Saumya Bhatnagar

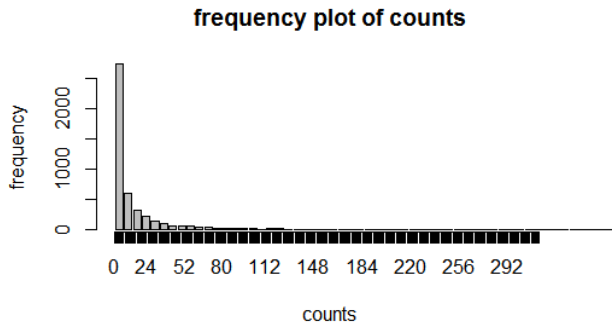
OBJECTIVE- The question I'm trying to answer here is: How the mean –depth, square of mean-depth, standard deviation and latitude effects the catching of Atlantic cod in a tow.

DATA- The data consists of the information of 4863 tows 4863 tows in the Gulf of Main and Georges Bank region during fall, 1970-2008, collected by the Northeast Fisheries Sciences Center (NEFSC).

If we look at the summary of the quantitative variables

mean.depth	mean.depthsq	std.dev	latitude
Min. :-1.5785	Min. :-0.64882	Min. :-0.7941	Min. :-2.1205
1st Qu.:-0.9439	1st Qu.:-0.50521	1st Qu.:-0.5630	1st Qu.:-0.9076
Median :-0.3091	Median :-0.23326	Median :-0.2324	Median :-0.2145
Mean :-0.1425	Mean :-0.04994	Mean : 0.1040	Mean :-0.1525
3rd Qu.: 0.5901	3rd Qu.: 0.16001	3rd Qu.: 0.2859	3rd Qu.: 0.5653
Max. : 4.3084	Max. :11.40413	Max. : 7.8382	Max. : 2.0381

The mean is slightly greater than median for all four variables, this means that all these four variables are slightly skewed to the right. If we look at the frequency plot of the variable ycount (which stores the counts of cod caught at a tow).



The Frequency plot shows that count data contains almost more than 2000 zeros for 4863 tows. This means we have so many zeros in our data.

MODEL-

Model1- The first model I chose is to check the effect of mean –depth, square of mean-depth, and standard deviation and latitude effects on the variable ybin (which is 1 if the Atlantic cod caught in a given tow and 0 otherwise). The model is:

$$log\frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1(mean - length) + \beta_2(mean - length)^2 + \beta_3\,(std - dev) + \beta_4(lattitude)$$

Where π_i= E($ybin_i$) and $ybin_i$~Bin(1,π_i)

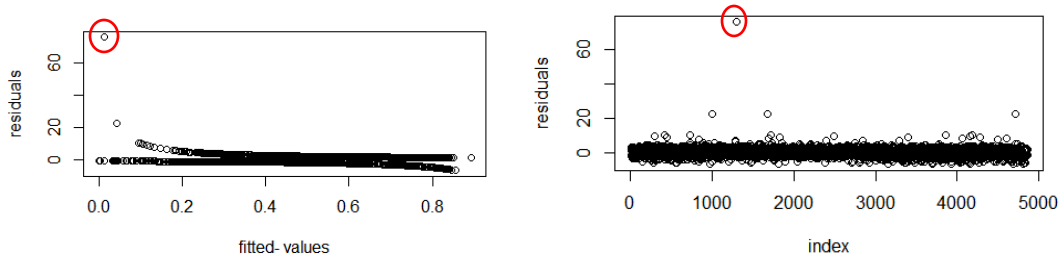
The results from R as follows:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.32120    0.03357  -9.568  < 2e-16 ***
mean_depth    -0.56922    0.04288 -13.274  < 2e-16 ***
mean_depth_sq -0.42054    0.06225  -6.756 1.42e-11 ***
stdev          0.30480    0.03196   9.538  < 2e-16 ***
lattitude      0.75203    0.03914  19.214  < 2e-16 ***
---

    Null deviance: 6660.0  on 4862  degrees of freedom
Residual deviance: 5966.7  on 4858  degrees of freedom
```

All the covariates in the model are significant. But before the model interpretation, let's check the model fitting. Although residual deviance is greater than df we can't comment on the model fitting because it is binary data.

Diagnostic



The plots of residuals with fitted values shows a pattern and hence that model is incorrect. Also the plot for residuals with index doesn't show a pattern but aren't random. Hence the model is incorrect. The possible reasons are that may be there are missing covariates, interaction and non-linear terms are missing. So we can try these different models. I'm not going further with this model.

Model2- The next model I tried is Poisson regression to check the effects of mean –depth, square of mean-depth, standard deviation and latitude effects on the counts of Atlantic cod in a tow. The model is

$$log\mu = \beta_0 + \beta_1(mean - length) + \beta_2(mean - length)^2 + \beta_3\,(std - dev) + \beta_4(lattitude)$$

Where μ_i= E ($ycount_i$) and ycount~ Poisson (μ_i)

The results from R

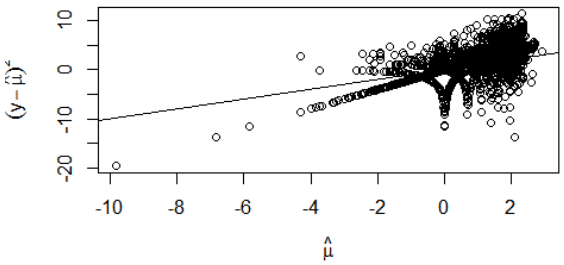
```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.721136   0.015021   48.01   <2e-16 ***
mean_depth    -1.345767   0.017079  -78.80   <2e-16 ***
mean_depth_sq -0.329299   0.022416  -14.69   <2e-16 ***
stdev          0.459639   0.007796   58.95   <2e-16 ***
lattitude      0.337777   0.007995   42.25   <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: 69355  on 4862  degrees of freedom
Residual deviance: 57879  on 4858  degrees of freedom
AIC: 64737
```

Here all the covariates are significant and if we compare the residual deviance with chi-square at significance level=0.05, then this model isn't a good fit. And if we look at the residual deviance the deviance is much greater than df, This could be the case of overdispersion.

Diagnostic- If we look at the plot of mean vs variance



There are so many observations which are above the straight line shows that variance is greater than mean. This shows that we have the case of overdispersion.

Model3-Earlier we have seen that there are many zeros in the count data, so this is zero inflated count data. The next model I tried is zero inflated negative binomial regression model to counter the overdispersion and zero inflated problem

Here if ycount=0, we don't know whether the Atlantic cod is present here or not. So for zero inflated negative binomial model which is a mixture of Bernoulli and negative binomial distribution.

$Y = V(1 - B)$, where $V \sim$ negative binomial (μ, α), $B \sim$ Bin $(1, p)$, and V and B are independent. Here if ybin=0 then Prob (ycount=0/ybin) =1 and if ybin=1, then Prob (ycount/ybin=1) = Negative-binomial (μ_i, α). Here α is overdispersion parameter.

The summary of the model is:

```
Count model coefficients (negbin with log link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.11607    0.03401  32.817  < 2e-16 ***
mean_depth    -1.29463    0.04039 -32.054  < 2e-16 ***
mean_depth_sq -0.15650    0.05406  -2.895  0.00379 **
stdev          0.47959    0.03785  12.669  < 2e-16 ***
lattitude     -0.08567    0.04088  -2.095  0.03613 *
Log(theta)    -0.93570    0.03742 -25.007  < 2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.82935    0.29447 -13.004  < 2e-16 ***
mean_depth    -0.72208    0.19455  -3.712 0.000206 ***
mean_depth_sq  0.25616    0.12063   2.123 0.033720 *
stdev          0.20534    0.09449   2.173 0.029763 *
lattitude     -3.42132    0.19008 -17.999  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 0.3923
Number of iterations in BFGS optimization: 23
Log-likelihood: -8680 on 11 Df
```

Here all the covariates are significant at significance level=0.05.To find out whether the zero inflated negative binomial model fits better than Poisson model, I did the vuong test for non-nested model.

```
Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
-------------------------------------------------------------
              Vuong z-statistic             H_A    p-value
Raw                   14.48313 model1 > model2 < 2.22e-16
AIC-corrected         14.47997 model1 > model2 < 2.22e-16
BIC-corrected         14.46972 model1 > model2 < 2.22e-16
```

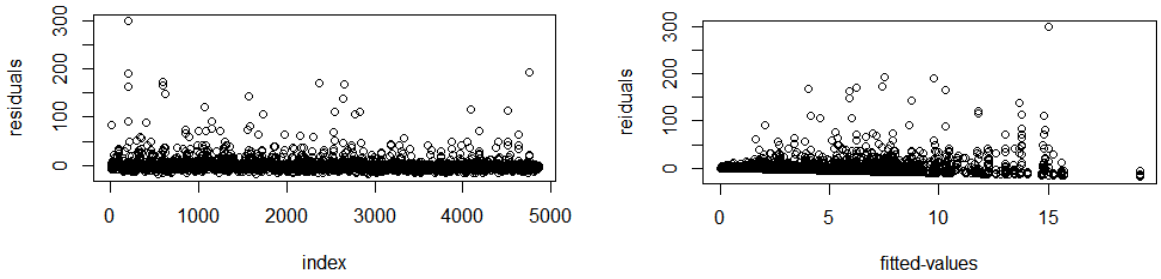
Here model-1 is zero inflated negative binomial and model-2 is Poisson model. The test result shows that zero inflated negative binomial fits better than Poisson model.

To test whether dispersion is significant or not. I did the likelihood ratio test

```
> 2*(logLik(mdl_inflnb)-logLik(mdl_poi))
'log Lik.' 47368.33 (df=11)
```

p-value is very low, hence I conclude that there is a significant amount of dispersion in the data.

Diagnosis



From the plot of residuals with index, although the plot doesn't have an apparent pattern but still residuals aren't random. From the plot of residuals with fitted values, the plot between residuals and fitted value doesn't have any pattern except the residuals are aggregated in a flat line. So this means something is wrong with this model which can be improved by adding additional covariates or transformation of covariates.

Interpretation – from the count part the negative coefficients for mean-depth and (mean-depth)^2 and latitude, shows that increase in depth and latitude decrease the log mean of catching a fish. And also increase in latitude decrease the log mean of catching fish. Also increase in standard deviation of the ocean depth increase the log mean of catching fish. From the binary model, negative coefficient of mean-depth and latitude shows that log odds of catching a fish decreases with increase in mean-depth and latitude and positive coefficient for (mean-depth)^2 and standard deviation for ocean depth shows that increase in log odds of catching a fish with increase in (mean-depth)^2 and standard deviation.

CONCLUSION- 1 from the exploratory data analysis, I saw that data is zero inflated. That is counts of number of fishes caught from a tow contains so many zeros.

2. First I fit binary logit model to the binary data and from the model diagnostic found that logistic regression model isn't a good fit. Further study can be done in this direction by adding interaction between covariates, transformation of covariates. The link function can be tested as well.

3. The second model I tried is Poisson regression model with log link. The results show that model isn't a good fit as there is a case of overdispersion in the data.

4. To counter the zero inflation and overdispersion, I applied zero inflated Negative Binomial model to the data and found that it's better than Poisson model and there is a significant overdispersion in data. Although this model is better than other models, the residual plots shows that there may be some missing covariates. Like this data is collected at different regions and at different time intervals. So temporal and spatial effects can be considered.

REFERENCE :- Xia Wang, Ming-Hui Chen, Rita C. Kuo, and Dipak K. Dey : Bayesian Spatial-Temporal Modeling Of Ecological Zero-Inflated Count Data.

Vuong's closeness test: Wikipedia.