Poisson Regression

Joel Cantero Priego & Ricard Meyerhofer Parra

24/11/2019

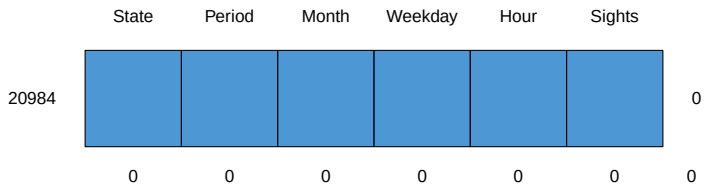
Introduction

In this assignment, we are going to use the UFO's National UFO Reporting Center dataset, it is available on UFO sightings in the United States in this millennium (till 2014).The database has a total of 52,813 reported sightings grouped into 20,984 observations, these have been classified according to 5 variables:

Variable name	Description	Values
State	USA State where the UFO was sighted	chr (50 states)
Period	Period of five years when it was sighted	int (1: 2000-2004, 2: 2005-2009, 3: 2010-2014)
Month	Month of the year when it was sighted	int (1:12)
Weekday	Day of the week when it was sighted	int (1: 7)
Hour	Time of the day when it was sighted.	int (1: 00-05, 2: 06-11, 3: 12-17,4: 18-23)
Sights	Number of sights recorded	int

First of all, we are going to convert all explanatory variables as factors (categorical variables).

Before we start doing the Exploratory Data Analysis, we have to check if there is some missing value in our dataset.



Thanks to **md.pattern** function from mice package, we can assume that our dataset is cleaned.

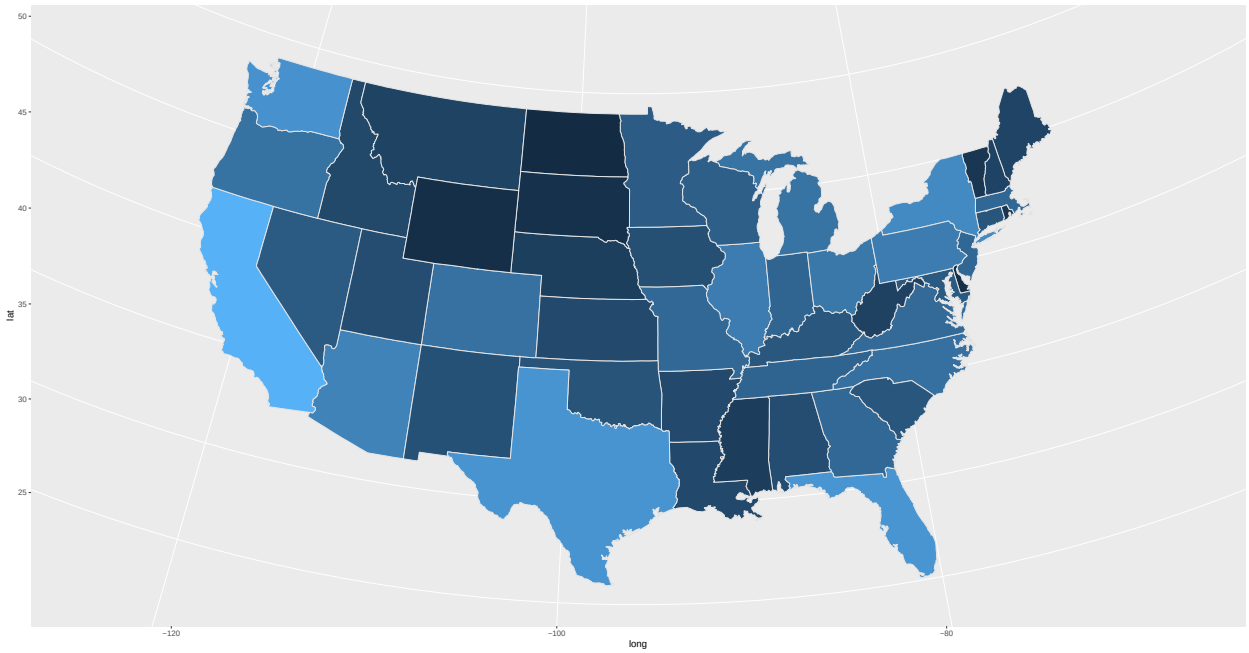
Exploratory Data Analysis

In this section we are going to focus in explaining the most interesting conclusions of our data, perform an univariate and multivariate analysis of the variables in order to find outliers and see how each of these variables is related with the sights.

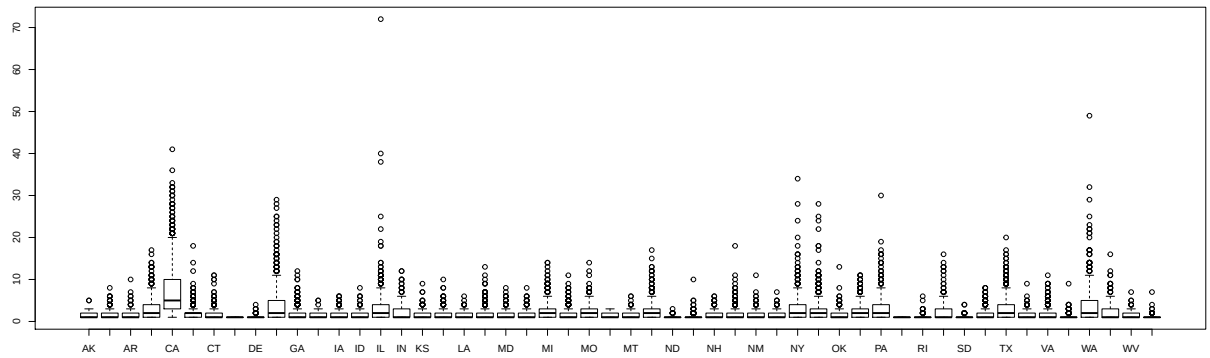
State

In State variable, we can observe that there is a certain imbalanced proportion in State proportion. We can identify some minority states like: PR (7), DC (7), ND (80), SD (123), WY (110) against some majority states like CA (976), FL (801), IL (644), NY (734)...

In order to see if there is any pattern we are going to display a map of the USA painted with the number of sights on each state.

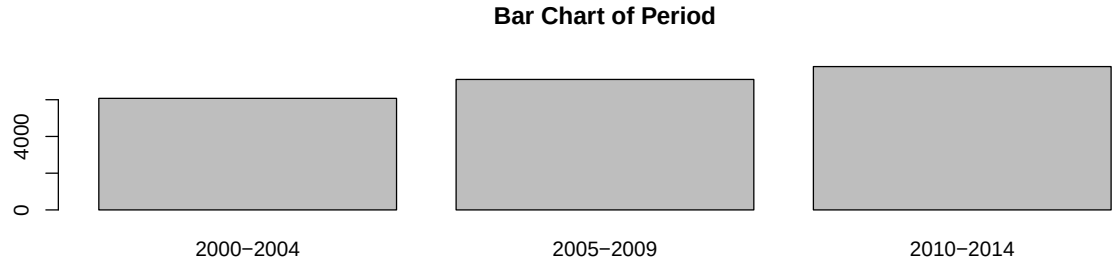


If we check it with a barplot, we can see that there is not a strong correlation between the year of release and the gross obtained from those years. We can observe that California state has a high mean of sights than the rest of states which could possibly also be due to its population among other factors. We can also observe some outliers values in IL State.



Period

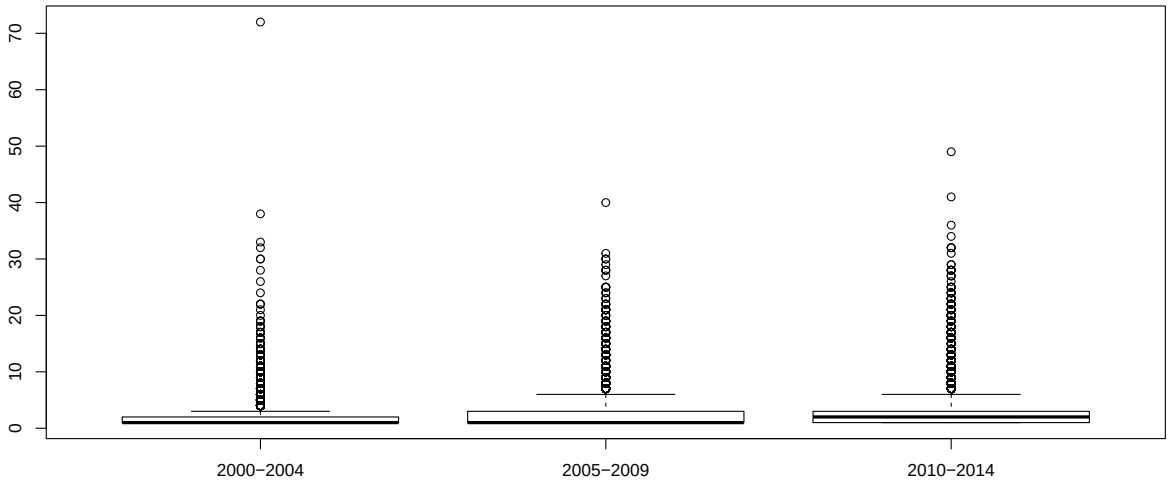
No problems for period variable, there is a certain expected balanced in period-years proportion even that we can see a growth in the number of observations.



Relationship between period and sights

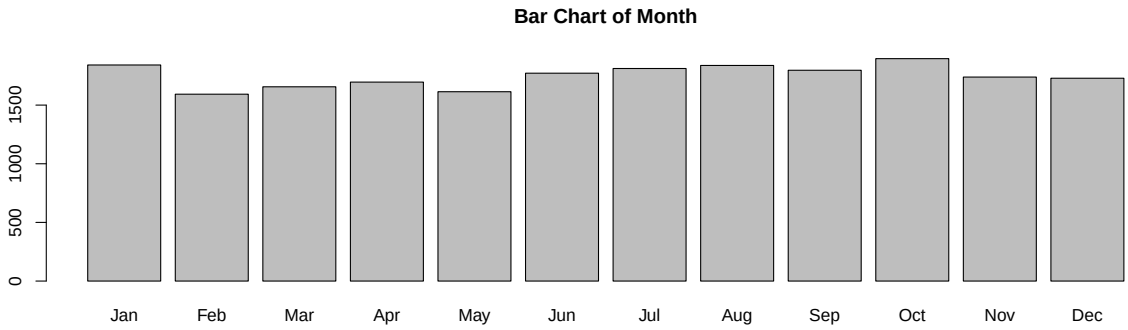
We can see that there is not a strong correlation between the period-year and the sights obtained. Even that 2010-2014 period is better than the previous ones the trend seems to be the same across the years.

We can look at some outlier point in 2000-2004 with 74 sights, that extreme value is the same we have observed at IL state.



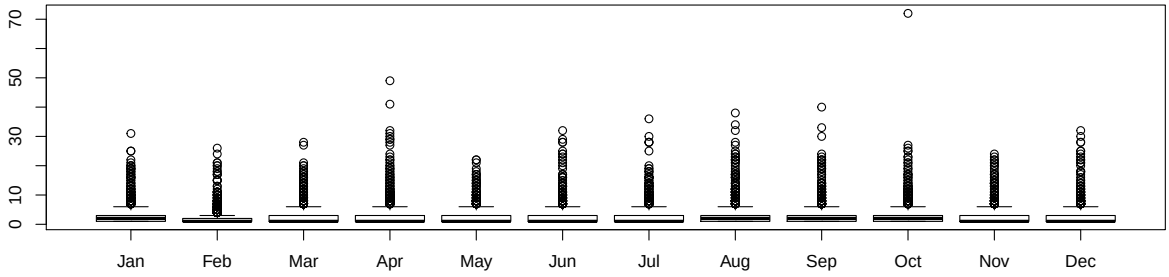
Month

No problems for month variable, there is a certain expected balanced in months proportion.



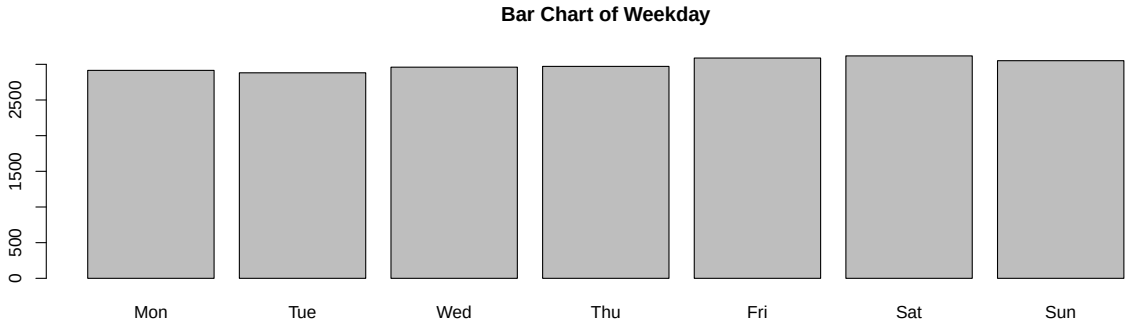
Relationship between months and sights

We can see that there is not a strong correlation between the month and the sights obtained from these observations. Even that there are some months better than others the trend seems to be the same across the year. We obtain the same outlier as the previous analysis.



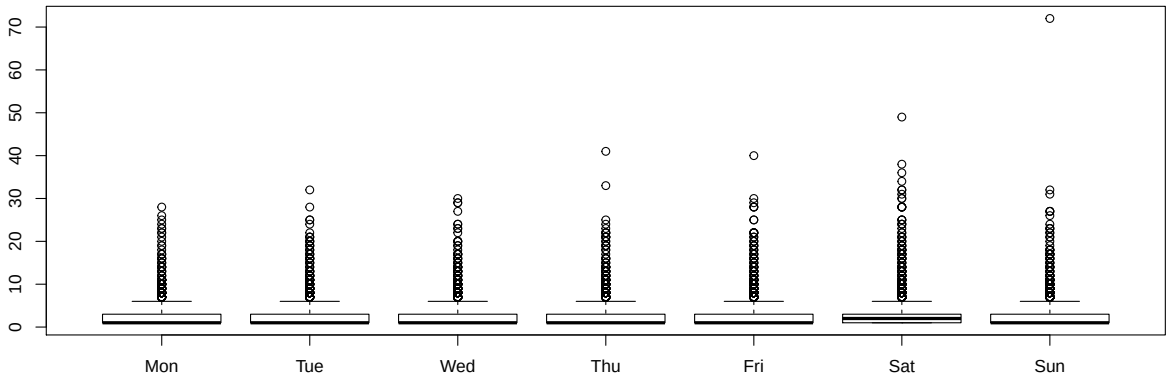
Weekday

No problems for Weekday variable, there is a certain expected balanced in weekday proportion.



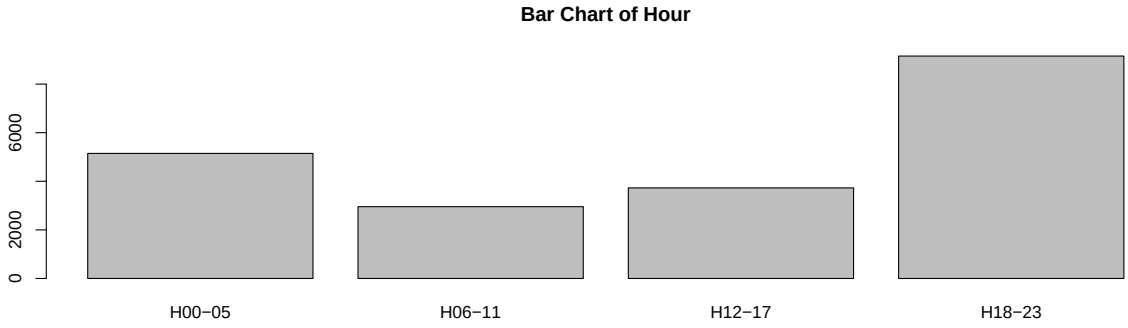
Relationship between weekdays and sights

We can see that there is not a strong correlation between the weekdays and the sights obtained from these observations. We obtain the same outlier as the previous analysis.



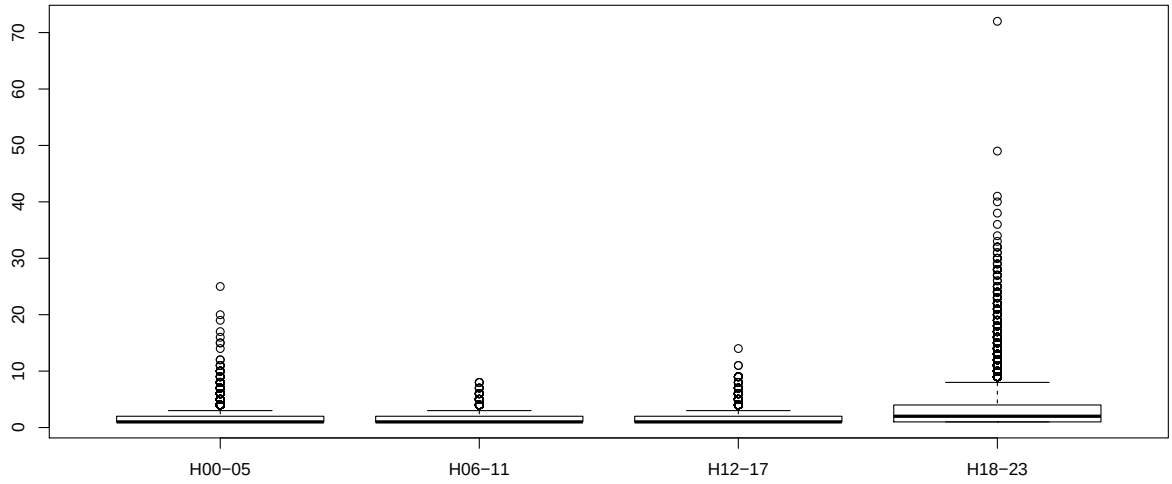
Hour

We can see a certain imbalanced proportion in hour period. There are more observations from 18-23 hours than the other hour periods.



Relationship between hours and sights

We can see that there is not a strong correlation between the hours and the sights obtained from these observations. However, we can see that 18-23 hour period the mean is higher than the other hour periods. We obtain the same outlier as the previous analysis.



Dealing with outliers

We have decided to remove the outlier we have detected on Exploratory Data Analysis. We think it is an extreme value and it will impact on our model performance.

Fitting the log-linear model without interactions.

We are going to build our generalized linear model as a log-linear model depending on all the variables with no interactions.

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
42556.46	20982	-34830.43	69808.86	70397.27	16455.81	20909

First Order Interactions

We evaluate possible first order interactions and we checked if it was significantly different to the complete model and if it had significant interactions. We found that again the model fits our data but that it has a smaller deviance than the full model.

Automatic Variable Selection process

We are going to use the stepwise procedure by using AIC and BIC criterions to choose our final model. We are going to use both criterias and starting from null and complete model in order to compare all models and pick the one that performs the best + is more legible and simpler.

BIC Forward

With the formula: Sights ~ State + Period + Month + Weekday + Hour + State:Hour + Period:Hour

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
42556.46	20982	-33807.1	68074.2	69903.03	14409.15	20753

AIC Forward

With the formula: Sights ~ State + Hour + Period + Month + Weekday + State:Hour + Hour:Period + State:Period + Hour:Month + Month:Weekday + Period:Month + Period:Weekday + Hour:Weekday

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
42556.46	20982	-33252.06	67462.12	71270.87	13299.07	20504

BIC Backward

Formula: Sights ~ State + Period + Month + Weekday + Hour + State:Hour + Period:Hour

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
42556.46	20982	-33807.1	68074.2	69903.03	14409.15	20753

AIC Backward

Formula: Sights ~ State + Period + Month + Weekday + Hour + State:Period + State:Hour + Period:Month + Period:Weekday + Period:Hour + Month:Weekday + Month:Hour + Weekday:Hour

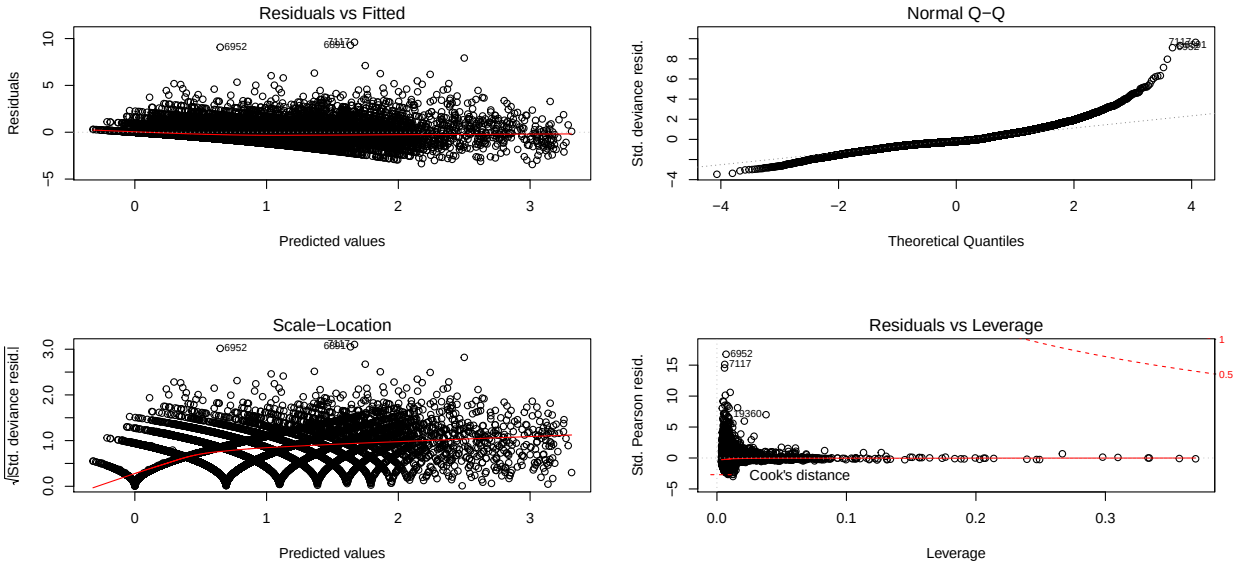
null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
42556.46	20982	-33252.06	67462.12	71270.87	13299.07	20504

We are going to finally chose the BIC forward model. We decide this model over the others because it is way simple and more understandable than the other models, specially AIC ones.

We perform an anova test in order to find if our results, are significant. Therefore, it will say us if we need to reject or not the hypothesis test.

```
anova(chosenModel, test = "Chisq")
anova(nullModel, chosenModel, test="Chisq")
```

Since we have p.value<0.05 we can refuse the null hypothesis which means that we pass the test and that our model is different than the null model.



In the previous plots we can see the following:

- Residuals vs fitted: There is assymetry in the distribution therefore the normality is not fulfilled.
- Normal Q-Q: We can see that the distribution follows the expected values along the normal line except for the right-hand tale. Therefore it looks like as in the previous, is not fulfilled.
- Scale-Location: We can see a distribution with a very linear standard deviance residuals along the fitted values, which is a good indicator. However, at the left-hand side, we can see a weird distribution
- Residuals vs Leverage: We can see that even we have some outliers that cause this weird performance in the previous plots, we can see that they don't seem to end.

This means that we should take a deeper look at the data for corrections and outlier detection which is out of the scope of this project.

Overdispersion analysis

This test states if the mean and variance are equal or not. If the constant c is > 0 it means that there is overdispersion, whereas if it is < 0 it means that there is underdispersion.

```
library(car)
library(AER)
dispersiontest(chosenModel)
```

```
##
##  Overdispersion test
##
## data:  chosenModel
## z = -7.4803, p-value = 1
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##  0.8062898
```

With trafo = NULL, we obtain a p-value of 1, which means that we cannot reject the null hypothesis, that tells that the true dispersion is not greater than 1. This means that there is underdispersion in the model.

```r
dispersiontest(chosenModel, trafo = 1)
```

```
##
##  Overdispersion test
##
## data:  chosenModel
## z = -7.4803, p-value = 1
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##      alpha
## -0.1937102
```

Using a value trafo of 1, the results show that we also cannot reject the null hypothesis, so the results are the same as before.

```r
dispersiontest(chosenModel, trafo = 2)
```

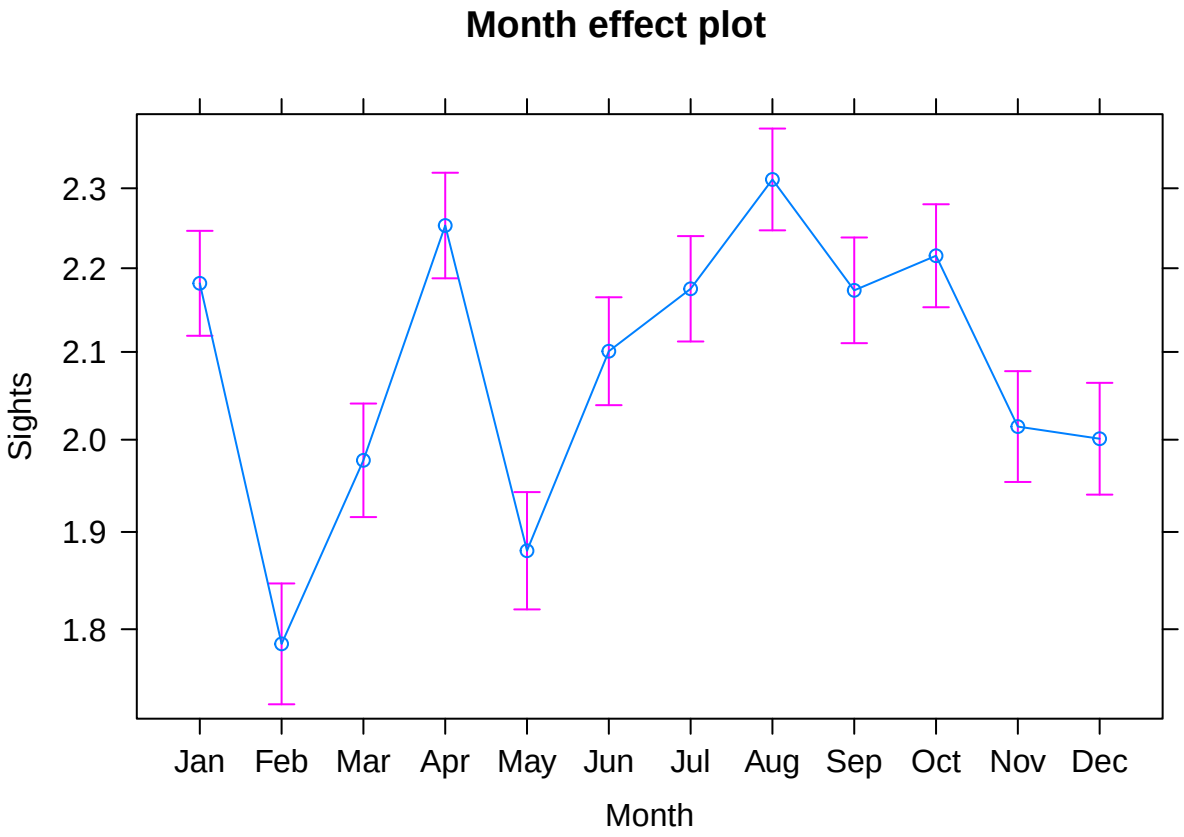
```
##
##  Overdispersion test
##
## data:  chosenModel
## z = 8.1391, p-value < 2.2e-16
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##      alpha
## 0.06134455
```

With a trafo value of 2, the results show that the null hypothesis can be rejected, as we have a p-value of 2.2e-16, which means that the alpha is greater than 0 and, therefore, there is overdispersion, but it is very close to 0.

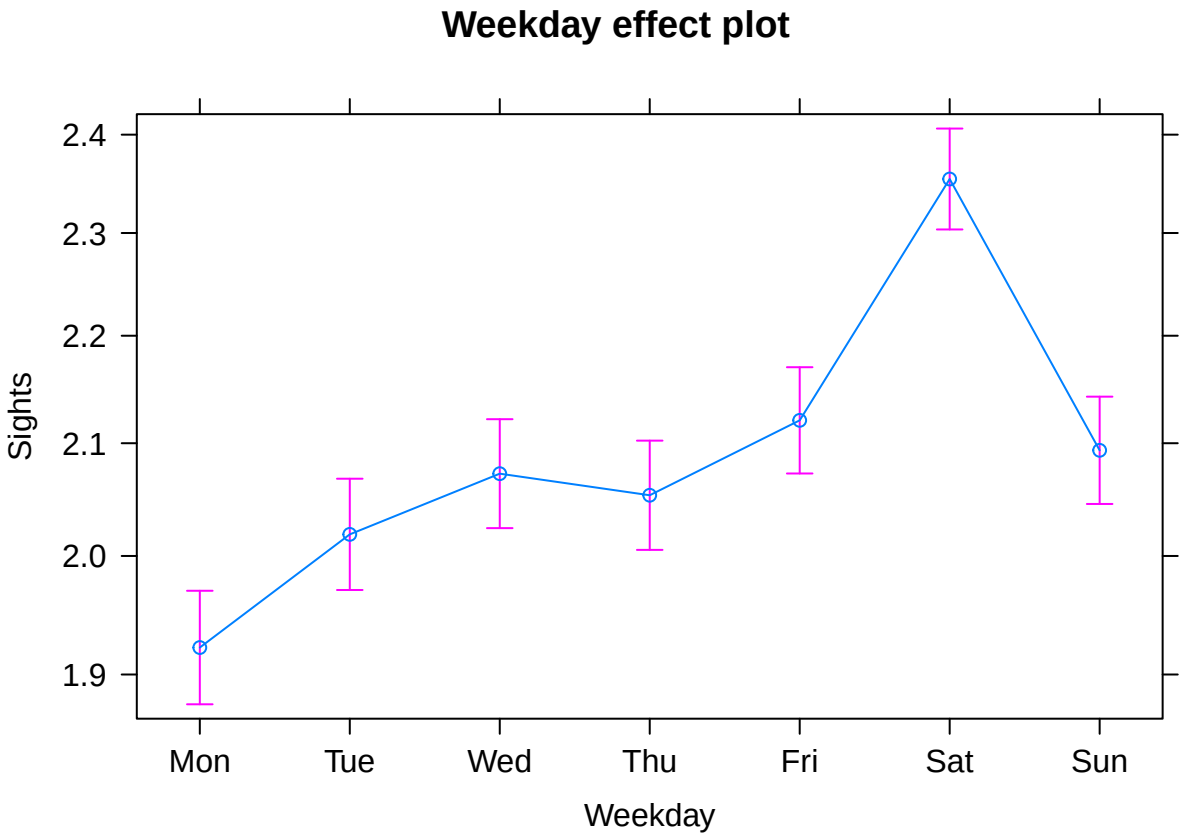
Interpretation

For the model interpretation, we will use the *effect* plots, taking into account the interaction terms. As we are building a Poisson regression model, we can not directly interpret the coefficients and response values, since they do not represent the response variable class but the value of the link function.

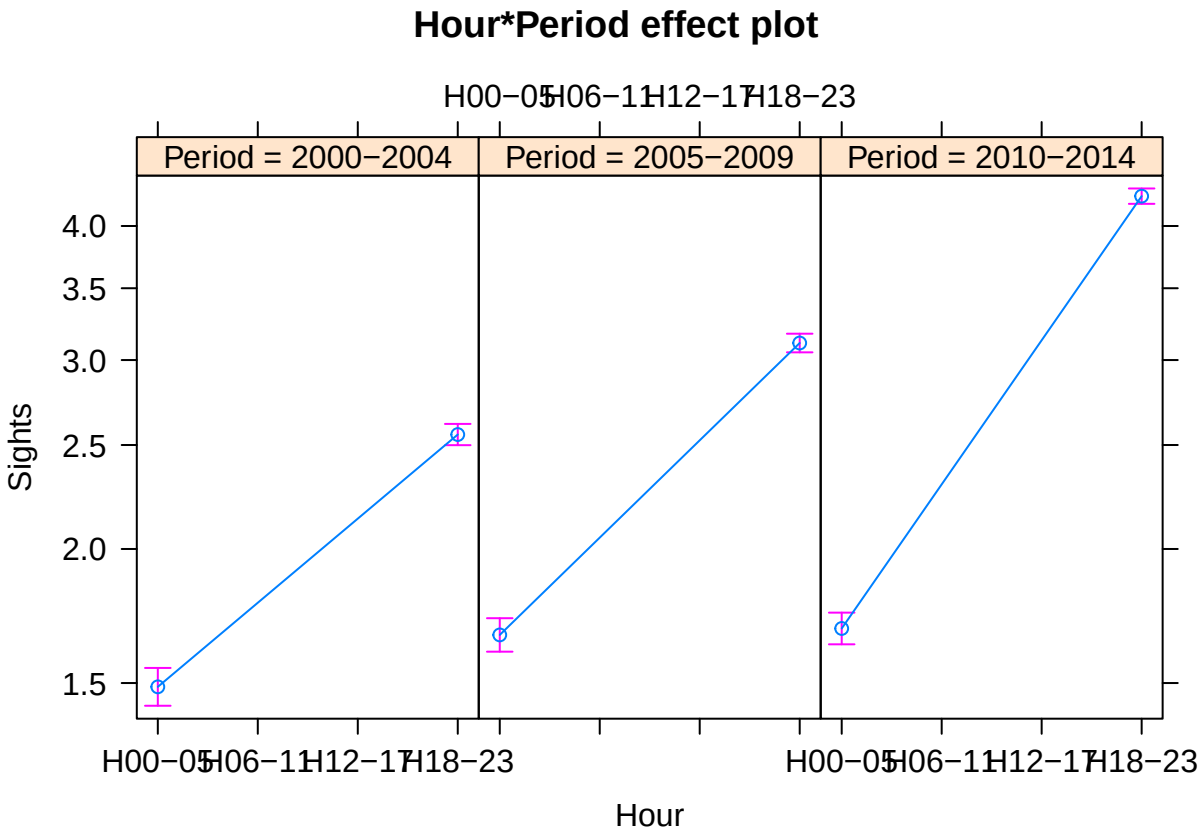
```r
library(effects)
m.effects <- effects::allEffects(chosenModel)
plot(m.effects$Month)
```



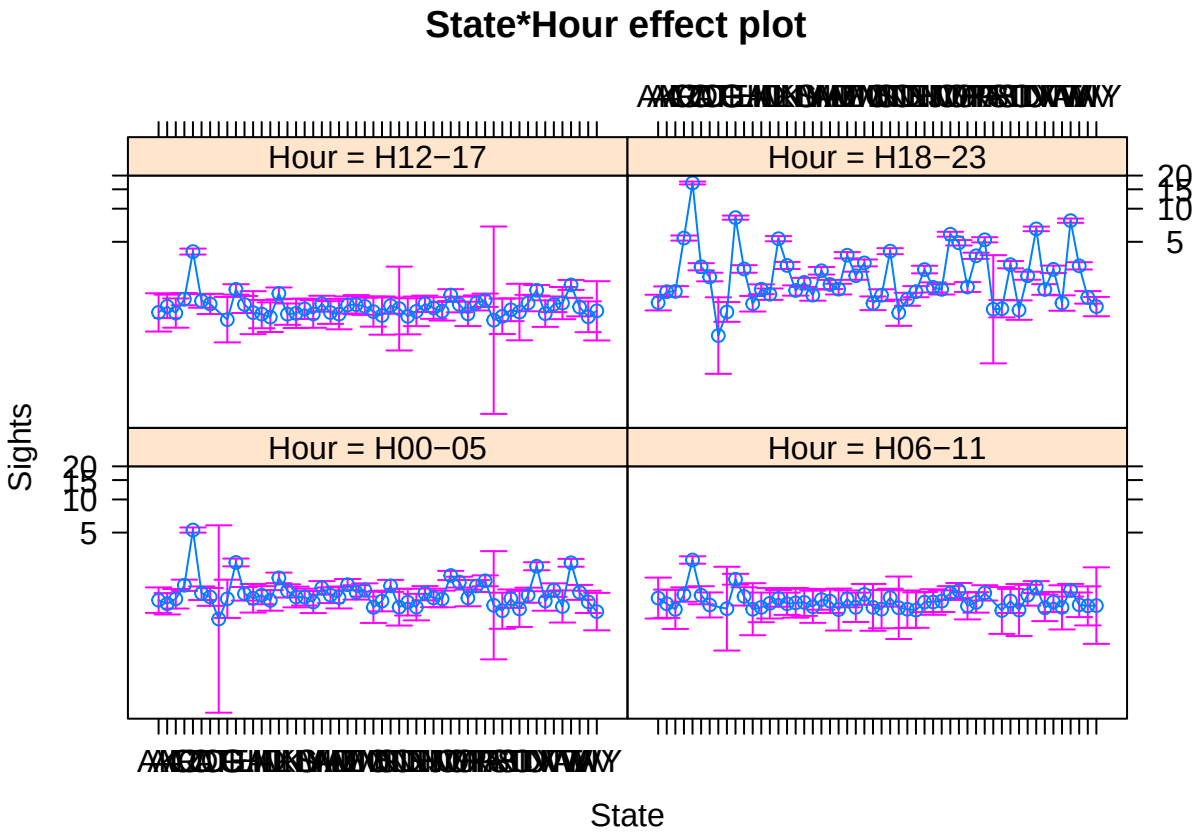

```
plot(m.effects$Weekday)
```




```
plot(m.effects$`Hour:Period`)
```




```
plot(m.effects$`State:Hour`)
```



We can see that there is some months differences in the reports, we can also see that the more on the weekends we are, the more aliens there is and also the more late the hour, the better and this happens across the years. So as we said before, we can see that the sighting rate per state does not seem to change significantly during the first three hour periods of the day. But from 18-23, we can see that there is more variance in the rate of sightings per state.