IDS 702

Poisson regression

Poisson distribution

- PMF: $Pr[X = x] = \dfrac{\lambda^x e^{-\lambda}}{x!}$
- Support: positive integers (includes 0)
- Parameter: Rate = λ = mean = variance (> 0)
- Count data, rates
- Sample mean is given by $\hat{\lambda} = \sum_{i=1}^{n} \dfrac{y_i}{n}$

Poisson distribution

A website has, on average, 28 visitors per hour. What is the probability that the website will have 34 visitors in an hour?

Data example

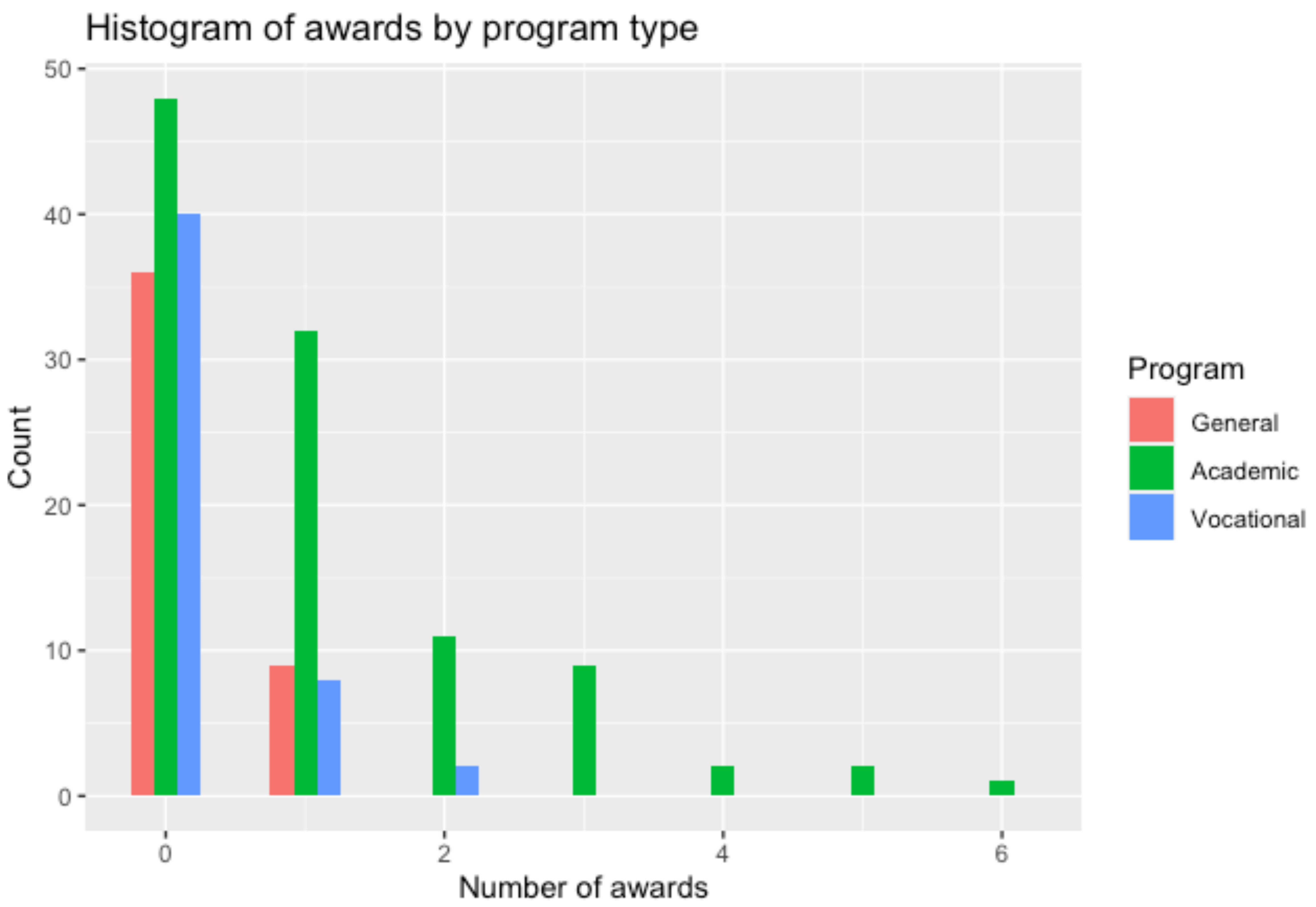
- Number of awards earned by students at a high school
- Predictors:
  - Type of program in which student is enrolled (vocational, general, academic)
  - Final math exam score

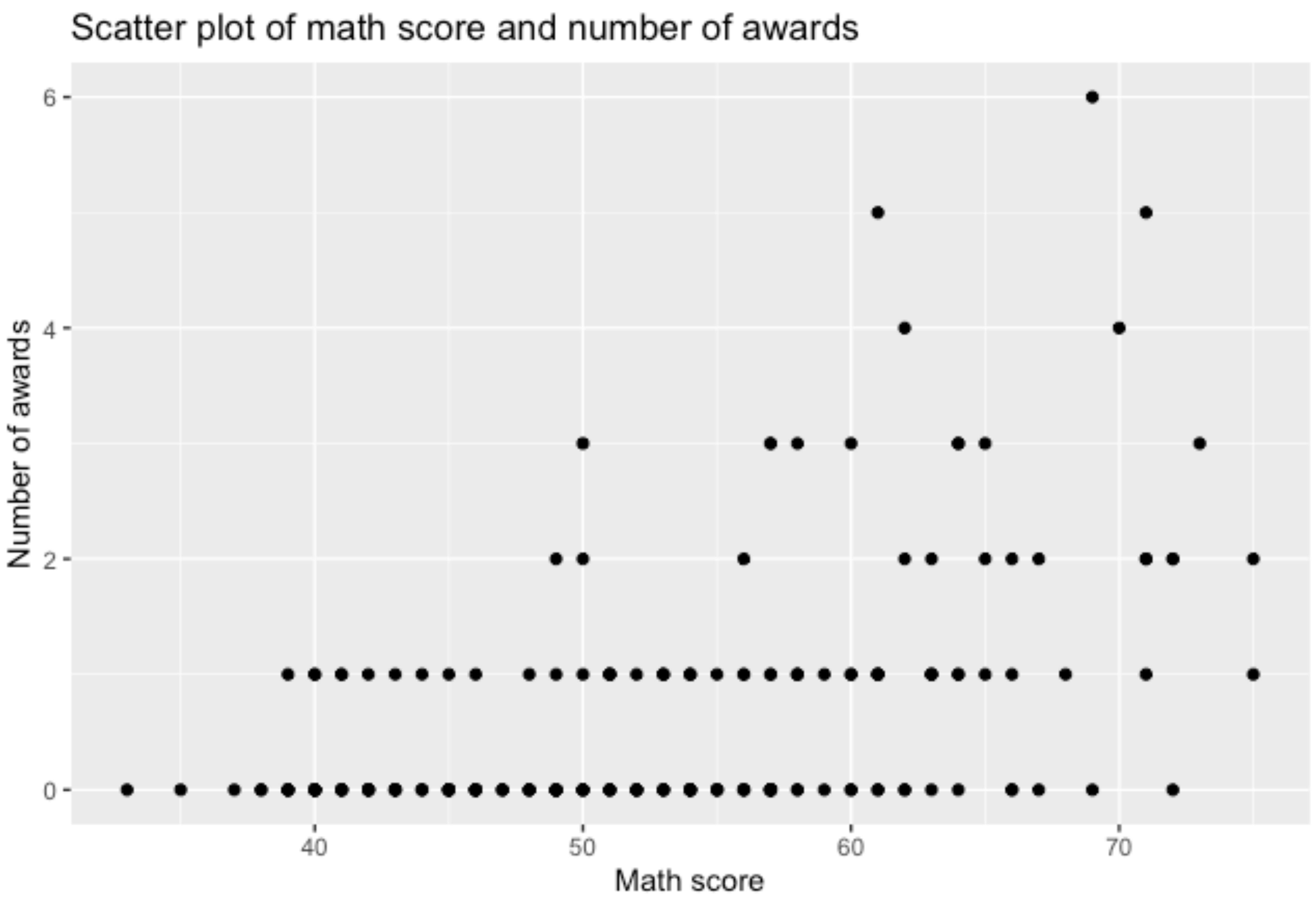
https://stats.oarc.ucla.edu/r/dae/poisson-regression/

Data example

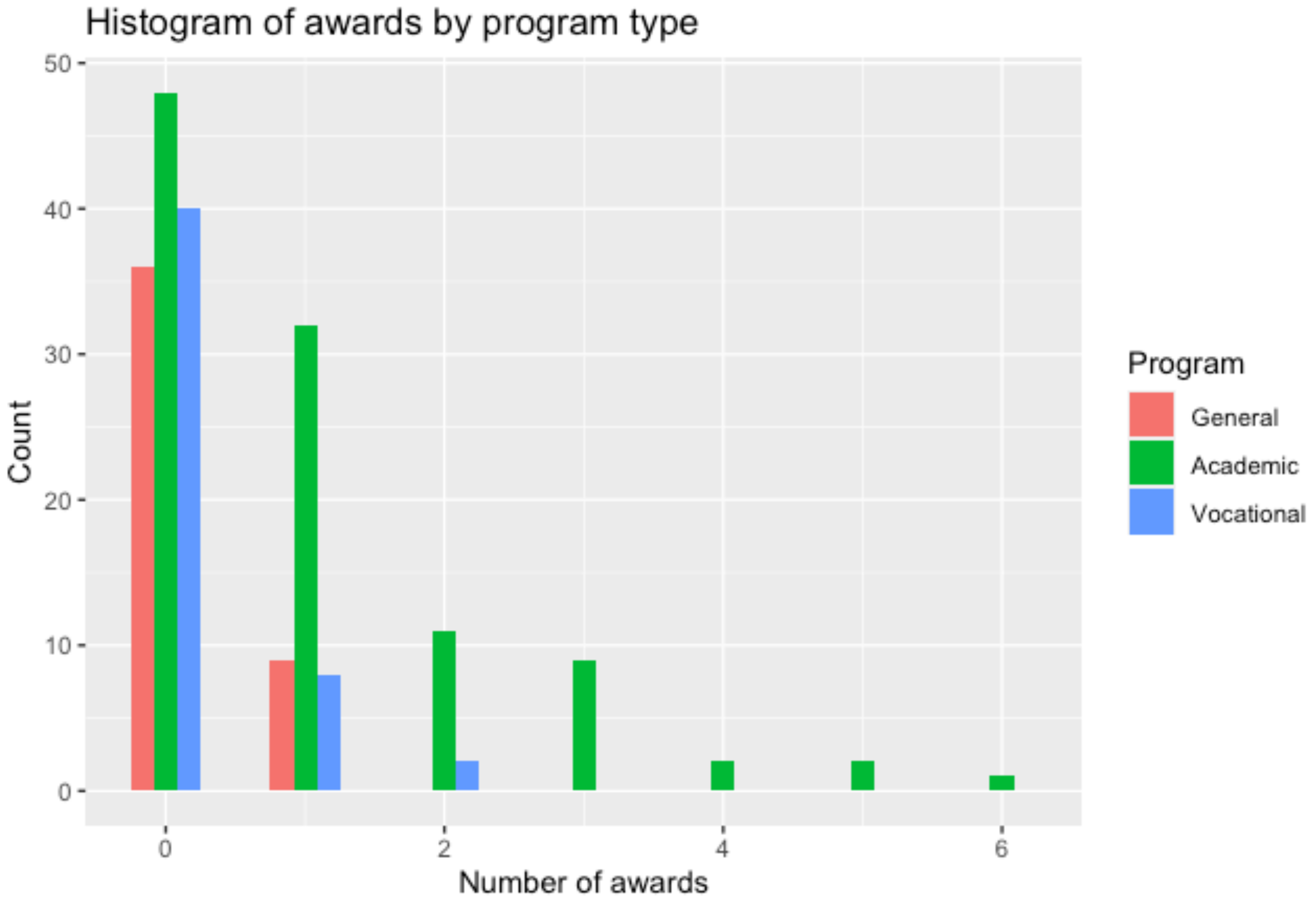
```
> pdat <- read.csv("https://stats.idre.ucla.edu/stat/data/poisson_sim.csv")
> str(pdat)
'data.frame':   200 obs. of  4 variables:
 $ id        : int  45 108 15 67 153 51 164 133 2 53 ...
 $ num_awards: int  0 0 0 0 0 0 0 0 0 0 ...
 $ prog      : int  3 1 3 3 3 1 3 3 3 3 ...
 $ math      : int  41 41 44 42 40 42 46 40 33 46 ...
> pdat$prog_fac <- factor(pdat$prog, levels=1:3, labels=c("General","Academic","Vocati
onal"))
> summary(pdat)
       id           num_awards         prog            math      
 Min.   :  1.00   Min.   :0.00   Min.   :1.000   Min.   :33.00  
 1st Qu.: 50.75   1st Qu.:0.00   1st Qu.:2.000   1st Qu.:45.00  
 Median :100.50   Median :0.00   Median :2.000   Median :52.00  
 Mean   :100.50   Mean   :0.63   Mean   :2.025   Mean   :52.65  
 3rd Qu.:150.25   3rd Qu.:1.00   3rd Qu.:2.250   3rd Qu.:59.00  
 Max.   :200.00   Max.   :6.00   Max.   :3.000   Max.   :75.00  
       prog_fac  
 General   : 45  
 Academic  :105  
 Vocational: 50  
```

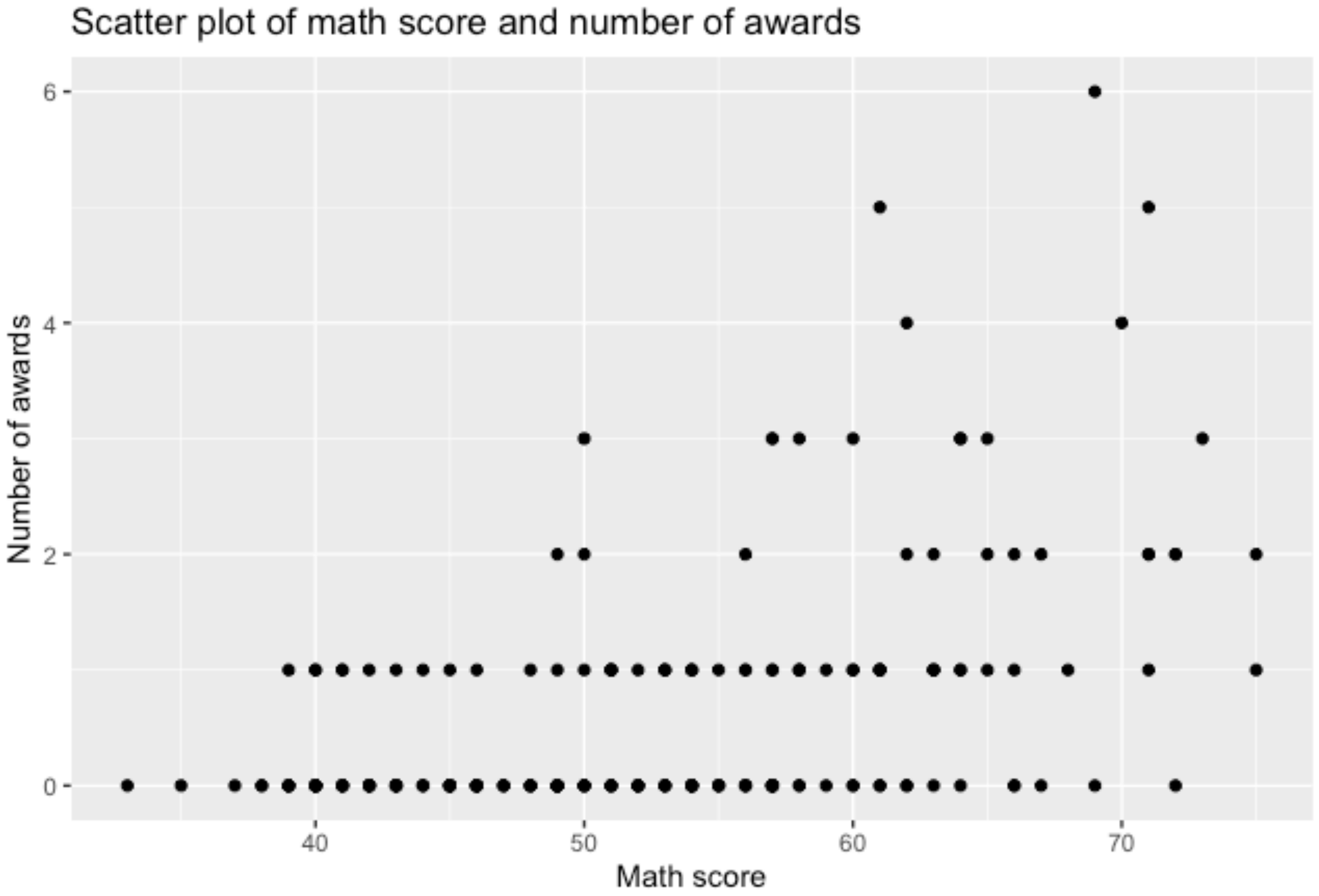
Data example





Data example





Why not just do a log transformation on the outcome and use linear regression?

Poisson regression setup

- We assume a poisson distribution for the outcome:

$$y_i | x_i \sim Poisson(\lambda_i), i = 1,...,n$$

- We need a link function that ensures $\lambda_i > 0$ at any value of x_i:
  $log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$

- Putting these pieces together, we have the poisson model:

$$log(E[y_i | x_i]) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, i = 1,...,n$$

Interpretation

- $log(E[y_i | x_i]) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$
- $\lambda_i = e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}$
- We can interpret the e^{β_j}s as multiplicative effects on the expected counts
    - Continuous x_j: the expected count of Y increases by a multiplicative factor of $e^{\hat{\beta}_j}$ when increasing x_j by one unit
    - Binary x_j: the expected count of Y increases by a multiplicative unit of $e^{\hat{\beta}_j}$ for the group with $x_j = 1$ compared to the group with $x_j = 0$

Implementation in R

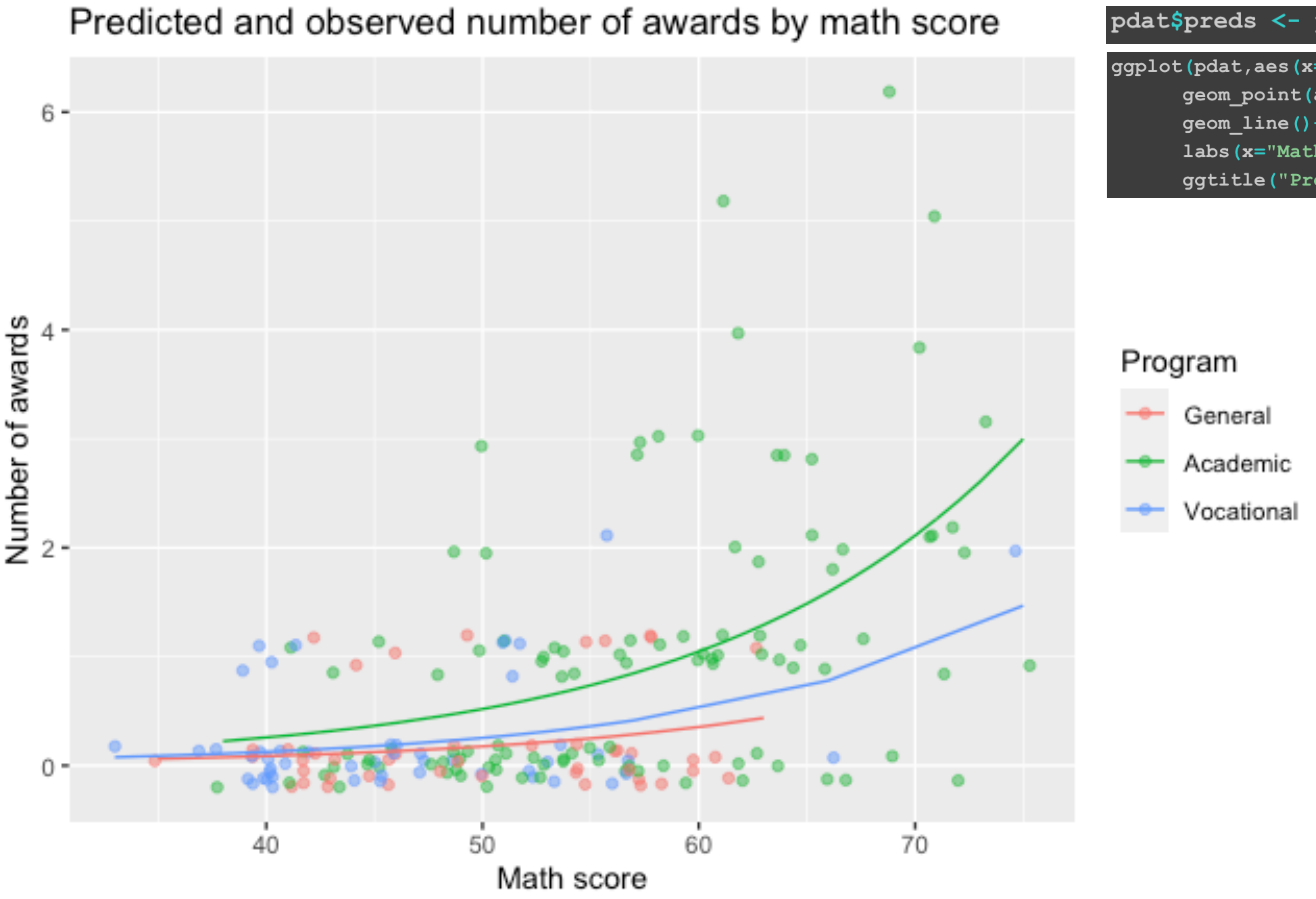
```
> poismod <- glm(num_awards~math+prog_fac,data=pdat,family="poisson")
> summary(poismod)

Call:
glm(formula = num_awards ~ math + prog_fac, family = "poisson",
    data = pdat)

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 ***
prog_facAcademic    1.08386    0.35825   3.025  0.00248 **
prog_facVocational  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
```

Interpretation

```
> cbind(exp(coef(poismod)),exp(confint(poismod)),summary(poismod)$coefficient[,4])
Waiting for profiling to be done...
                                     2.5 %    97.5 %
(Intercept)         0.00526263 0.001400301 0.0185869 1.601368e-15
math                1.07267164 1.050742170 1.0953533 3.625010e-11
prog_facAcademic    2.95606545 1.545030949 6.3972554 2.483032e-03
prog_facVocational  1.44745846 0.612501440 3.5467192 4.017859e-01
```

Controlling for program, the expected number of awards increases 7% per one unit increase in math final exam score. We are 95% confident that the true percent increase is between 5 and 9%. Math score is statistically significantly associated with number of awards ($p<.001$).

Plotting predictions and observed counts




```
pdat$preds <- predict(poismod,type="response")
```

```
ggplot(pdat,aes(x=math,y=preds,colour=prog_fac))+
      geom_point(aes(y=num_awards),alpha=.5,position=position_jitter(h=.2))+
      geom_line()+
      labs(x="Math score",y="Number of awards",colour="Program")+
      ggtitle("Predicted and observed number of awards by math score")
```

Model assessment

- Assumptions
- Model fit
- Predictions

Overdispersion

- The poisson model assumes that the mean and variance are equal. That is, as the mean number of counts grows, so does the variance.
- If this assumption does not hold, you might consider other modeling options such as negative binomial regression.

```
> library(AER)
> dispersiontest(poismod)

	Overdispersion test

data:  poismod
z = 0.53224, p-value = 0.2973
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  1.047254
```