on average if the hsgrade is zero [ßg higher than control group]. Model 1: for the control group (group "NO"), the mean risk of dropout is estimated to be ß0 on average if both the income and hsgrade are zero. Meanwhile , for the group "YES", the mean risk of dropout is estimated to be (ß0+ ßg) on average if both the income and hsgrade are zero [ßg higher than control group].]

- (5pts) For each model, if we estimated the intercept coefficient ß0, would that estimate be based on an interpolation or an extrapolation? Explain in 1-2 sentences. (**Hint**: For this question, it may be useful to look back at your scatter plots from the previous questions.)

[To answer this question we need to see whether we have the income and hsgrade variables zero or not. If we had data in range of zero, then it is interpolation, otherwise it is extrapolation. In terms of income, we do have data with income in the range of 0 (interpolation), but in terms of grades, we don't have data with hsgrade in the range of 0 (extrapolation). Hence the model A since only relies on income it would be interpolation, the model B only relies on the hsgrade then it is extrapolation. The model 1 since relies on both hsgrade and income, because hsgrade is extrapolation the model 1 also would be extrapolation.]

- (5pts) For Model B and Model 1 above, write out the interpretation of the high school grade coefficient ßh for each model.

[For model B, the dropout risk is estimated to increase by ßh for every one-unit (one-grade) increase in hsgrade. For model 1, the dropout risk is estimated to increase by ßh for every one-unit (one-grade) increase in hsgrade (also for this interpenetration, the income should not be changed).]

4. (16pts) Now we'll explore Model 1 in Question. For this part, answer the following questions.

- (4pts) First, write code to run the linear regression corresponding to Model 1 in Question 3. (**Hint**: This model includes `income`, `hsgrade`, and `grant` as explanatory variables, with no interactions.) To do this, fill in the template code below:

Hide

```
linReg1 = lm(dropout ~ grant + income + hsgrade, data = italyData)
summary(linReg1)
```

```
##
## Call:
## lm(formula = dropout ~ grant + income + hsgrade, data = italyData)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.265827 -0.054843 -0.008354  0.055854  0.273258
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.392e-01  1.432e-01   2.369 0.018471 *
## grantyes     2.316e-02  1.422e-02   1.629 0.104369
## income      -6.402e-06  1.787e-06  -3.584 0.000396 ***
## hsgrade      1.203e-03  1.860e-03   0.647 0.518111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09263 on 295 degrees of freedom
## Multiple R-squared:  0.1244, Adjusted R-squared:  0.1155
## F-statistic: 13.97 on 3 and 295 DF,  p-value: 1.53e-08
```

with "…" appropriately filled in. Then, you can use the `summary()` function to look at the output for this linear regression.

- (3pts) Now write (and turn in) code that provides the 95% confidence interval for each coefficient in Model 1. (**Hint**: Use the `confint()` function; see Lab5 or the Lecture4 R Demo.)

Hide

```
confint(linReg1)
```

```
##                     2.5 %        97.5 %
## (Intercept)  5.743070e-02  6.209513e-01
## grantyes    -4.819919e-03  5.114580e-02
## income      -9.918287e-06 -2.886477e-06
## hsgrade     -2.456541e-03  4.863007e-03
```

- (4pts) Using the regression output for this model, assess whether receiving the grant raises or lowers students' risk of dropping out of college, on average. In particular, provide a p-value, point estimate, and confidence interval for the effect of receiving the grant (compared to not receiving the grant) based on this model. What is your scientific conclusion from this analysis, and how does it compare to the conclusion you made in Question 1?

[Based on the Regression1 results the estimate for the group yes is 2.316e-02 higher than the intercept of the group no. however the p-value is 0.104369 (above 0.05) this shows that we do not have enough evidence to reject the null hypothesis of ßg=0. Or in ither words the value that we received for ßg, 2.316e-02, has the possibility of being 0 for the population's true ßg. Hence we cannot necessarily say that group yes (the people who received the grant) would have higher risk of dropout in comparison with those who did not receive grant. Moreover, the 95% CI shows that the population true ßg falls in the (-4.819919e-03 5.114580e-02) range that includes 0. Hence, compared to people who did not receive the grant, people who received the grant can have a lower dropout as 4.819919e-03 lower or higher dropout risk as 5.114580e-02 higher. It is contradictory to the conclusion made in Q1, however this contradiction is understandable because in Q1 we did not consider other variables such as income and grades.]

- (5pts) Finally, using the linear regression model you ran above, write out (and turn in) the prediction equation for estimating the mean outcome for a student based on their grant status, income level, and high school grades. (**Hint**: This equation should contain specific numbers.) Notice that the estimate for the income coefficient is very small (note that -6.402e-6 = -0.000006402). [More generally, in R, the notation "Ke-N" means K*10^-N - i.e., you should move the decimal point N places to the left.] After writing out your prediction equation, use it to estimate the mean outcome for the following students:

Student A: No grant, high school grades = 80, income = 0

Student B: No grant, high school grades = 80, income = 15,000

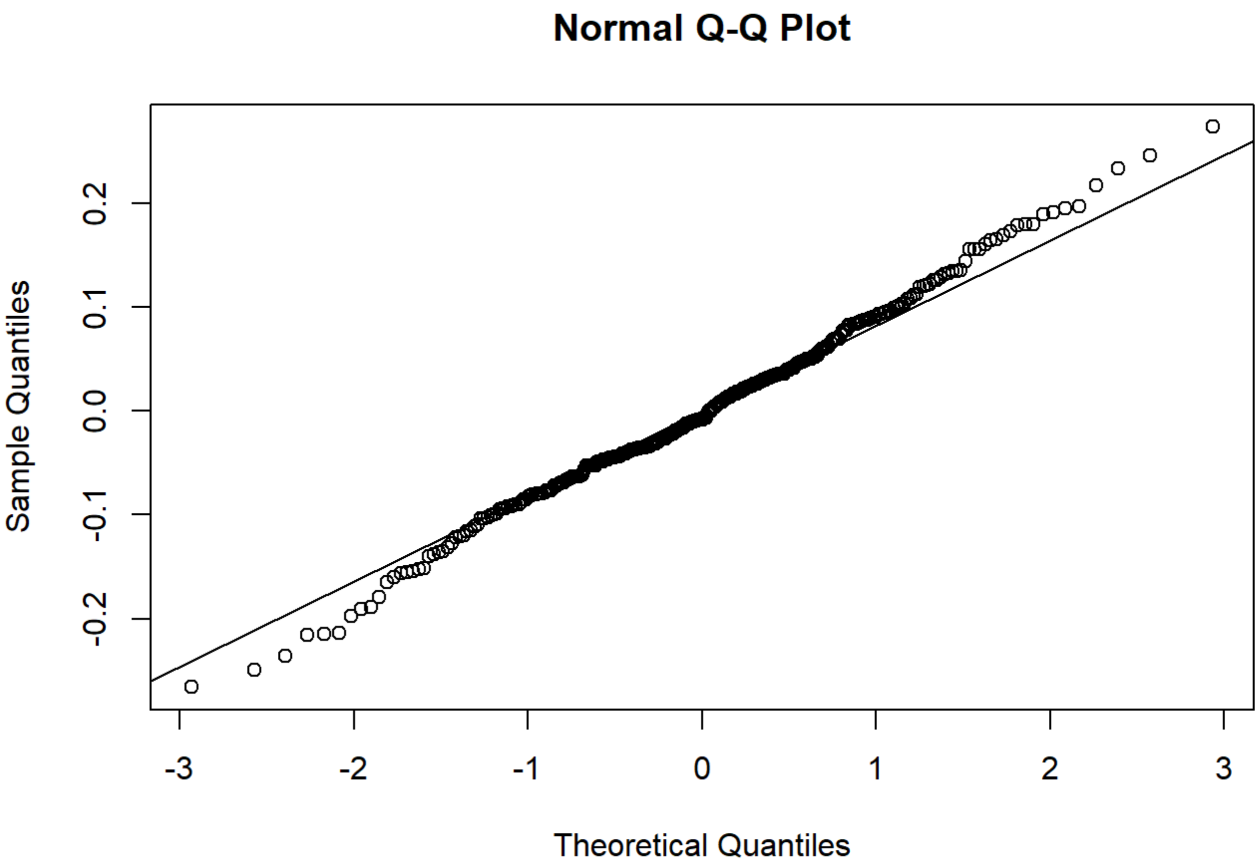
[Risk of dropout = ß0 + ßi * income + ßh * hsgrade + ßg * Ig = 0 .3392 - 0.000006402 * income + 0.001203 * hsgrade + 0.02316 * Ig Mean drop out risk (no grant) = 0 .3392 - 0.000006402 * income + 0.001203 * hsgrade Mean drop out risk (yes grant) = [0 .3392+ 0.02316] - 0.000006402 * income + 0.001203 * hsgrade = 0.36236- 0.000006402 * income + 0.001203 * hsgrade Student A: 0 .3392 - 0.000006402 * 0 + 0.001203 * 80= 0.43544 Student B: 0 .3392 - 0.000006402 * 15000 + 0.001203 * 80= 0.33941]

[Note that the income ranges from 0 to 15,000 in the data. The point of this question is to see that, even though estimated coefficients may be very small, they can make sizable differences (depending on the scale of the variables' measurements).]

5. (10pts) Using **residual analysis plots**, assess the Normal, Equal Variance, and Linearity assumptions for Model 1 from the previous two questions. In your answer, state whether you think each assumption is Definitely Plausible, Somewhat Plausible, Somewhat Not Plausible, or Definitely Not Plausible, as well as what (qualitative/visual) evidence you used from the residual analysis plots to arrive at each conclusion. For the equal variance and linearity assumptions, assess these assumptions within each treatment group (i.e., within the "received grant" group and the "didn't receive grant" group) by coloring the points by the `grant` variable (you can do this using the `grantColor` variable you made in Question 2).

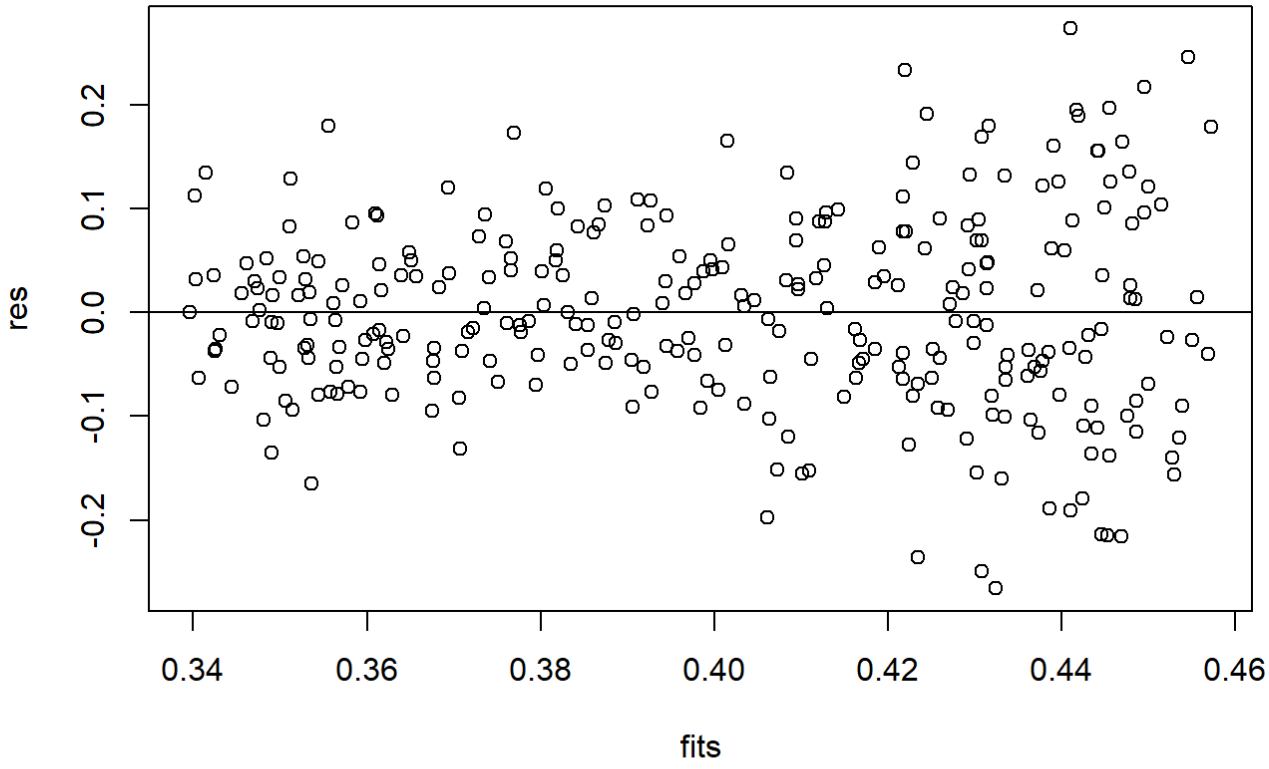
Hide

```
res = residuals(linReg1)
fits = fitted(linReg1)
#Here's a quantile-normal plot of the residuals:
qqnorm(res)
qqline(res)
```



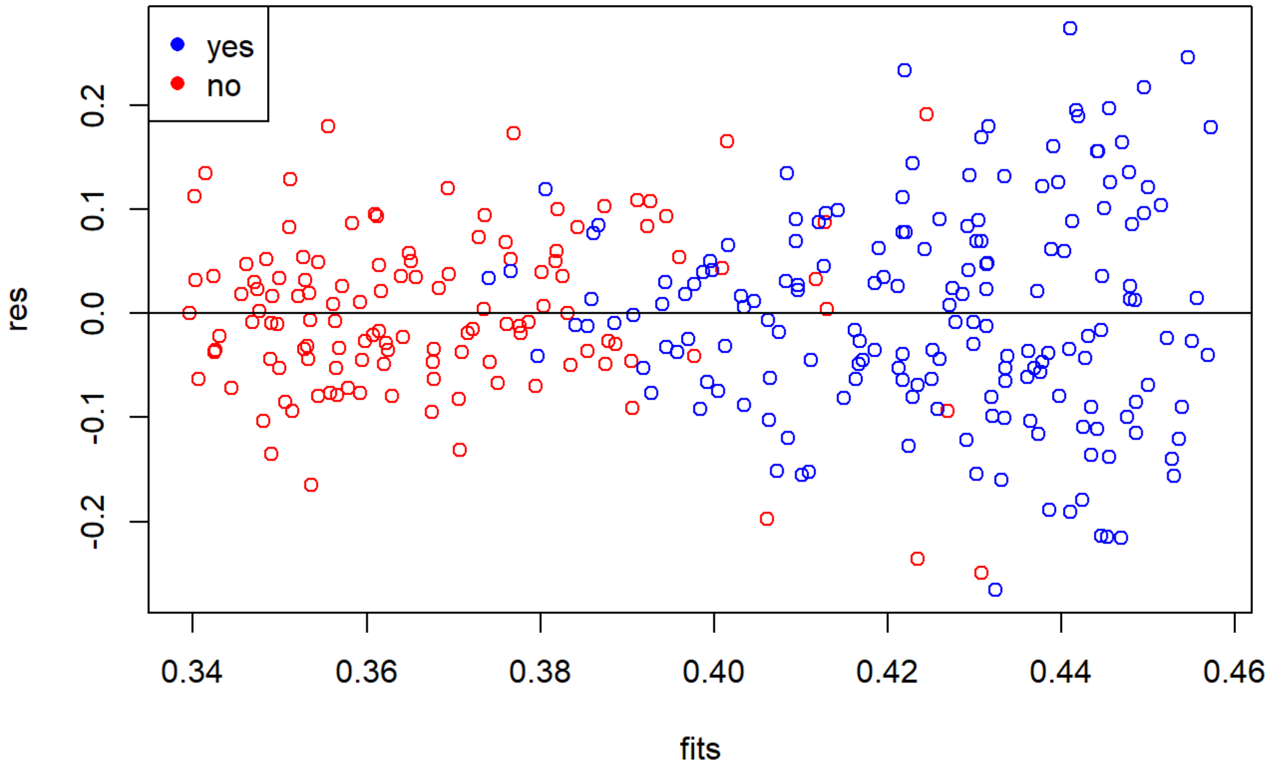
Hide

```
#And here’s a residual-vs-fit plot:
plot(fits, res)
abline(h=0)
```



Hide

```
#And here’s a residual-vs-fit plot based on grant:
plot(fits, res, col = grantColor)
abline(h=0)
legend("topleft",
legend = c("yes", "no"),
col = c("blue", "red"),
pch = 16)
```



[The Q-Q plot shows that normality is Definitely Plausible because most of the data especially in the middle are on the line. Just some data on the right and left are not on the line which is not a violation of normality.]

6. (10pts) Now we'll consider the following two linear models:

- **Model 2**: `dropout` is Normally distributed with mean equal to:

ß0 + ßi * income + ßh * hsgrade + ßg * Ig + ßhg * hsgrade * Ig.

- **Model 3**: `dropout` is Normally distributed with mean equal to:

ß0 + ßi * income + ßh * hsgrade + ßg * Ig + ßhg * hsgrade * Ig + ßig * income * Ig

Again, Ig = 1 for students who receive the grant and Ig = 0 for students who do not receive the grant. For this part, answer the following two questions.

- (6pts) For each model, write out the interpretation of ßi and ßh (i.e., the coefficients for `income` and `hsgrade` , respectively). In your answer, be sure to state the type of students these parameters are describing for each model.

[Model 2, for those who did not receive the grant: the interpretation of ßi shows that the dropout risk is estimated to increase by ßi for every one-unit increase in income when the hsgrade is fixed. the interpretation of ßh shows that the dropout risk is estimated to increase by ßh for every one-unit increase in hsgrade when the income is fixed. Model 2, for those who did receive the grant: the interpretation of ßi is same as those who did not receive the grant. the interpretation of ßh shows that the dropout risk is estimated to increase by ßh+ßhg for every one-unit increase in hsgrade when the income is fixed. Model 3, for those who did not receive the grant: the interpretation of ßi

and ßh is similar to the interpretation of ßi and ßh for Model 2, for those who did not receive the grant. Model 3, for those who did receive the grant: the interpretation of ßh is the same the interpretation of ßh in Model 2, for those who did receive the grant. the interpretation of ßi shows that the dropout risk is estimated to increase by ßi+ ßig for every one-unit increase in income when the hsgrade is fixed.]

- (4pts) For each model in the previous question, does the interpretation of the intercept change, as compared to the interpretation of the intercept for Model 1 in Question 3? Explain your answer in 1-3 sentences.

[no it wont change, the intercept onterpretaion for the modelu 2 and 3 is same as the module 1. the reason is that no specific category is added to the modeuls 2 and 3 that would change the intercept. the module 2 and 3 are more related to the interactions of grant groups with the the income and grade variables. so it is more related to the interpretation of slop not the intercept]

7. (10pts) Now we'll run the linear regression models corresponding to Model 2 and to Model 3 in the previous question. For this part, answer the following two questions.

- (6pts) First, write code to run the linear regressions for Model 2 and Model 3. To do this, fill in the template code below:

Hide

```
#running Model 2
linReg2 = lm(dropout ~ grant * hsgrade + income, data = italyData)
summary(linReg2)
```

```
##
## Call:
## lm(formula = dropout ~ grant * hsgrade + income, data = italyData)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.264906 -0.056468 -0.004162  0.059663  0.279049
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       7.031e-01  2.545e-01   2.763 0.006097 **
## grantyes         -4.811e-01  2.923e-01  -1.645 0.100937
## hsgrade          -3.333e-03  3.215e-03  -1.037 0.300731
## income           -6.098e-06  1.789e-06  -3.408 0.000745 ***
## grantyes:hsgrade  6.269e-03  3.631e-03   1.727 0.085263 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09232 on 294 degrees of freedom
## Multiple R-squared:  0.1332, Adjusted R-squared:  0.1214
## F-statistic: 11.29 on 4 and 294 DF,  p-value: 1.54e-08
```

Hide

```
#running Model 3
linReg3 = lm(dropout ~ grant * hsgrade + grant* income, data = italyData)
summary(linReg3)
```

```
##
## Call:
## lm(formula = dropout ~ grant * hsgrade + grant * income, data = italyData)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.266292 -0.056961 -0.005918  0.060885  0.276629
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       7.768e-01  2.678e-01   2.901   0.0040 **
## grantyes         -5.912e-01  3.177e-01  -1.861   0.0637 .
## hsgrade          -4.495e-03  3.472e-03  -1.295   0.1964
## income           -4.190e-06  2.795e-06  -1.499   0.1349
## grantyes:hsgrade  7.974e-03  4.108e-03   1.941   0.0532 .
## grantyes:income  -3.233e-06  3.639e-06  -0.888   0.3750
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09235 on 293 degrees of freedom
## Multiple R-squared:  0.1355, Adjusted R-squared:  0.1208
## F-statistic: 9.187 on 5 and 293 DF,  p-value: 3.896e-08
```

with "…" appropriately filled in. Then, you can use the `summary()` function to look at the output for this linear regression. For this part, all you need to do is provide the code for `linReg2` and `linReg3` (and print out the output using the `summary()` function).

- (4pts) To make sure you're running the correct models, you should have gotten the following R-squared values for Models 1, 2, and 3 respectively: 0.1244, 0.1332, 0.1355. Also, you should have gotten the following Adjusted R-squared values for Models 1, 2, and 3 respectively: 0.1155, 0.1214, 0.1208. (Thus, if you did not get these numbers, you should go back and check your code.) For this part, answer the following: How should we interpret the fact that, when going from Model 2 to Model 3, the R-squared goes up but the Adjusted R-squared goes down?

[Whenever a variables interaction with the outcome is added to the regression model, the multiple R-squared would increase. However if this addition is not relevant (addition of explanatory variables that are unassociated with the outcome), the adjusted R-squared would penalized for that addition by being decreased. Here the model 3's adjusted R-squared is less than module 2's. so It shows that the interaction between grant and income was not that relevant to be added to the interactive regression model or the income variable is not associated with the dropout risk.]

8. (11pts) Now we'll consider a prediction equation for Model 3. For this part, answer the following questions.

- (4pts) Write out (and turn in) the prediction equation for Model 3 (using specific numbers).

[Mean drop out risk = ß0 + ßi * income + ßh * hsgrade + ßg * Ig + ßhg * hsgrade * Ig + ßig * income * Ig Mean drop out risk (no grant) = ß0 + ßi * income + ßh * hsgrade = 0.7768 - 0.00000419* income - 0.004495* hsgrade
Mean drop out risk (yes grant) = ß0 + ßi * income + ßh * hsgrade + ßg + ßhg * hsgrade + ßig * income = 0.1856 - 0.000007423* income + 0.003479* hsgrade]

- (5pts) Based on your prediction equation, compute this model's **estimate of the treatment effect** (i.e., the effect of receiving the grant versus not receiving the grant) for these types students:
  - Student A: High school grades = 70, income = 5000
  - Student B: High school grades = 70, income = 10000
  - Student C: High school grades = 85, income = 5000
  - Student D: High school grades = 85, income = 10000

Note that, for each student, you'll have to compute *two* numbers: Your estimate of `dropout` when they receive the grant and your estimate when they do not receive the grant. Then, your estimate of the treatment effect for that student will be the difference between those two numbers. (**Hint**: For this problem, it can be useful to first simplify your prediction equation for the case where a student receives the grant, and then for the case where a student does not receive the grant. Then, the treatment effect will be the difference between those two prediction equations.)

[Student A if receives the grant: 0.1856 - 0.000007423* 5000 + 0.003479* 70 = 0.392015 Student A if does not receive the grant: 0.7768 - 0.00000419* 5000 - 0.004495* 70 = 0.4412 Effect of treatment : Student A if receives the grant - Student A if does not receive the grant: 0.392015-0.4412 = -0.049185]

- (2pts) Based on your estimates above, for what types of students does the grant appear to be helpful? For what types of students does the grant appear to be unhelpful?

[based on the estimates above, the grant's effectivness was more related to the highschool grades rather than income. it seems that when the grades are lower (75) the grant is not effective (negative effect), however when the gardes are higher (85) the grant seems to have a positive effect]

9. (8pts) At this point, you should have successfully defined `linReg1` , `linReg2` , and `linReg3` . (`linReg1` should have been defined in Question 4; `linReg2` and `linReg3` should have been defined in Question 7.) We'll now compare these three models. For this part, answer the following two questions.

- (4pts) First, uncomment the following code (which will only run if you've defined `linReg1` , `linReg2` , and `linReg3`):

Hide

```
anova(linReg1, linReg2, linReg3)
```

```
## Analysis of Variance Table
##
## Model 1: dropout ~ grant + income + hsgrade
## Model 2: dropout ~ grant * hsgrade + income
## Model 3: dropout ~ grant * hsgrade + grant * income
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    295 2.5310
## 2    294 2.5056  1 0.0254112 2.9795 0.08538 .
## 3    293 2.4989  1 0.0067318 0.7893 0.37503
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

You should see a table with three rows and two p-values in the "Pr(>F)" column. Based on this output, which of the three models (Model 1, Model 2, or Model 3) is most preferable? Explain your reasoning in 1-2 sentences.

[the last P-value on the third line which is for the comparison of module 2 over 3, is 0.37503 (above 0.05). therefore module 2 is preferred over the module 3. it shows that the added interaction between grant and income is not associated enough with the outcome compared to the module2. The first p-value on the second line which is for the comparison of the module 1 over 2, is 0.08538 (above 0.05). therefore module 1 is preferred over the module 2. it shows that the added interaction between grant and grade is not associated enough with the outcome compared to the module1. Therefore module 1, the additive module is preferred most among these three.]

- (4pts) Look back at the p-value in the third row of the above table; it should be 0.375. Now look at `summary(linReg3)` (after you successfully defined `linReg3`). You should find another p-value equal to 0.375. Why is it equal to the p-value in the third row of the above table?

[It is equal to the p-value for the slop in terms of grantyes:income. The null hypothesis is that ßig=0 and the alternative hypothesis is that ßig ≠ 0. Since the p-value is higher than 0.05, then we fail to reject the null hypothesis. Hence, it shows that interaction between grant and income is not relevant and not associated with the risk of dropout. Since in module 3 this income variable is added, the p-value in the anova is also 0.375 that is above 0.05]

10. (5pts, ONLY REQUIRED FOR 36-749 STUDENTS; BONUS QUESTION FOR 36-309 STUDENTS) After you finish this homework, imagine that you are meeting with Italy's Ministry of Education, University, and Research (https://en.wikipedia.org/wiki/Ministry_of_Education,_University_and_Research_(Italy)). The ministry asks you two questions (in English, not Italian):

- Do the university financial grants significantly affect first-year students' risk of dropping out of university?
- Do the effects of the university financial grants significantly vary for different students (e.g., for income levels and high school grade levels)?

Using what you did in this homework, answer these two questions in 2-3 sentences. Assume that the ministry does not understand what you mean by "p-value" or "confidence interval" (they speak English, but they don't speak statistics!) So, you will have to use more ubiquitous language (but still using statistical arguments to back up your claims). Also, be careful about making causal claims versus claims about association.

[for the first question, we dont have enough evidence to necessarily say that financial grants significantly affect the first-year students' risk of dropping out of university. since based on module 1, whaich is selceted to be the most preferred model in this case, it is clear that the estimate for the group yes is 2.316e-02 higher than the intercept of the group no. however the p-value is 0.104369 (above 0.05) this shows that we do not have enough evidence to reject the null hypothesis of ßg=0. Or in other words the value that we received for ßg, 2.316e-02, has the possibility of being 0 for the population's true ßg. Hence we cannot necessarily say that group yes (the people who received the grant) would have higher risk of dropout in comparison with those who did not receive grant. Moreover, the 95% CI shows that the population true ßg falls in the (-4.819919e-03 5.114580e-02) range that includes 0. Hence, compared to people who did not receive the grant, people who received the grant can have a lower dropout as 4.819919e-03 lower or higher dropout risk as 5.114580e-02 higher in comparison with the intercept of the group no. also in terms of second question which i would also refer to the module 1, I would say we have enough evidence to say that the effects of the university financial grants significantly vary for different students in terms of income, as the P-value is 0.000396 (less than 0.05), hence in the regression model the slope or relevenace of the income would not be 0. the opposite of the income situation can be interpretted for the high school grade levels, since the p-value is 0.518111 (above 0.05). and we dont have enough evidence to say that the slope or relevenace of the grade to the effect of the university financial grants on dropout level would not be 0]