Computer exercise 4
Poisson Regression

Please write your names and "personal identification numbers" here. During the exercise fill in the blanks marked by black bullets and answer the posed questions. To pass the exercise, all questions should be answered and handed in to the computer exercise supervisor.

•

All necessary files are downloadable from the course home page
http://www.math.chalmers.se/Stat/Grundutb/CTH/mve240/0809/files/data.zip.
Please download the data.zip file and uncompress it at the directory you plan to use for the computer exercises.

When dealing with two or more variables, the functional relation between the variables is often of interest. For count data, one model that is frequently used is the Poisson regression model and applications are found in most sciences: technology, medicine etc. The Poisson regression model is also implemented in many packages for statistical analysis of data.

In this computer exercise you will learn more about:

(1) The Poisson regression model and how to estimate the model parameters

(2) Model selection, i.e. the number of explanatory variables to use

(3) Confidence intervals and the delta method

1 Preparatory exercises

1. Read the instructions for the computer exercise and chapter 7.1-7.3 and 8.3 in the book.
2. Try to explain the difference between linear regression and Poisson regression.

•

2 Road accident data

The Swedish Road Administration is the national authority that has the overall responsibility for the entire road transport system. One main issue is road safetybility and continuous work to improve road safety is performed. From their internet site `http://www.vv.se` it is possible to obtain a number of different statistics about road accidents[1]. We will in this exercise use

[1] Another good source for all kinds of statistics about transport and communications is the Swedish Institute For Transport and Communications Analysis, `http://www.sika-institute.se/`.

traffic accident data from years 1950-2004, given in Årsdata. 1950-2004.xls. The data is used to fit a Poisson regression model to the number of people perished in traffic accidents, cf. Example 7.16 in the book. The estimated model is then used to predict the expected number of perished year 2016.
You are encourage to use more recent data which includes the years 2005-2008. This can probably be found at

`http://www.vv.se/Startsida-foretag/Trafiken/Skade--och-olycksdata/\\`
`Skade---olycksstatistik1/Nationell-statistik/Arsdata-fran-1950/`

Download the file,

`http://www.vv.se/PageFiles/8407/arsdata_1950_2008.xls`

open it and replace '1994*' by '1994'. Then save it in your personal matlab directory. Finally import it to the Matlab workspace with the command

```
>> data = xlsread('arsdata_1950_2008_3.xls');
>> Nyear=2008;
```

Otherwise use the data from years 1950-2004

```
>> data = xlsread('Årsdata. 1950-2004.xls');
>> Nyear=2004;
```

The variable data now consists of 9 columns but we are only interested in columns $\{1, 2, 5, 6\}$, i.e. {year, number of people killed, number of cars, amount of sold petrol}. We store the data in a structure array

```
>> traffic = struct('year',data(:,1),'killed',data(:,2),'cars',data(:,5),...
             'petrol',data(:,6));
```

Plot the number of people killed each year

```
>> plot(traffic.year, traffic.killed, 'o')
```

Try also plotting the number of people killed vs. number of cars and the petrol consumption. Do you se any connection?

From the plot it can be seen that the trend of increasing number of people killed is broken around year 1965. And from year 1970 the number starts to decrease. Why did the number of people killed increase in years 1950-1965? What was the reason for the brake of the increasing trend? (Hint: right-side driving (1967), front seat-belts in new cars (1969), mandatory use of front seat-belts (1975)).

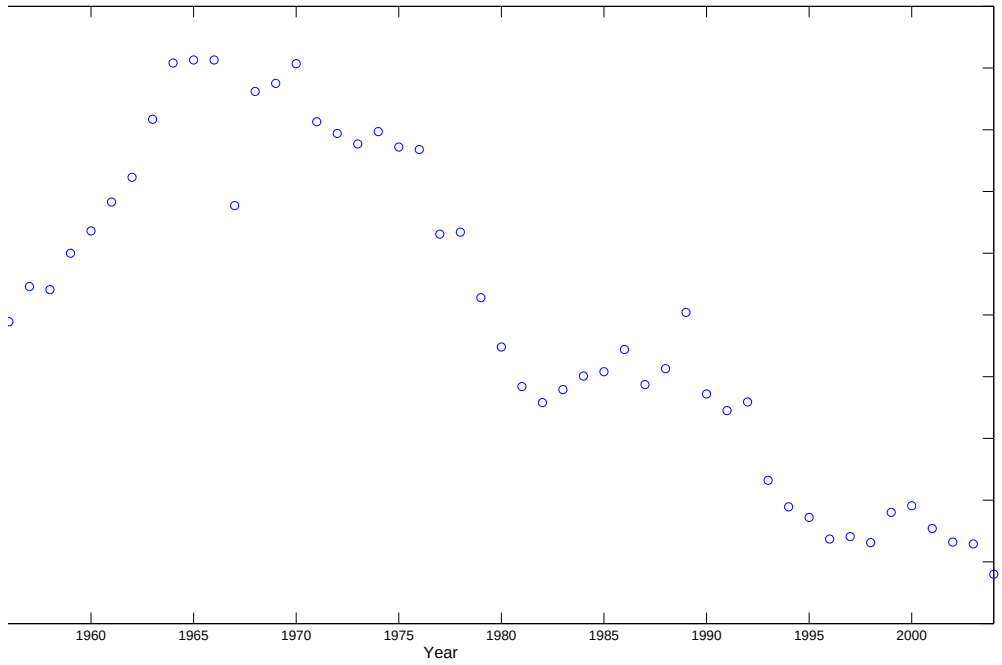

Figure 1: The number of people killed per year in road accidents in Sweden from year 1950 to year 2004. (Source: The Swedish Road Administration.)

3 The Poisson regression model

Lets say we have a sequence of count data, n_i, $i = 1, \ldots, k$, for some event, i.e. the number of perished in traffic accidents in a year. This count data is assumed to be observations from random variables $N_i \in \text{Po}(\mu_i)$, (called responses or dependent variables) with mean value $\mu_i = \mu_i(x_{i1}, \ldots, x_{ip})$. The variables, $x_{i1}, \ldots, x_{ip}$, are called explanatory variables[2] and are assumed to measure factors that influence the count data.
We restrict μ_i to be a log-linear function[3],

$$\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}) \tag{1}$$

And thus the probability that $N_i = n$ is,

$$\mathsf{P}(N_i = n) = \frac{e^{-\mu_i}(\mu_i)^n}{n!} = \frac{e^{-e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}}(e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}})^n}{n!}, \; n = 0, 1, 2, \ldots. \tag{2}$$

3.1 Estimating model parameters $\beta_0, \ldots, \beta_p$

To simplify the notation we introduce $x_{i0} = 1$ and can now write (1) as,

$$\mathsf{E}[N_i] = \mu_i = \exp(\sum_{j=0}^{p} \beta_j x_{ij}), \tag{3}$$

where $N_i \in \text{Po}(\mu_i)$ for $i = 1, \ldots, k$.

[2]Several other names exist in the literature: independent variables, regressor variables, predictor variables.
[3]Sometimes the model incorporates an extra term t_i: $\mu_i = t_i \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_2 x_{ip})$.

The likelihood function is calculated as,

$$L(\beta) = \prod_{i=1}^{k} \mathsf{P}(N_i = n_i) = \prod_{i=1}^{k} \frac{\mu_i^{n_i}}{n_i!} e^{-\mu_i}. \tag{4}$$

where $\mu_i = \mu_i(\beta_{\mathbf{p}})$ is a function of $\beta_{\mathbf{p}} = (\beta_0, \ldots, \beta_p)$. The ML-estimates $\beta^*_{\mathbf{p}} = (\beta^*_0, \ldots, \beta^*_p)$ are the values of β that maximize the likelihood function $L(\beta)$. Often it is easier to maximize the log-likelihood function,

$$l(\beta) = -\sum_{i=1}^{k} \log(n_i!) + \sum_{i=1}^{k} n_i \log(\mu_i) - \sum_{i=1}^{k} \mu_i. \tag{5}$$

By setting the first order derivates of the log-likelihood equal to zero, we get a system of $(p+1)$ non-linear equations in β_j,

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{k} \frac{\partial \mu_i}{\partial \beta_j} (\frac{n_i}{\mu_i} - 1) = \sum_{i=1}^{k} (n_i - \mu_i) x_{ij} = 0, \; j = 0, \ldots, p. \tag{6}$$

Usually, the equation system must be solved with some numerical method, e.g. the Newton-Raphson algorithm, cf. section 7.3.3. in the course book. This is also the method implemented in the function `poiss_regress`. Use the command `>> type poiss_regress` to see the code.

4 Poisson regression of traffic data

We will now try to fit the Poisson regression model to the traffic data of the number of people killed in road accidents. Above, we could see that there was a break in the trend of increasing number people killed around year 1965-1975, mainly because of the improvement in car safety due to the use of safety belts. Because of this it seems reasonable to fit our model to data starting from year 1975. Why?

•

```
>> traffic = struct('year',data(26:end,1),'killed',data(26:end,2),...
             'cars',data(26:end,5),'petrol',data(26:end,6));
```

Which are the explanatory variables? And the response?

•

Redraw the plot from above for the reduced data set

```
>> plot(traffic.year,traffic.killed,'o')
>> figure(1), hold on
```

We start the analysis with one explanatory variable, `traffic.year`. What is the variable year supposed to measure?

•

```
>> X1 = [traffic.year-mean(traffic.year)];
>> n = traffic.killed;
>> beta1 = poiss_regress(X1,n,1e-6);
>> my_fit = glmval(beta1, X1,'log');
>> plot(traffic.year, my_fit, 'b-')
```

What is your estimate of β? Convince yourself that this is the solution to (6). Judging from the plot, is this model sufficient to describe the number of people killed in traffic accidents?

•

Although this simple model seems to capture the overall trend, adding further explanatory variables may improve the fit. Thus, we try adding the number of cars as a variable in our model.

```
>> X2 = [traffic.year-mean(traffic.year), traffic.cars-mean(traffic.cars)];
>> beta2 = poiss_regress(X2,n,1e-6);
>> my_fit = glmval(beta2, X2,'log');
>> plot(traffic.year, my_fit, 'g-')
```

Have your estimates β_0^* and β_1^* changed? Does the number of cars improve the fit?

•

It seems reasonable also to add the quantity of sold petrol as this would reflect the total mileage of all cars[4].

```
>> X3 = [traffic.year-mean(traffic.year), traffic.cars-mean(traffic.cars),...
        traffic.petrol-mean(traffic.petrol)];
>> beta3 = poiss_regress(X3,n,1e-6);
>> my_fit = glmval(beta3, X3,'log');
>> plot(traffic.year, my_fit, 'r-')
```

Have your estimates of β changed now? Use the command `format long` to display more digits. Which model do you choose?

•

Poisson regression model belongs to a class of models called generalized linear models. In a generalized linear model (GLM), the mean of the response, μ, is modeled as a monotonic (non-linear) transformation of a linear function of the explanatory variables, $g(\beta_0 + \beta_1 x_1 + \beta_2 x_2 ...)$. The inverse of the transformation function g is called the " canonical link function". In Poisson regression this function is the log function, but in other GLM's different link functions are used, see `doc glmfit` for a list of supported link functions in the Matlab function `glmfit`[5]. Also, the response may take different distributions, such as the normal or the binomial distribution.

[4] Assuming that the mean fuel consumption of a car has been constant over the years - a 1970 year model of a Volvo used about 10l per 100km which is approximately the same as for the 2000 year model. Of course, the year 2000 model has more than twice the horsepower.

[5] `glmfit` uses a method called weighted least squares to compute the β estimates.

4.1 Model selection - Deviance

It is not always easy to decide, just by looking at the plot, which model to choose. Even though adding more variables improves the fit, it also increases the uncertainty of the estimates. One method to choose complexity of the model is to use the deviance and a hypothesis test.
Let $\beta^*_{\mathbf{p}} = \{\beta^*_0, \beta^*_1, \ldots, \beta^*_p\}$ be the ML-estimates of the model parameters $\{\beta_0, \beta_1, \ldots, \beta_p\}$ of the full model with p explanatory variables and $\beta^*_{\mathbf{q}}$ the estimates of a simpler model where only q ($q < p$) of the explanatory variables have been used. Then for large k, and under suitable regularity conditions, the deviance

$$\text{DEV} = 2 \cdot (l(\beta^*_{\mathbf{p}}) - l((\beta^*_{\mathbf{q}}))) \tag{7}$$

is approximately $\chi^2(p-q)$ distributed if the less complex model is true. Thus, it is possible to test if the simpler model can be rejected compared to the full model. How? (use `chi2inv` to get the quantiles of the χ^2 distribution)

- $\chi^2_\alpha =$

The deviance for model 3 compared to model 2 is calculated as

```
>> DEV2 = 2*traffic.killed'*([ones(length(traffic.killed),1),X3]*beta3-...
          [ones(length(traffic.killed),1),X2]*beta2)
```

Is the improvement with model 3 significant compared to model 2?

-

Repeat the test for model 2 against model 1 and also model 3 against model 1? Which model do you choose?

5 Prediction

Now we want to use our model to predict the expected number of perished in traffic accidents ten years from now, i.e. year 2016 . In order to do this we first must have an estimate of the number of cars that year. Start by plotting the number of cars vs. year,

```
>> figure(2)
>> plot(traffic.year, traffic.cars, 'o')
>> hold on
```

We will here use a simple linear model for the number of cars, y_i, year x_i

$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i \tag{8}$$

where the errors, $\varepsilon_i \in \text{N}(0, (\sigma_\varepsilon)^2)$, are assumed to be independent and identically distributed. This is called a linear regression model. It is possible to estimate the parameters with the maximum likelihood method similar as for the Poisson regression model above. What is the likelihood function?

- $L(\beta) =$

In Matlab, the function `regress` computes the least-squares (LS) estimates of the linear regression model. In the case of ε_i being normally distributed, the LS method is equivalent to the ML method with exactly the same estimates.

```
>> phat = regress(traffic.cars,[ones(length(traffic.cars),1) [1975:Nyear]'])
>> plot(1975:2016, phat(1)+phat(2)*[1975:2016],'r')
>> cars_2016=phat(1)+phat(2)*2016;
```

Evaluate the fit by looking at the residuals.

```
>> res = traffic.cars-(phat(1)+phat(2)*traffic.year);
>> figure(3), plot(traffic.year,res,'o')
>> figure(4), normplot(res)
```

Do the residuals conform to the requirements of the model errors ε_i?

-

However, for our purpose this rough estimate of the number of cars year 2016 is sufficient. The expected number of perished can now be predicted using (1),

```
>> x = [1 2016-mean(traffic.year) cars_2016-mean(traffic.cars) ]'
>> my_2016 = exp(beta2'*x)    %----- Model 2 -----
```

Is the prediction reasonable? Is i possible to predict the number of perished for, say, year 2100?

-

5.1 Optional: Confidence interval - Delta method

Under some regularity conditions, Theorem 4.4 in the course book gives that the ML-estimates, β^*, are asymptotically multivariate normal distributed. Based on this and with the use of Taylor expansion it can be shown that also the error distribution is asymptotically normal with zero mean and variance equal to $(\sigma_\epsilon^*)^2$,

$$\mathcal{E} = \mu(\beta) - \mu(\beta^*) \in \mathrm{AsN}(0, (\sigma_\varepsilon^*)^2). \tag{9}$$

Thus we can construct an interval that with approximately $1-\alpha$ confidence contains $\mu(\beta)$,

$$I_\mu = [\mu(\beta^*) + \lambda_{1-\alpha/2} \cdot \sigma_\varepsilon^*,\ \ \mu(\beta^*) + \lambda_{\alpha/2} \cdot \sigma_\varepsilon^*] \tag{10}$$

The standard deviation of the error, σ_ε^*, can be computed using Gauss' formula,

$$(\sigma_\epsilon^*)^2 \approx \nabla\mu(\beta^*)^\mathsf{T}\Sigma^*\nabla\mu(\beta^*) = (\frac{\partial\mu}{\partial\beta_0}, \ldots, \frac{\partial\mu}{\partial\beta_d})^\mathsf{T}\Sigma^*(\frac{\partial\mu}{\partial\beta_0}, \ldots, \frac{\partial\mu}{\partial\beta_d}) \tag{11}$$

where $\Sigma^* = [(\sigma_{ij}^*)^2] = [-\ddot{l}(\beta^*)]^{-1}$ is the covariance matrix of the estimated $d+1$ model parameters, i.e. $(\sigma_{ij}^*)^2 = \mathsf{Cov}(\beta_i^*, \beta_j^*)$ with $0 \le i, j \le d$.

```
>> nabla = my_2016*x
>> sigma_star = covm(X2,beta2)
>> sd2 = nabla'*sigma_star*nabla
>> ci = [my_2016-1.64*sqrt(sd2) my_2016+1.64*sqrt(sd2)]
```

What is the approximate confidence of the interval?

- $1 - \alpha =$

5.2 Prediction interval

Note that in the previous subsection we derived an approximative confidence interval for the expected number $\mu_i = \mathsf{E}[N_i]$ of persons killed in traffic accidents year i, $(i = 2016)$. Obviously the actual number of killed in traffic year i will be, with high probability, different from the expected number. In our model we assumed that N_i is Poisson distributed with mean μ_i. However, since μ_i is surely bigger than 10 we can say that $N_i \approx N(\mu_i, \mu_i)$ or equivalently

$$N_i = \mu_i + \epsilon_i,$$

where $\epsilon_i \approx N(0, \mu_i^*)$, by means of the normal approximation of the Poisson distribution. Now using (9) we can write that

$$N_i = \mu_i^* + \ \mathcal{E} + \epsilon_i,$$

and assume that $\mathcal{E}$ and ϵ_i are approximately independent and normally distributed, viz.

$$N_i \approx N(\mu_i^*, (\sigma_\varepsilon^*)^2 + \mu_i^*). \tag{12}$$

The prediction interval predicts the range of variability of the variable of interest in our case N_i. Using (12) we can write down the approximative $1 - \alpha$ prediction interval

$$I_{N_i} = \left[\mu_i^* + \lambda_{1-\alpha/2} \cdot \sqrt{(\sigma_\varepsilon^*)^2 + \mu_i^*},\ \ \mu_i^* - \lambda_{\alpha/2} \cdot \sqrt{(\sigma_\varepsilon^*)^2 + \mu_i^*}\right]. \tag{13}$$

Compute the prediction interval for N_{2009}

```
>> cars_2009=phat(1)+phat(2)*2009;
>> x = [1 2009-mean(traffic.year) cars_2009-mean(traffic.cars) ]'
>> my_2009 = exp(beta2'*x)    %----- Model 2 -----
>> nabla = my_2009*x
>> sigma_star = covm(X2,beta2)
>> sd2 = nabla'*sigma_star*nabla
>> ci = [my_2009-1.64*sqrt(sd2+my_2009) my_2009+1.64*sqrt(sd2+my_2009)]
```

Does the observed value for year 2009 (approx 355) lie within the interval?

-