# STOR 455 Exam #2

**50 points**

**Directions:** This exam is open books, notes, internet, and all things other than direct communication with others. The *LAdata.csv* dataset is needed to complete the exam. This dataset can be imported from the web address below or from the csv file, also attached in this Sakai assignment. You should complete the exam in this R Notebook, including all code, plots, and explanations. For your submission, you should knit the notebook and submit it as a pdf to Gradescope. If you are unable to knit your exam, you should submit the RMD file to Sakai under the ‘Unable to Knit’ tab. The dataset can be found at GitHub at the address below:

<https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/LAdata.csv>

**Should STOR have Undergraduate Learning Assitants? (YES! Starting Fall 2022…)**

Large introductory STEM courses historically have high failure rates, and failing such courses often leads students to change majors or even drop out of college. Instructional innovations such as the Learning Assistant model can influence this trend by changing institutional norms. In collaboration with faculty who teach large-enrollment introductory STEM courses, undergraduate learning assistants (LAs) use research-based instructional strategies designed to encourage active student engagement and elicit student thinking. These instructional innovations help students master the types of skills necessary for college success such as critical thinking and defending ideas. A study was conducted to investigate the relationship between exposure to LA support in large introductory STEM courses and general failure rates in introductory courses at University of Colorado Boulder.

Alzen, J.L., Langdon, L.S. & Otero, V.K. (2018) A logistic regression investigation of the relationship between the Learning Assistant model and failure rates in introductory STEM courses. *International Journal of STEM Education*, *56*(5). <https://doi.org/10.1186/s40594-018-0152-1>

The *LAdata.csv* dataset represent a subset of the variables examined in this study and a random sample of the data for students in one course, MATH 1300.

| **Variables** | **Descriptions** |
| --- | --- |
| *la\_stud* | Did the student’s course have a learning assistant? (1=yes; 0=no) |
| *sex* | Identified sex of the student (1=male; 0=female) |
| *nonwhite* | Does the student identify as not white? (1=yes; 0=no) |
| *first.gen* | Is the student a first generation college student? (1=yes; 0=no) |
| *finaid\_ever* | Has the student ever receive financial aid? (1=yes; 0=no) |
| *act\_new* | ACT score for the student |
| *hs\_gpa* | High school GPA for the student |
| *credits\_entry* | College credits earned before entering the university |
| *grade* | Grade in the course on a 0.0 to 4.0 scale |
| *fail* | Did the student fail the course? (1=yes; 0=no) |

library(readr)

LAdata <- read\_csv('https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/LAdata.csv')

1. Use the *LAdata.csv* dataset to construct a model to predict if students will *fail* a class using *la\_stud*, *sex*, *nonwhite*, *first.gen*, as well as the interactions between *la\_stud* with each of *sex*, *nonwhite* and *first.gen* as predictors (ie 3 interactions of two variables, with *la\_stud* and each one of the other predictors). Include a summary of this model. *6 pts*

mod1 = glm(fail ~

la\_stud +

nonwhite +

first.gen +

sex +

la\_stud\*nonwhite +

la\_stud\*first.gen +

la\_stud\*sex,

data=LAdata,

family=binomial)

summary(mod1)

##

## Call:

## glm(formula = fail ~ la\_stud + nonwhite + first.gen + sex + la\_stud \*

## nonwhite + la\_stud \* first.gen + la\_stud \* sex, family = binomial,

## data = LAdata)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -0.9828 -0.6412 -0.5567 -0.3909 2.3218

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -1.22376 0.20397 -6.000 1.98e-09 \*\*\*

## la\_stud -0.44374 0.25337 -1.751 0.0799 .

## nonwhite 0.74711 0.34294 2.179 0.0294 \*

## first.gen -0.09224 0.40773 -0.226 0.8210

## sex -1.30957 0.32798 -3.993 6.53e-05 \*\*\*

## la\_stud:nonwhite -0.55691 0.40915 -1.361 0.1735

## la\_stud:first.gen 0.70932 0.47646 1.489 0.1366

## la\_stud:sex 0.91348 0.38946 2.346 0.0190 \*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 959.54 on 1068 degrees of freedom

## Residual deviance: 926.12 on 1061 degrees of freedom

## AIC: 942.12

##

## Number of Fisher Scoring iterations: 5

1. Conduct a hypothesis test at the 0.05 significance level to determine the effectiveness of the *la\_stud* terms in the model constructed in question 1. Cite your hypotheses, p-value, and conclusion in context. *8pts*

*H*0:*β*1=*β*5=*β*6=*β*7=0

*HA*:*At* *least* *one* *of* *β*1,*β*5,*β*6,*β*7≠0

Since the p-value is small (0.04286 < 0.05), there is statistically significant evidence to claim that at least one of the la\_stud terms have a nonzero coefficient.

mod2 = glm(fail ~

nonwhite +

first.gen +

sex,

data=LAdata,

family=binomial)

anova(mod2, mod1, test='Chisq')

|  |
| --- |
|  |

|  | **Resid. Df**  **<dbl>** | **Resid. Dev**  **<dbl>** | **Df**  **<dbl>** | **Deviance**  **<dbl>** | **Pr(>Chi)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- |
| 1 | 1065 | 935.9784 | NA | NA | NA |
| 2 | 1061 | 926.1187 | 4 | 9.859697 | 0.04285856 |

2 rows

# Can also be done this way

G = summary(mod2)$deviance - summary(mod1)$deviance

1 - pchisq(G, 4)

## [1] 0.04285856

1. For non first generation nonwhite female students, what does the model from question 1 predict will be the probability that these students will **pass** the course for courses **with** a learning assistant? What does the model from question 1 predict will be the probability that these students will **pass** the course for courses **without** a learning assistant? *8pts*

s1 = data.frame(sex=0, nonwhite=1, first.gen=0, la\_stud=1)

s2 = data.frame(sex=0, nonwhite=1, first.gen=0, la\_stud=0)

#Probability non first generation nonwhite female student passes with LA

1 - predict(mod1, s1, type='response')

## 1

## 0.8141637

#Probability non first generation nonwhite female student passes without LA

1 - predict(mod1, s2, type='response')

## 1

## 0.6169563

1. Construct a model to predict students’ *grade* in the course using all of the variables (except for *fail*, *hs\_gpa*, and *act\_new*) and including the interactions between *la\_stud* and all of the other variables (except for *fail*, *hs\_gpa*, and *act\_new* and as with question 1, the interactions between *la\_stud* and each one of the other predictors). You should not include transformations, nor a residual analysis. Include a summary of this model. *5 pts*

mod4 = lm(grade~. + la\_stud\*., data=LAdata[-c(6, 7, 10)])

summary(mod4)

##

## Call:

## lm(formula = grade ~ . + la\_stud \* ., data = LAdata[-c(6, 7,

## 10)])

##

## Residuals:

## Min 1Q Median 3Q Max

## -3.2440 -0.6333 0.1551 0.8389 2.2176

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 2.3955419 0.1124618 21.301 < 2e-16 \*\*\*

## la\_stud 0.0026514 0.1343917 0.020 0.984264

## sex 0.4466620 0.1224744 3.647 0.000278 \*\*\*

## nonwhite -0.3225878 0.1495213 -2.157 0.031194 \*

## first.gen 0.1253382 0.1706271 0.735 0.462762

## finaid\_ever -0.4158908 0.1284067 -3.239 0.001237 \*\*

## credits\_entry 0.0272103 0.0084656 3.214 0.001348 \*\*

## la\_stud:sex -0.2341253 0.1478221 -1.584 0.113531

## la\_stud:nonwhite 0.1872411 0.1767231 1.060 0.289606

## la\_stud:first.gen -0.3123343 0.2060254 -1.516 0.129819

## la\_stud:finaid\_ever 0.3500524 0.1542170 2.270 0.023416 \*

## la\_stud:credits\_entry 0.0006063 0.0097250 0.062 0.950300

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 1.106 on 1057 degrees of freedom

## Multiple R-squared: 0.08483, Adjusted R-squared: 0.0753

## F-statistic: 8.907 on 11 and 1057 DF, p-value: 2.621e-15

1. Using **only** the summary output of the model constructed in question 4, What does the coefficient of the *la\_stud* variable (not the interactions, nor the test statistic or p-value) tell you about the relationship between the *la\_stud* and *grade* variables in this model, as well as what specific students this coefficient applies to? *5pts*

The la\_stud coefficient (0.0026514) in the increase in GPA for students in a course with a learning assistant, specifically for students that are female (sex=0), not nonwhite (nonwhite=0), not first generation college students (first.gen=0), have not had financial aid ever (finaid\_ever=0), and have no college credits before entering the university (credits\_entry=0).

1. Perform a **backwards** model selection method (for the lowest Mallow’s Cp) using all of the terms in the model that you constructed in question 4 as possible predictors. Construct this best model and include a summary of this model. *6pts*

MSE = (summary(mod4)$sigma)^2

mod5 = step(mod4, scale=MSE, trace=FALSE)

summary(mod5)

##

## Call:

## lm(formula = grade ~ la\_stud + sex + nonwhite + finaid\_ever +

## credits\_entry + la\_stud:sex + la\_stud:finaid\_ever, data = LAdata[-c(6,

## 7, 10)])

##

## Residuals:

## Min 1Q Median 3Q Max

## -3.1515 -0.6342 0.1717 0.8420 2.2277

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 2.382222 0.108682 21.919 < 2e-16 \*\*\*

## la\_stud 0.019275 0.127257 0.151 0.879639

## sex 0.446036 0.121615 3.668 0.000257 \*\*\*

## nonwhite -0.200140 0.078110 -2.562 0.010536 \*

## finaid\_ever -0.409754 0.122709 -3.339 0.000869 \*\*\*

## credits\_entry 0.028039 0.004157 6.744 2.52e-11 \*\*\*

## la\_stud:sex -0.241861 0.146793 -1.648 0.099725 .

## la\_stud:finaid\_ever 0.314556 0.146807 2.143 0.032369 \*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 1.105 on 1061 degrees of freedom

## Multiple R-squared: 0.08155, Adjusted R-squared: 0.07549

## F-statistic: 13.46 on 7 and 1061 DF, p-value: < 2.2e-16

1. Conduct a hypothesis test at the 0.05 significance level to determine the effectiveness of the *la\_stud* terms in the model constructed in question 6. Cite your hypotheses, p-value, and conclusion in context. *8pts*

*H*0:*β*1=*β*6=*β*7=0

*HA*:*At* *least* *one* *of* *β*1,*β*6,*β*7≠0

Since the p-value is small (0.0447 < 0.05), there is statistically significant evidence to claim that at least one of the la\_stud terms have a nonzero coefficient.

mod7 = lm(grade~sex+nonwhite+finaid\_ever+credits\_entry, data=LAdata)

anova(mod7, mod5)

|  |
| --- |
|  |

|  | **Res.Df**  **<dbl>** | **RSS**  **<dbl>** | **Df**  **<dbl>** | **Sum of Sq**  **<dbl>** | **F**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 1064 | 1306.490 | NA | NA | NA | NA |
| 2 | 1061 | 1296.602 | 3 | 9.887836 | 2.697048 | 0.0447025 |

2 rows

1. In question 6 I specifically asked you to perform a backwards model selection method rather than a forward or stepwise method. Using the backwards method, *la\_stud* terms were included in your best model. If you had used a forwards or stepwise method, you would find that *la\_stud* terms would **not** be included in your best models. Using the summary of the model from question 6, and knowledge of how the backwards, forwards, and stepwise procedures determine the best model, why would the *la\_stud* terms be included in the backwards model selection output, but not in the forward and stepwise model selection output? *4pts*

backwards will first look to remove the interaction terms and not the individual LA\_stud terms. Since the interaction terms are not all removed from the model in backwards selection, it is never an option to remove the LA\_stud term. For forwards and stepwise, you begin with not predictors. THe interaction terms are not considered unless the individual terms are first added to the model. LA\_stud alone has a high p-value and was never added to the model, so the interaction terms were never considered.

step(mod4, scale=MSE)

## Start: AIC=12

## grade ~ la\_stud + sex + nonwhite + first.gen + finaid\_ever +

## credits\_entry + la\_stud \* (la\_stud + sex + nonwhite + first.gen +

## finaid\_ever + credits\_entry)

##

## Df Sum of Sq RSS Cp

## - la\_stud:credits\_entry 1 0.0048 1292.0 10.004

## - la\_stud:nonwhite 1 1.3721 1293.3 11.123

## <none> 1292.0 12.000

## - la\_stud:first.gen 1 2.8092 1294.8 12.298

## - la\_stud:sex 1 3.0662 1295.0 12.508

## - la\_stud:finaid\_ever 1 6.2977 1298.3 15.152

##

## Step: AIC=10

## grade ~ la\_stud + sex + nonwhite + first.gen + finaid\_ever +

## credits\_entry + la\_stud:sex + la\_stud:nonwhite + la\_stud:first.gen +

## la\_stud:finaid\_ever

##

## Df Sum of Sq RSS Cp

## - la\_stud:nonwhite 1 1.369 1293.3 9.1243

## <none> 1292.0 10.0039

## - la\_stud:first.gen 1 2.830 1294.8 10.3189

## - la\_stud:sex 1 3.075 1295.1 10.5199

## - la\_stud:finaid\_ever 1 6.305 1298.3 13.1617

## - credits\_entry 1 53.907 1345.9 52.1067

##

## Step: AIC=9.12

## grade ~ la\_stud + sex + nonwhite + first.gen + finaid\_ever +

## credits\_entry + la\_stud:sex + la\_stud:first.gen + la\_stud:finaid\_ever

##

## Df Sum of Sq RSS Cp

## - la\_stud:first.gen 1 2.113 1295.5 8.8532

## <none> 1293.3 9.1243

## - la\_stud:sex 1 3.061 1296.4 9.6289

## - nonwhite 1 6.838 1300.2 12.7189

## - la\_stud:finaid\_ever 1 7.168 1300.5 12.9887

## - credits\_entry 1 54.251 1347.6 51.5084

##

## Step: AIC=8.85

## grade ~ la\_stud + sex + nonwhite + first.gen + finaid\_ever +

## credits\_entry + la\_stud:sex + la\_stud:finaid\_ever

##

## Df Sum of Sq RSS Cp

## - first.gen 1 1.134 1296.6 7.7812

## <none> 1295.5 8.8532

## - la\_stud:sex 1 3.322 1298.8 9.5709

## - la\_stud:finaid\_ever 1 5.588 1301.1 11.4249

## - nonwhite 1 6.607 1302.1 12.2582

## - credits\_entry 1 54.412 1349.9 51.3691

##

## Step: AIC=7.78

## grade ~ la\_stud + sex + nonwhite + finaid\_ever + credits\_entry +

## la\_stud:sex + la\_stud:finaid\_ever

##

## Df Sum of Sq RSS Cp

## <none> 1296.6 7.7812

## - la\_stud:sex 1 3.317 1299.9 8.4953

## - la\_stud:finaid\_ever 1 5.610 1302.2 10.3711

## - nonwhite 1 8.023 1304.6 12.3450

## - credits\_entry 1 55.584 1352.2 51.2557

##

## Call:

## lm(formula = grade ~ la\_stud + sex + nonwhite + finaid\_ever +

## credits\_entry + la\_stud:sex + la\_stud:finaid\_ever, data = LAdata[-c(6,

## 7, 10)])

##

## Coefficients:

## (Intercept) la\_stud sex

## 2.38222 0.01927 0.44604

## nonwhite finaid\_ever credits\_entry

## -0.20014 -0.40975 0.02804

## la\_stud:sex la\_stud:finaid\_ever

## -0.24186 0.31456

none = lm(grade~1, data=LAdata[-c(6, 7, 10)])

mod8 = step(none, scope = list(upper=mod4), direction='both')

## Start: AIC=299.28

## grade ~ 1

##

## Df Sum of Sq RSS AIC

## + credits\_entry 1 64.112 1347.6 251.60

## + sex 1 26.302 1385.4 281.18

## + nonwhite 1 12.199 1399.5 292.01

## + finaid\_ever 1 11.344 1400.4 292.66

## + first.gen 1 7.351 1404.4 295.70

## <none> 1411.7 299.28

## + la\_stud 1 2.234 1409.5 299.59

##

## Step: AIC=251.6

## grade ~ credits\_entry

##

## Df Sum of Sq RSS AIC

## + sex 1 20.158 1327.5 237.49

## + finaid\_ever 1 13.418 1334.2 242.90

## + nonwhite 1 11.980 1335.6 244.06

## + first.gen 1 5.499 1342.1 249.23

## <none> 1347.6 251.60

## + la\_stud 1 0.628 1347.0 253.10

## - credits\_entry 1 64.112 1411.7 299.28

##

## Step: AIC=237.49

## grade ~ credits\_entry + sex

##

## Df Sum of Sq RSS AIC

## + finaid\_ever 1 13.289 1314.2 228.73

## + nonwhite 1 11.216 1316.2 230.42

## + first.gen 1 6.512 1321.0 234.23

## <none> 1327.5 237.49

## + la\_stud 1 1.259 1326.2 238.47

## - sex 1 20.158 1347.6 251.60

## - credits\_entry 1 57.968 1385.4 281.18

##

## Step: AIC=228.73

## grade ~ credits\_entry + sex + finaid\_ever

##

## Df Sum of Sq RSS AIC

## + nonwhite 1 7.686 1306.5 224.46

## + first.gen 1 2.550 1311.6 228.66

## <none> 1314.2 228.73

## + la\_stud 1 0.876 1313.3 230.02

## - finaid\_ever 1 13.289 1327.5 237.49

## - sex 1 20.029 1334.2 242.90

## - credits\_entry 1 59.944 1374.1 274.42

##

## Step: AIC=224.46

## grade ~ credits\_entry + sex + finaid\_ever + nonwhite

##

## Df Sum of Sq RSS AIC

## <none> 1306.5 224.46

## + first.gen 1 1.157 1305.3 225.52

## + la\_stud 1 1.079 1305.4 225.58

## - nonwhite 1 7.686 1314.2 228.73

## - finaid\_ever 1 9.759 1316.2 230.42

## - sex 1 19.404 1325.9 238.22

## - credits\_entry 1 59.589 1366.1 270.14