Regression Models

Master in Statistics for Data Sciences

20th December 2023

Name/s:

A We will then consider data on lung cancer occurrence among Spanish men between 1968-1971. The data.frame contains 24 observations corresponding to the following variables:

- `Cases`: Number of cases of lung cancer
- `Pop`: Population in each city for each age group
- `Age`: Categorical variable with 6 categories: Age40-54, Age55-59, Age60-64, Age65-69, Age70-74, Age75+
- `City`: Categorical variable with 4 categories: A,B,C y D (corresponding to the 4 cities)

Let the number of lung cancer cases Y_i, $i = 1, \ldots, 24$ be a Poisson random variable with mean μ_i; t_i is the corresponding population, and η_i is the linear predictor

Given the following output:

```
model1 <- glm(Cases ~ offset(log(Pop))+Age+City , family=poisson, data=cancer)
summary(model1)
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.6321     0.2003 -28.125  < 2e-16 ***
Age55-59     1.1010     0.2483   4.434 9.23e-06 ***
Age60-64     1.5186     0.2316   6.556 5.53e-11 ***
Age65-69     1.7677     0.2294   7.704 1.31e-14 ***
Age70-74     1.8569     0.2353   7.891 3.00e-15 ***
Age75+       1.4197     0.2503   5.672 1.41e-08 ***
CityB       -0.3301     0.1815  -1.818   0.0690 .
CityC       -0.3715     0.1878  -1.978   0.0479 *
CityD       -0.2723     0.1879  -1.450   0.1472
---
Null deviance: 129.908  on 23  degrees of freedom
Residual deviance:  23.447  on 15  degrees of freedom
```

(a) (1 point) Interpret the coefficient of `CityC`.

For people in the same age group, the incidence (rate of cases) of lung cancer in City C is $0.69 = e^{-0.371}$ times smaller than in city A

(b) (1 point) What would be the expected number of cases, per 100000 inhabitants in city A and each group 60-64?

$$\hat{\mu} = e^{-5.632+1.518} \times 100000 \approx 1634$$

(c) (1 point) Which age group has the largest incidence rate? Which city has the lowest?

- Age group with largest incidence: 70-74
- City with lowest incidence : City C

(d) (1 point) Explain how would you carry out a goodness of fit test for this model (indicate how would you calculate the p-value)

Use the deviance test:

H_0: The model is good enough
H_1: The model is not good enough

The test statistic would be: $Deviance_{model1} \approx \chi^2_{15}$

The p-value is calculated as: $Pr(\chi^2_5 > 23.44)$

Now we fit the following model:

```
model2 <- glm(Cases ~ offset(log(Pop))+Age*City, family=poisson, data=cancer)
anova(model2,test="Chisq")
Terms added sequentially (first to last)

          Df  Deviance   Resid. Df Resid. Dev Pr(>Chi)
NULL                         23    129.908
Age        5  101.601         18     28.307  < 2e-16 ***
City       3    4.859         15     23.447  0.18241
Age:City  15   23.447          0      0.000  0.07509 .
```

e) (1 point) Is the difference in cancer rate between age groups 55-59 and 60-64 is constant in all cities?. Justify your answer

The answer is yes, since the interaction term is not significant

g) (1 point) Which model is best?. Explain how you chose it
The best model is the one with the variable Age only.

First we compare the models with and without interaction:

H_0: Model with Age+City
H_1: Model with Age+City+Age:City

The p-value of this test is $Pr(\chi^2_{15} > 23.44) = 0.07$, therefore, we do not reject the null hypothesis

Now we compare:

H_0: Model with Age
H_1: Model with Age+City

The p-value of this test is $Pr(\chi^2_3 > 4.859) = 0.18$, therefore, we do not reject the null hypothesis

Finally, we compare :

H_0: Null model
H_1: Model with Age

he p-value of this test is $Pr(\chi^2_5 > 101.01) \approx 0$, therefore we reject the null hypothesis, and the best model s the model that includes Age

B. Consider worldwide airline fatalities for the period 1976?1985. Data are given in the table below, where we find number of annual fatal accidents, number of annual passenger deaths and the annual number of recorded passenger miles (100 million). Note that the volume of air traffic nearly doubled over this 10-year period.

Year	Fatal accidents	Passenger deaths	Passenger miles (10^8)
1976	24	734	3863
1977	25	516	4300
1978	31	754	5027
1979	31	877	5481
1980	22	814	5814
1981	21	362	6033
1982	26	764	5877
1983	20	809	6223
1984	16	223	7433
1985	22	1066	7107

Two models are fitted

```
summary(m1)
Call:
glm(formula = fatals ~ Year + offset(log(miles)), family = poisson,
data = flights)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 201.32999 45.62333 4.413 1.02e-05 ***
time         -0.10442 0.02304 -4.532 5.84e-06 ***

Null deviance: 26.1320 on 9 degrees of freedom
Residual deviance: 5.4551 on 8 degrees of freedom
AIC: 59.424

summary(m2)
Call:
glm(formula = deaths ~ Year + offset(log(miles)), family = poisson,
data = flights)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 117.767873 8.420250 13.99 <2e-16 ***
time         -0.060522 0.004252 -14.23 <2e-16 ***

Null deviance: 1253.6 on 9 degrees of freedom
Residual deviance: 1051.4 on 8 degrees of freedom
```

1. (1 point) By how much has increased/decreased the rate of accidents (per 100 million miles) between 1980 and 1983?

   The model fitted for accidents is:

   $$log(\hat{\mu}) = 201.32 - 0.104 \times Year + log(miles)$$

   Then

   $$\hat{rate} = 201.32 - 0.104 \times Year$$

and so:

$$\begin{aligned}
\hat{r}ate_{1980} &= e^{201.32-0.104\times 1980} \\
\hat{r}ate_{1983} &= e^{201.32-0.104\times 1983} \\
\hat{r}ate_{1983} &= e^{-0.104\times 3} \times \hat{r}ate_{1980} \\
\hat{r}ate_{1983} &= 0.73 \times \hat{r}ate_{1980}
\end{aligned}$$

2. (1 point) What is the difference in the predicted number of deaths between 1978 and 1981?

   The model fitted for deaths is:

$$log(\hat{\mu}) = 117.75 - 0.064 \times Year + log(miles)$$

   and so:

$$\begin{aligned}
\widehat{r}ate_{1978} &= e^{117.77-0.06\times 1978} \\
\hat{d}eaths_{1978} &= e^{117.77-0.06\times 1978} \times 5027 \approx 2023 \\
\hat{r}ate_{1981} &= e^{117.77-0.06\times 1981} \\
\hat{d}eaths_{1981} &= e^{117.77-0.06\times 1981} \times 6033 \approx 2028
\end{aligned}$$

   So, the difference is 5 deaths

3. (1 point) Investigate goodness of fit for the models using the following quantiles (in brackets are the corresponding degrees of freedom): $\chi^2_{0.05}(1) = 3.84$, $\chi^2_{0.05}(8) = 15.51$, $\chi^2_{0.05}(10) = 18.31$. If a model is not working well, try to think of an explanation (in terms of Poisson processes and aircraft sizes)

   For model $m1$ the residual deviance $D = 5.4551 < 15.51 = \chi^2_{0.05}(8)$. Hence, we have a good fit.
   For model $m2$ the residual deviance $D = 1051.4 > 15.51$, therefore, we have a bad fit,

   Model $m2$ does not fit well due to overdispersion. The the residual deviance for the passenger death model is very large compared to the degrees of freedom. This large degree of overdispersion is due to compounding with the aircraft size, as each fatal accident leads to some multiple number of deaths. So, while the fatal accident data is consistent with an underlying Poisson process giving Poisson counts for the number of fatal accidents each year, the number of passenger deaths each year is a Poisson sum of random variables from the aircraft size distribution, which is not Poisson distributed