Report Assessment 3

Joana Levtcheva, CID 01252821

Data Summary

The data contains 48 observations of two variables: the response variable `y` (number of lesions counted on each egg) with nonnegative integer values in the range $[5, 259]$, and the covariate `x` (the log_2 of the dilution factor of the virus concentration) with nonnegative integer values $0, 1, 2, 3, 4$.

```
##        y                x        
##  Min.   :  5.00   Min.   :0.000  
##  1st Qu.: 18.75   1st Qu.:1.000  
##  Median : 48.50   Median :2.000  
##  Mean   : 71.04   Mean   :2.083  
##  3rd Qu.:115.25   3rd Qu.:3.000  
##  Max.   :259.00   Max.   :4.000
```

Exploratory Data Analysis

```
##   x    mean        sd
## 1 0 186.625 42.203377
## 2 1 104.700 25.833011
## 3 2  50.800 18.984204
## 4 3  27.100 13.963842
## 5 4   9.100  4.201851
```

`Fig. 1` represents a scatter plot of `x` vs `y`. `Fig. 2` shows the mean values and the corresponding standard deviation of each group of `y` values corresponding to a value of `x`.

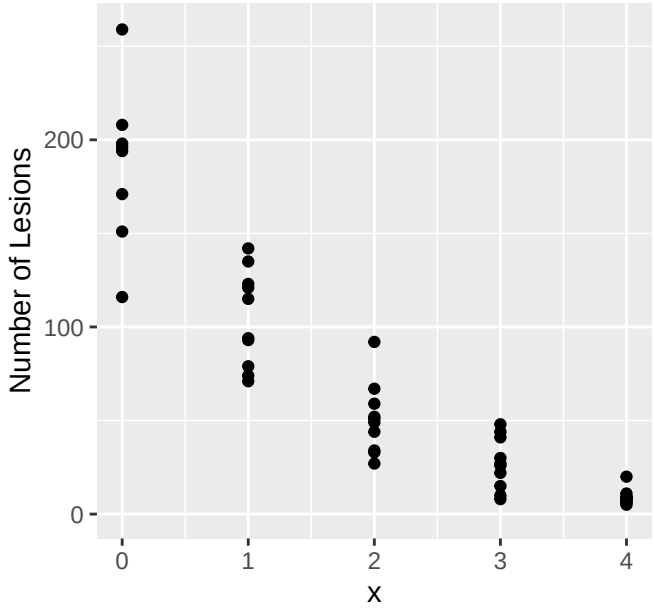


Fig. 1

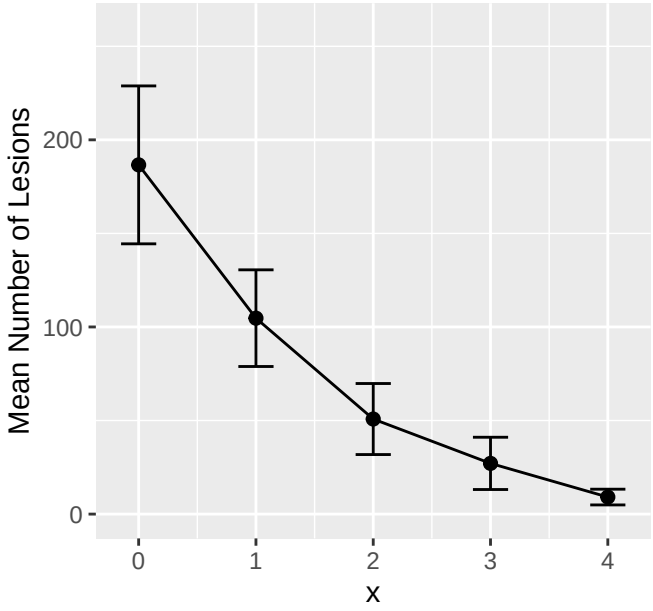


Fig. 2

`Fig. 1` shows there are 5 groups of `y` values corresponding to each value of `x`, all containing nonnegative whole numbers. From `Fig. 2` it is visible that the variance is not constant among the groups, and the larger the mean of `y`. Moreover, the means (and variances) are not "lying" on a line, rather they seem to be having an exponential relationship.

Fitting a Linear Model

Let's start by fitting a linear model, where the number of lesions will be explained by the log_2 of the dilution factor. We can easily outline a few problems:

- From the Linear Model plot is clear that for value of `x` nearing 4 the prediction would be negative. But the number of lesions is meaningful when it's a nonnegative whole number.
- In a linear model the response should be continous-valued but the response `y` is integer-valued
- A linear model assumes a constant variance but in Fig. 2 we already showed that the variance is not constant.

Poisson Generalized Linearised Model

The Poisson distribution is usually used to model counts. Suppose that a random variable `Y` takes on nonnegative integer values, $Y \in 0, 1, 2, ...$ If Y follows a Poisson distribution then

$$Pr(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!} \text{for} k = 0, 1, 2, ...,$$

where $\lambda > 0$ is the expected value of Y and also $\lambda = E(Y) = Var(Y)$.

In the context of our setting, we expect the mean number of lesions $\lambda = E(y)$ to vary as a function of the log_2 of the dilution factor. We consider the following model for the mean $\lambda = \lambda(x) = E(y)$:

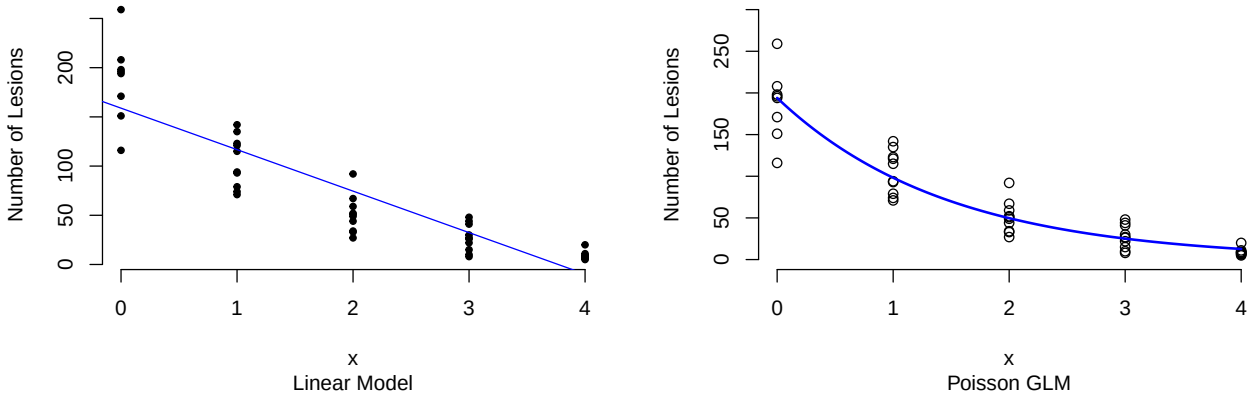
$$log\lambda(x) = \beta_0 + \beta_1 x$$

or equivalently

$$\lambda(x) = e^{\beta_0 + \beta_1 x},$$

where β_0, β_1 are parameters to be estimated. The above two settings define the Poisson generalized linear model for our case.

In our context, it makes sense to use a log link function because the data set has a log relationship, more specifically, in each setup we are decreasing the variola virus concentration twofold and as the final result is strongly correlated to the initial concentration of viruses we can expect that the relationship will be kept in our dataset. Also, the Poisson GLM will ensure there will be no negative response values.



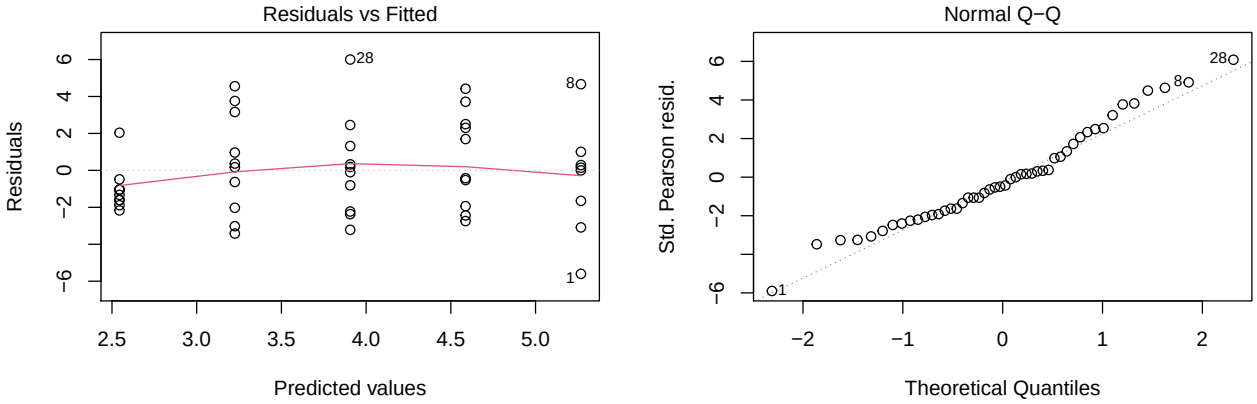
Fitting a Poisson GLM An increase in x by one unit is associated with a change in $E(y) = \lambda$ by a factor of e^{β_1}.

Fitting the Poisson GLM shows that when $x = 0$, or equivalently the dilution factor is 1, the expected number of lesions is ~194. Every additional twofold dilution is associated with a change in mean number of lesions by a factor of 0.506, or on average 50.6% as many lesions will occur relative to when the dilution factor is 1.

Under the Poisson model, we have $\lambda = E(Y) = Var(Y)$, therefore we have assumed that the mean number of lesions for a given dilution factor equals the variance of the number of lesions for that dilution factor. Looking back at `Fig. 2` the variance appears to be much higher than the mean. Therefore, overdispersion is present (which is also confirmed by the inflated z-values in the summary of the model fit). One way to deal with the overdispersion might be to fit a quasi-GLM.

There are no negative predicted values because the Poisson model only allows nonnegative values.

We can take a look at the diagnostic plots for our Poisson GLM:



From the first plot, we have that our fitted values are not systematically above or systematically below the observed values, and the magnitudes of our residuals are stable across fitted values

Fromk the second plot, we can see the largest residuals are larger than what we would expect under the fitted model, and the smallest residiauls are also larger than what we would expect. This, again, indicates our data is probably overdispersed as compared to our fitted model.

Summary

In conclusion, the dilution factor and the number of lesions counted on each egg are reversely proportional. The final number of lesions will decrease as many times as we increase the dilution factor. More specifically, when the dilution factor is 1, the expected number of lesions counted on an egg is 194, and for every additional twofold dilution, on average, 50.6% as many lesions will occur compared to when the dilution factor is 1.