Analysis of Award Allocation for High School Students

McKade Thomas

12/16/2020

Abstract

Number of Scholastic Awards Given to Students Based on Program Type and Math Score

At a certain high school, students were given awards throughout the year based on scholastic achievement. Our statistical findings suggest that of the three programs types these students were enrolled in, academic students received the most awards, implicating a potential bias towards awarding these students more frequently than other programs. Methods used to obtain these results included analyzing a random sample of 200 students who were asked to report their program, number of awards received, and score on a particular math assessment. Using several Poisson regression models with varying numbers of model coefficients, term significance as well as analysis of deviance was compared for each model. The results of the modeling showed that both math score and program type had a significant effect on predicting number of awards, though no interaction term between these two was found to be significant in any of the models. Of the three programs, academic enrollment yielded the highest number of awards based on the sample.

Table of Contents

Introduction

At a certain high school, the school board is interested in knowing what factors were affecting the number of awards their students would receive. They randomly sampled 200 students and included explanatory variables of the program the student was in as well as a score they received on a generalized math assessment. The question of intent is as follows: are the factors of program and math score important in predicting the number of awards students receive ("Mathematics and Statistics Help (Mash)" 2020)?

Statistically, the following hypotheses, first informally then formally, will be assessed to answer this question (using an alpha level of .05):

Null Hypotheses:

- There is no significant difference in mean number of awards received among students in the three programs or based on math score. There is also no significant interaction between program type and math score for mean number of awards.
- Ho (program): Mu1 = Mu2 = Mu3
- Ho (math score): Mu1 = Mu2 = . . . = MuN
- Ho (program X math score): Program does not depend on math score (and conversely)

Alternative Hypotheses:

- There is a significant difference in mean number of awards received by students based on their program type and/or their math score. Or, and in addition, there is a significant interaction effect between math score and program type for the mean number of awards received by students at this high school.
- Ha (program): Not (Mu1 = Mu2 = Mu3)
- Ha (math score): Not (Mu1 = Mu2 = . . . = MuN)
- Ha (program X math score): Program does depend on math score (and/or conversely)

The goal of this report is to establish the findings of these hypotheses and produce prediction metrics for number of awards based on the explanatory variables, with comments on the effectiveness of these predictions.

Explanation and Exploration of the Data

The methods used to analyze this problem include a Poisson analysis using generalized linear models as well as analysis of variance analyzing both main and interaction effects of these predictor variables. As previously mentioned, an alpha level of 0.05 will be used in determining the significance of either the main or interaction effects of the explanatory variables on the response: number of awards earned.

In gathering the data, 200 students were randomly sampled with three primary factors being recorded for each student: the number of awards which is a quantitative response variable, the program the student is in separated into three categories (General, Academic,

or Vocational) which serves as a categorical explanatory variable, and the score they received on a particular math test which serves as a quantitative explanatory variable.

In exploring the data, it appears that all students who received more than 3 awards were enrolled in an academic program (Figure 1). Further, it appears that students that were enrolled in the Academic program seemed to receive more awards in general than the other two programs. It is also important to note that all students who received more than three awards had math scores higher than 60, and a trend can possibly be depicted of higher math scores equating to more awards received (Figure 2).

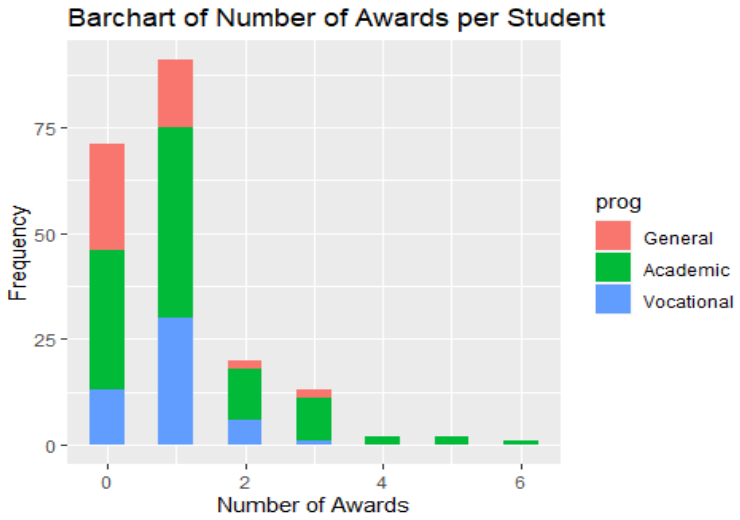

Figure 1: Number of awards received per student of a sample of 200 students at a certain high school based on program enrollment.

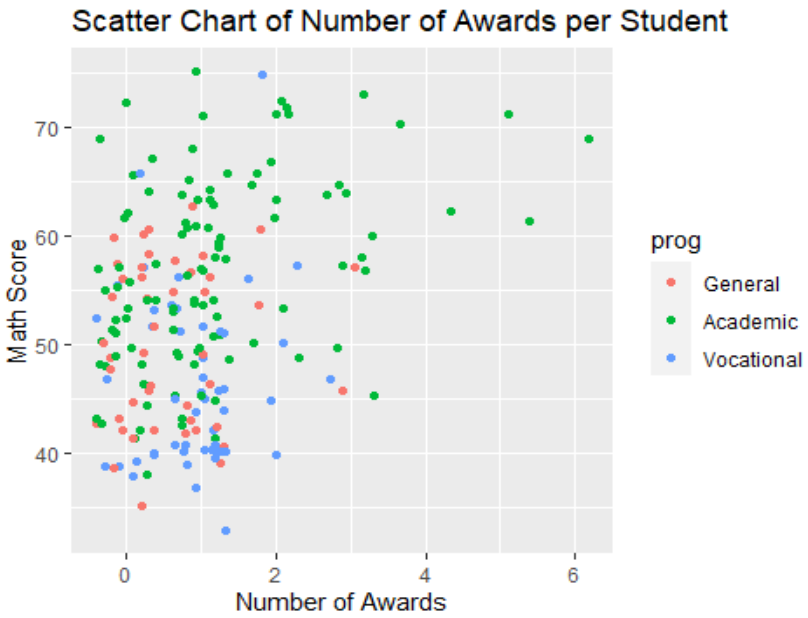

Figure 2: Number of awards received per student of a sample of 200 students at a certain high school based on math score on a particular assessment and program type.

These important features in the data lead us to believe that there may be an interaction between math scores and program enrollment in producing the most awards received. This would lead to the inference that perhaps students enrolled in a particular program also tend to score higher on this particular math assessment. In viewing this interaction, it does appear that students who scored well on the highest on the math assessment were also usually enrolled in the academic program, the program that also seemingly accounted for students who received the highest number of awards among the three programs (Figure 3).

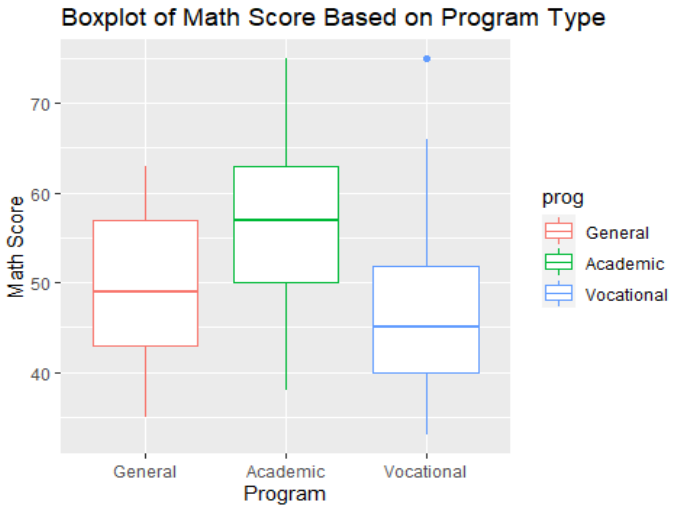

Figure 3: Math Scores on a particular assessment from a sample of 200 students at a certain high school based on their program type.

Methodology for Analysis

A common type of analysis for count data (which is what our response variable number of awards is for this dataset) is Poisson regression, which will help us know which of our explanatory variables, program and math score, is the best predictor. It should also be noted that Poisson regression is best for events that occur less often, like number of accidents at a certain intersection in a year, as opposed to events occurring more often thus having a larger mean which may follow a normal distribution more closely. This makes our response variable ideal for a Poisson regression since it can occur only a few times a year for each student. Assumptions that we must make about our data in order to fit a model such as this for Poisson regression include the following and will be addressed in a later section of the report (Glen, n.d.):

- The response variable is count data
- All counts are positive
- The mean and the variance of the counts are equal
- Each observation is independent

The method we will use for Poisson regression will involve creating a Poisson generalized linear model with a natural log link function, which forces all predictor values to be positive, a great trait for count data such as number of awards (Lillis, n.d.). After fitting a generalized linear model to our data, we will be able to examine the effect that each of our explanatory variables may have in predicting number of awards students will receive.

Modeling the Data

As aforementioned, the model chosen for analysis of the data is a Poisson regression model fit with a generalized linear model with natural log link function, a very common type of method used for count data. This model was selected because all assumptions appear to have been satisfied for using this method and our data is count data with low occurrences:

- The assumptions of our response variable being positive count data is met with number of awards being used as the y-variable.
- The mean of number of awards was 0.97 with variance 1.09. Because these values are very similar, we will assume that the condition of equal mean and variance has been met*.
- Because the data was collected randomly, we will assume that there is independence between each observation in the data set.

With all assumptions satisfied for performing the evaluation, we can now examine the details, effectiveness, and implications of the model. Because it was discovered in the examination of the data that the academic program seems to be producing both the highest number of awards and math scores, the data was re-leveled so that each coefficient in the model is compared to it.

**Note: Had there been concern about mean and variance for the count data not being equivalent, we could have performed a transformation on the data. Because a square root and log transformation did not produce a mean and variance that were closer in equivalence, no transformation was performed. The choice to exclude a transformation was also based on a decision to allow for the easiest of interpretations of the results for the school board, so that changes to both programs and math instruction could be made efficiently.*

Results

Two models were used that both compare each level of program and math score to the mean number of awards received by those in the academic program, which based on the exploration of the data was concluded to produce both the greatest number of awards and highest math scores. Thus, each coefficient is determined to be significant based on how it performs compared to the factor of being in the academic program ("POISSON Regression | R Data Analysis Examples," n.d.).

The first model included main effects for both explanatory variables, math score and program, without an interaction effect (see Table 1 and 2). Based on a plot of the residuals, the model seems to have performed an analysis properly given our underlying assumptions

mentioned previously, though we do have some concern about values predicted on the higher end of number of awards, perhaps due to high variance (see Figure 4). The results suggest that being in an academic program and having a high math score will both significantly improve the number of awards a student will receive (at an $\alpha = .05$). Enrollment in a general program also proved to be significant, though it had less of an effect than math score or academic programs.

Factor	**Model Coefficient (Back Transformed)**	**Std. Error**	**P-Value**
Program: Academic	0.14	0.50	<0.001
Program: General	0.64	0.23	0.04
Program: Vocational	1.12	0.19	0.57
Math Score	1.04	0.01	<0.001

Table 1: Summary of the Poisson generalized linear model with main effects math score and program type used as explanatory variables for the response variable number of awards given to students of a sample of 200 from a certain high school.

Factor	**DF**	**Deviance**	**Residual DF**	**Residual Deviance**
NULL			199	228.83
Program	2	12,733	197	216.10
Math Score	1	18.053	196	198.05

Table 2: Analysis of Deviance of the Poisson generalized linear model with main effects math score and program type used as explanatory variables for the response variable number of awards given to students of a sample of 200 from a certain high school.

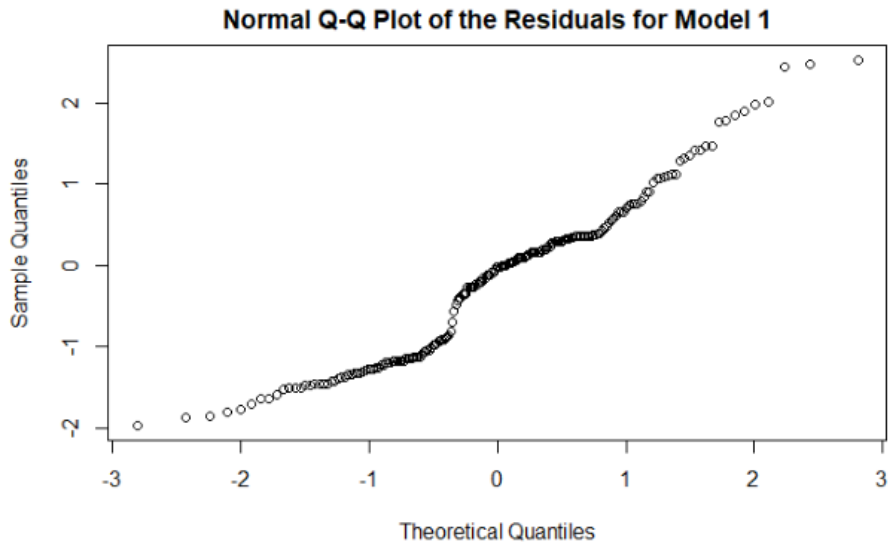

Figure 4: A plot of the residuals for a Poisson regression model with main effects of math score and program type for the response variable number of awards from a sample of 200 students at a particular high school.

Using a second model that included all main and interaction effects between the two explanatory variables, none of the interaction effects had significance in the model while both math and being in an academic program still came back significant, suggesting there may not be a significant interaction between the two explanatory variables, when compared with the effect of solely being in an academic program (see Table 3 and 4). Based on a plot of the residuals, the model also seems to have performed an analysis properly given our underlying assumptions mentioned previously, though concerns exist similar to those of the first model (see Figure 5).

Factor	**Model Coefficient (Back Transformed)**	**Std. Error**	**P-Value**
Program: Academic	0.07	0.65	<0.001
Program: General	2.89	1.53	0.49
Program: Vocational	7.56	1.08	0.06
Math Score	1.05	0.01	<0.001
General:Math Interaction	0.97	0.03	0.34
Vocational:Math Interaction	0.96	0.02	0.08

Table 3: Summary of the Poisson generalized linear model with main effects math score and program type and all interaction effects, used as explanatory variables for the response variable number of awards given to students of a sample of 200 from a certain high school.

Factor	DF	Deviance	Residual DF	Residual Deviance
NULL			199	228.83
Program	2	12,733	197	216.10
Math Score	1	18.053	196	198.05
Program:Math (Interaction)	2	3.691	194	194.35

Table 4: Analysis of Deviance of the Poisson generalized linear model with main effects math score and program type and all interaction effects, used as explanatory variables for the response variable number of awards given to students of a sample of 200 from a certain high school.

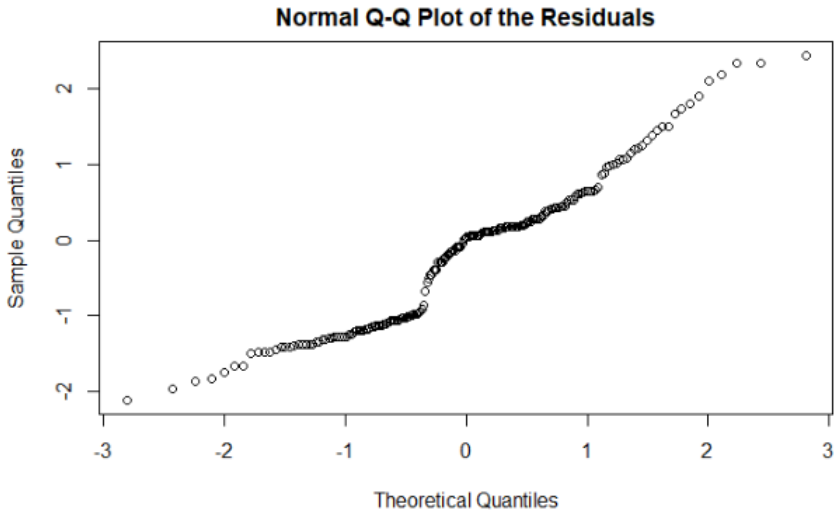

Figure 5: A plot of the residuals for a Poisson regression model with main effects of math score and program type as well as the interaction effect for these two factors for the response variable number of awards from a sample of 200 students at a particular high school.

Based on the model coefficients, it appears that both the program a student is in as well as their performance on the math evaluation have significant, positive effects on the number of awards they will receive. Specifically, it appears that a student being in the academic program will improve their chances of receiving an award more than both the other

programs as well as having a better math score. It also appears that just improving their math score also has a positive effect on number of awards received as well.

It does not appear that there is a significant interaction effect between math and program. This is due to the fact that p-value was greater than a α of .05 for all interaction effects in the second model and each time the interaction was added to the model it decreased the significance of the other terms.

It should also be noted that in analyzing all of these models and coefficients, a back-transformation was performed by exponentiating the model coefficients. This is because the Poisson generalized linear model uses a natural logarithm link function. To properly analyze the model coefficients, we need to reverse this transformation through exponentiation.

Summary of Conclusions

In considering what will help students receive more awards, the number one factor for improving number of awards received is having a student enroll in the academic program instead of the vocational or general, though general was preferred over vocational (See Figure 4). This presents several factors for consideration by the school board.

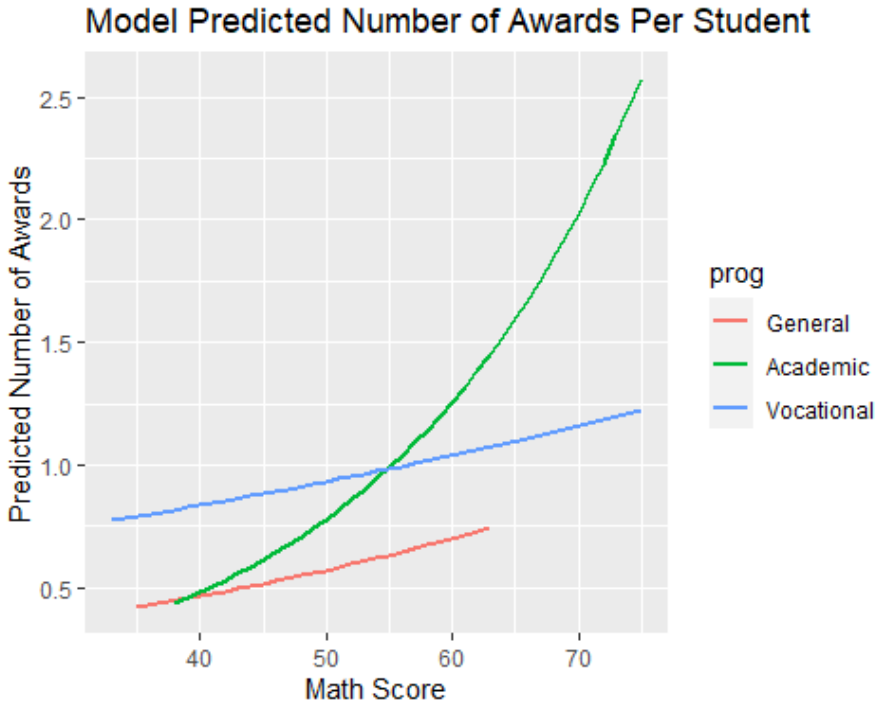

Figure 4: Predicted number of awards a student will receive at a certain high school based on a Poisson regression model with explanatory variables math score and program enrollment.

First, is it possible that bias exists towards students on an academic program track, awarding them more often than those in other programs? If the board would like to

encourage higher enrollment in other programs and award those students as well as those in the academic program, an alternative award system may need to be put into effect.

On the other hand, if awards are in fact seen as the desired result and measure of success for all students regardless of program, encouraging more students to enroll in the academic program would perhaps solicit higher numbers of awards for the student body.

In addition to the academic program yielding the highest increase in number of awards, another factor that should be considered would be to help the students improve their math score as this also had a positive effect (see Figure 4). For all program types, a higher math score led to a higher average number of awards. Thus, increased math proficiency, when considered in the context of the math assessment administered, would also increase the number of awards on average for the student body.

Both options, enrolling in an academic program and improving student math scores, would positively affect the number of awards the student is likely to receive. Thus, the school board should focus on these two solutions when looking for how to improve this metric at the high school. However, they should also be aware of bias towards giving more awards to those in academic programs, warranting a possible deflation of performance for those in the other two programs.

Improvement of the Study

Going forward, it would be nice to have a larger sample size in order to ensure that the factors that were found to be significant are indeed good predictors for the number of awards the student will receive. It would also be interesting to note what other areas of scholastic performance might have an effect such as reading comprehension or writing ability.

A question I still have for the researcher is what type of awards these are. Being able to factor in different types of awards could also add another layer to this analysis that would perhaps give more accurate predictions based on which type of award the school board would like to see given out. This could lead to an improved award system that would be unique to each program type, a system that could possibly be derived from the results of another study with award type as a factor.

Bibliography

Glen, Stephanie. n.d. “Poisson Regression.” https://www.statisticshowto.com/poisson-regression/.

Lillis, David. n.d. “Generalized Linear Models in R, Part 6: Poisson Regression for Count Variables.” https://www.theanalysisfactor.com/generalized-linear-models-in-r-part-6-poisson-regression-count-variables/.

“Mathematics and Statistics Help (Mash).” 2020. The University of Scheffield. 2020. https://www.sheffield.ac.uk/mash/statistics/datasets.

“POISSON Regression | R Data Analysis Examples.” n.d. UCLA: Statistical Consulting Group. https://stats.idre.ucla.edu/r/dae/poisson-regression/.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Xie, Yihui. 2014. “Knitr: A Comprehensive Tool for Reproducible Research in R.” In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook.

Appendix

```
library(ggplot2)
library(knitr)
library(utils)
library(rmarkdown)
# url for data: https://www.sheffield.ac.uk/mash/statistics/datasets
awards <- read.csv("C:/Users/McKade/Documents/BYU Coursework/Fall
2020/Stat435/Binomial Analysis/Awards_R.csv",header=TRUE)



#Cast char to factors
awards$prog <- factor(awards$prog, levels=1:3, labels=c("General",
"Academic", "Vocational"))

awards$id <- factor(awards$id)
str(awards)
#Examine the Effect of Program
ggplot(awards, aes(fill=prog, x=num_awards)) +
    geom_histogram(position="stack",binwidth=.5) +
    labs(title="Barchart of Number of Awards per Student",
         x = "Number of Awards", y = "Frequency")
#Examine the effect of math
a_plot <- ggplot(awards, aes(x=num_awards, y=math,color=prog))
m_p <- a_plot + geom_point(position=position_jitter())
m_p <- m_p + labs(title="Scatter Chart of Number of Awards per Student",
         x = "Number of Awards", y = "Math Score")

m_p
#Look at interaction Effect
ggplot(awards, aes(x = prog, y = math, color = prog)) + geom_boxplot() +
  labs(title="Boxplot of Math Score Based on Program Type",
x="Program",y="Math Score")
mean(awards$num_awards)
var(awards$num_awards)

mean(sqrt(awards$num_awards))
var(sqrt(awards$num_awards))

ggplot(data = awards, aes(x = prog, y = sqrt(math), color = prog)) +
geom_boxplot() + labs(title="Distribution of Math Scores from a particular
assessment on a total of 200 students seperated by which learning program
they were in (1, 2, or 3)", x="Program",y="Math Score")
## Poisson Genearlized Linear Model
#Re-level the data to compare all to program 2
awards$prog_rel <- relevel(awards$prog, ref = "Academic")

#Model with main effects only
pois_model1 <- glm(num_awards ~ prog_rel+math, data = awards, family =
"poisson")
```

```
anova(pois_model1)
summary(pois_model1)
exp(pois_model1$coefficients)
residuals1 <- resid(pois_model1)
qqnorm(residuals1)
#Model with all main effects/interactions
pois_model2 <- glm(num_awards ~ prog_rel+math+(prog_rel*math), data = awards,
family = poisson)
anova(pois_model2)
summary(pois_model2)
exp(pois_model2$coefficients)
residuals2 <- resid(pois_model2)
qqnorm(residuals2)
#Get predictions
aw.pred1 <- predict(pois_model2,type="response")

#Compare num_awards to predicted values
ggplot(awards, aes(x=math, y=aw.pred1,color=prog)) +geom_line(size=1) +
    labs(title="Model Predicted Number of Awards Per Student",
         x = "Math Score", y = "Predicted Number of Awards")
knitr::write_bib(c('knitr','stringr'), '', width = 60)
```