Tutorial: GLM & AIC

BIOL 414/514

Quantitative Analysis of Vertebrate Populations

Learning Objectives:

1. Learn to run generalized linear models in R
2. Improve understanding of regression outputs
3. Develop skills in predicting expected outcomes from regression models
4. Compare models using AIC and differentiate the best model

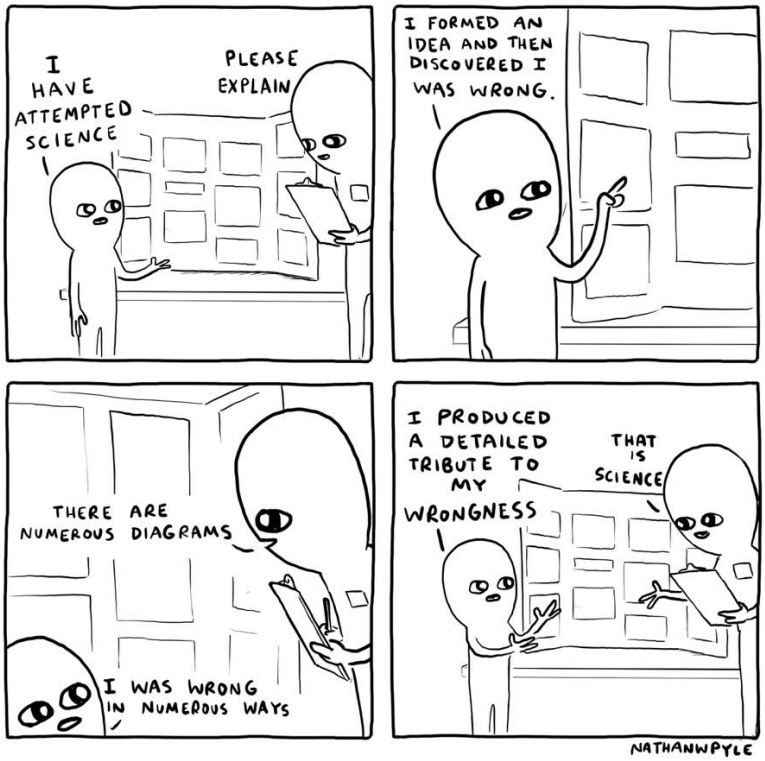


Figure 1: All models are wrong, but some are useful. - George Box

Introduction

Remember our assumptions of linear models:

- Linearity

- Homogeneity of variance (Homoscedasticity)
- Normally distributed error
- Minimal multicollinearity among independent variables (if multiple X)
- Independence of observations (no autocorrelation)

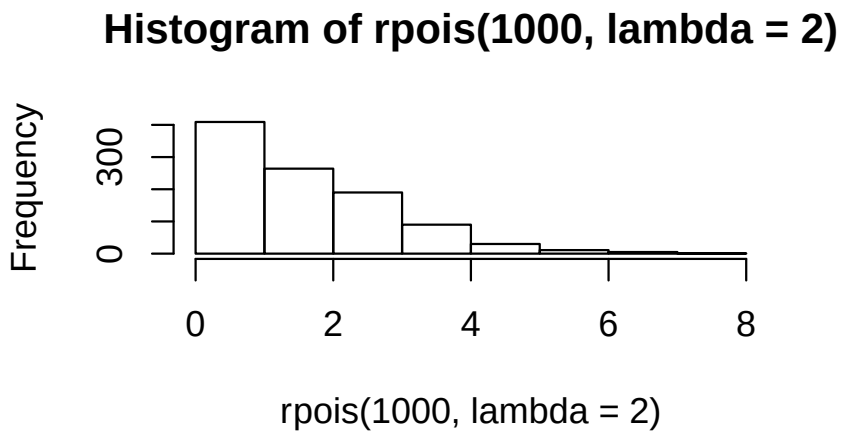
The (ecological) processes that generate different patterns and our observational processes in sampling to generate representative data of these patterns do not always meet the assumptions of linear regression. In fact, you will find that in ecology we almost always violate these assumptions. These deviations from the requirements of the model mean that we have to use different methods to analyze the data. Generalized linear models represent an approach to handling particular violations of linear models (that assume normal distributions).

For example, if we are counting the number of animals in plots across the landscape (or parasites in plant or animal tissue samples), we know that the counts must be integers and they must be zero or greater. A normal (Gaussian) distribution can range from negative infinity to infinity and is continuous so values can be any real number, not just integers. Luckily, we have the Poisson distribution at our disposal, which has the properties we are looking for. The Poisson ranges from 0 to infinity and is discrete meaning that it is comprised of whole numbers (integers).

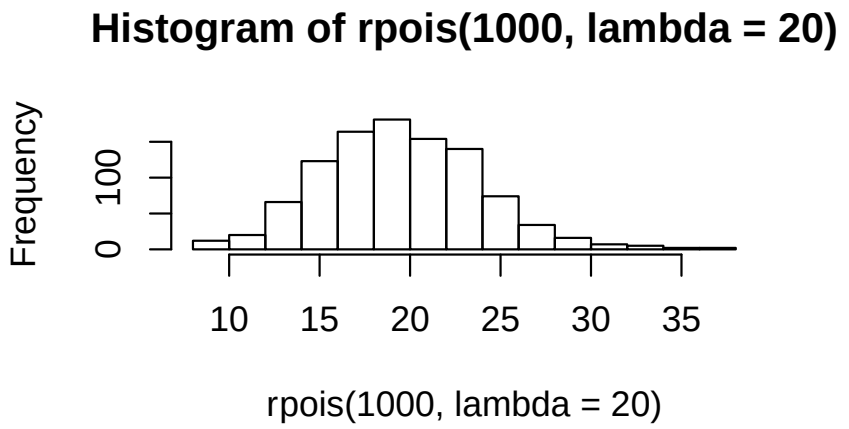
Poisson distribution. discrete. range: 0 - ∞. mean = variance

The Poisson is defined by a single parameter, λ which is the Greek letter lambda. This is both the mean and the variance. Below are some examples of Poisson distributions with different values of λ. You can run this code and see how the distribution changes with different values.

```
hist(rpois(1000, lambda = 2))
```




```
hist(rpois(1000, lambda = 20))
```



So this fixes a potential problem with our modeling distributional assumptions. However, typically when counting things our error gets larger as our counts get larger. For example, if we were counting a flock of 10 snow geese and our error rate was 10% then we'd only be off by 1 goose. However if the flock was 100 geese and we had the same error rate, we would be off by 10 geese. If the flock was 1000 geese (*go to Bombay Hook NWR in Delaware in the winter to see this*), we could over or under count by 100 geese. Therefore, when using a Poisson distribtion for a Generalized Linear Model (GLM), we typically use a log link. This link serves two purposes. It makes it so the expected error grows with the increasing counts and links the linear predictor model that can range into negative values to the Poisson distribution.

1. Distribution: $C_i \sim Poisson(\lambda_i)$
2. Link Function: $log(\lambda_i)$
3. Linear Predictor: $log(\lambda_i) = \alpha + \beta X_i$

So even if $\alpha + \beta X_i = -3$ then this is equal to the $log(\lambda_i)$ and therefore $\lambda_i = 0.0498$, which is still above 0.

GLM with Count Data

So let's try this with real data. These are real counts of brook trout across the eastern US from electrofishing.

```
trout <- read.csv("Data/Trout/regional_occupancy_data.csv",
                  stringsAsFactors = FALSE)
str(trout)
```

```
## 'data.frame':    13161 obs. of  5 variables:
##  $ featureid: int  20625575 20625640 20625683 20626130 20626158 20626179 20626219 20626320 20626437 206
##  $ species  : chr  "BKT" "BKT" "BKT" "BKT" ...
##  $ catch    : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ year_min : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ year_max : int  NA NA NA NA NA NA NA NA NA NA ...
```

These are the environmental and landscape covariates that might be useful for our regression.

```
covariates <- readRDS("Data/Trout/covariates.rds")
str(covariates)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    13142 obs. of  9 variables:
##  $ featureid     : num  20625575 20625640 20625683 20626130 20626158 ...
##  $ forest        : num  89.2 45.5 40.1 39.8 38.8 ...
##  $ zone          : chr  "upstream" "upstream" "upstream" "upstream" ...
##  $ ann_tmax_c    : num  17.1 16.9 17.1 16.8 16.7 ...
##  $ AreaSqKM      : num  28.96 26.54 4.48 201.56 186.76 ...
##  $ winter_prcp_mm: num  262 307 304 307 309 ...
##  $ spring_prcp_mm: num  314 345 342 351 352 ...
##  $ summer_prcp_mm: num  291 293 287 313 314 ...
##  $ fall_prcp_mm  : num  250 261 258 263 264 ...
##  - attr(*, "vars")=List of 1
##   ..$ : symbol featureid
```

Select covariates and join with trout data

```
library(dplyr)

df_trout <- trout %>%
  left_join(covariates) %>%
  filter(!is.na(catch))
```

`left_join()` to combine datasets based on common column names

Run GLM

Let's see if the amount of forest cover in the catchment (drainage) affects the number of observed trout.

```
glm1 <- glm(catch ~ forest, data = df_trout,
            family = poisson(link = "log"))
summary(glm1)
```

```
##
## Call:
## glm(formula = catch ~ forest, family = poisson(link = "log"),
##     data = df_trout)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
##  -9.461   -5.617   -3.927   -1.619  286.955
##
## Coefficients:
##               Estimate Std. Error z value
## (Intercept) -0.4164261  0.0131320  -31.71
## forest       0.0421767  0.0001548  272.46
##             Pr(>|z|)
## (Intercept)   <2e-16 ***
## forest        <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: 937234  on 13135  degrees of freedom
## Residual deviance: 842684  on 13134  degrees of freedom
##   (25 observations deleted due to missingness)
## AIC: 873189
##
## Number of Fisher Scoring iterations: 8
```

It appears there is a significant positiv affect of forest on trout counts. However, with 13,000+ data points it's not hard to reject the null hypothesis. The effect size is more important in this case and we can see that it's a positive moderate effect of $e^0.042$ or roughly 1 more trout per 100 m stream reach (sample distance) for every 1% increase in forest cover. Even though 0.042 is a small value, it's on the log scale and if we had a 30% increase in forest cover, it would actually add a lot of trout to a stream.

Effect size is at least as important as a *p*-value and must be put in the context of the data and species.

Model Comparison

A model with just forest cover is probably a bit overly simplistic, especially with 13,000 data points from which to draw inference. Let's try a few other models and do some model comparison using AIC_c.

vignette("AICcmodavg")
help("AICcmodavg")

```
library(AICcmodavg)

glm2 <- glm(catch ~ forest + ann_tmax_c + winter_prcp_mm,
            data = df_trout, family = poisson(link = "log"))

glm3 <- glm(catch ~ forest + ann_tmax_c + winter_prcp_mm +
              AreaSqKM * winter_prcp_mm,
```

```
            data = df_trout, family = poisson(link = "log"))

# make list of models
models <- list(glm1, glm2, glm3)

# names of models
mod_names <- c("forest",
               "forest-temp-precip",
               "Area interaction")

# makde AIC table
aictab(cand.set = models, modnames = mod_names)
```

	Modnames	K	AICc	Delta_AICc	ModelLik	AICcWt	LL	Cum.Wt
3	Area interaction	6	837212.9	0.00	1	1	-418600.5	1
2	forest-temp-precip	4	865170.0	27957.01	0	0	-432581.0	1
1	forest	2	873189.2	35976.30	0	0	-436592.6	1

What we see is that the ΔAIC_c is greater than 2 between the top model and the next best model so we can conclude that the model complex model with the interaction between area and precipitation (representing flow) is the best model so we should look at the model results.

```
summary(glm3)
```

```
##
## Call:
## glm(formula = catch ~ forest + ann_tmax_c + winter_prcp_mm +
##     AreaSqKM * winter_prcp_mm, family = poisson(link = "log"),
##     data = df_trout)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -11.890   -5.469   -3.573   -1.029  278.512
##
## Coefficients:
##                             Estimate
## (Intercept)                7.093e-01
## forest                     3.947e-02
## ann_tmax_c                 4.414e-02
## winter_prcp_mm            -5.392e-03
## AreaSqKM                  -4.591e-03
## winter_prcp_mm:AreaSqKM -1.215e-06
```

```
##                          Std. Error z value
## (Intercept)               2.365e-02  29.996
## forest                    1.564e-04 252.360
## ann_tmax_c                9.920e-04  44.491
## winter_prcp_mm            6.891e-05 -78.249
## AreaSqKM                  3.211e-04 -14.298
## winter_prcp_mm:AreaSqKM   1.341e-06  -0.906
##                          Pr(>|z|)
## (Intercept)                <2e-16 ***
## forest                     <2e-16 ***
## ann_tmax_c                 <2e-16 ***
## winter_prcp_mm             <2e-16 ***
## AreaSqKM                   <2e-16 ***
## winter_prcp_mm:AreaSqKM    0.365
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 937234  on 13135  degrees of freedom
## Residual deviance: 806700  on 13130  degrees of freedom
##   (25 observations deleted due to missingness)
## AIC: 837213
##
## Number of Fisher Scoring iterations: 9
```

DO NOT mix model selection (AIC) and null hypothesis testing (p-values)

Ignore the p-values since we did model selection and focus on the effect sizes and model fit or predicitive ability (hold out data for validation). We can expect more trout with more forest and higher temperatures but fewer with more winter rain (scouring of eggs and young of the year in the cobble) and larger drainage areas.

Logistic Regression (GLM with binomial distribution)

If we have any yes/no or presence/absence data, we can do a logistic regression, which is just another name of a GLM with a binomial distribution. We typically use a logit link with the binomial distribution to keep the range of values between zero and one.

1. Distribution: $Y \sim binomial(p, N)$
2. Link Function: $log(\frac{(p)}{1-p})$
3. Linear Predictor:$log(\frac{(p)}{1-p}) = \alpha + \beta X_i$

Let's pretend we just had presence-absence data for the trout and run a model. First, I will convert any counts > 0 to 1.

```r
df_trout$pres <- ifelse(df_trout$catch > 0, 1, df_trout$catch)

glm4 <- glm(pres ~ forest +
              ann_tmax_c +
              winter_prcp_mm +
              AreaSqKM * winter_prcp_mm,
            data = df_trout,
            family = binomial(link = "logit"))

summary(glm4)
```

```
##
## Call:
## glm(formula = pres ~ forest + ann_tmax_c + winter_prcp_mm + AreaSqKM *
##     winter_prcp_mm, family = binomial(link = "logit"), data = df_trout)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8649  -0.8208   0.3419   0.7940   5.5661
##
## Coefficients:
##                            Estimate
## (Intercept)              8.037e+00
## forest                   2.164e-02
## ann_tmax_c              -6.791e-01
## winter_prcp_mm           2.913e-04
## AreaSqKM                -3.839e-03
## winter_prcp_mm:AreaSqKM  5.801e-06
##                          Std. Error z value
## (Intercept)              2.488e-01  32.306
## forest                   1.309e-03  16.532
## ann_tmax_c               1.565e-02 -43.405
## winter_prcp_mm           6.936e-04   0.420
## AreaSqKM                 1.133e-03  -3.387
## winter_prcp_mm:AreaSqKM  4.649e-06   1.248
##                          Pr(>|z|)
## (Intercept)               < 2e-16 ***
## forest                    < 2e-16 ***
## ann_tmax_c                < 2e-16 ***
## winter_prcp_mm           0.674477
## AreaSqKM                 0.000705 ***
## winter_prcp_mm:AreaSqKM 0.212135
## ---
## Signif. codes:
```

```
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17975  on 13135  degrees of freedom
## Residual deviance: 12815  on 13130  degrees of freedom
##   (25 observations deleted due to missingness)
## AIC: 12827
##
## Number of Fisher Scoring iterations: 8
```