COMM 581 - Assignment #09
Poisson Regression

Name: _____Gurpal Bisra_____ **Total: 20 marks**
Due date: Monday Nov. 21, 2015 (11pm)

Background:

At a high school, the administration is trying to determine which variables explain the number of awards each student receives. Number of awards is the response variable. There are two possible explanatory variables: math final exam score (continuous) and program (categorical with 3 levels: 1 = "General", 2 = "Academic" and 3 = "Vocational"). Based on the results of your analyses, the administration would like your opinion about how to distribute 5 new scholarships among the programs.

Questions

Include captions for each of your graphs (Figure 1, Figure 2, etc.) describing what is shown in the graph and how different category levels are represented. (**1 mark**)

Please submit your R script file for this assignment as part of your assignment PDF. Clearly label each model that you used in the assignment. (**1 mark**)

1. What are your predictions (with explanation) regarding the relationships between explanatory variables and the response variable? (**1 mark**)

I would expect a sigmoidal curve between the response variable and the explanatory variables. The response variable is num.awards (i.e. the number of awards a student receives). First, I would imagine the variable num.awards plotted against the continuous explanatory variable math (i.e. the final math exam score a student achieves) would appear sigmoidal because one would be required to achieve a specific high mark in order to get an award. Another explanatory variable, suggested for predicting the number of awards a student receives, is program (i.e. categorical with 3 levels: 1 = "General", 2 = "Academic" and 3 = "Vocational"). I would imagine that students in the academic program enjoy being challenged and would likely be more competitive to achieve a higher mark. While students in “Vocational” program may achieve a good mark too, they may be there due to lowered academic performance; consequently, they may be less likely to receive as many awards when their entire set of course marks are evaluated.

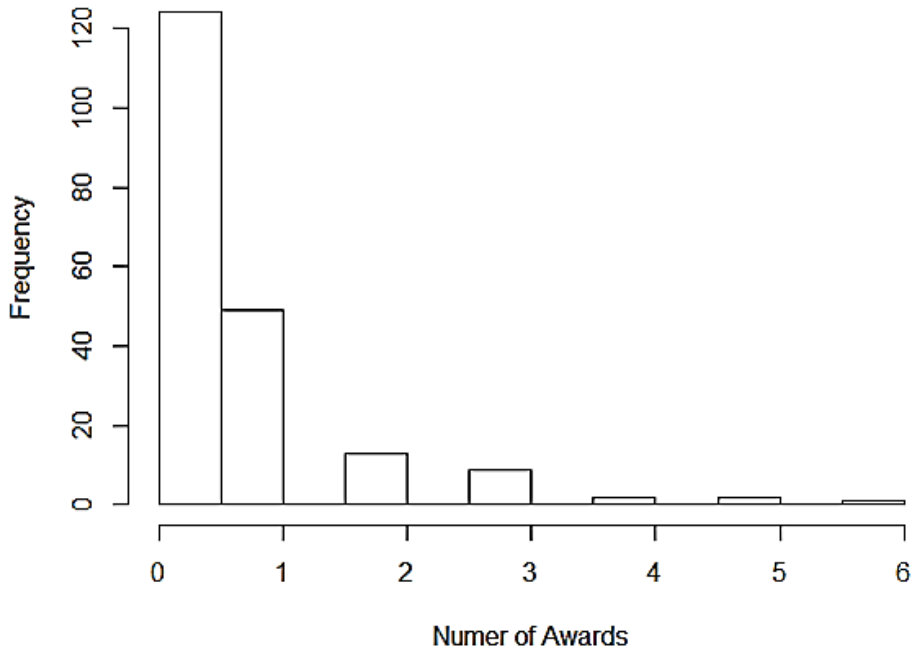
2. Import the data and change program to a factor using the factor command (change the labels to make them more informative).

```
> mydata <- read.csv("awards_data.csv", header=TRUE)
> str(mydata)
'data.frame':   200 obs. of  4 variables:
 $ id         : int  1 2 3 4 5 6 7 8 9 10 ...
 $ num.awards: int  0 0 0 1 1 0 1 0 0 1 ...
 $ prog       : int  3 3 2 2 2 2 2 2 3 1 ...
 $ math       : int  40 33 48 41 43 46 59 52 52 49 ...
> mydata$program <- as.factor(mydata$prog)
> str(mydata)
'data.frame':   200 obs. of  5 variables:
 $ id         : int  1 2 3 4 5 6 7 8 9 10 ...
 $ num.awards: int  0 0 0 1 1 0 1 0 0 1 ...
 $ prog       : int  3 3 2 2 2 2 2 2 3 1 ...
 $ math       : int  40 33 48 41 43 46 59 52 52 49 ...
 $ program    : Factor w/ 3 levels "1","2","3": 3 3 2 2 2 2 2 2 3 1 ...
```

3. Create a histogram of the responses. Do you think this follows a Poisson distribution? Why or why not? (**1 mark**)

A histogram of the responses is shown in Figure 1 below. The histogram appears be to follow a Poisson distribution since the experiment is counting the number of rare discrete instances, which occur independently of each other, a student receives an award over a period of time. The histogram is right-skewed and appears to have a $\lambda \sim 0.63$ (i.e. mean of num.awards values).

Figure 1: Histogram of the number of awards and appears to follow the Poisson distribution with $\lambda = 0.63$. The histogram appears right-skewed, unimodal and not symmetric.

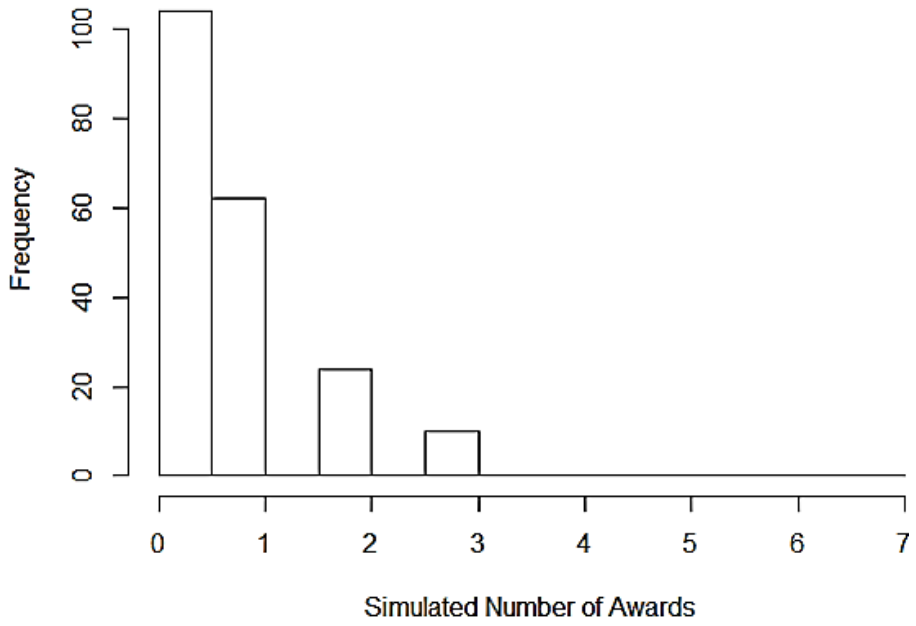


In order to verify the histogram does follow a Poisson distribution, I simulated data that follows a Poisson distribution to compare to my data. First, I determined I have 200 observations and the mean number of awards is 0.63 as follows:

```
> mean(mydata$num.awards)
[1] 0.63
> nrow(mydata)
[1] 200
```

Next, I simulated the data to see if the distribution appears Poisson. My simulated histogram is shown in Figure 2 below. I determined there are fewer zeros, more counts of students with one award, and the histogram still appears right-skewed. I did not predict a perfect Poisson distribution because the simulation produces different counts each time is run.

Figure 2: Simulated histogram of the number of awards and appears to follow the Poisson distribution with $\lambda = 0.63$. The histogram appears right-skewed, unimodal and not symmetric.



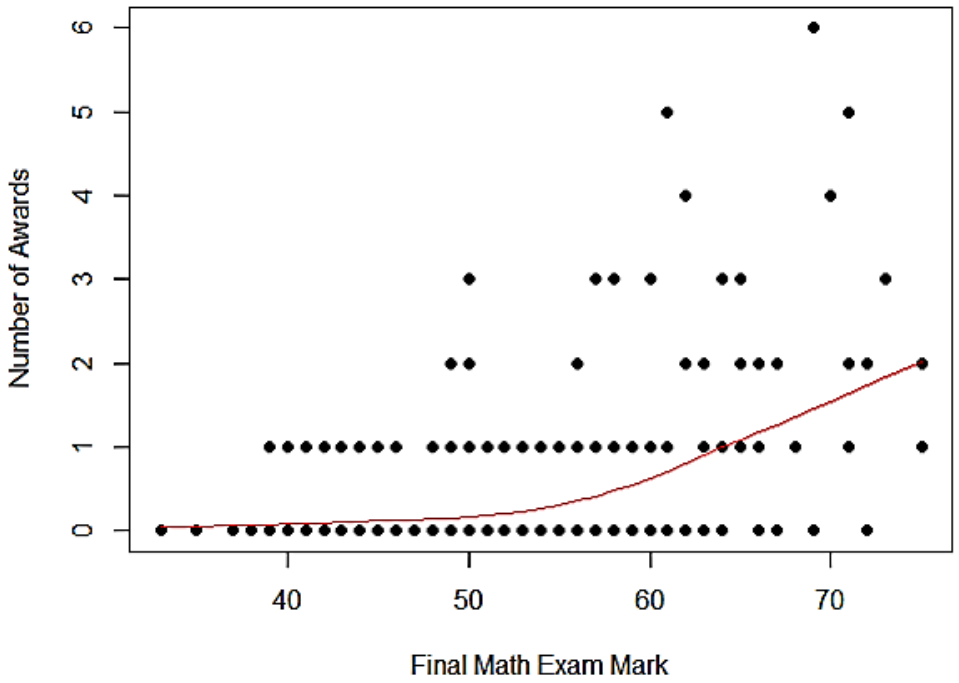
Most Importantly, I determined the variance is not equal to the mean. Therefore, the histogram might not follow a Poisson distribution even though it looks like one.

```
> var(mydata$num.awards)
[1] 1.108643
```

4. Create a scatterplot of the response variable against the continuous explanatory variable with a lowess line. Does this indicate that math score would be a good explanatory variable? (**1 mark**)

A scatterplot of the response variable, the number of awards, plotted against the continuous explanatory variable with a lowess line is depicted in Figure 3 below. While the scatterplot does logically show that higher math marks indicates more awards, it would be very difficult to fit a multiple linear regression model to fit the data well while meeting the required assumptions. Therefore, since the data is so scattered, it seems difficult to conclude the math score is a good explanatory variable from visual observation alone.

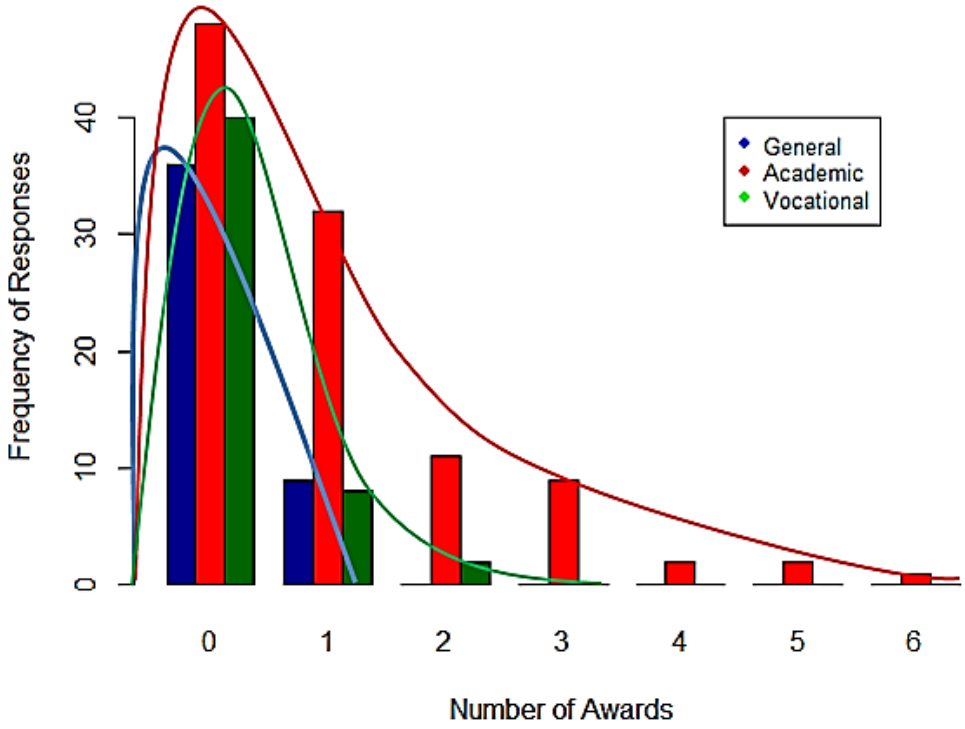
Figure 3: The number of awards given to a student plotted against students' final math exam marks. The line does not fit the data well.



5. Create a grouped bar plot that shows the frequency of different responses, grouped by program. Does it look like each category level follows a Poisson distribution? In the caption, explain how each category level is represented on the graph. (**1 mark**)

A grouped bar plot which shows the frequency of different responses, grouped by program types (i.e. "General", "Academic" and "Vocational") is shown in Figure 4 below. Each of the three program levels appears to follow a Poisson distribution if I were to draw a line through the top of each of the same colored bars.

Figure 4: The frequency of responses plotted against the number of awards appears to follow a Poisson distribution for each of the program level types: (1) "general" as indicated in blue; (2) "academic" as shown in red; and (3) "vocational" as depicted in green.



6. What is the mean of awards received per student for the different programs (general, academic, vocational)? Use the tapply function for this. Does this indicate that program would be a good explanatory variable? (**0.5 marks**)

The mean number of awards received per student for the different program is: (1) 0.2 for students in each of the "general" and "vocational" programs; and (2) 1.00 for students the "academic" program. These values do indicate the program is a good explanatory variable because students in the "academic" program enjoy being challenged and would likely be more competitive to achieve a higher mark. Such higher marks would lead to receive more awards.

```
> prog.means <- tapply(mydata$num.awards, mydata$prog, FUN = "mean")
> prog.means
   1    2    3
0.20 1.00 0.24
```

7. Fit a null model. What is the value of the intercept? Backtransform this value using the appropriate backtransformation for Poisson regression. What does this value represent? (**0.5 marks**)

My R code for fitting the null model is found directly below. I determined the value of the intercept to be -0.46204. The back-transformed value represents the mean value of the null model given the num.awards (i.e. the mean of the num.awards response variable ~ 0.63).

```
> z.null <- glm(num.awards ~ 1, data=mydata, family="poisson"(link="log"))
> summary(z.null)

Call:
glm(formula = num.awards ~ 1, family = poisson(link = "log"),
    data = mydata)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.123  -1.123  -1.123   0.429   4.038

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.46204    0.08909  -5.186 2.14e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 287.67  on 199  degrees of freedom
Residual deviance: 287.67  on 199  degrees of freedom
AIC: 465.73

Number of Fisher Scoring iterations: 6

> exp(-0.46204)
[1] 0.6299971
```

8. Fit a model with math score as the explanatory variable (Model A). Test the significance of the regression using a likelihood ratio test (include all four steps of your hypothesis test). Write the full calculation for the likelihood ratio test statistic (based on log likelihood values from R). (**1 mark**)

My R code for fitting a model with math score as an explanatory variable (Model A) is found directly below.

```
> z.1 <- glm(num.awards ~ math, data=mydata, family="poisson"(link="log"))
> summary(z.1)

Call:
glm(formula = num.awards ~ math, family = poisson(link = "log"),
    data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1853  -0.9070  -0.6001   0.3246   2.9529

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.333532   0.591261  -9.021   <2e-16 ***
math         0.086166   0.009679   8.902   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 287.67  on 199  degrees of freedom
Residual deviance: 204.02  on 198  degrees of freedom
AIC: 384.08

Number of Fisher Scoring iterations: 6
```

My testing of the significance of the regression using a likelihood ratio test is shown below:

Step 1:
H0: No difference between the two models (restriction is justified -> use simpler model)

H1: The two models have significantly different likelihoods (restriction is not justified -> use more complex model)

Step 2: Calculate the G-statistic

```
> -2*(logLik(z.null)-logLik(z.1))
'log Lik.' 83.65093 (df=1)

> # test the significance of the model (compare to null model)
> anova(z.1, z.null, test="Chi")
Analysis of Deviance Table

Model 1: num.awards ~ math
Model 2: num.awards ~ 1
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       198     204.02
2       199     287.67 -1  -83.651 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Step 3: Compare the Chi-statistic

p-value = 2.2e-16 < a = 0.05

$X^2_{1,1-0.05}$ = 3.84 < G = 83.65093

Step 4: Therefore, the regression is significant. I will reject the null hypothesis and so go with the more complex model which includes math marks as an explanatory variable.

9. Fit a model with math score and program as explanatory variables (Model B); include the interaction term. Test the significance of the regression using a likelihood ratio test. Test the significance of the regression using a likelihood ratio test (include all four steps of your hypothesis test). Write the full calculation for the likelihood ratio test statistic (based on log likelihood values from R). (**1 mark**)

My R code for fitting a model with math score and program as explanatory variables (Model B) is found directly below.

```
> z.2 <- glm(num.awards ~ math + program + math*program, data=mydata, family="poisson"(link="log"))
> summary(z.2)

Call:
glm(formula = num.awards ~ math + program + math * program, family = poisson(link = "log"),
    data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2295  -0.7958  -0.5298   0.2528   2.6826

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.86179    2.49317  -1.549    0.121
math          0.04400    0.04721   0.932    0.351
program2     -0.44107    2.60299  -0.169    0.865
program3     -0.84473    2.86990  -0.294    0.768
math:program2  0.02841    0.04870   0.583    0.560
math:program3  0.02290    0.05421   0.422    0.673

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 287.67  on 199  degrees of freedom
Residual deviance: 189.10  on 194  degrees of freedom
AIC: 377.16

Number of Fisher Scoring iterations: 6
```

My testing of the significance of the regression using a likelihood ratio test is shown below:

Step 1:
H0: No difference between the two models (restriction is justified -> use simpler model)

H1: The two models have significantly different likelihoods (restriction is not justified -> use more complex model)

Step 2: Calculate the G-statistic

```
> -2*(logLik(z.null)-logLik(z.1))
'log Lik.' 83.65093 (df=1)

> # test the significance of the model (compare to null model)
> anova(z.1, z.null, test="Chi")
Analysis of Deviance Table

Model 1: num.awards ~ math
Model 2: num.awards ~ 1
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       198     204.02
2       199     287.67 -1  -83.651 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Step 3: Compare the Chi-statistic

p-value = 2.2e-16 < a = 0.05

$X^2_{1,1-0.05}$ = 3.84 < G = 83.65093

Step 4: Therefore, the regression is significant. I will reject the null hypothesis and so go with the more complex model which includes math marks as an explanatory variable.

10. Test each variable in Model B using likelihood ratio tests (include all four steps of your hypothesis test). Write the full calculation for the likelihood ratio test statistic (based on log likelihood values from R). What do you conclude about the explanatory variables in Model B? Is model B a good model? (**2 marks**)

In order to test each variable in Model B using the likelihood ratio tests, I first wrote out the various models I used:

z.1 <- glm(num.awards ~ math, data=mydata, family="poisson"(link="log"))

y = bo + b1*math (df = 1)

z.2 <- glm(num.awards ~ math + program + math*program, data=mydata, family="poisson"(link="log"))

y = bo + b1*prog.1 + b2*prog.2 + b3*math + b4*prog.1*math + b5*prog.2*math (df = 5)

	x1	x2 (i.e. dummy variables)
level 3	1	0
level 2	0	1
level 1	0	0

z.3 <- glm(num.awards ~ program, data=mydata, family="poisson"(link="log"))

y = bo + b1*prog.1 + b2*prog.2 (df = 3)

First, I compared models z.1 and z.2 to test the significance of variable program.

Step 1:

H0: No difference between the two models (restriction is justified -> use simpler model)

H1: The two models have significantly different likelihoods (restriction is not justified -> use more complex model)

Step 2: Calculate the G-statistic

```
> z.3 <- glm(num.awards ~ program, data=mydata, family="poisson"(link="log"))
> -2*(logLik(z.1)-logLik(z.2))
'log Lik.' 14.91968 (df=2)
> anova(z.2, z.1, test="Chi")
Analysis of Deviance Table

Model 1: num.awards ~ math + program + math * program
Model 2: num.awards ~ math
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       194     189.10
2       198     204.02 -4   -14.92 0.004871 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Step 3: Compare the Chi-statistic

p-value = 0.004871 < a = 0.05

$X^2_{2,1-0.05}$ = 5.99 < G = 14.91968

Step 4: Therefore, the more complex model is significantly better. I reject the null hypothesis and the so go with more complex model which includes the variable math mark in model.

Next, I compared models z.2 and z.3 to test the significance of variable math.

Step 1:

H0: No difference between the two models (restriction is justified -> use simpler model)

H1: The two models have significantly different likelihoods (restriction is not justified -> use more complex model)

Step 2: Calculate the G-statistic

```
> -2*(logLik(z.3)-logLik(z.2))
'log Lik.' 45.35836 (df=3)
> anova(z.2, z.3, test="Chi")
Analysis of Deviance Table

Model 1: num.awards ~ math + program + math * program
Model 2: num.awards ~ program
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       194     189.10
2       197     234.46 -3  -45.358 7.764e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Step 3: Compare the Chi-statistic

p-value = 7.764e-10 < a = 0.05

$X^2_{3,1-0.05}$ = 7.81 < G = 45.35836

Step 4: Therefore, the more complex model is significantly better. I reject the null hypothesis and the so go with more complex model which includes the variable program in model.

11. Fit a model with math score and program as explanatory variables excluding the interaction term (Model C). Test the significance of the interaction term using a likelihood ratio test comparing model B and C. (**1 mark**)

In order to test the model with math and program as explanatory variables, I fitted a model which excludes the interaction term (Model C) to test the significance of the interaction term.

z.4 <- glm(num.awards ~ math + program, data=mydata, family="poisson"(link="log"))

y = bo + b1*prog.1 + b2*prog.2 + b3*math (df = 3)

	x1	x2
level 3	1	0
level 2	0	1
level 1	0	0

Step 1:

H0: No difference between the two models (restriction is justified -> use simpler model)

H1: The two models have significantly different likelihoods (restriction is not justified -> use more complex model)

Step 2: Calculate the G-statistic

```
> z.4 <- glm(num.awards ~ math + program, data=mydata, family="poisson"(link="log"))
> -2*(logLik(z.4)-logLik(z.2))
'log Lik.' 0.3480014 (df=4)
> anova(z.2, z.4, test="Chi")
Analysis of Deviance Table

Model 1: num.awards ~ math + program + math * program
Model 2: num.awards ~ math + program
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       194     189.10
2       196     189.45 -2   -0.348   0.8403
```

Step 3: Compare the Chi-statistic

p-value = 0.8503 > a = 0.05

$X^2_{4,1-0.05}$ = 9.49 > G = 0.3480014

Step 4: Therefore, the more complex model is significantly better. I reject the null hypothesis and the so go with more complex model which includes the variable math mark in model.

12. Write the equation of the final model with the values of the co-efficients. (**1 mark**)

The final model, which contains three degrees of freedom, with values of the coefficients is:

$$\ln(\widehat{num.awards}) = -5.24712 + 0.07015*\text{math} + 1.08386*x_2 + 0.36981*x_1$$

where the categorical variables takes on the following 2 dummy variables:

	x1	x2
level 3 ("Vocational") =	1	0
level 2 ("Academic") =	0	1
level 1 ("General") =	0	0

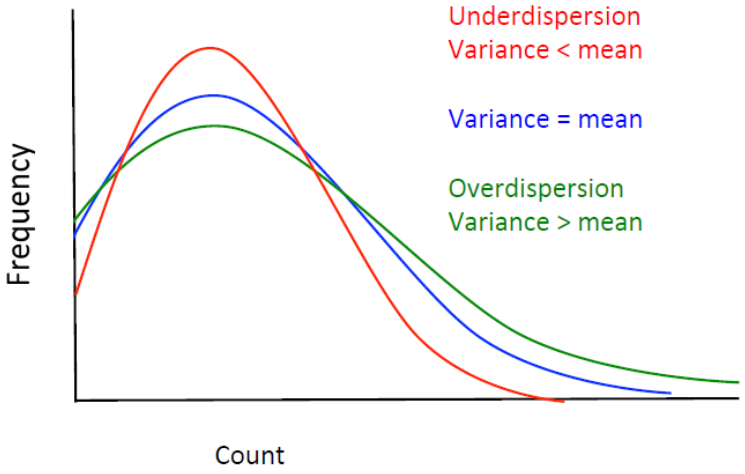
These coefficients were obtained by using the summary command on my final model. Interestingly, the p-value for the "Vocational" level for the program variable is greater than 0.05. Consequently, the program levels of "Vocational" and "General" are not statistically different.

```
> summary(z.4)

Call:
glm(formula = num.awards ~ math + program, family = poisson(link = "log"),
    data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2043  -0.8436  -0.5106   0.2558   2.6796

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.24712    0.65845  -7.969 1.60e-15 ***
math         0.07015    0.01060   6.619 3.63e-11 ***
program2     1.08386    0.35825   3.025  0.00248 **
program3     0.36981    0.44107   0.838  0.40179
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 287.67  on 199  degrees of freedom
Residual deviance: 189.45  on 196  degrees of freedom
AIC: 373.5

Number of Fisher Scoring iterations: 6
```

13. Define the terms overdispersion and underdispersion (use graphs or sketches to supplement your explanation). (**1 mark**)

The Poisson distribution, used to model count data, is unique in that the mean is equal to the variance. This means that changing one value cannot occur independently of the other. Overdispersion occurs when there is a presence of greater variability (statistical dispersion) in a data set than would be expected based on a given statistical model. More specifically, overdispersion occurs when the variance is greater than the mean Residual Deviance/residual df > 1.5. When this occurs, a negative binomial distribution results. In contrast, underdispersion means there is less variation in the data than predicted. This can be detected when Residual Deviance/residual df values are much less than 1. Overdispersion and underdispersion can be visually observed in Figure 5.

Figure 5: Graph of overdispersion, underdispersion, and their relation to the variance of a Poisson distriubution.



14. Calculate the residual deviance / degrees of freedom. Do you see any evidence for overdispersion or underdispersion in your final model? (**1 mark**)

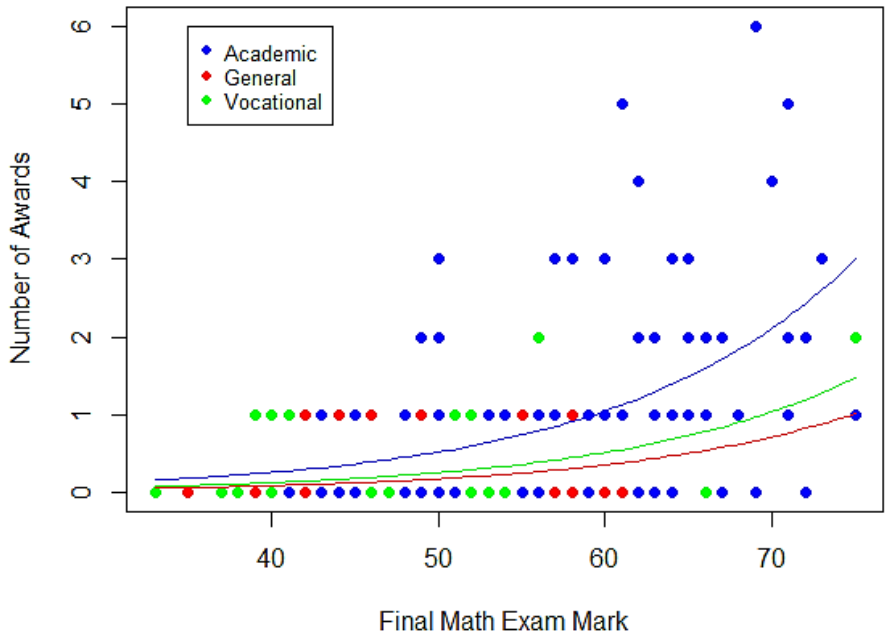
In my final model, which I have denoted as z.4, I calculated my $\frac{residual\ deviance}{degrees\ of\ freedom}$ to be 0.9665816. Since my value is less than 1.5, there is no evidence of overdispersion. Given that values displaying underdispersion have values well below 1, there is not enough evidence underdispersion is evident either.

```
> summary(z.4)

Call:
glm(formula = num.awards ~ math + program, family = poisson(link = "log"),
    data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2043  -0.8436  -0.5106   0.2558   2.6796

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.24712    0.65845  -7.969 1.60e-15 ***
math         0.07015    0.01060   6.619 3.63e-11 ***
program2     1.08386    0.35825   3.025  0.00248 **
program3     0.36981    0.44107   0.838  0.40179
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 287.67  on 199  degrees of freedom
Residual deviance: 189.45  on 196  degrees of freedom
AIC: 373.5

Number of Fisher Scoring iterations: 6

> residual.deviance.df = 189.45/196
> residual.deviance.df
[1] 0.9665816
```

15. Create a scatterplot of the response variable against the math score with program shown in different colors. Add the lines (in the matching color) of the model fit. (**1 mark**)

A scatterplot of the response variable, the number of awards, against the math score with program levels shown in different colors (i.e. "General", "Academic" and "Vocational") is shown in Figure 6 below. Lines of the model fit, in corresponding matching colors, and a legend were added.

Figure 6: The number of awards plotted against students' final math exam marks.



16. What are your observations from the graph? How do these observations relate to the coefficients from the model? (**2 marks**)

When we have an intercept only model and back transform the intercept, we get the mean value for the group "number of awards." However, this is not the case when we have more than an intercept (i.e. adding the continuous variable final exam math mark). We have found a model that is better than the intercept-only model so it is better than just predicting the mean alone.

I observe students in the academic program appear to obtain more awards than both students in the general or vocational programs. Furthermore, students in the vocational program appear to obtain more awards than students in the general program for final math exam marks below 55. The coefficients for the model were obtained by using the summary command on my final model. Interestingly, the p-value for the "Vocational" level for the program variable is greater than 0.05. Consequently, the program levels of "Vocational" and "General" are not statistically different. The coefficients indicate the following: (1) the incidence rate for the program level "Academic" is 1.08 times higher than the rate for program level "General"; (2) the incidence rate for the program level "Vocational" is 0.37 times higher than the rate for program level "General"; and (3) the percent change in the rate of awards given increases by 7.015% for every increase in the final math exam mark.

```
> summary(z.4)

Call:
glm(formula = num.awards ~ math + program, family = poisson(link = "log"),
    data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2043  -0.8436  -0.5106   0.2558   2.6796

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.24712    0.65845  -7.969 1.60e-15 ***
math         0.07015    0.01060   6.619 3.63e-11 ***
program2     1.08386    0.35825   3.025  0.00248 **
program3     0.36981    0.44107   0.838  0.40179
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 287.67  on 199  degrees of freedom
Residual deviance: 189.45  on 196  degrees of freedom
AIC: 373.5

Number of Fisher Scoring iterations: 6
```

17. The school is wondering about adding 5 new scholarships. Give your opinion about how the scholarships should be distributed among the programs. Discuss: how the program(s) you chose would benefit the most from the additional scholarships, and which criteria should be used for determining who receives the scholarship. **Justify your opinion based on the results of your analyses and how you define the primary objective when giving scholarships**. (**2 marks**)

I believe scholarships should go to individuals who exhibit financial need, high academic marks, and will pursue studies that the awards can help fund. For instance, a student who achieves > 70 as a final math exam mark appears to already receive a number of awards. Such awards might already provide adequate financial assistance to pursue further schooling. This may make them have a smaller financial need than students who obtain < 70 as their final exam math mark. Based on Figure 6 above, vocational and general program students generally receive marks < 60. On average, they receive fewer awards, which may be financial scholarships, and may consequently exhibit higher financial needs. Next, I believe the awards should be allocated based on the program type as follows: (1) 2 awards should be given to students in the "General" program; (2) 2 awards to students in the "Vocational" program; and (3) 1 award given to a student in the "Academic" program. This would ensure not all five awards end up going to students who already receive several awards (i.e. students in the "Academic" program). Also, one's final math exam mark might not be the most important factor to consider. For instance, someone who wants to pursue psychology or fine arts may not require a high math mark to be successful in such programs. Therefore, it is important to consider the student's preferred area of future studies and their highest marks in related course. Thus, the awards should be allocated to students with the highest academic marks related to their future area of study, given they have an unmet financial need, based the aforementioned program considerations.

R Code:

```
1  ##################################
2  # ASSIGNMENT 9 - POISSON REGRESSION
3  # Instructor: Martha Essak
4  # Gurpal Bisra
5  # Student Number: 69295061
6  # Nov. 21, 2016
7  ##################################
8
9  # 1. What are your predictions (with explanation) regarding the relationships between explanato
   ry variables and the response variable? (1 mark)
10
11 #-------------------------------------------------------------------------------------------
12
13 ######################
14 # IMPORT DATA
15 ######################
16
17 # 2. Import the data and change program to a factor using the factor command (change the labels
   to make them more informative).
18
19 mydata <- read.csv("awards_data.csv", header=TRUE)
20
21 str(mydata)
22 # > str(mydata)
23 # 'data.frame': 200 obs. of  4 variables:
24 # $ id         : int  1 2 3 4 5 6 7 8 9 10 ...
25 # $ num.awards: int  0 0 0 1 1 0 1 0 0 1 ...
26 # $ prog       : int  3 3 2 2 2 2 2 2 3 1 ...
27 # $ math       : int  40 33 48 41 43 46 59 52 52 49 ...
28
29 mydata$program <- as.factor(mydata$prog)
30 str(mydata)
31 # 'data.frame': 200 obs. of  5 variables:
32 # $ id         : int  1 2 3 4 5 6 7 8 9 10 ...
33 # $ num.awards: int  0 0 0 1 1 0 1 0 0 1 ...
34 # $ prog       : int  3 3 2 2 2 2 2 2 3 1 ...
35 # $ math       : int  40 33 48 41 43 46 59 52 52 49 ...
36 # $ program    : Factor w/ 3 levels "1","2","3": 3 3 2 2 2 2 2 2 3 1 ...
37
38 #-------------------------------------------------------------------------------------------
39
40 # 3. Create a histogram of the responses. Do you think this follows a Poisson distribution? Why
   or why not? (1 mark)
41
42 # response variable
43 hist(mydata$num.awards, main = "Histogram of Awards Data", xlab="Number of Awards", ylab
   ="Frequency")
44
45 # Do you think the data follows a Poisson distribution?
46
47 # You could simulate data that follows a Poisson distribution to check.
48 mean(mydata$num.awards) # 0.63
49 nrow(mydata) # 200
50 # have 200 observations
51
52 # now, simulate data to see if histogram is Poisson
53 poisson.sim.data <- rpois(200, 0.63)
54 hist(poisson.sim.data, breaks=seq(0, 7, by=0.5), ylim=c(0,110), main = "Histogram of Awards
```

```r

# now, fewer zeros
# more values in middle
# not perfect poisson distribution, was randomly generated (changes each time)
# gives idea of what our data should look like
# now, looks right-skewed, but variance might be larger than mean so use goodness of fit

# For our data, is the mean equal to the variance?
var(mydata$num.awards) # 1.108643

#-----------------------------------------------------------------------------------------------

#####################
# DATA VISUALIZATION
#####################

# 4. Create a scatterplot of the response variable against the continuous explanatory variable with a lowess line. Does this indicate that math score would be a good explanatory variable? (1 mark)

# What is the potential relationship between carapace width and satellite males?
plot(num.awards ~ math, data=mydata, pch=16, main = "Number of Awards Vs. Final Math Exam Mark", xlab="Final Math Exam Mark", ylab="Number of Awards")
lines(lowess(mydata$math, mydata$num.awards, delta = 0.1), col = "red")
# higher math marks indicate higher number of awards
# since the data is scattered, it seems as though it's hard to conclude the math score is a good explanatory variable
# it would be difficult to create a MLR as the assumptions would be hard to meet

#-----------------------------------------------------------------------------------------------

# 5. Create a grouped bar plot that shows the frequency of different responses, grouped by program. Does it look like each category level follows a Poisson distribution? In the caption, explain how each category level is represented on the graph. (1 mark)

# Plot the response variable against program (a categorical variable)

barplot(table(mydata$program, mydata$num.awards), beside=TRUE, col=c("darkblue","red", "darkgreen"), main = "Number of Students Vs. Number of Awards", xlab="Number of Awards", ylab ="Frequency of Responses")
legend(20, 40, legend=c("General", "Academic", "Vocational"),
       col=c("Blue", "Red", "Green"), pch=16:16, cex=0.8)
# yes, the response variable against the programs (i.e. 1, 2, 3) do appear as Poission distributions.

#-----------------------------------------------------------------------------------------------

# 6. What is the mean of awards received per student for the different programs (general, academic, vocational)? Use the tapply function for this. Does this indicate that program would be a good explanatory variable? (0.5 marks)

tapply(mydata$num.awards, mydata$program, mean)
# 1    2    3
# 0.20 1.00 0.24

prog.means <- tapply(mydata$num.awards, mydata$prog, FUN = "mean")
```

```r
prog.means

# yes, it would be a good explanatory variable because academic is expected to be the highest

#-----------------------------------------------------------------------------------------------

# 7. Fit a null model. What is the value of the intercept? Backtransform this value using the
appropriate backtransformation for Poisson regression. What does this value represent? (0.5
marks)

###############
# MODELS
###############

# null model
mean(mydata$num.awards) # 0.63

z.null <- glm(num.awards ~ 1, data=mydata, family="poisson"(link="log"))
summary(z.null)

# intercept = -0.46204
exp(-0.46204) # 0.6299971

# the value represents the mean value of the null model given the num.awards
# the vaule is ~ the mean of num.awards

#-----------------------------------------------------------------------------------------------

# 8. Fit a model with math score as the explanatory variable (Model A). Test the significance
of the regression using a likelihood ratio test (include all four steps of your hypothesis test
). Write the full calculation for the likelihood ratio test statistic (based on log likelihood
values from R). (1 mark)

# create a model using math score
plot(num.awards ~ math, data=mydata, pch=16)
lines(lowess(mydata$math, mydata$num.awards))

z.1 <- glm(num.awards ~ math, data=mydata, family="poisson"(link="log"))
summary(z.1)
# here, see the residual variance and dfs****

# test the significance of the model (compare to null model)
anova(z.1, z.null, test="Chi")
# p = 2.2e-16
# significant, so better than null model, so candidate of x possible

# Step 1:
# H0: No difference between the two models (restriction is justified -> use simpler model)
# H1: The two models have significantly different likelihoods (restriction is not justified ->
use more complex model)

# Step 2: Calculate the G-statistic
-2*(logLik(z.null)-logLik(z.1))
# 'log Lik.' 83.65093 (df=1)

# Step 3: Compare the Chi-statistic
```

```
# p-value = 2.2e-16 < a = 0.05
# X^1(1,1-0.05) = 3.84 < G = 83.65093

# Step 4: Therefore, the regression is significant. Iw ill reject the null hypothesis and so go with the more complex model which includes math marks as an explanatory variable.

#-----------------------------------------------------------------------------------------------

# 9. Fit a model with math score and program as explanatory variables (Model B); include the interaction term. Test the significance of the regression using a likelihood ratio test. Test the significance of the regression using a likelihood ratio test (include all four steps of your hypothesis test). Write the full calculation for the likelihood ratio test statistic (based on log likelihood values from R). (1 mark)

z.2 <- glm(num.awards ~ math + program + math*program, data=mydata, family="poisson"(link="log"))
summary(z.2)
# here, see the residual variance and dfs****

# test the significance of the model (compare to null model)
anova(z.2, z.null, test="Chi")
# p = 2.2e-16
# significant, so better than null model, so candidate of x possible

# Step 1: Hypothesis
# H0: No difference between the two models (restriction is justified -> use simpler model)
# H1: The two models have significantly different likelihoods (restriction is not justified -> use more complex model)

# Step 2: Calculate the G-statistic
-2*(logLik(z.null)-logLik(z.2))
# 'log Lik.' 98.57062 (df=1)

# Step 3: Compare the Chi-statistic
# p-value = 2.2e-16 < a = 0.05
# X^1(3,1-0.05) = 7.81 < G = 98.57062

# Step 4: Therefore, the regression is significant. Iw ill reject the null hypothesis and so go with the more complex model which includes math marks, one's program type, and the interaction between math and one's program type as the explanatory variable.s

#-----------------------------------------------------------------------------------------------

# 10. Test each variable in Model B using likelihood ratio tests (include all four steps of your hypothesis test). Write the full calculation for the likelihood ratio test statistic (based on log likelihood values from R). What do you conclude about the explanatory variables in Model B? Is model B a good model? (2 marks)

# z.1 <- glm(num.awards ~ math, data=mydata, family="poisson"(link="log"))
# y = bo + b1*math  (df = 1)
# z.2 <- glm(num.awards ~ math + program + math*program, data=mydata, family="poisson"(link="log"))
# y = bo + b1*prog.1 + b2*prog.2 + b3*math + b4*prog.1*math + b5*prog.2*math    (df = 5)
#           x1 x2
# level 3   1  0
# level 2   0  1
# level 1   0  0
```

```
z.3 <- glm(num.awards ~ program, data=mydata, family="poisson"(link="log"))
# y = bo + b1*prog.1 + b2*prog.2    (df = 3)

### COMPARING MODELS 1 AND 2

# HO: No difference between the two models (restriction is justified so use simpler model)
# H1: The two models have significantly different likelihoods (restriction is not justified os
use more complex model)

# Step 2: Calculate G-statistic
-2*(logLik(z.1)-logLik(z.2))
# 'log Lik.' 14.91968

# Step 3: Compare to Chi-statistic
anova(z.2, z.1, test="Chi")
# p-value = 0.004871) < (a = 0.05)
# (X^2,4,1-a) = 9.49 < 14.91968

# Step 4: Therefore, the more complex model is significantly better. I reject the null
hypothesis and the so go with more complex model which includes the variable math mark in model
.

### COMPARING MODELS 3 AND 2

# Step 1: Hypothesis
# HO: No difference between the two models (restriction is justified so use simpler model)
# H1: The two models have significantly different likelihoods (restriction is not justified so
use more complex model)

# Step 2: Calculate G-statistic
-2*(logLik(z.3)-logLik(z.2))
# 'log Lik.' 45.35836 (df=3)
# the df carried over so you cannot predict the df value, so ignore this df
# still just count the number of terms you drop

# Step 3: Compare to Chi-statistic
anova(z.2, z.3, test="Chi")
# p-value = 0.004871) < (a = 0.05)
# (X^2,3,1-a) = 7.81 < 45.35836

# Step 4: Therefore, the more complex model is significantly better. I reject the null
hypothesis and the so go with more complex model which includes the variable program in model.

#-------------------------------------------------------------------------------------------

# 11. Fit a model with math score and program as explanatory variables excluding the interactio
n term (Model C). Test the significance of the interaction term using a likelihood ratio test
comparing model B and C. (1 mark)

z.4 <- glm(num.awards ~ math + program, data=mydata, family="poisson"(link="log"))
# y = bo + b1*prog.1 + b2*prog.2 + b3*math (df = 3)
#           x1 x2
# level 3   1  0
# level 2   0  1
# level 1   0  0

### COMPARING MODELS 4 AND 2

```

```
# Step 1: Hypothesis
# H0: No difference between the two models (restriction is justified so use simpler model)
# H1: The two models have significantly different likelihoods (restriction is not justified so use more complex model)

# Step 2: Calculate G-statistic
-2*(logLik(z.4)-logLik(z.2))
# 'log Lik.' 0.3480014 (df=4)
# the df carried over so you cannot predict the df value, so ignore this df
# still just count the number of terms you drop

# Step 3: Compare to Chi-statistic
anova(z.2, z.4, test="Chi")
# p-value = 0.8403) > (a = 0.05)
# (X^1,4,1-a) = 3.84 > 0.3480014

# Step 4: Therefore, the less complex model is significantly better. I fail to reject the null hypothesis and the so go with less complex model which includes the variables math mark and program in model but not the interaction term.

#-----------------------------------------------------------------------------------------------

# 12. Write the equation of the final model with the values of the co-efficients. (1 mark)

# num.awards = bo + b1*prog.1 + b2*prog.2 + b3*math (df = 3)
#           x1 x2
# level 3   1  0
# level 2   0  1
# level 1   0  0
summary(z.4)
# Coefficients:
# Estimate Std. Error z value Pr(>|z|)
#    (Intercept) -5.24712    0.65845  -7.969 1.60e-15 ***
#    math         0.07015    0.01060   6.619 3.63e-11 ***
#    program2     1.08386    0.35825   3.025  0.00248 **
#    program3     0.36981    0.44107   0.838  0.40179

# Equation: ln(num.awards)_hat = -5.24712 + 0.07015*math + 1.08386*prog.2 + 0.36981*prog.3
# programs 1 and 3 are not statistically different

#-----------------------------------------------------------------------------------------------

# 13. Define the terms overdispersion and underdispersion (use graphs or sketches to supplement your explanation). (1 mark)

# slide 13

#-----------------------------------------------------------------------------------------------

# 14. Calculate the residual deviance / degrees of freedom. Do you see any evidence for overdispersion or underdispersion in your final model? (1 mark)

# Does the model have overdispersion?
summary(z.4)
# (Dispersion parameter for poisson family taken to be 1)
# Null deviance: 287.67  on 199  degrees of freedom
# Residual deviance: 189.45  on 196  degrees of freedom
# AIC: 373.5
```

```
residual.deviance.df = 189.45/196
residual.deviance.df
# [1] 0.9665816
# Residual deviance / df < 1.5
# therefore, our value is < 1.5 so we do not have evidence for overdispersion

# regarding underdispersion, we seek values that well below 1, here it is not low enough

#-----------------------------------------------------------------------------------------------

# 15. Create a scatterplot of the response variable against the math score with program shown
in different colors. Add the lines (in the matching color) of the model fit. (1 mark)

############################
# PLOTS TO SHOW MODEL FIT
############################

# PLOT WITH MODEL FIT

xnew <- seq(min(mydata$math), max(mydata$math), length.out = 100)
xnew

# bind rows of 3 levels which must be data frames
xnew.2 <- rbind(as.data.frame(xnew), as.data.frame(xnew), as.data.frame(xnew))

# now assign program number to each section
programs <- as.data.frame(c(rep(1, 100), rep(2,100), rep(3,100)))

# now column bind
new.data <- cbind(xnew.2, programs)
# now give fancy column names
names(new.data) <- c("math", "program")
new.data$program <- as.factor(new.data$program)

ynew <- predict(z.4, data.frame(new.data), type="response")
ynew

new.data.2 <- cbind(new.data, ynew)

plot(num.awards ~ math, data=mydata, pch=16)

# Subset the new data into the different colors
mydata.col.1 <- subset(new.data.2, new.data.2$program == "1")
mydata.col.2 <- subset(new.data.2, new.data.2$program == "2")
mydata.col.3 <- subset(new.data.2, new.data.2$program == "3")

# Plot the data and model fit with color coding
plot(num.awards ~ math, data=mydata, pch=16, col = c("red", "blue", "green")[as.factor(mydata$p
rogram)], main = "Number of Awards Vs. Final Math Exam Mark", xlab="Final Math Exam Mark", ylab
="Number of Awards")
legend(35, 6, legend=c("Academic", "General", "Vocational"),
       col=c("Blue", "Red", "Green"), pch=16:16, cex=0.8)
lines(mydata.col.1$math, mydata.col.1$ynew, lty=1, col="red")
lines(mydata.col.2$math, mydata.col.2$ynew, lty=1, col="blue")
lines(mydata.col.3$math, mydata.col.3$ynew, lty=1, col="green")

#-----------------------------------------------------------------------------------------------

# 16. What are your observations from the graph? How do these observations relate to the co
-efficients from the model? (2 marks)

#-----------------------------------------------------------------------------------------------

# 17. The school is wondering about adding 5 new scholarships. Give your opinion about how the
scholarships should be distributed among the programs. Discuss: how the program(s) you chose
would benefit the most from the additional scholarships, and which criteria should be used for
determining who receives the scholarship. Justify your opinion based on the results of your
analyses and how you define the primary objective when giving scholarships. (2 marks)
```