Lesson 13

Clark

Today we're going to start talking about models for when our response variable is a count (integer). Recall, if we knew the number of trials and we were modeling the number of successes we could use:

However, sometimes we are in a different situation. For example, let's say we wanted to build out a model for the number of burglaries that occur in a given city. We *could* consider the number of buildings in the city as the number of trials and define a success if the building had been burgled, however this isn't typically how we would think about the data. To motivate the Poisson distribution (which will be the first GLM response variable we will consider) let's first start with a binomial density and let's assume that $p = \frac{\mu}{n}$. We can write:

And we see here after we let $n \to \infty$ we have the density function for a Poisson random variable. Therefore, we can think about the Poisson distribution as being the limiting distribution of the binommial when the number of trails increases indefinitely and μ is the expected value of the number of successes from the trials.

We can also think of μ as the rate parameter. That is, μ gives us the expected number of events we would observe in the time period we are conducting our observation.

We can write the Poisson in exponential dispersion family form and we get derive the canonical link

From here we see that it is natural to place structure on $\log(\lambda)$. We also see that for a Poisson $\phi = 1$ meaning we have a fixed variance to mean ratio. In fact, it is quite fixed.

```
library(faraway)
library(tidyverse)
data(gala)
gala_df <- gala %>%
  select(-Endemics)

gala_df$Species
```

```
##  [1]  58  31   3  25   2  18  24  10   8   2  97  93  58   5  40 347  51   2 104
## [20] 108  12  70 280 237 444  62 285  44  16  21
```

If we look at building a model for the number of species it really wouldn't make sense to build out a binomial regression model as we don't have an idea of the *number of trials* for each observation. If we did have a column for *possible number of new species* we could use that model, but that doesn't really make sense. While we *could* build out a linear regression model it might not make sense here. Why?

Let's build out a Poisson regresion model using elevation and Scruz as our predictors.

```
pois_glm <- glm(Species ~ Scruz + Elevation, data=gala_df,
                family=poisson)

summary(pois_glm)
```

```
##
## Call:
## glm(formula = Species ~ Scruz + Elevation, family = poisson,
##     data = gala_df)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.928e+00  3.627e-02  108.32   <2e-16 ***
## Scruz       -5.560e-03  4.702e-04  -11.82   <2e-16 ***
## Elevation    1.440e-03  3.175e-05   45.34   <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: 3510.7  on 29  degrees of freedom
## Residual deviance: 1654.8  on 27  degrees of freedom
## AIC: 1821.6
##
## Number of Fisher Scoring iterations: 5
```

Note that since we are using a member of the exponential dispersion family for our response variable we know that our parameters are estimated via maximum likelihood. Therefore, we can conduct our test against the null model examining deviance

```r
chi_stat <- 3510-1654
1-pchisq(chi_stat,2)
```

```
## [1] 0
```

We can test for the significance of `Scruz` through:

```r
smaller_mod <- glm(Species ~ Elevation, data=gala_df,
                   family=poisson)

anova(smaller_mod,pois_glm,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Species ~ Elevation
## Model 2: Species ~ Scruz + Elevation
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        28     1826.2
## 2        27     1654.8  1   171.37 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

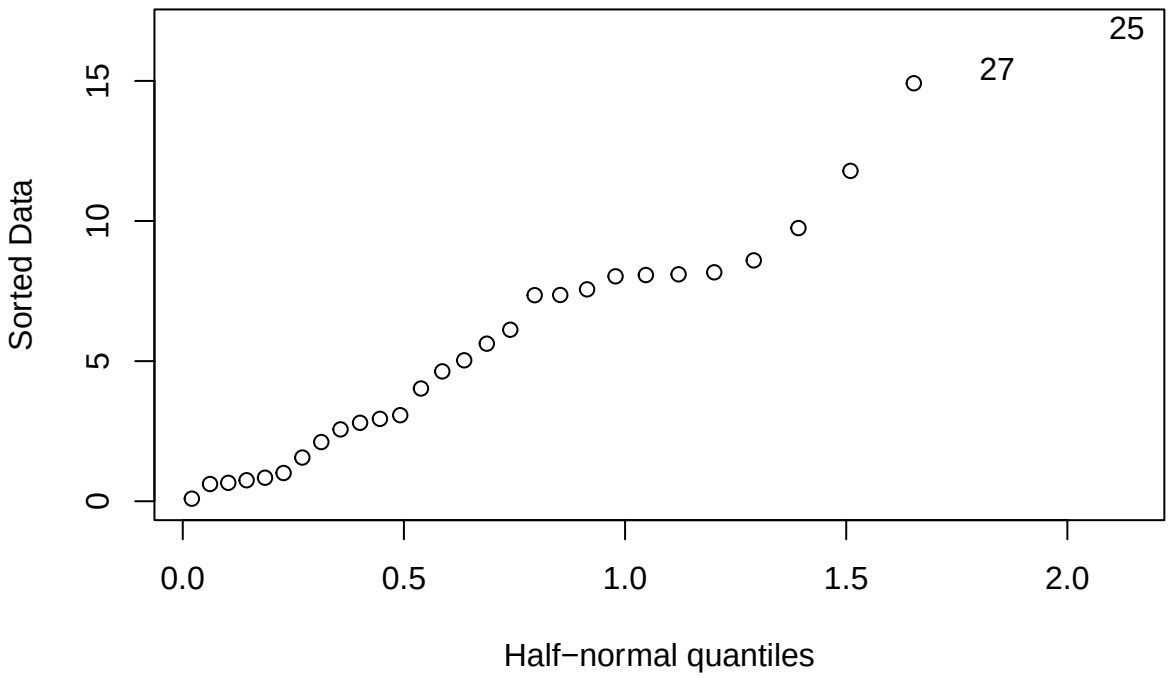
and we can conduct a goodness of fit test comparing our model to the saturated model

```r
1-pchisq(deviance(pois_glm),df.residual(pois_glm))
```

```
## [1] 0
```

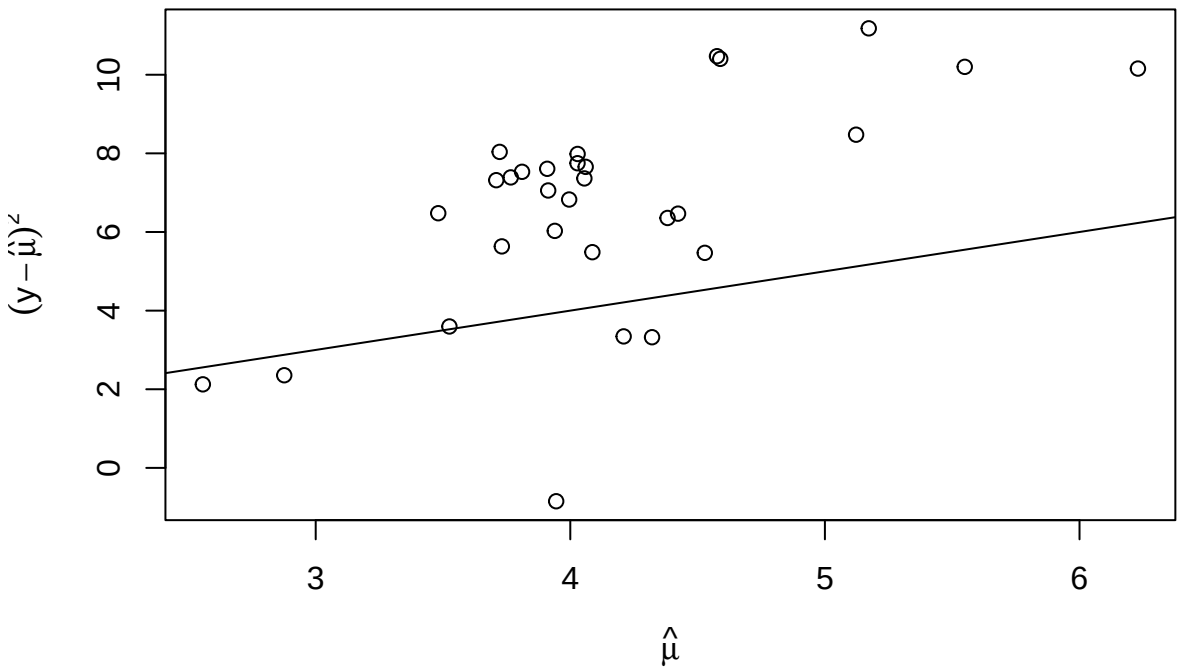
What should we do now?

```r
halfnorm(residuals(pois_glm))
```



Nothing jumps out.

```
plot(log(fitted(pois_glm)),log((gala_df$Species-fitted(pois_glm))^2),
     xlab=expression(hat(mu)), ylab=expression((y-hat(mu))^2))
abline(0,1)
```



From here, we see that our variance appears to be bigger than our mean, especially for larger values of μ. To further see this, we can estimate ϕ. Remember that by our model we are fixing ϕ to be 1

```
deviance(pois_glm) / df.residual(pois_glm)
```

```
## [1] 61.28803
```

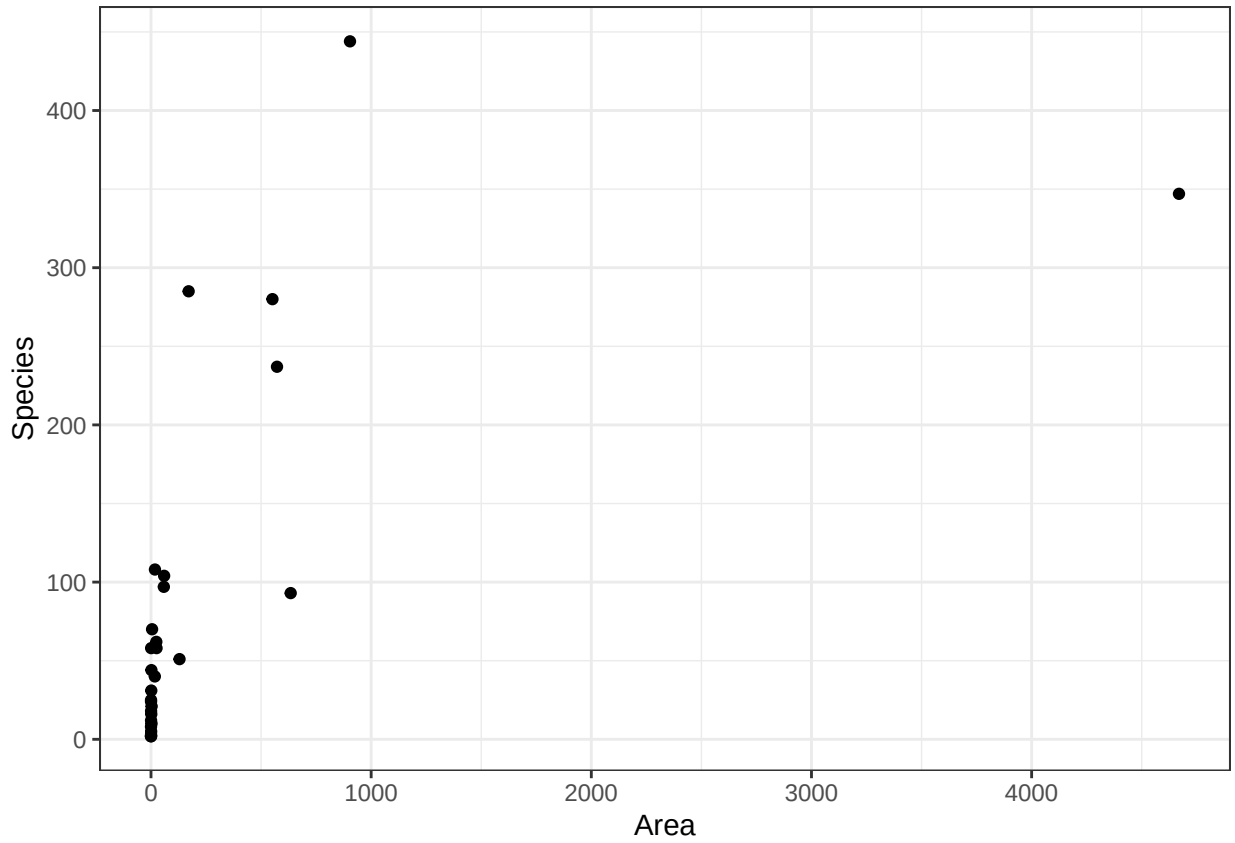
```
sum(residuals(pois_glm, type = "pearson")^2)/pois_glm$df.res
```

```
## [1] 65.80715
```

Ugh. Well, how did we address this for Binomial models?

We will do the same here (revisit next week...) Another possibility that we didn't have available to consider when we used a binomial is that different regions in the galpagos have different areas

```
gala_df %>% ggplot(aes(x=Area,y=Species)) +
  geom_point() + theme_bw()
```



While this may not fix everything, before we move to a quasi-Poisson we might want to consider adding an Offset to our model. That is, instead of building a model for $\log(\mu)$ we will build a model for $\log(\frac{\mu}{A})$ where A is the area of the region we are looking at. Note, this same thought process can be used if we are looking at the number of events over time and each observation took place for a different amount of time. For instance, perhaps I observed the number of crimes in one block for 2 weeks and I observed the number of crimes in another place for 3 weeks, here I would build a model for $\log(\frac{\mu}{T})$ where T is time.

The model becomes:

```
rate_glm <- glm(Species ~ offset(log(Area)) + Scruz +
                Elevation, data=gala_df,
                family=poisson)
```

We can see that this doesn't really fix the issue though.