*Report on Analysis of Poh Data*

Introduction  
The data has two separate text files which contain variables of the code of year and semester (Cohort), Year, Semester, whether students taking it again, SLT and Non SLT Group, Pass or Fail Result, number of students in that category (Count), Sex, Age, ATAR, students Type. Many of variables are self-explanatory. The aim of the analysis is to find the evidence that the SLT program affects student performance.

Methodology  
The data were checked for any missing values and categorical variable (Year) was converted to factor. Since Poh data are summed up by Cohort, we need to create new data for each student. After we replicates the Poh data we can merge with Poh2 which is demographic data for each student. Then, we change Result variable Pass and Fail to 1 and 0 for analysis. Since the mean of a response is a proportion, generalised linear model was performed on the data. Final model for the data is selected after removing the nonsignificant variable with the largest p-value until only significant variables are left. Then, we plot the model to see if there are any violation of assumptions.

Results  
The final model equation is log (odds of passing unit) = 0.8409 + 0.4006Cohort2013s1i + 0.3022Cohort2014s2i + 0.6564Cohort2017s1i + 0.5406Cohort2017s2i – 1.1414Repeati + 1.0178Groupi. Diagnostic plot (residuals vs fitted) of the model showed residuals tend to increase when logit increase, other than that, no particular violations were observed.

Discussion  
Based on the generalized linear model, student performance is influenced by SLT program.  
In other words log odds of passing unit increase if student took SLT program. According to the model, the odds of student passing the unit  
- about 2.77 times higher if student is SLT group  
- about 1.49 times higher if student took unit in 2013 S1  
- about 1.35 times higher if student took unit in 2014 S2  
- about 1.93 times higher if student took unit in 2017 S1  
- about 1.71 times higher if student took unit in 2017 S2  
- about 3.13 times lower if student retrying for the unit.

SLT students perform better than non-SLT students. Other than this, what was unusual was that 2017 was particularly easy, and students retaking the course showed particularly low performance.  
It is difficult to analyse data and make decision about hypothesis due to different levels of difficulty or teaching from semester to semester, and there are so many other variables besides SLT. If there are only SLT variables, it will be a more accurate answer to the hypothesis.

*Report on Analysis of Space Shuttle Data*

Introduction  
The data has two variables: temperature at launch time in degrees Fahrenheit (Temperature) and O-rings failure of the launch (Failure). The aim of the analysis is to find the evidence that temperature at launch time would affect the O-ring failure.

Methodology  
The data were checked for any missing values. Then, we change Failure variable Yes and No to 1 and 0 for analysis. Plot between temperature and failure (Fig. 1) was used to skim the pattern. Generalised linear model was performed on the data. Then, we plot the model to see if there are any violation of assumptions.

Results  
The plot (Fig. 1) shows possibility of relation between temperature and failure. The final model equation is log (odds of failing launch) = 10.8754 – 0.1713Temperaturei. The diagnostic plot of the model (Fig. 2), Q-Q plot does not show special pattern. However, residuals vs fitted shows that random pattern. Also, the residuals vs leverage plot shows some randomness as well. But they are all in acceptable boundary, so the model does not violate any assumption.

Discussion  
Based on the final model, the failure of space shuttle is statistically influenced by temperature at launch time. According to the model, the odds of failure decrease as temperature increase. We would expect every one-degree change makes 0.8426 multiplicative change in the odds of failure. For example, expected probability of O-rings failure at 31-degree Fahrenheit is 0.9962 (99.6%).

However, the prediction is not reliable because of small sample size. We only have 24 sample cases for space shuttle launch. Sample size less than 30 would cause bigger chance in error therefore, it needs more sample. Also, there would be more factor (variables) which can affect on failure than temperature, so it would not be accurate if we only focus on only one variable.

**Report 1 Appendix**

Sean Park

2021 9 15

Poh = **read.table**("Poh1.txt", header=T, sep = "\t", stringsAsFactors = TRUE)  
Poh**$**Year = **factor**(Poh**$**Year)

ind = **rep**(1**:NROW**(Poh), Poh**$**Count)  
NewPoh = **data.frame**(Poh**$**Cohort[ind],Poh**$**Year[ind],Poh**$**Semester[ind],Poh**$**Repeat[ind],Poh**$**Group[ind],Poh**$**Result[ind])  
**colnames**(NewPoh) = **c**("Cohort","Year", "Semester","Repeat","Group","Result")

**library**(plyr)

## Warning: package 'plyr' was built under R version 4.0.5

NewPoh**$**Result = **revalue**(NewPoh**$**Result, **c**(Pass = 1, Fail = 0))

Poh2 = **read.table**("Poh2.txt", header=T, sep = "\t", stringsAsFactors = TRUE)

FinPoh = **data.frame**(NewPoh,Poh2)

FinPoh.glm1 = **glm**(Result**~**., data= FinPoh, family = **binomial**(link=logit))

**summary**(FinPoh.glm1)

##   
## Call:  
## glm(formula = Result ~ ., family = binomial(link = logit), data = FinPoh)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2749 -1.1094 0.7379 0.8141 1.3515   
##   
## Coefficients: (7 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.159292 0.669163 0.238 0.811845   
## Cohort2013S1 0.396927 0.116135 3.418 0.000631 \*\*\*  
## Cohort2013S2 0.051531 0.128978 0.400 0.689498   
## Cohort2014S1 0.182420 0.114413 1.594 0.110847   
## Cohort2014S2 0.298846 0.127514 2.344 0.019096 \*   
## Cohort2015S1 0.104916 0.113711 0.923 0.356187   
## Cohort2015S2 0.247295 0.131718 1.877 0.060455 .   
## Cohort2016S1 0.059213 0.112711 0.525 0.599340   
## Cohort2016S2 0.097400 0.131924 0.738 0.460328   
## Cohort2017S1 0.650915 0.123740 5.260 1.44e-07 \*\*\*  
## Cohort2017S2 0.537307 0.148529 3.618 0.000297 \*\*\*  
## Cohort2018S1 0.104694 0.121273 0.863 0.387977   
## Cohort2018S2 0.070243 0.132508 0.530 0.596036   
## Year2013 NA NA NA NA   
## Year2014 NA NA NA NA   
## Year2015 NA NA NA NA   
## Year2016 NA NA NA NA   
## Year2017 NA NA NA NA   
## Year2018 NA NA NA NA   
## SemesterS2 NA NA NA NA   
## RepeatY -1.141406 0.058466 -19.522 < 2e-16 \*\*\*  
## GroupSLT 1.019576 0.114518 8.903 < 2e-16 \*\*\*  
## SexMale -0.053919 0.045242 -1.192 0.233336   
## Age -0.004249 0.010863 -0.391 0.695684   
## ATAR 0.009320 0.007367 1.265 0.205840   
## TypeInternational -0.014774 0.049876 -0.296 0.767075   
## TypeNZ -0.018821 0.103381 -0.182 0.855543   
## TypePR 0.084729 0.106085 0.799 0.424468   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 12528 on 10399 degrees of freedom  
## Residual deviance: 11991 on 10379 degrees of freedom  
## AIC: 12033  
##   
## Number of Fisher Scoring iterations: 4

**step**(FinPoh.glm1, test = "LRT")

## Start: AIC=12033.39  
## Result ~ Cohort + Year + Semester + Repeat + Group + Sex + Age +   
## ATAR + Type  
##   
##   
## Step: AIC=12033.39  
## Result ~ Cohort + Year + Repeat + Group + Sex + Age + ATAR +   
## Type  
##   
##   
## Step: AIC=12033.39  
## Result ~ Cohort + Repeat + Group + Sex + Age + ATAR + Type  
##   
## Df Deviance AIC LRT Pr(>Chi)   
## - Type 3 11992 12028 0.86 0.8353   
## - Age 1 11992 12032 0.15 0.6958   
## - Sex 1 11993 12033 1.42 0.2336   
## - ATAR 1 11993 12033 1.60 0.2058   
## <none> 11991 12033   
## - Cohort 12 12062 12080 70.38 2.725e-10 \*\*\*  
## - Group 1 12085 12125 93.68 < 2.2e-16 \*\*\*  
## - Repeat 1 12371 12411 379.33 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Step: AIC=12028.25  
## Result ~ Cohort + Repeat + Group + Sex + Age + ATAR  
##   
## Df Deviance AIC LRT Pr(>Chi)   
## - Age 1 11992 12026 0.16 0.6934   
## - Sex 1 11994 12028 1.40 0.2362   
## - ATAR 1 11994 12028 1.62 0.2024   
## <none> 11992 12028   
## - Cohort 12 12063 12075 70.37 2.726e-10 \*\*\*  
## - Group 1 12086 12120 93.40 < 2.2e-16 \*\*\*  
## - Repeat 1 12372 12406 379.98 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Step: AIC=12026.41  
## Result ~ Cohort + Repeat + Group + Sex + ATAR  
##   
## Df Deviance AIC LRT Pr(>Chi)   
## - Sex 1 11994 12026 1.40 0.236   
## - ATAR 1 11994 12026 1.63 0.202   
## <none> 11992 12026   
## - Cohort 12 12063 12073 70.38 2.715e-10 \*\*\*  
## - Group 1 12086 12118 93.37 < 2.2e-16 \*\*\*  
## - Repeat 1 12372 12404 379.88 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Step: AIC=12025.81  
## Result ~ Cohort + Repeat + Group + ATAR  
##   
## Df Deviance AIC LRT Pr(>Chi)   
## - ATAR 1 11995 12025 1.60 0.2059   
## <none> 11994 12026   
## - Cohort 12 12064 12072 70.46 2.623e-10 \*\*\*  
## - Group 1 12087 12117 93.32 < 2.2e-16 \*\*\*  
## - Repeat 1 12373 12403 379.34 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Step: AIC=12025.41  
## Result ~ Cohort + Repeat + Group  
##   
## Df Deviance AIC LRT Pr(>Chi)   
## <none> 11995 12025   
## - Cohort 12 12066 12072 70.83 2.24e-10 \*\*\*  
## - Group 1 12089 12117 93.40 < 2.2e-16 \*\*\*  
## - Repeat 1 12375 12403 379.78 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##   
## Call: glm(formula = Result ~ Cohort + Repeat + Group, family = binomial(link = logit),   
## data = FinPoh)  
##   
## Coefficients:  
## (Intercept) Cohort2013S1 Cohort2013S2 Cohort2014S1 Cohort2014S2   
## 0.84087 0.40059 0.05712 0.18785 0.30218   
## Cohort2015S1 Cohort2015S2 Cohort2016S1 Cohort2016S2 Cohort2017S1   
## 0.10932 0.25294 0.06143 0.09893 0.65636   
## Cohort2017S2 Cohort2018S1 Cohort2018S2 RepeatY GroupSLT   
## 0.54056 0.10794 0.07677 -1.14145 1.01784   
##   
## Degrees of Freedom: 10399 Total (i.e. Null); 10385 Residual  
## Null Deviance: 12530   
## Residual Deviance: 12000 AIC: 12030

FinPoh.glm2 = **glm**(Result**~** Cohort **+** Repeat **+** Group, data= FinPoh, family = **binomial**(link=logit))  
**summary**(FinPoh.glm2)

##   
## Call:  
## glm(formula = Result ~ Cohort + Repeat + Group, family = binomial(link = logit),   
## data = FinPoh)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2772 -1.0973 0.7440 0.8122 1.3074   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.84087 0.09429 8.918 < 2e-16 \*\*\*  
## Cohort2013S1 0.40059 0.11608 3.451 0.000559 \*\*\*  
## Cohort2013S2 0.05712 0.12888 0.443 0.657615   
## Cohort2014S1 0.18785 0.11434 1.643 0.100406   
## Cohort2014S2 0.30218 0.12742 2.371 0.017720 \*   
## Cohort2015S1 0.10932 0.11361 0.962 0.335936   
## Cohort2015S2 0.25294 0.13163 1.922 0.054650 .   
## Cohort2016S1 0.06143 0.11266 0.545 0.585580   
## Cohort2016S2 0.09893 0.13181 0.751 0.452895   
## Cohort2017S1 0.65636 0.12368 5.307 1.12e-07 \*\*\*  
## Cohort2017S2 0.54056 0.14841 3.642 0.000270 \*\*\*  
## Cohort2018S1 0.10794 0.12122 0.890 0.373217   
## Cohort2018S2 0.07677 0.13239 0.580 0.562021   
## RepeatY -1.14145 0.05843 -19.534 < 2e-16 \*\*\*  
## GroupSLT 1.01784 0.11449 8.890 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 12528 on 10399 degrees of freedom  
## Residual deviance: 11995 on 10385 degrees of freedom  
## AIC: 12025  
##   
## Number of Fisher Scoring iterations: 4

figure(1)

**plot**(FinPoh.glm2)

Chart, scatter chart

Description automatically generatedChart

Description automatically generatedChart, scatter chart

Description automatically generatedScatter chart

Description automatically generated with medium confidence

**Report 2 Appendix**

Sean Park

2021 9 17

library(Sleuth3)

**library**(Sleuth3)

## Warning: package 'Sleuth3' was built under R version 4.0.5

Space = ex2011  
**summary**(Space)

## Temperature Failure   
## Min. :53.00 No :17   
## 1st Qu.:67.00 Yes: 7   
## Median :70.00   
## Mean :69.92   
## 3rd Qu.:75.25   
## Max. :81.00

**library**(plyr)

## Warning: package 'plyr' was built under R version 4.0.5

Space**$**Failure = **revalue**(Space**$**Failure, **c**(Yes = 1, No = 0))

figure(1)

**plot**(Space)

Chart, histogram, scatter chart

Description automatically generated

Space.glm = **glm**(Failure**~**Temperature, data= Space, family = **binomial**(link=logit))  
**summary**(Space.glm)

##   
## Call:  
## glm(formula = Failure ~ Temperature, family = binomial(link = logit),   
## data = Space)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.2125 -0.8253 -0.4706 0.5907 2.0512   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 10.87535 5.70291 1.907 0.0565 .  
## Temperature -0.17132 0.08344 -2.053 0.0400 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 28.975 on 23 degrees of freedom  
## Residual deviance: 23.030 on 22 degrees of freedom  
## AIC: 27.03  
##   
## Number of Fisher Scoring iterations: 4

Figure 2

**plot**(Space.glm)

Chart, line chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generated

prob.fail = **exp**(10.87535 **-** 0.17132**\***31) **/** (1 **+** **exp**(10.87535-0.17132**\***31))  
prob.fail

## [1] 0.9961829