UPPSALA
UNIVERSITET

U.U.D.M. Project Report 2016:43

Statistical analysis of trends in climate indicators by means of counts

Erik Jansson

Examensarbete i matematik, 15 hp
Handledare: Jesper Rydén
Examinator: Jörgen Östensson
Oktober 2016

Department of Mathematics
Uppsala University

Statistical analysis of trends in climate indicators by means of counts

Uppsala University
Department of Mathematics
Degree Project C in Mathematics, 15c
Erik Jansson
Erik.jansson.4820@student.uu.se

Supervisor
Jesper Rydén, Lecturer
Uppsala University
Deptartment of Mathematics

October 19, 2016

Abstract

Modelling climate change is an important way to learn what impact our way of living have on this planet. In this report we examine the trend in the climate indicator; *number of very wet days* with data from the last half a century 1961-2011 in five major Swedish cities; *Jönköping, Luleå, Göteborg, Malmö* and the capital city; *Stockholm*. In the first four cities we come to the conclusion that there is a significant, increasing, trend in the *number of very wet days*. In the capital city the increasing trend was shown *not* to be significant.

Key words: *poisson regression, climate indicators, count data, generalized linear models, time series, exponential family.*

Figure 1: A map of Sweden with the geographical location of the 5 cites of interest indicated.

Contents

1 Introduction

In climatic science, studies of extremes are of great interest. Climate indicators such as the number of *very wet days*, *heat waves* or *cold spells* are studied[1]. In this report we are interested in investigating if the number of very wet days per year have significantly change over the last half a century, 1961 to 2011.

We look at No. of very wet days from data presented by ECA&D[2] where the No. of very wet days is defined as the No. of days above the 95-percentile of the daily precipitation amount at wet days (≥ 1 mm).

We will use *generalized linear models* to model this count data which can be assumed to follow the Poisson distribution[3]. We test the assumptions for Poisson distributed variables i.e. equidispersion, $\mu = \sigma^2$. Check for autocorrelation among the error term so we can rule out the possibility of a time dependent relation in the yearly data.

2 Theory

2.1 Generalized linear models

To model count response variables we need to relax the assumption that the response variabels are independently normally distributed, and allow the variables to belong to any member of the exponential family of distributions such as Normal, Poisson, gamma and binomial distributions[4].

If we look at the *general linear model*, not to be confused with *generalized linear models*, we model it by $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$ where $\mathbf{Y}$ is a linear function of the independent variables $\mathbf{X}$ and residuals $\mathbf{e}$, assumed to be independent, normally distributed with the same variance.

The *generalized linear models* is a generalization of the general linear models where we relax the assumption that $\mathbf{Y}$ is normally distributed with constant variance and allow ourselves to look at other distributions from the exponential family. Here, instead of directly modelling $\mu = \mathbb{E}[\mathbf{Y}]$ we need to introduce a function of the linear predictors $\mathbf{X}\beta$, a function $g(\mu) = \mathbf{X}\beta$ where this function $g(\mu)$ is called the link function. Thus we need to specify

1. the distribution,
2. the link function $g(.)$,
3. the linear predictors $\mathbf{X}\beta$

to be able to use the *generalized linear models*.

[1] Rydén [1]

[2] European Climate Assessment & Dataset: `http://eca.knmi.nl/download/millennium/millennium.php`

[3] Nelder et al. [2]

[4] Olsson [3]

2.2 The exponential family

A general class of distributions is the exponential family of distributions, distributions such as Normal, Poisson, gamma and binomial are special cases of this family of distribution. It can be denoted[5]

$$f(y;\theta,\phi) = exp\left[\frac{(y\theta - b(\theta))}{a(\phi)} + c(y,\phi)\right] \tag{1}$$

where a, b, c, and θ are some functions. The function $b(.)$ has a special meaning in the *generalized linear model*. It describes the relationship between the mean and the variance in the model. This can be shown by the *Maximum Likelihood* estimation of the parameters[6].

2.2.1 Poisson distribution

If $Y \sim Po(\mu)$ with intensity parameter $\mu > 0$ and t as the duration of time as the event occurs then Y has the probability density function,

$$\mathbb{P}(Y = y) = \frac{e^{-\mu}\mu^y}{y!} \quad y = 0, 1, 2, \ldots \tag{2}$$

assumed $t = 1$ and for Poisson distributed variables we also assume that $\mathbb{E}[y_i|\mathbf{x_i}] = V[y_i|\mathbf{x_i}] = \mu$ also referred to as *equidispersion*. This is a special case of the exponential family. If we rewrite the *pdf* as

$$\frac{e^{-\mu}\mu^y}{y!} = \exp\left[y\log(\mu) - \mu - \log(y!)\right]$$

we see that this is the same as (1) where

$$\begin{aligned} \theta &= \log(\mu) \\ b(\theta) &= \exp(\theta) \\ c(y,\phi) &= -\log(y!) \\ a(\phi) &= 1 \end{aligned}$$

thus $f(y;\mu) = \exp\left[y\theta - \exp(\theta) - \log(y!)\right]$.

2.2.2 Negative Binomial distribution

If Y is Negative Binomial distributed with parameters $\alpha \geq 0$ and $\theta \geq 0$, $Y \sim Negbin(\alpha, \theta)$ the *probability density function* is then given by

$$\mathbb{P}(Y = k) = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)\Gamma(k+1)}\left(\frac{1}{1+\theta}\right)^{\alpha}\left(\frac{\theta}{1+\theta}\right)^{k}, \quad k = 0, 1, 2, \ldots \tag{3}$$

where the gamma function is defined as $\Gamma(s) = \int_0^\infty z^{s-1}e^{-z}dz, \quad s > 0$.

When we suspect overdispersion is present in the Poisson distributed model

[5] Olsson [3]
[6] Olsson [3]

another model is needed to handle this. An alternative model is then the *Negative Binomial*. One can also show that the Negative Binomial distribution converges to the Poisson distribution[7] so it is a natural choice in dealing with overdispersed Poisson distributed models. The Negative binomial has the same expected mean, $\mathbb{E}[y_i|\mathbf{x_i}] = \mu$ but a scaled variance $V[y_i|\mathbf{x_i}] = \mu + \alpha\mu$ where α is the scale or *dispersion* parameter. In some literatures this model is referred to as the *Negative Binomial 1* or **NB1**[8] for short.

For the Negative binomial distribution we define (1) with

$$\begin{aligned} \theta &= \log\left(1 - \frac{r}{\mu}\right) \\ b(\theta) &= r\left(\theta - \log\left(1 - e^{\theta}\right)\right) \\ c(y, \phi) &= \log\binom{y-1}{r-1} \\ a(\phi) &= 1 \end{aligned}$$

where y is number of trails until we have r recorded successes.

2.3 Link function $g(\mu)$

The link function $g(\mu)$ is used to establish a relationship between the count response Y and the linear predictors $X_1, \ldots, X_n$ in a *generalized linear model*. The link function is chosen based upon the type of data in the model, in our case count data i.e. *discrete* data.

The general form of the function is $g(\mu) = X\beta$. For *Poisson* data we have the link function,

$$g(\mu) = log(\mu)$$

and *Negative Binomial* has a similar link function,

$$g(\mu) = \log\left(1 - \frac{r}{\mu}\right)$$

and as we see in both these cases, the link function is the same as θ from the *Exponential family* therefore it is in some sense a natural choise and is called the *canonical link*[9].

2.4 Goodness-of-fit test

2.4.1 Pearson's χ^2 statistic

For a model y_i with mean μ_i and variance ω_i we estimate the mean and variance and define the *Pearson's χ^2 statistic* as,

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\omega}_i} \sim \chi^2_{df} \tag{4}$$

[7] Winkelmann [4]
[8] Cameron et al. [5]
[9] Olsson [3]

with expected value $E[X^2] = n$ since $E[(y_i - \mu_i)^2/\omega_i] = 1$ if we have correctly specified the mean and variance. This is used to evaluate the model fit to the data. The residual sum of squares used in the normal *linear models* are analogous to Pearson's χ^2. It can also be shown that there is minimal difference between X^2 and the *residual deviance*[10], so they can be used interchangably.

2.4.2 Dispersion test

Equidispersion is a crucial condition for the Poisson distribution and failure of equidispersion leads to the conclusion that another distribution is needed to model the data. A standard model for dealing with overdispersion is the **NB1**[11] with *scaled* variance function $V[y_i|\mathbf{x_i}] = \mu + \alpha\mu$.

We test,

$$H_0 : \alpha = 0, \quad H_1 : \alpha > 0$$

for test of overdispersion. Where α is tested against the corresponding t-statistic which under the null hypothesis is asymtotically standard normal and α is estimated by an auxiliary ordinary least squares regression[12].

A two-sided test is performed by assuming the *alternative hypothesis* $\mathbf{H_1} : \alpha \neq 0$ and to test for *underdispersion* we assume the alternative to be $\mathbf{H_1} : \alpha < 0$.

2.5 Ljung-Box test

A general test of independence in the time series is the Ljung-Box, also called Ljung-Box-Pierce, test which takes into consideration the magnitude of the *autocorrelation* functions as a group. The test statistic is given by,

$$Q = n(n+2)\sum_{h=1}^{H} \frac{\hat{\rho}^2(h)}{n-h} \sim \chi^2_{H-p-q} \tag{5}$$

under the *null hypothesis* of model adequacy, where n large, h is the individual lag, H the choosen group size. For more details see [8].

2.6 Autocorrelation

Given a random variable X, the *autocorrelation function*[13] or **ACF** for short is defined as,

$$\rho(s,t) = \frac{\mathbb{E}\left[(X_s - \mu_s)(X_t - \mu_t\right]}{\sigma_s \sigma_t},$$

where $\rho \in (-1, 1)$ and $\mathbb{E}\left[(X_s - \mu_s)(X_t - \mu_t\right] = Cov(X_s, X_t)$. The function is a measure of the correlation between values in the series at time lag s and lag t.

This is an important tool to rule out the possiblity of *dependence* in the time series count data of interest.

[10] Faraway [6]
[11] Cameron et al. [5]
[12] Cameron et al. [5]
[13] Shumway et al. [7]

3 Model evaluation and Testing

3.1 The models

From the *ECA&D* dataset we have chosen 5 different places in Sweden. All dispersed in the country from north to south, east and west coast and one in the middle of the country. The places have been selected to represent Sweden as a whole. The places we choose to fit to our regression models are,

1 Jönköping Airport (1410)

2 Luleå (5779)

3 Göteborg (0462)

4 Malmö (3509)

5 Stockholm (0010)

the four digits after the name represents the last four digits from the *ECA&D* dataset: *indexR95p00xxxx.txt*

3.2 Data handling

Each dataset begins in the $1960's$ and spans until somewhere between $2011 - 2014$ with a few missing value in each dataset. We could have used some clever method to handle missing value, for example to impute the mean or impute a random value from the dataset but this requires a deeper understanding of these kinds of methods which is not within the scope of this paper. Therefore we have chosen to remove the missing value from the dataset entirely. The datapoints are also divided by 100 to scale the data to the actual No. of very wet days.

3.3 Model Testing

We base our regression model on the assumption that the data are *Poisson* distributed random variables. To test this we perform a *goodness-of-fit test*, testing the *null hypothesis* that the data are *Poisson* distributed against the *alternative hypothesis* that the count data does *not* follow the Poisson distribution.

Table 1: Maximum likelihood based test for testing the hypothesis of data having a *Poisson* distribution with the alternative of *not* being Poisson distributed. If the *p-value* is *less than* 0.05 we reject the *null hypothesis* in favour for the alternative.

Data	χ^2	**Degrees of Freedom**	**p-value**
Jönköping	11.54	9	0.24
Luleå	13.95	10	0.17
Göteborg	19.41	10	0.035*
Malmö	9.22	8	0.32
Stockholm	10.66	11	0.47

As table [1] shows the conclusion can be drawn to reject the hypothesis about *Poisson distribution* for the *Göteborg* data set. We fit it to a *Negative Binomial*

and perform the same likelihood based test with resulting *p-value* 0.0644 > 0.05 thus we cannot reject the hypothesis that the data set *Göteborg* follows a *Negative Binomial* distribution. In table [2] we see that *Malmö* data suffers from *underdispersion* so we assume that the *Negative Binomial* distribution will handle this. With *p-value* 0.224 > 0.05 we cannot reject the *null hypothesis* in the likelihood based test which assumes *Malmö* data follows a *Negative Binomial* distribution.

3.3.1 Time Series Plots

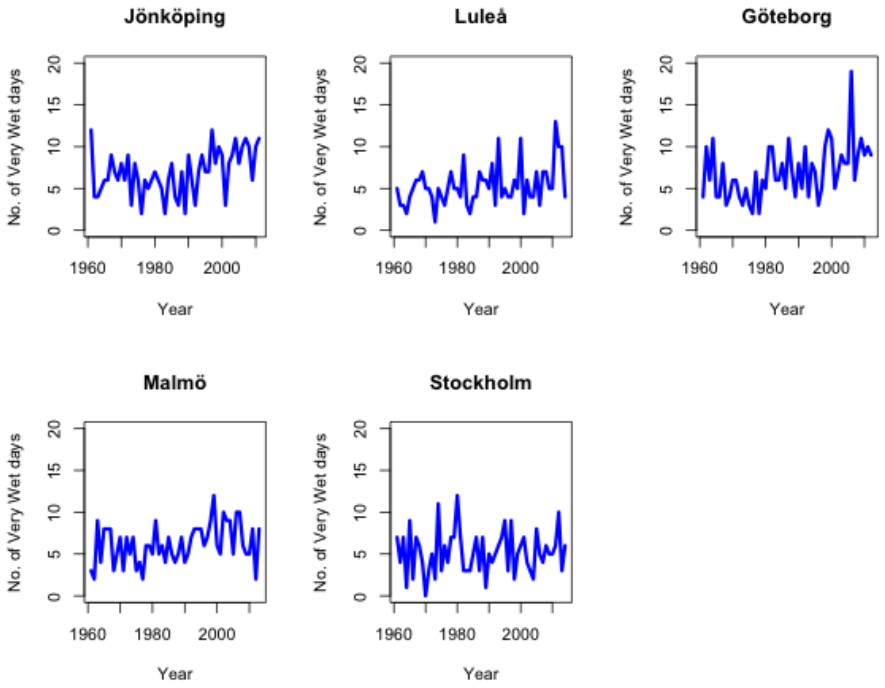

Figure 2: Time Series plot of Number of Very Wet days in Jönköping, Luleå, Göteborg, Malmö and Stockholm

From Figure 2 we might suspect a slightly increasing trend in the *number of very wet days* in *Jönköping*, *Luleå* and *Göteborg* from the 1980's and onwards. In the time series plot of *Malmö* and *Stockholm* no clear pattern can be seen.

3.3.2 Equidispersion

Under the assumption that the data is from the Poisson distribution we need to test the properties of the Poisson distribution i.e. *equidispersion*. In table [1] we can see *mean* and *variance* of each data set, assumed to be equal under the properties of the Poisson distribution. We look at the ratio between the *mean* and *variance* as an indicator of over- or underdispersion. In model 3 and

5 we suspect *overdispersion* whilst in model 4 we might suspect *underdispersion*.

The quotient $\frac{resid.dev}{df}$ is assumed to be 1 if the data is *equidispersed*. After performing a dispersion test, table [1], testing for *equidispersion* we see that we *cannot* reject the *null hypothesis* of *equidispersion* in any case except in model 4: *Malmö*. Where we on the 5% significance level reject *null hypothesis* in favour of the *alternative hypothesis* of *underdispersion* i.e. variance less than the mean.

This leads to a change in model for *Malmö* data to the *Negative Binomial*. A surprising conclusion is that model 3: *Göteborg* is assumed to be *equidispersed* even though there is a huge difference in the mean and variance. This might be because of the sparsity of that data set. Since sets with sparse data might produce deviance values which are unusually low compared to degrees of freedom, thus *goodness-of-fit tests* can produce misleading results in these cases[14].

Table 2: *The mean, variance, quotient of Residual Deviance and degree of freedom, alternative hypothesis and p-value for that hypothesis for each model.*

Data	$\mathbb{E}[\mathbf{Y_t}]$	$\mathbf{Var(Y_t)}$	**Resid.dev/df**	**Alt. Hypothesis**	**p-value**
Jönköping	6.90	7.05	0.96	$\alpha \neq 0$	0.511
Luleå	5.33	5.92	0.916	$\alpha \neq 0$	0.479
Göteborg	6.98	10.13	1.131	$\alpha > 0$	0.292
Malmö	6.20	5.39	0.832	$\alpha < 0$	0.037 *
Stockholm	5.20	6.54	1.38	$\alpha > 0$	0.148

3.3.3 Independence

An important notion of regression analysis is that the residuals of the fitted data are independent therefore we plot the *autocorrelation* function for each data set to see if any dependent structure can be detected. A significant result, $\rho(s,t) > 1.96/\sqrt{n}$ or $\rho(s,t) < -1.96/\sqrt{n}$, would indicate a dependence between Y_s and Y_t where $s \neq t$ and n is the number of data points.

In figure [3], depicting the *autocorrelation functions* for each data sets, we see that we have a significant *autocorrelation* for the *Jönköping* data set at lag 3. This could be because of some phenomenon recurring every third year, like the *El-Niño*[15] effect, but we would expect to see the same effect on at least one of the other data sets. Since we are not seeing any patterns in the other data sets we regard will this as a pure random effect and disregard it in our continued analysis.

In the rest of the data sets we have significance at some higher lags as well, this is most likely due to *mass significance* i.e. when performing enough of these tests it is likely that at least one would give a false positive reading. To

[14] Nelder et al. [2]

[15] "An irregular climate pattern that occurs over a period of three to seven years, which influences ocean and atmospheric currents across the tropical zone of the Pacific Ocean."- Gorse et. al. [9]

manage this problem of *mass significance* we will not be taking these higher lags into consideration. Also the *Ljung-Box* test, table [3], which tests independence in the time series as the *null hypothesis* cannot be rejected for any of the 5 data sets. This confirms our assumption of that each of the five data sets are independent.

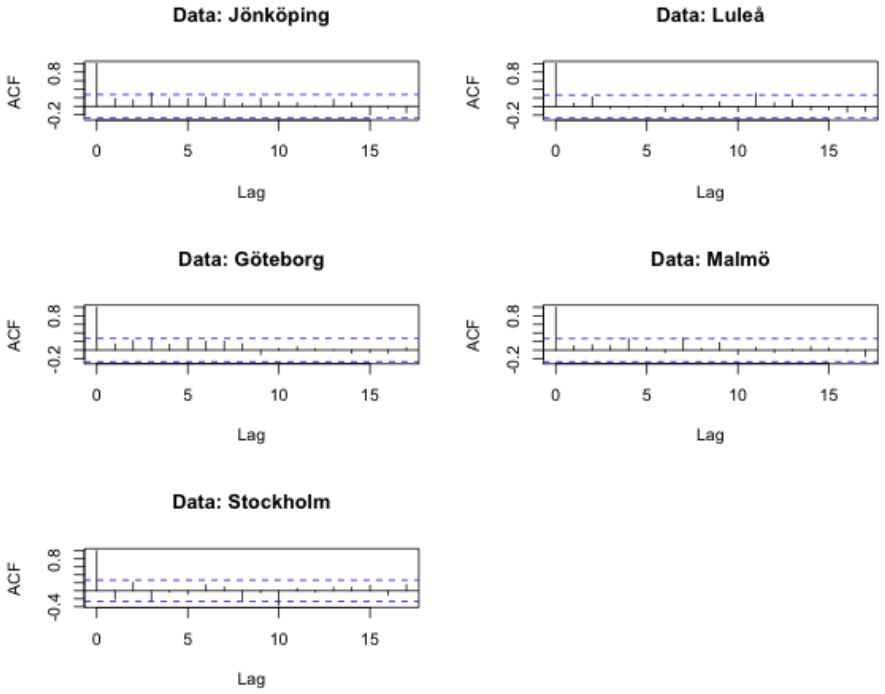

Figure 3: The *autocorrelation function*, ACF, for each of the five data sets.

Table 3: The Ljung-Box test for independence in a given time series.

Data	**Test statistic: χ^2**	**p-value**
Jönköping	1.70	0.191
Luleå	0.32	0.571
Göteborg	1.09	0.294
Malmö	0.53	0.463
Stockholm	2.57	0.108

3.4 Regression analysis

After removing the missing data points from the datasets we begin by fitting a *Poisson regression* model to each of the five models. We fit $Y_t = e^{\beta_0+\beta_1 t}$ to each model where Y_t is the *number of very wet days* and t is the year.

We continue by testing the assumption for data being *Poisson distributed* i.e.

Table 4: Summary of the regression analysis for the 5 data sets.

Data	**Est.** β_1	**Std. Err.**	**p-value**
Jönköping	0.0100	0.0036	0.00584^{**}
Luleå	0.0106	0.0038	0.00515^{**}
Göteborg	0.0135	0.0037	0.000268^{***}
Malmö	0.0078	0.0036	0.0292^{*}
Stockholm	0.0018	0.0038	0.62

equidispersion and *goodness-of-fit test*. If we find the assumptions being fulfilled we move on with analyzing the results.

If not we change the model to the *negative binomial* to deal with the problem of *over- and underdispersion* and then start analyzing the results.

The conditional mean Y_t is then modeled according to,

$$\mu = \mathbb{E}[Y|t] = e^{\beta_0 + \beta_1 t} \tag{6}$$

for all five data sets. As we can see in table [4] time t is a significant factor in the *number of very wet days* with the exception for the *Stockholm* data. We can also conclude that the *number of very wet days* has a positive, increasing, trend since $\beta_1 > 0$.

The interpretation of β_1 is that as time t increases by one unit (*year*), Y_t or the *number of very wet days* increase by e^{β_1} units (*days*).

A typical way of detecting *model misspecification* in the case of *linear regression* is residual analysis, where a residual has a clear definition of begin the the difference between fitted and actual value. In the case of *generalized linear models* there is no unique definition of a residual. We have chosen not to delve deeper into the field of residual analysis in this paper. Some examples of different residuals are the *Pearson*, *Deviance*, *Anscombe* and the *Generalized residual*[16].

4 Discussion

In this report we examined the *number of very wet days* in five major cities in Sweden; *Jönköping, Luleå, Göteborg, Malmö* and *Stockholm*, to see if we could find any significant trend over the last half a century 1961 to 2011. In four out of the five cities we got a significant result that the *number of very wet days* are increasing with time. There could be many reasons for this increase in the *number of very wet days* but something that has changed dramatically over the last 50 years is the level of *carbon dioxide,* CO_2 in the athmosphere.

The measurements are regarded as count variables so special statistical methods where employed; *Generalized Linear Models* and especially *Poisson regression*, to deal with this particular data type. As the measurements are measured over time, special consideration where taken to make sure the data was independent

[16] Cameron et al. [5]

of time. The *Ljung-Box* test is essentially a test which assumes continuous data and we are dealing with discrete datasets, so this could be a source of discrepency in the results. Further investigation is needed to conclude if this is a problem.

A continuation of this project would be to dig deeper into the reason why *Stockholm* data did not have a significant trend as the rest of the data. Could this be because of where the city is geographically located, Swedish east coast, or could there be some other explanation. Another thing to investigate is the positive trend in relation to *carbon emission* into the athmosphere, to see if this is a contributing factor to the increase in *number of very wet days*.

References

[1] Rydén, J. (2015). *A statistical analysis of trends for warm and cold spells in Uppsala by means of counts* Physical Geography, 97, 431-436. *DOI: 10.1111/geoa.12083*

[2] McCullagh, P., Nelder, J.A. (1989). *Generalized Linear Models* 2:nd ed. CRC.

[3] Olsson, U (2002). *Generalized Linear Models: An applied approach.* Lund, Studentlitteratur.

[4] Winkelmann, R. (2008). *Econometric analysis of count data* 5:th ed. Springer.

[5] Cameron, A.C., Trevedi, P.K. (2013). *Regression Analysis of Count Data* 2:nd ed. New York, Cambridge University Press.

[6] Faraway, J. J. (2006). *Extending the Linear Model with R.* Chapman & Hall/CRC.

[7] Shumway, R.H.,Stoffer, D.S (2011). *Time Series Analysis and Its Application: With R Examples* 3:rd ed. Springer.

[8] Ljung, G., Box, G. (1978). *On a Measure of Lack of Fit in Time Series Models.* Biometrika, 65(2), 297-303.

[9] Gorse, C., Johnston, D., Pritchard, M. (2012). *A Dictionary of Construction, Surveying and Civil Engineering* Oxford University Press.

A Regression analysis

A.1 Model 1: Jönköping

```
Call:
glm(formula = V3 ~ V2, family = "poisson", data = jonkoping_flygplats)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.24508  -0.49514   0.01301   0.54073   2.48722

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -18.034621   7.246661  -2.489  0.01282 *
V2            0.010048   0.003645   2.757  0.00584 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 54.886  on 50  degrees of freedom
Residual deviance: 47.236  on 49  degrees of freedom
AIC: 240.37

Number of Fisher Scoring iterations: 4
```

A.2 Model 2: Luleå

```
Call:
glm(formula = V3 ~ V2, family = "poisson", data = lulea)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.00057  -0.71684  -0.06252   0.36645   2.12712

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -19.535625   7.586501  -2.575  0.01002 *
V2            0.010665   0.003812   2.798  0.00515 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 55.552  on 53  degrees of freedom
Residual deviance: 47.661  on 52  degrees of freedom
AIC: 237.9

Number of Fisher Scoring iterations: 4
```

A.3 Model 3: Göteborg

A.3.1 Assuming Poisson distribution

```
Call:
glm(formula = V3 ~ V2, family = "poisson", data = goteborg_a)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9607  -0.7185  -0.2130   0.4594   2.9292

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -25.135130   7.045096  -3.568  0.00036 ***
V2            0.013621   0.003541   3.847  0.00012 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 71.534  on 51  degrees of freedom
Residual deviance: 56.552  on 50  degrees of freedom
AIC: 253.24

Number of Fisher Scoring iterations: 4
```

A.3.2 Assuming Negative Binomial distribution

```
Call:
glm.nb(formula = V3 ~ V2, data = goteborg_a, init.theta = 68.94226543,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8800  -0.6924  -0.1995   0.4379   2.7115

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -24.981378   7.393249  -3.379 0.000728 ***
V2            0.013543   0.003716   3.644 0.000268 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(68.9423) family taken to be 1)

    Null deviance: 64.807  on 51  degrees of freedom
Residual deviance: 51.286  on 50  degrees of freedom
AIC: 254.96

Number of Fisher Scoring iterations: 1
```

```
             Theta:  69
         Std. Err.:  140

 2 x log-likelihood:  -248.957
```

A.4 Model 4: Malmö

A.4.1 Assuming Poisson distribution

```
Call:
glm(formula = V3 ~ V2, family = "poisson", data = malmo_a)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.39236  -0.75476   0.01188   0.62485   1.80838

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.867903   7.199514  -1.926   0.0541 .
V2            0.007895   0.003620   2.181   0.0292 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 47.238  on 52  degrees of freedom
Residual deviance: 42.461  on 51  degrees of freedom
AIC: 238.06

Number of Fisher Scoring iterations: 4
```

A.4.2 Assuming Negative Binomial distribution

```
Call:
glm.nb(formula = V3 ~ V2, data = malmo_a, init.theta = 118837.4245,
    link = log)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.39230  -0.75474   0.01188   0.62483   1.80831

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.867920   7.199711  -1.926   0.0541 .
V2            0.007895   0.003620   2.181   0.0292 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(118837.4) family taken to be 1)
```

```
    Null deviance: 47.236  on 52  degrees of freedom
Residual deviance: 42.459  on 51  degrees of freedom
AIC: 240.06

Number of Fisher Scoring iterations: 1


              Theta:  118837
          Std. Err.:  2534629
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -234.064
```

A.5 Model 5: Stockholm

```
Call:
glm(formula = V3 ~ V2, family = "poisson", data = sthlm)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1722  -1.0205  -0.1040   0.7532   2.5808

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.122351   7.611204  -0.279     0.78
V2           0.001898   0.003829   0.496     0.62

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 72.066  on 53  degrees of freedom
Residual deviance: 71.820  on 52  degrees of freedom
AIC: 257.16

Number of Fisher Scoring iterations: 5
```

B R Code

```
# set working directory
setwd("/Users/Erik/Desktop/Kandidatuppsats/ECA_indexR95p/")

#loading packages
library(graphics)
# loading package 'AER' for dispersion test
library(AER)
# loading package 'vcd' for Goodness-of-fit test
library(vcd)
# loading package 'MASS' for GLM NB
library(MASS)
```

```
# read data
jonkoping_flygplats = read.table("indexR95p001410.txt", skip = 30)[c('V2','V3')] # 1961-20
lulea = read.table("indexR95p005779.txt", skip = 30)[c('V2','V3')] # 1961-2011
goteborg_a = read.table("indexR95p000462.txt", skip = 30)[c('V2','V3')] # 1961-2012
malmo_a = read.table("indexR95p003509.txt", skip = 30)[c('V2','V3')] # 1961-2013
sthlm = read.table("indexR95p000010.txt", skip = 30)[c('V2','V3')] #1961-2014



# change all missing values, coded -999999, to NA
jonkoping_flygplats[jonkoping_flygplats<0] = NA
lulea[lulea<0] = NA
goteborg_a[goteborg_a<0] = NA
malmo_a[malmo_a<0] = NA
sthlm[sthlm<0] = NA


# removing all missing values from dataset
jonkoping_flygplats = na.omit(jonkoping_flygplats)
lulea = na.omit(lulea)
goteborg_a = na.omit(goteborg_a)
malmo_a = na.omit(malmo_a)
sthlm = na.omit(sthlm)


# scaling the data back by 1/100 to get the accual counts
jonkoping_flygplats$V3 = jonkoping_flygplats$V3*0.01
lulea$V3 = lulea$V3*0.01
goteborg_a$V3 = goteborg_a$V3*0.01
malmo_a$V3 = malmo_a$V3*0.01
sthlm$V3 = sthlm$V3*0.01

# time series plot of the data
par(mfrow=c(2,3))
plot(jonkoping_flygplats$V2,jonkoping_flygplats$V3,ylim=c(0,20), main="Jönköping",
type="l", col="blue", lwd=2, xlab="Year", ylab="No. of Very Wet days")
plot(lulea$V2,lulea$V3, type="l", main="Luleå",
ylim=c(0,20), xlab="Year", col="blue", lwd=2, ylab="No. of Very Wet days")
plot(goteborg_a$V2, goteborg_a$V3, type="l", ylim=c(0,20), main="Göteborg",
col="blue", lwd=2, xlab="Year", ylab="No. of Very Wet days")
plot(malmo_a$V2, malmo_a$V3, type="l", ylim=c(0,20), main="Malmö",
col="blue", lwd=2, xlab="Year", ylab="No. of Very Wet days")
plot(sthlm$V2,sthlm$V3,type="l",ylim=c(0,20),main="Stockholm",
col="blue", lwd=2, xlab="Year", ylab="No. of Very Wet days")
par(mfrow=c(1,1))

#regression upon the data: '# very wet days' = A + B*year
# assuming the data begin Poisson distributed
m1 = glm(V3 ~ V2, family = "poisson", data = jonkoping_flygplats)
```

```
m2 = glm(V3 ~ V2, family = "poisson", data = lulea)
m3 = glm(V3 ~ V2, family = "poisson", data = goteborg_a)
m4 = glm(V3 ~ V2, family = "poisson", data = malmo_a)
m5 = glm(V3 ~ V2, family = "poisson", data = sthlm)



# Goodness-of-fit test
gf1 = goodfit(jonkoping_flygplats$V3, type = "poisson", method = "ML")
gf2 = goodfit(lulea$V3, type = "poisson", method = "ML")
gf3 = goodfit(goteborg_a$V3, type = "poisson", method = "ML")
gf4 = goodfit(malmo_a$V3, type = "poisson", method = "ML")
gf5 = goodfit(sthlm$V3, type = "poisson", method = "ML")

# Negbin
m3.nb <- glm.nb(V3 ~ V2, data = goteborg_a)
gf3.nb <- goodfit(goteborg_a$V3,type = "nbinomial", method = "ML")
m4.nb <- glm.nb(V3 ~ V2,data = malmo_a)
gf4.nb <- goodfit(malmo_a$V3,type="nbinomial", method = "ML")


# plot of GOF-test
plot(gf1, main = "Count data vs Poisson distr. for Jonkopings airport");grid()
plot(gf2, main = "Count data vs Poisson distr. for Lulea");grid()
plot(gf3, main = "Count data vs Poisson distr. for Goteborg A");grid()
plot(gf4, main = "Count data vs Poisson distr. for Malmo A");grid()
plot(gf5, main = "Count data vs Poisson distr. for Stockholm");grid()

# Ljung-Box test for independence
library(stats)
lb1 <- Box.test(jonkoping_flygplats$V3,type = "Ljung-Box");
lb2 <- Box.test(lulea$V3,type = "Ljung-Box");
lb3 <- Box.test(goteborg_a$V3,type = "Ljung-Box");
lb4 <- Box.test(malmo_a$V3,type = "Ljung-Box");
lb5 <- Box.test(sthlm$V3,type = "Ljung-Box");

# autocorrelation
par(mfrow=c(3,2))
acf(jonkoping_flygplats$V3,main="Data: Jönköping")
acf(lulea$V3,main="Data: Luleå")
acf(goteborg_a$V3,main="Data: Göteborg")
acf(malmo_a$V3,main="Data: Malmö")
acf(sthlm$V3,main="Data: Stockholm")
par(mfrow=c(1,1))
```