22s:152 Applied Linear Regression

Chapter 15 Section 2:
Poisson Regression

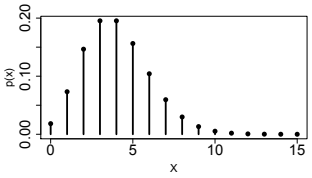
RECALL: The Poisson distribution

Let Y be distributed as a Poisson random variable with the single parameter λ.

$$P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!} \quad \text{for} \quad y \in \{0, 1, 2, 3, 4, \ldots\}$$

Y is a discrete random variable that has probability mass only on the whole numbers.

Suppose $\lambda = 4$, how is Y distributed?



If $Y \sim \text{Poisson}(\lambda)$,

$$\boldsymbol{E(Y) = \lambda} \quad \text{and} \quad \boldsymbol{V(Y) = \lambda}.$$

The mean and the variance are equal.

The variance is tied to the mean.

Suppose the mean of this Y depends on the covariate X...

For example, let Y represent the number of accidents at an intersection, and X represent a characteristic of the intersection (like number of lanes).

If $E[Y|X = 2] < E[Y|X = 4]$, then

$$V[Y|X = 2] < V[Y|X = 4].$$

Or...

if the mean increases with X,
so does the variance.

Poisson Regression

- When the response variable is a count following a Poisson distribution with a mean that depends on the covariates, we can fit a Poisson regression model.

- Poisson Regression model:

$$\ln(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$$

Or, $Y_i \sim Poisson(\lambda_i)$ where

$$\lambda_i = e^{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}}$$

and the Poisson parameter λ_i depends on the covariates of each observation (so, each observation can have its own mean).

Thus, the mean depends on the covariates, and the variance depends on the covariates.

- The Poisson regression model is another **GENERALIZED LINEAR MODEL**.

- Instead of a logit function of the Bernoulli parameter π_i (logistic regression), we use a $\log_e$ function of the Poisson parameter λ_i.

$$\lambda_i > 0 \quad \Rightarrow \quad -\infty < \ln(\lambda_i) < \infty$$

- The logit function in the logistic model and the $\log_e$ function in the Poisson model are called the *link* functions for these generalized linear models.

- In this modeling, we assume the $\ln(\lambda_i)$ is linearly related to the independent variables. And that the mean and variance are equal for a given λ_i (as we're using Poisson).

- An iterative process is used to solve the likelihood equations and get maximum likelihood estimates.

Example: Mating of elephants

There is competition for female mates between young and old male elephants.

Because male elephants continue to grow throughout their lives, older elephants are larger and tend to be more successful at mating.

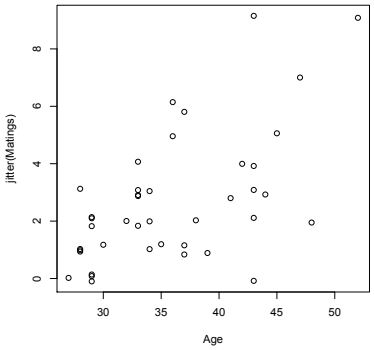
J.H. Poole, Mate Guarding, Reproductive Success and Female Choice in African Elephants, Animal Behavior 37 (1989): 842-49.

Variables:

Response: Matings (number of mates)
Predictor: Age of male elephant (years)

First, let's look at a scatterplot of the data.

```
> plot(Age,jitter(Matings))
```

It looks like the number of mates tends to be higher for older elephants, AND there seems to be more variability in the number of mates as age increases.

Elephants of age 30 have between 0 and 4 mates.
Elephants of age 45 have between 0 and 9 mates.

If the dispersion increases with the mean for a count response, then Poisson regression may be a good modeling choice.

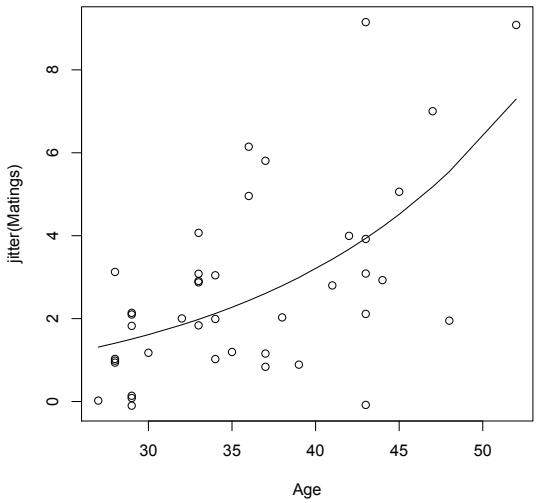
$$log(\lambda_i) = \hat{\beta}_0 + \hat{\beta}_1 x$$

```
> glm.out=glm(Matings~Age,family=poisson)

## Get and plot the fitted mean structure:
> coeff=coef(glm.out)
> coeff
(Intercept)         Age
-1.57955278  0.06859472

> xvalues=sort(Age)
> log.means=coeff[1]+coeff[2]*xvalues

## Un-log the values to get to the lambdas:
> mean.values=exp(log.means)
> lines(xvalues,mean.values)
```

The Poisson-regression model is a nonlinear model for the expected response.

What is the fitted Poisson model for an elephant of 30 years?

```
> lambda=exp(coeff[1]+coeff[2]*30)
> lambda
   1.613311
```

mean number of mates = 1.6
variance in number of mates = 1.6

What is the fitted Poisson model for an elephant of 45 years?

```
> lambda=exp(coeff[1]+coeff[2]*45)
> lambda
   4.514117
```

mean number of mates = 4.5
variance in number of mates = 4.5

We can test for significant effects as before.

```
> summary(glm.out)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.57955    0.54595  -2.893  0.00381 **
Age          0.06859    0.01378   4.979  6.4e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 75.372  on 40  degrees of freedom
Residual deviance: 51.176  on 39  degrees of freedom
AIC: 156.62

Number of Fisher Scoring iterations: 5
```

Age is a significant predictor of the number of mates for an elephant.

Since the coefficient is positive, the expected number of mates increases with *Age*.

Parameter interpretation
{one covariate, $\ln(\lambda_i) = \beta_0 + \beta_1 X_i$)}

Interpretation of β_0:
e^{β_0} is the mean of the Poisson distribution when $X = 0$.

Interpretation of β_1:
Increasing X by 1 unit, has multiplicative effect on the mean of the Poisson by e^{β_1},

$$\frac{\lambda_{(x+1)}}{\lambda_{(x)}} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = \frac{e^{\beta_0}\, e^{\beta_1 x}\, e^{\beta_1}}{e^{\beta_0}\, e^{\beta_1 x}} = e^{\beta_1}$$

$$\Rightarrow \lambda_{(x+1)} = \lambda_{(x)} e^{\beta_1}$$

– If $\beta_1 > 0$, then the expected count increases as X increases

– If $\beta_1 < 0$, then the expected count decreases as X increases.

For the elephant data,

Interpretation of β_0:
Not meaningful in the context of the data as Age=0 is not meaningful, and is out of the range of the data.

Interpretation of β_1:
An increase of 1 year in age increases the expected number of elephant mates by a multiplicative factor of $e^{0.06859} \approx 1.07$

• **Example**: Days absent at high school

School administrators study the attendance behavior of high school juniors at two schools.

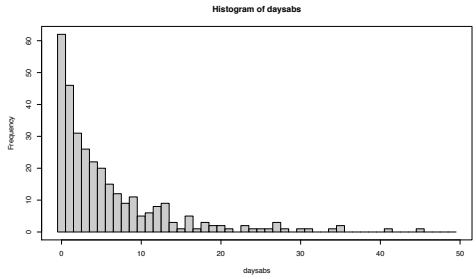
VARIABLES:

daysabs	days absent
math	standardized test scores
langarts	standardized test scores
male	gender (1=male, 0=female)

```
> attach(p)
> head(p)

    id school male      math langarts daysabs
1 1001      1    1 56.988830 42.45086       4
2 1002      1    1 37.094160 46.82059       4
3 1003      1    0 32.275460 43.56657       2
4 1004      1    0 29.056720 43.56657       3
5 1005      1    0  6.748048 27.24847       3
6 1006      1    0 61.654280 48.41482      13
```


A histogram of all counts:



A model with math, langarts, and male:

```
> glm.out.1=glm(daysabs~math+langarts+male,
                                    family=poisson)
> summary(glm.out.1)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.687666   0.072651  36.994  < 2e-16 ***
math        -0.003523   0.001821  -1.934   0.0531 .
langarts    -0.012152   0.001835  -6.623 3.52e-11 ***
male        -0.400921   0.048412  -8.281  < 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: 2409.8  on 315  degrees of freedom
Residual deviance: 2234.5  on 312  degrees of freedom
AIC: 3103.9

Number of Fisher Scoring iterations: 6
```

```
## Performing the global null hypothesis:

> glm.out.null=glm(daysabs~1, family=poisson,data=p)

> chisq.st=glm.out.null$deviance-glm.out.1$deviance
> pchisq(chisq.st,3,lower.tail=FALSE)
[1] 9.245878e-38
```

Very significant. At least one of the predictors in the model helps explain days absent.

Since *math* was not significant, we will fit the simpler model:

```
> glm.out.2=glm(daysabs~langarts+male, family=poisson)
> summary(glm.out.2)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.646976   0.069776  37.935   <2e-16 ***
langarts    -0.014670   0.001293 -11.342   <2e-16 ***
male        -0.409353   0.048219  -8.489   <2e-16 ***
```

For a student with an average language arts score, what affect does gender have?

```
> ave.lang=mean(langarts); ave.lang
[1] 50.06379

> lambda.female=exp(2.6470-0.0147*ave.lang + 0)
> lambda.female
[1] 6.760107

> lambda.male=exp(2.6470-0.0147*ave.lang + -0.4094)
> lambda.male
[1] 4.489039
```

For the average language arts student, a male is expected to miss ~ 2.3 fewer days.

As the coefficient on *langarts* is negative, there is a tendency for student with higher language arts scores to miss fewer days.

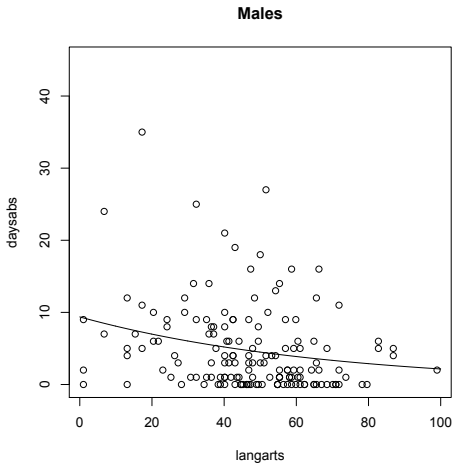
For a given sex...
a 1-unit increase in *langarts* coincides with a decrease in the mean number of days missed by a multiplicative factor of $e^{-0.0147} = 0.9854$

a 10-unit increase in *langarts* coincides with a decrease in the mean number of days missed by a multiplicative factor of $e^{-0.0147*10} = 0.8633$

We can plot the fitted curve for each sex...

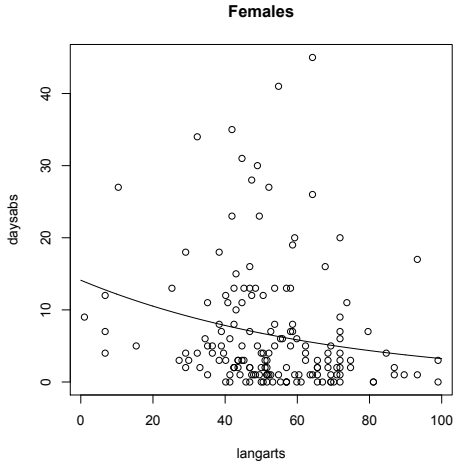

```
## Subset to only the male data points:
> male.data=p[male==1,]
> plot(male.data$daysabs~male.data$langarts,ylim=c(0,45)
       xlab="langarts",ylab="daysabs",main="Males")

## Add the fitted curve:
> xvalues=seq(0,100,.1)
> mean.curve=exp(2.6469765-0.0147*xvalues-0.4094)
> lines(xvalues,mean.curve)
```



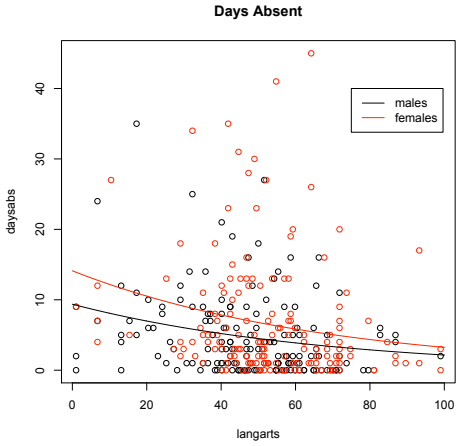

```
## Subset to only the female data points:
> female.data=p[male==0,]
> plot(female.data$daysabs~female.data$langarts,
       xlab="langarts",ylab="daysabs",main="Females")

## Add the fitted curve:
> xvalues=seq(0,100,.1)
> mean.curve=exp(2.6469765-0.0147*xvalues+0)
> lines(xvalues,mean.curve)
```



Overlaid plots:

```
> plot(male.data$daysabs~male.data$langarts,
     xlab="langarts",ylab="daysabs",main="Days Absent",
     ylim=c(0,45))
> points(female.data$daysabs~female.data$langarts,col=2)
> mean.curve.m=exp(2.6469765-0.0147*xvalues-0.4094 )
> lines(xvalues,mean.curve.m)
> mean.curve.f=exp(2.6469765-0.0147*xvalues+0 )
> lines(xvalues,mean.curve.f,col=2)
> legend(75,40,c("males","females"),col=c(1,2),lty=c(1,1)
```



In some cases, the variance may not be quite equal to the mean, but they are still related.

There could be *overdispersion* (variance is greater than the mean) or *underdispersion* (variance is less than the mean). Shown as

$$V(Y_i|\mu_i) = \phi\mu_i$$

where ϕ is the *dispersion parameter*.

If $\phi > 1$, then the variance of Y increases more rapidly than the mean.

You can fit this model using a *quasi-likelihood*.

The estimated coefficients will actually be the same as in the Poisson model, but the standard errors (and tests) will more appropriately reflect the extra variability in the data.

Example: Days Absent at high school

The original fit of the data:

```
> glm.out.1=glm(daysabs~math+langarts+male,
                                  family=poisson)
> summary(glm.out.1)
.
.
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2409.8  on 315  degrees of freedom
Residual deviance: 2234.5  on 312  degrees of freedom
AIC: 3103.9

Number of Fisher Scoring iterations: 6
```

If the mean and variance were equal, the residual deviance should be approximately equal to the *df* for error.

Resid. deviance is 2234.5 on 312 *df* for error.

Thus, it looks like we have *overdispersion*.

Re-fit the model allowing for overdispersion:

```
> glm.out.od=glm(daysabs~math+langarts+male,
                               family=quasipoisson)
> summary(glm.out.od)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.687666   0.216646  12.406  < 2e-16 ***
math        -0.003523   0.005431  -0.649  0.51701
langarts    -0.012152   0.005471  -2.221  0.02707 *
male        -0.400921   0.144365  -2.777  0.00582 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for quasipoisson family
                                taken to be 8.892352)

    Null deviance: 2409.8  on 315  degrees of freedom
Residual deviance: 2234.5  on 312  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```


The estimates are the same, but the standard errors are different (and the p-values are different).

Notice that the Dispersion parameter is 8.892352, and it was 1 before (i.e. mean equal to variance before).

References for the days absent data...

Long, J. S. 1997. Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage Publications.

Zeileis, A., Kleiber, C. and Jackman, S. Regression Models for Count Data in R.

Everitt, B. S. and Hothorn, T. A Handbook of Statistical Analyses Using R.

For the elephant data,

The original fit of the data:

```
> glm.out=glm(Matings~Age,family=poisson)
> summary(glm.out)
.
.
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 75.372  on 40  degrees of freedom
Residual deviance: 51.176  on 39  degrees of freedom
AIC: 156.62

Number of Fisher Scoring iterations: 5
```

Resid. deviance is close to the *df* for error (so, no overdispersion).

If we re-fit allowing for overdispersion (next page), it suggests that the original model with mean=variance fits reasonably well...

the standard errors don't change much,
the Dispersion parameter is 1.16

Model allowing for overdispersion:

```
> glm.out.od=glm(Matings~Age,family=quasipoisson)
> summary(glm.out.od)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.57955    0.58846  -2.684   0.0106 *
Age          0.06859    0.01485   4.619 4.13e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for quasipoisson family
                              taken to be 1.161794)

    Null deviance: 75.372  on 40  degrees of freedom
Residual deviance: 51.176  on 39  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```


- **Example**: Ear infection in swimmers

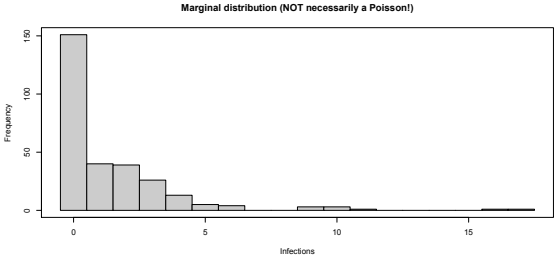
Count response: *Infections*
Number of ear infections

Categorical predictors:
Swimmer: Frequent (0) or Occasional (1)
Location: Beach (0) or Non-beach (1)

```
> attach(ear.inf)
> head(ear.inf)
  Swimmer Location Infections
1   Occas NonBeach          0
2   Occas NonBeach          0
3   Occas NonBeach          0
4   Occas NonBeach          0
5   Occas NonBeach          0
6   Occas NonBeach          0

## Switch to sum-to-zero constraints (like 2-way ANOVA):
> contrasts(Swimmer)=contr.sum(levels(Swimmer))
> contrasts(Location)=contr.sum(levels(Location))

> hist(Infections,
       breaks=seq(0,18)-.5,col="grey80")
```

We will use the 'family=poisson' option:

```
> glm.out =glm(Infections~Swimmer + Location +
                              Swimmer:Location,family=poisson)
> summary(glm.out)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.25825    0.05355   4.822 1.42e-06 ***
Swimmer1           -0.29193    0.05355  -5.451 5.00e-08 ***
Location1          -0.23438    0.05355  -4.377 1.20e-05 ***
Swimmer1:Location1  0.06893    0.05355   1.287    0.198
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 824.51  on 286  degrees of freedom
Residual deviance: 763.00  on 283  degrees of freedom
AIC: 1143.4

Number of Fisher Scoring iterations: 6
```

The model (two-way ANOVA flavor):

$$ln(\lambda_{ij}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$
$$\text{for } i = 1, 2 \text{ and } j = 1, 2$$

Overall null test:

$$H_0 : \alpha_i = \beta_j = (\alpha\beta)_{ij} = 0 \text{ for all } i, j$$

```
> chisq.stat=824.51-763.00
> pchisq(chisq.stat,3,lower.tail=FALSE)
[1] 2.796289e-13
```

We reject H_0 and conclude at least one of the terms in the model is significant.

From the previous summary output, we see that the interaction is not significant, I will refit the simpler model to continue.


```
> glm.out.2 =glm(Infections~Swimmer + Location,
                                    family=poisson)
> summary(glm.out.2)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.25505    0.05376   4.745 2.09e-06 ***
Swimmer1    -0.30652    0.05249  -5.839 5.24e-09 ***
Location1   -0.25437    0.05140  -4.948 7.49e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 824.51  on 286  degrees of freedom
Residual deviance: 764.65  on 284  degrees of freedom
AIC: 1143.0
```

Both factors are significant in the additive model.

Checking the way the factors are coded:

```
> contrasts(Swimmer)
      [,1]
Freq     1
Occas   -1
```



```
> contrasts(Location)
         [,1]
Beach       1
NonBeach   -1
```

There are 4 distinct groups (or cells), each with its own expected number of ear infections.

```
> levels(Swimmer)
[1] "Freq"  "Occas"

> levels(Location)
[1] "Beach"    "NonBeach"
```

```
> glm.out.2$coefficients
(Intercept)    Swimmer1   Location1
  0.2550506  -0.3065178  -0.2543653
```

The `Swimmer1` ('Freq') coefficient or $\hat{\alpha}_1$ is negative, so frequent swimmers tend to have fewer ear infections (and occasional swimmers tend to have more infections).

The `Location1` ('Beach') coefficient or $\hat{\beta}_1$ is negative, so beach swimmers tend to have fewer ear infections (and non-beach swimmers tend to have more infections).

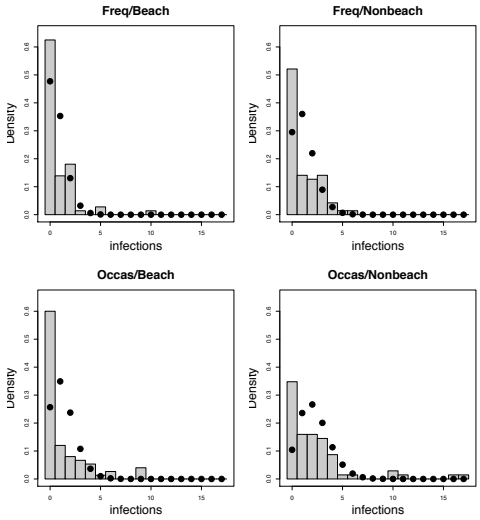
In other words, occasional non-beach swimmers tend to have the most ear infections.

(Note that this is an observational study and people with more ear infections may just choose to swim less).

Hand D.J., Daly F., Lunn A.D., McConway K.J., Ostrowski E. (1994). A Handbook of Small Data Sets. London: Chapman & Hall. Data set 328

Group	Poisson mean
Frequent Swimmers at the Beach	$\lambda = e^{(\hat{\mu}+\hat{\alpha}_1+\hat{\beta}_1)} = 0.74$
Frequent Swimmers not at the Beach	$\lambda = e^{(\hat{\mu}+\hat{\alpha}_1+\hat{\beta}_2)} = 1.22$
Occasional Swimmers at the beach	$\lambda = e^{(\hat{\mu}+\hat{\alpha}_2+\hat{\beta}_1)} = 1.36$
Occasional Swimmers not at the beach	$\lambda = e^{(\hat{\mu}+\hat{\alpha}_2+\hat{\beta}_2)} = 2.26$

When your predictors are categorical, you can look at the observed distribution of counts compared to the predicted distribution of counts based on the fitted Poisson distribution (or, similarly, the observed relative frequencies compared to the fitted probabilities) for each group separately...



In this case, though the predictors do seem to impact the distribution of ear infections, the Poisson may not be a good fit due to the large number of zeros.

In this type of situation...

... we can fit a zero-inflated Poisson (ZIP) regression.

This type of model accounts for the increased number of zeros.

MORE ON THIS IN THE NEXT SECTION OF NOTES.