MLDS 401/IEMS 404-1 (Fall 2023): Lab 8 – 11/20/2023

9.2 (MLE equations for exponential and Poisson distributions) Derive the equations for finding the MLE of the regression parameter vector $\boldsymbol{\beta}$ if the response variable distribution is exponential or Poisson, and show that they have the form (9.6).

For the Poisson distribution the log-likelihood function is

$$L = \prod_{i=1}^{n} \left[\frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \right],$$

where $\ln \lambda_i = \boldsymbol{x}_i' \boldsymbol{\beta}$ or $\lambda_i = \exp(\boldsymbol{x}_i' \boldsymbol{\beta})$. Hence

$$\ln L = -\sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n} y_i \ln \lambda_i - \sum_{i=1}^{n} \ln y_i!.$$

Set the derivative of $\ln L$ w.r.t. $\boldsymbol{\beta}$ equal to zero. Using the fact that $d(\ln \lambda_i)/d\boldsymbol{\beta} = d(\boldsymbol{x}_i' \boldsymbol{\beta})/d\boldsymbol{\beta} = \boldsymbol{x}_i$ and $d\lambda_i/d\boldsymbol{\beta} = d(\exp(\boldsymbol{x}_i' \boldsymbol{\beta}))/d\boldsymbol{\beta} = \exp(\boldsymbol{x}_i' \boldsymbol{\beta}) \boldsymbol{x}_i = \lambda_i \boldsymbol{x}_i$, we get

$$\frac{d \ln L}{d\boldsymbol{\beta}} = -\sum_{i=1}^{n} \lambda_i \boldsymbol{x}_i + \sum_{i=1}^{n} y_i \boldsymbol{x}_i = \mathbf{0}.$$

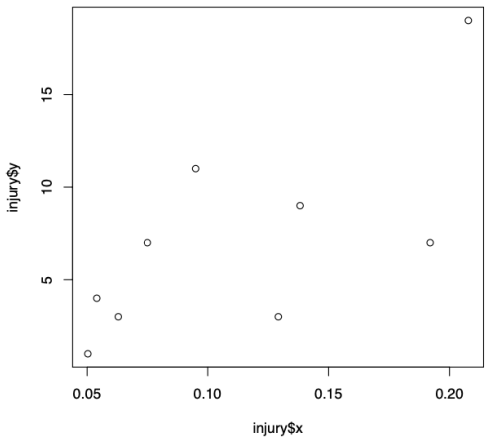
Noting that $\lambda_i = \mu_i$ for the Poisson distribution, the above equation is equivalent to $\boldsymbol{X}' \boldsymbol{\mu} = \boldsymbol{X}' \boldsymbol{y}$.

Thus both equations are of the form (9.6).

9.3 (Airline injury incidents) The file `Airline-Injury.csv` contains data from Chatterjee and Hadi (2012) on the number of injury incidents and the proportion of the total number of flights out of New York for 9 airlines.

Fit three different models to predict the number of injury incidents (y) as a function of the proportion of the total number of flights (x): (1) simple linear regression without any transformation of y, (2) simple linear regression with square-root transformation of y and (3) Poisson regression. Calculate the SSE for each model. Which model do you prefer and why?

(a) As the proportion of flights increases, the number of injuries increases. The variance seems to increase with the number of injuries.

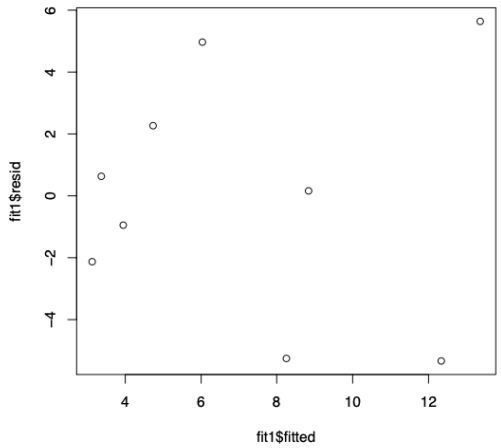


(b) **(1) Simple linear regression without transformation:**

Here is the output. The SSE = 123.5302. The residual plot exhibits increasing variance with fitted values.

```
lm(formula = y ~ x, data = injury)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3351 -2.1281  0.1605  2.2670  5.6382

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1402     3.1412  -0.045   0.9657
x            64.9755    25.1959   2.579   0.0365 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.201 on 7 degrees of freedom
Multiple R-squared:  0.4872,    Adjusted R-squared:  0.4139
F-statistic:  6.65 on 1 and 7 DF,  p-value: 0.03654

> SSE1 = sum((injury$y-fit1$fitted)^2) # SSE
> SSE1
[1] 123.5302
```

The residual plot is as follows.

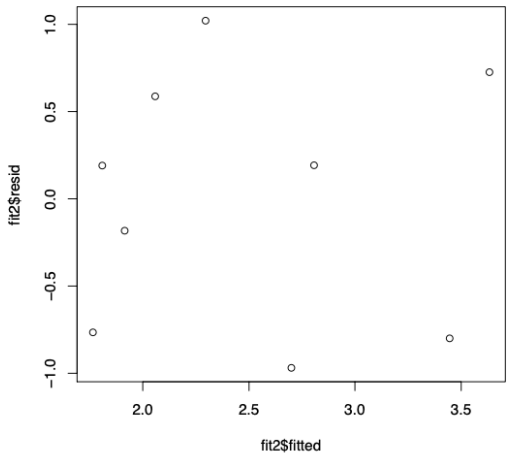


(2) Simple linear regression with square-root transformation:

Here is the output. The SSE = 123.0247, which is slightly less than the previous SSE. The residual plot still exhibits increasing variance with fitted values.

```
Call:
lm(formula = sqrt(y) ~ x, data = injury)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9690 -0.7655  0.1906  0.5874  1.0211

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.1692     0.5783   2.022   0.0829 .
x            11.8564     4.6382   2.556   0.0378 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7733 on 7 degrees of freedom
Multiple R-squared:  0.4828,    Adjusted R-squared:  0.4089
F-statistic: 6.535 on 1 and 7 DF,  p-value: 0.03776

> SSE2 = sum((injury$y-fit2$fitted^2)^2) # SSE
> SSE2
[1] 123.0247
```

The residual plot is as follows.

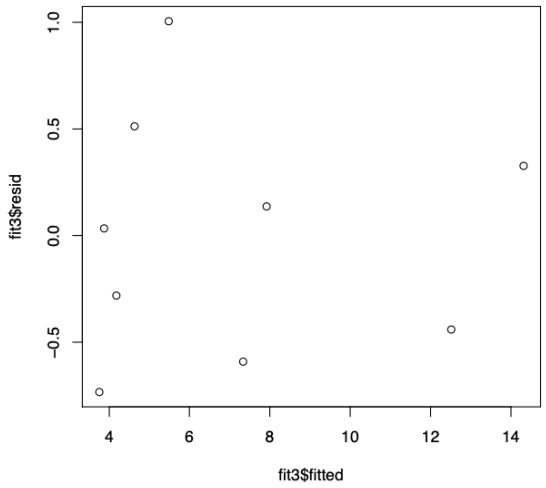


(3) Poisson regression:

Here is the output. The SSE = 117.3472, which is significantly less than the previous SSE. The residual plot does not exhibit increasing variance with fitted values. So this fit is the best of the three fits.

```
Call:
glm(formula = y ~ x, family = poisson(log), data = injury)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.81894  -1.69082   0.06495   1.02407   2.06811

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.8945     0.3265   2.739  0.00615 **
x             8.5018     2.1575   3.941 8.13e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 31.859  on 8  degrees of freedom
Residual deviance: 16.291  on 7  degrees of freedom
AIC: 52.251

Number of Fisher Scoring iterations: 5

> SSE3 = sum((injury$y-fit3$fitted)^2) # SSE
> SSE3
[1] 117.3472
```

The residual plot is as follows.



Question 4 (4 pts.): A Poisson regression model is developed to estimate the number of bugs in a software (y) as a function of four variables: x_1 = total number of lines of code (in thousands), x_2 = number of programmers in the team, x_3 = indicator variable for whether the team is domestic ($x_3 = 0$) or global ($x_3 = 1$) and x_4 = the complexity of the software rated on a scale of 1 (least) to 10 (most). The estimated coefficients for the full model are

$$\hat{\beta}_0 = 0.500, \hat{\beta}_1 = 0.001, \hat{\beta}_2 = 0.005, \hat{\beta}_3 = 0.010, \hat{\beta}_4 = 0.0030.$$

Suppose a particular software rated 10 on complexity has a million lines of code, and a team of 20 global programmers is developing it. What is the probability that the software will be bug-free?

Answer:

$$\ln(\widehat{\lambda}) = 0.500 + 0.001 \times 1000 + 0.005 \times 20 + 0.010 \times 1 + 0.0030 \times 10 = 1.64.$$

So, $\widehat{\lambda} = e^{1.64} = 5.155$ and probability of zero bugs equal to $e^{-5.155} = 0.0058$.

Question 9 (4 pts.): Refer to the previous question. Because there were excess zeros, a Zero Inflated Poisson (ZIP) model was fitted to predict the number of absences as a function of gender, math score and program. There were 3 programs and the mean number of absences by program were 10.650 (prog=1), 6.934 (prog=2) and 2.673 (prog=3). The output is shown below.

```
zeroinfl(formula = daysabs ~ factor(prog) + math | factor(gender), data = absdays)

Count model coefficients (poisson with log link):
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.591796   0.061496  42.146  < 2e-16 ***
factor(prog)2 -0.304307   0.056742  -5.363 8.18e-08 ***
factor(prog)3 -0.969535   0.082076 -11.813  < 2e-16 ***
math          -0.005302   0.000934  -5.676 1.38e-08 ***

Zero-inflation model coefficients (binomial with logit link):
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.0246     0.2638  -7.674 1.67e-14 ***
factor(gender)male   0.7689     0.3303   2.328   0.0199 *
```

Predict the number of absent days for a male student with math = 50 who is in prog=2.

Answer: For the Poisson regression part the expected number of zero absences is

$$\ln \widehat{\lambda} = 2.5918 - 0.3043 - 0.0054 \times 50 = 2.0225, \quad \text{hence} \quad \widehat{\lambda} = e^{2.0225} = 7.5572.$$

Next for the logisitc regression part

$$\ln\left(\frac{\widehat{p}}{1-\widehat{p}}\right) = -2.0246 + 0.7689 = -1.2557, \quad \text{hence} \quad \widehat{p} = \frac{e^{-1.2557}}{1+e^{-1.2557}} = 0.2217,$$

where $\widehat{p}$ is the estimated probability of 0 absences. Combining the two models we get the estimated number of zero absences $= 7.5572 \times (1 - 0.2217) = 5.8818$..