ASM Homework 3

Generalized Linear Model for UFO data

Maria Gkotsopoulou & Ricard Monge Calvo & Amalia Vradi

21/11/2019

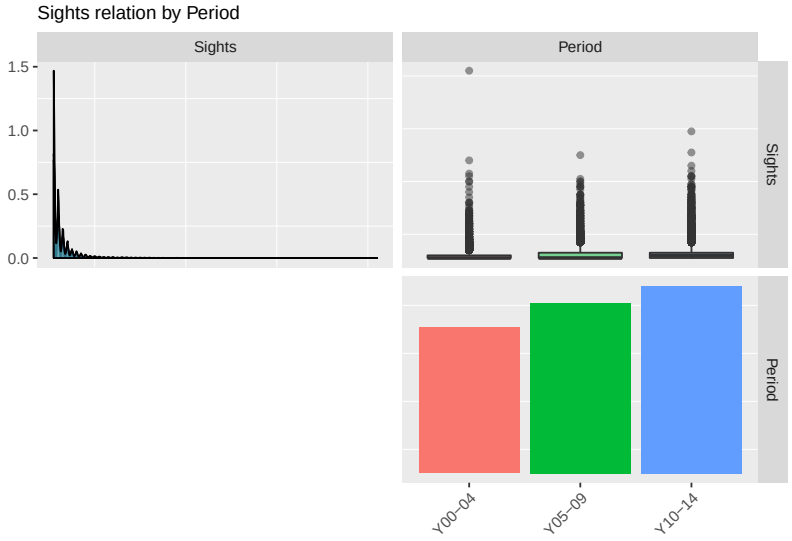
The aim of this work is to explore the Generalized Linear Model for count data by working with the UFO data set, which compiles the UFO sightings in the USA from 2000 to 2014 depending on state, period, month, weekday of sightings.

Exploratory Data Analysis

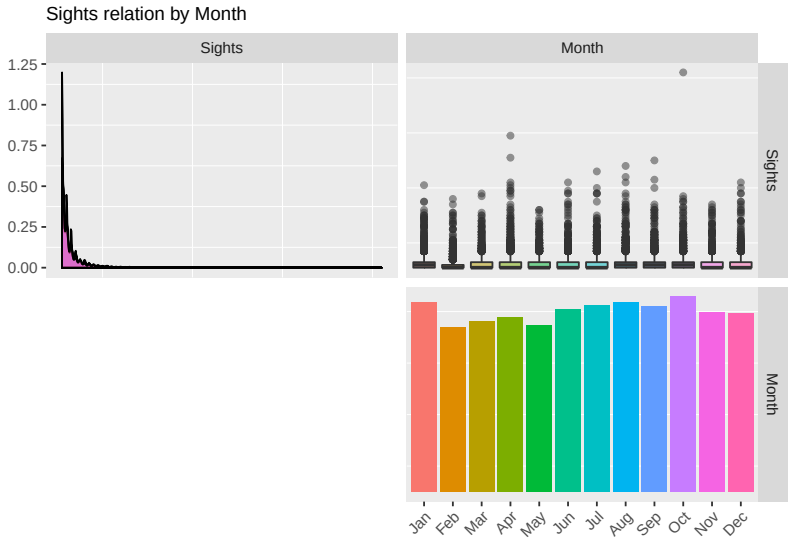
We see all predictors in the *UFO* dataset are categorical variables. In order to better interpret them, we change their labels to more understandable texts. For the year periods, *Period* variable, we put the letter Y followed by the last two digits of the year (as in Y00-04 for the period 2000-2004). For the *Month* variable we use the english month acronyms (Jan, Feb, etc). Finally for the *Weekday* variable we use the first two letters of the english weekday.

Note that the *Hour* is already in a similar format to our year variable, with an H letter followed by the hours of the period (as in H18-23).

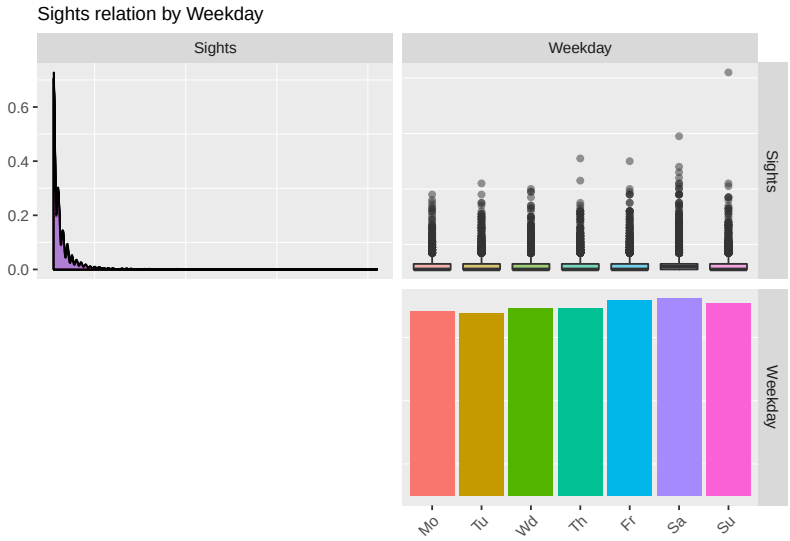
To see the relations between these factors and the response variable *Sights* we look at the boxplots.



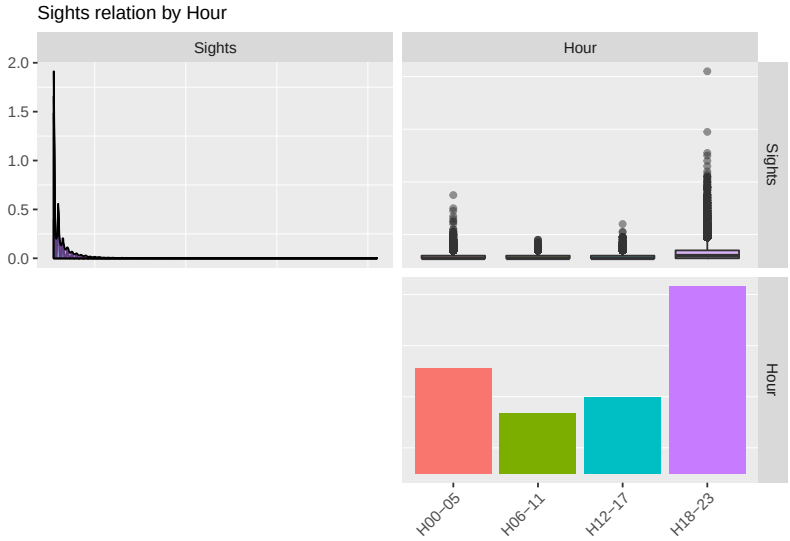
We see an increase in sightings as the years go by. As the distribution of sights has high kurtosis and it's asymmetric to the right, it has a long tail that does not fascilitate the comparison across the boxplots. Nevertheless, we do not see a difference in the means nor the variance.



We don't appreciate a trend for the months.



There is a slight tendency of higher values as the week goes by. Nevertheless, due to the long tail of the distribution, once again, we can't appreciate a difference in the means.



We see a difference in the mean for the case of the later hours of the day. In addition, it is also the period that has a longer tail.

Log-linear model

We start building our generalized linear model as a *log-linear* model depending on all the predictors with no interactions.

Table 1: Model Summary

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
42900.51	20983	-34964.23	70076.46	70664.87	16717.3	20910

```
## Residual Deviance: 16717.3
```

```
## Critical value (5%): 21247.5
```

```
## P-value: 1
```

We see that residual deviance is much lower than the critical value, assuming a χ^2 distribution of 2.091×10^4 degrees of freedom, and the p-value is 1. Therefore, we accept the Null hypoteshis that the model fits the data.

Nevertheless, we check for the significance of the coefficients corresponding to categorical variables by looking at the Anova test:

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: Sights
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                    20983      42901
## State   51  11979.7     20932      30921 < 2.2e-16 ***
## Period   2    960.4     20930      29960 < 2.2e-16 ***
## Month   11    229.4     20919      29731 < 2.2e-16 ***
## Weekday  6    143.3     20913      29588 < 2.2e-16 ***
## Hour     3  12870.3     20910      16717 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is clear that all predictors are significant, as expected from the previous goodness-of-fit test.

First Order Interactions

In order to better capture the relations between predictors and response variable, we evaluate the model with possible first order interactions. As before, we look the significance of the categorical variables together with the interaction terms.

As fiting a model with all interactions (without quadratic terms) takes an unviable amount of time, we fit a model with the interaction for each variable at a time, and check whether it's significantly different to the first full model (without interactions) and whether it has significant factors and interactions.

All independent interactions seem to be significant, and their associated models are significantly different from the full model, so we build the complete model with all the first-order interactions.

Table 2: Model Summary

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
42900.51	20983	-32871.91	68385.82	78889.77	12532.65	19663

```
## Analysis of Deviance Table
##
## Model 1: Sights ~ State + Period + Month + Weekday + Hour
## Model 2: Sights ~ (State + Period + Month + Weekday + Hour) * (State +
##     Period + Month + Weekday + Hour) - State * State - Period *
##     Period - Month * Month - Weekday * Weekday - Hour * Hour
##   Resid. Df Resid. Dev   Df Deviance  Pr(>Chi)
## 1     20910      16717
## 2     19663      12533 1247   4184.6 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: Sights
##
## Terms added sequentially (first to last)
##
##
##                 Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                            20983      42901
## State:Period   151  13342.7     20832      29558 < 2.2e-16 ***
## State:Month    560    963.4     20272      28594 < 2.2e-16 ***
## State:Weekday  302    463.8     19970      28131 5.813e-09 ***
## State:Hour     150  14534.4     19820      13596 < 2.2e-16 ***
## Period:Month    22    137.5     19798      13459 < 2.2e-16 ***
## Period:Weekday  12     95.1     19786      13364 4.970e-15 ***
## Period:Hour      6    404.7     19780      12959 < 2.2e-16 ***
## Month:Weekday   66    196.3     19714      12763 7.265e-15 ***
## Month:Hour      33    169.8     19681      12593 < 2.2e-16 ***
## Weekday:Hour    18     60.2     19663      12533 1.889e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Residual Deviance: 12532.65
```

```
## Critical value (5%): 19990.32
```

```
## P-value: 1
```

We have again a deviance much lower than the critical value, with a p-value of 1, so the model fits our data as before. Nonetheless, the deviance of the interactions model is lower than the full model. Together with the fact that the anova test between models gives a significant p-value, we can conclude the first order interactions capture better the relations in the data.

Automatic Variable Selection process

We use the stepwise procedure, by using the *AIC* & *BIC* criterion, to select our final model. Since our objective is the interpretability of the model we choose as starting point the null model, in contrast to starting form the complete. We place as an upper bound the previous complete model with first order interaction terms.

```
## AIC model formula
```

```
## Sights ~ State:Hour + Hour:Period + Month:Weekday + State:Period +
##     Hour:Month + Period:Month + Period:Weekday + Hour:Weekday
```

```
## BIC model formula
```

```
## Sights ~ State:Hour + Hour:Period + Period:Weekday + Period:Month
```

Table 3: Models Summary

model	AIC	BIC
Complete_interactions	68385.82	78889.77
stepwise_AIC	67723.76	71532.54
stepwise_BIC	68186.72	70285.92

Since the models are nested we use the *Anova* test to see if they are significantly different.

```
## Analysis of Deviance Table
##
## Model 1: Sights ~ State:Hour + Hour:Period + Month:Weekday + State:Period +
##     Hour:Month + Period:Month + Period:Weekday + Hour:Weekday
## Model 2: Sights ~ State:Hour + Hour:Period + Period:Weekday + Period:Month
##   Resid. Df Resid. Dev   Df Deviance  Pr(>Chi)
## 1     20505      13555
## 2     20720      14448 -215  -892.95 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The models are significantly different. This is due to the difference in criteria, penalizing more additional parameters with *BIC*. As expected, the *AIC* selected model has a better *AIC* measure while the *BIC* selected model has better *BIC* measure. In order to better interpret it, we choose the *BIC* which is the more succint model based on the parsimony criteria.

```
## Residual Deviance: 14447.55
```

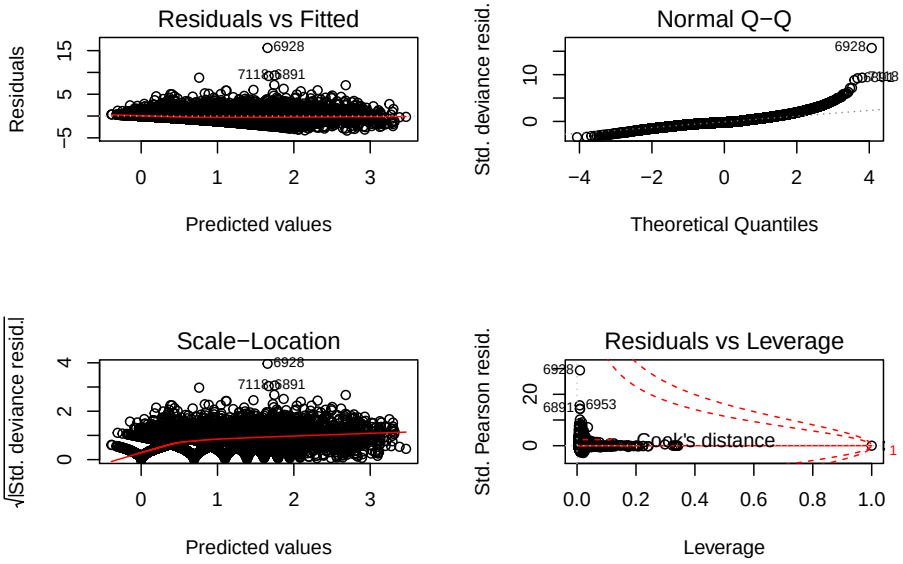
```
## Critical value (5%): 21055.97
```

```
## P-value: 1
```

Again our model seems to fit the data.

Model validation

After selecting our final model, we check the assumptions by looking at the following plots:

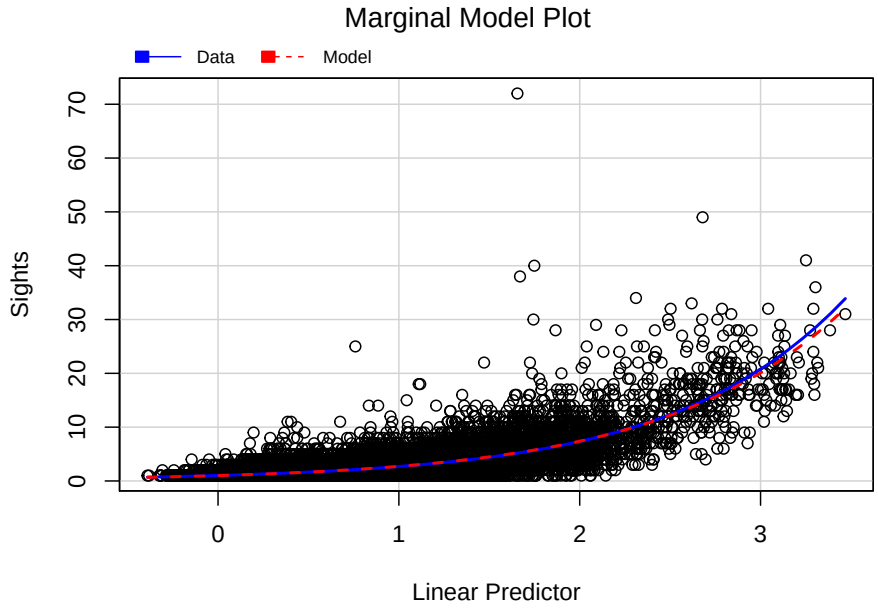


From the *Normal Q-Q* plot we see that there is assymetry in the distribution and we can conclude that normality of the residuals is not met.

Furthermore, from the *Scale-Location* plot, we seek to validate the assumption of homoskedasticity, which does not seem to hold in our case. Furthermore, we see that there are some patterns in the plot which could indicate quadratic relations between the predictors and the response. For the purposes of our assignment, we will not explore this any further.

Finally, from the *Residuals-Fitted* plot we see that the residuals don't seem to have any kind of tendency. However, they are not randomly distributed across the y axis. This could again indicate the need for other corrections in the predictors or response variables. For the purposes of our assignment, we will not explore this any further.

On the other hand, we can see the marginal residual plots to compare the mean values of the data and our model:



We see the data and model's mean values coincide quite well.

Overdispersion diagnosis

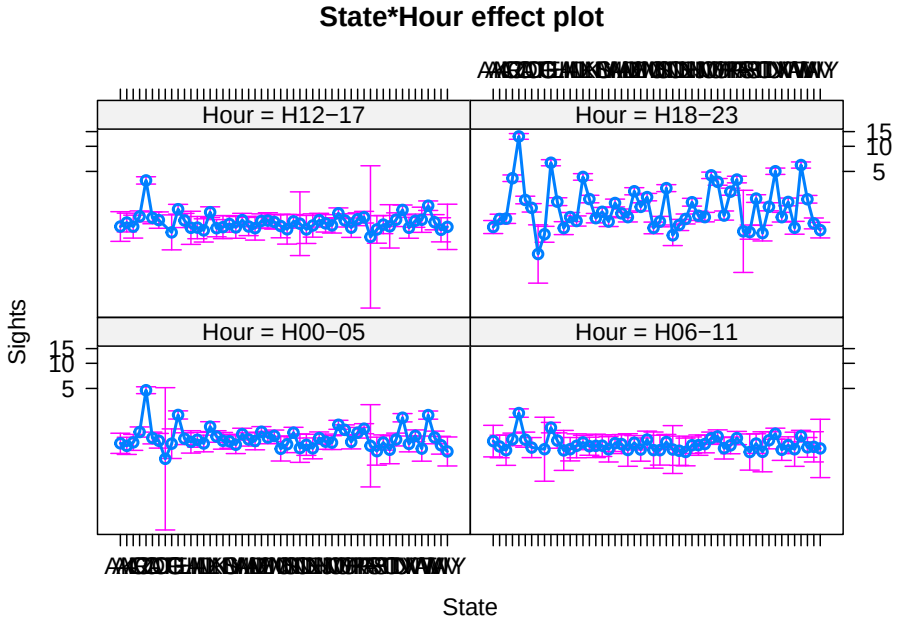
To complete the discussion about the goodness of our model, we perform a dispersion test.

```
##
##  Overdispersion test
##
## data:  m.final
## z = -3.7308, p-value = 0.9999
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##  0.8264732
```

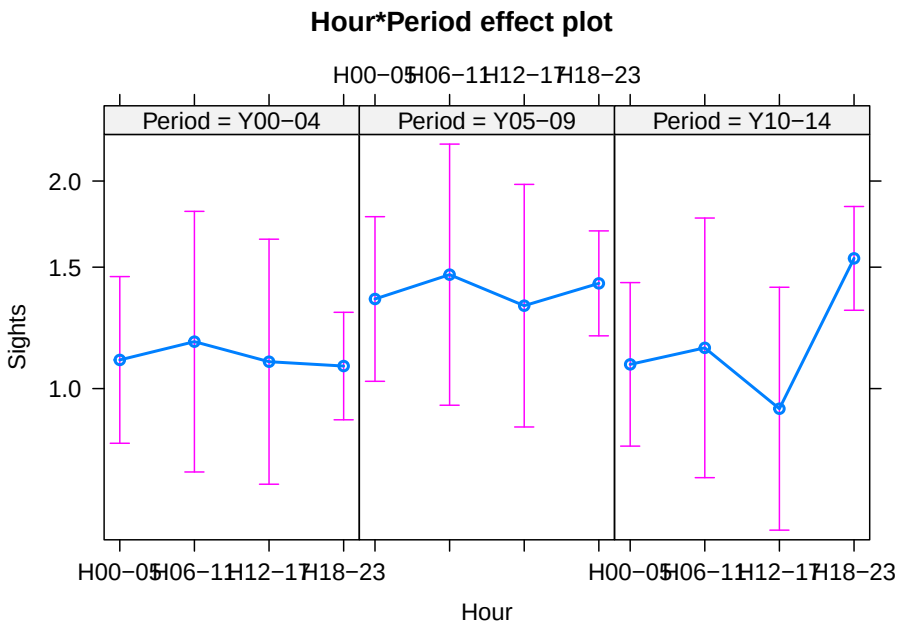
Although the overdispersion value is smaller than 1, according to the p-value we can not reject the null hypothesis of equidispersion. This confirms the goodness of our model.

Model interpretation

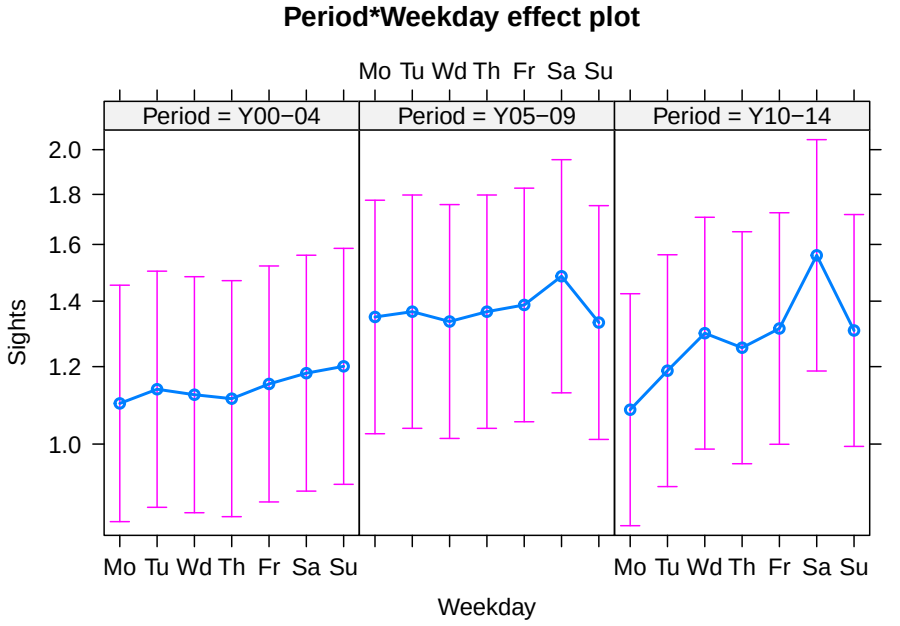
For the model interpretation, we will use the *effect* plots, taking into account the interaction terms. As we are building a Poisson regression model, we can not directly interpret the coefficients and response values, since they do not represent the response variable class but the value of the link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



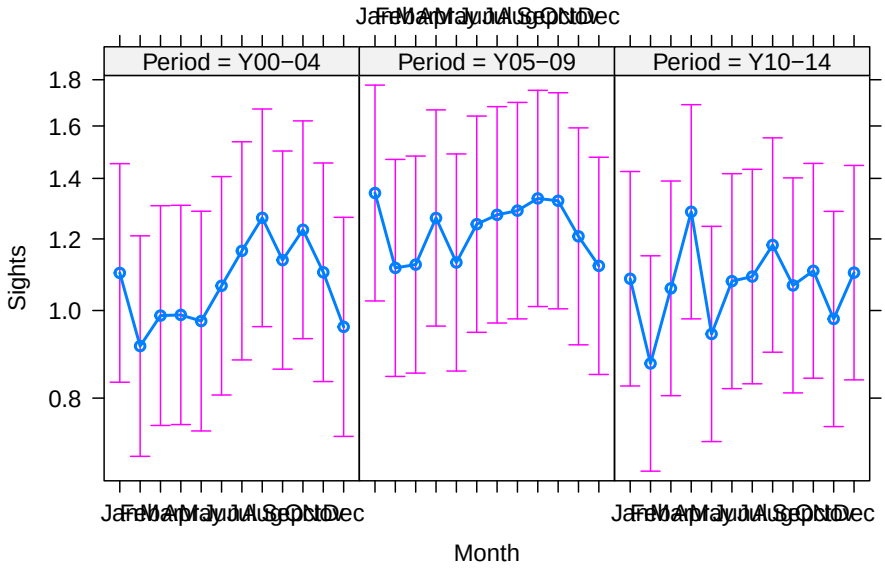
From this effects plots, we can see that the sighting rate per state does not seem to change significantly during the first three hour periods of the day. However, for the last period, from 18 to 23, we have more variance in the rate of sightings per state. This could be due to more people spending more time outdoors.



In this case, we see the mean hour period sighting rate for year period is higher for years 2005-2009. In addition, for year period 2010-2014 we have much higher sighting rate for the hour period from 18 to 23. One reason for this increase could be the bigger availability of *UFO* reporting devices like mobile phones in this year periods.



Again we see that in years 2005-2009 we have a higher mean of sighting rates per weekday. Furthermore, we see an increase towards the end of the week, when probably more people do not work.



Like in the previous plots, the mean for 2005-2009 is slightly higher, but we do not appreciate any other pattern.