Homework 4 P1 & P2

Amy Ly

2/9/2023

Problem 1 Context

The datset utilized involves car accidents at a particular T-intersection. The dataset has the following features:

- y = number of accidents
- presence = an indicator variable representing whether a refuge lane was present (1) or not (0)
- left = how many cars are turning left
- ADT = volume of traffic at the T-intersection
- length = length of the refuge lane

Problem 1

Consider the model with prescence as the only predictor and log(leftxADT) as the offset. Find the predicted value for the average number of accidents for a T-intersection with a refuge lane and log(leftxADT)=15.

```r
accident <- read.table("accident.txt")
names(accident)=c("y", "presence", "left","ADT", "length")

mod <- glm(y~presence +offset(log(left*ADT)), data=accident, family=poisson(link="log"))

theta <- mod$coef[1]+mod$coef[2]*1 + 15
lambda_hat <- exp(theta) # predicted average

V <- vcov(mod)
z <- qnorm(0.95)

g.prime <- c(1, 1)

theta.sd <- sqrt(g.prime%*%V%*%g.prime)

CI <- exp(c(theta-z*theta.sd, theta +z*theta.sd))
```

The predicted value is 1.5582137 accidents with a 95% confidence interval of [3.3955295, 6.645684] accidents.

Problem 2 Context

The dataset utilized records the number of cancer deaths among survivors of the atomic bombs dropped on Japan during WWII, categorized by time (years) after the bomb occured and by the amount of radiation exposure that the survivors received from the blast. Additionally, information about person-years (in the 100s) at risk was included and is defined as the sum total of all years spent by all persons in a category.

Suppose the mean number of cancer deaths in each cell is Poisson with mean $\mu =$ risk x rate, where risk is the person-years at risk and rate is the rate of cancer deaths per person per year. It is desired to described this rate in terms of the amount of radiation, adjusting for the effects of time after exposure.

Data Wrangling

```
df <- read.table("HW4-CancerDeathsData.txt", comment="", header=TRUE)
# to convert the txt file to csv
#write.csv(tab, "HW4-CancerDeathsData.csv", row.names=FALSE, quote=FALSE)

x <- df$YearsAfter

years <- as.numeric(unlist(strsplit(x, "to")))

#take the difference between years/2 and then add to the lower bound year
df$Time <- diff(years)[c(TRUE, FALSE)]/2 + years[c(TRUE, FALSE)]

df$YearsAfter <- as.factor(df$YearsAfter)
```

Part a

Using log(risk) as an offset, fit the Poisson log-linear regression model with time after blast treated as a factor (with seven levels) and with rads treated as a numerical covariate.

```
mod2 <- glm(Deaths~Exposure +YearsAfter + offset(log(AtRisk)), data=df, family=poisson(link="log"))

mod2
```

```
##
## Call:  glm(formula = Deaths ~ Exposure + YearsAfter + offset(log(AtRisk)),
##     family = poisson(link = "log"), data = df)
##
## Coefficients:
##      (Intercept)          Exposure  YearsAfter12to15  YearsAfter16to19
##        -3.214552          0.001832          0.551699          1.248244
## YearsAfter20to23  YearsAfter24to27  YearsAfter28to31   YearsAfter8to11
##         1.403878          1.736657          2.031144          0.233327
##
## Degrees of Freedom: 41 Total (i.e. Null);  34 Residual
## Null Deviance:       335.7
## Residual Deviance: 50.11     AIC: 215.9
```

Exposure is the number of rads that a person is exposed to. For every 1 increase in rads, the rate of cancer deaths on a log scale changes by a multiplicative factor after adjusting for the effects of time after exposure.

Part b

Try the same model as in part a, but have time be the calculated midpoint of each interval and include log(time) as a numerical explanatory variable.

```
mod3 <- glm(Deaths~Exposure + log(Time) + offset(log(AtRisk)), data=df, family=poisson(link="log"))

anova(mod2, mod3, test="LRT")
```

```
## Analysis of Deviance Table
## 
## Model 1: Deaths ~ Exposure + YearsAfter + offset(log(AtRisk))
## Model 2: Deaths ~ Exposure + log(Time) + offset(log(AtRisk))
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)    
## 1        34     50.106                          
## 2        39     77.077 -5  -26.971 5.78e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

According to the LRT, there is evidence to suggest that the more complex model is a better fit (p-value = 5.78eˆ-0.5). We should include log(Time) as a covariate.

Part c

Try fitting a model that includes the interaction of log(time) and exposure.

```
mod4 <- glm(Deaths~Exposure*log(Time) + offset(log(AtRisk)), data=df, family=poisson(link="log"))

anova(mod3, mod4, test = "LRT")
```

```
## Analysis of Deviance Table
## 
## Model 1: Deaths ~ Exposure + log(Time) + offset(log(AtRisk))
## Model 2: Deaths ~ Exposure * log(Time) + offset(log(AtRisk))
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        39     77.077                     
## 2        38     75.957  1   1.1195     0.29
```

According to the LRT, there is little evidence to suggest that the interaction between Exposure and log(time) is significant (p-value = 0.29).

Part d

Based on a good-fitting model, make a statement about the effect of the radiation exposure on the number of cancer deaths per person per year.

First I need to find a good model.

```
mod_a <- glm(Deaths~Exposure, data=df, family=poisson(link="log"))
mod_a$aic
```

```
## [1] 583.9174
```

```
# with offset
mod_b <- glm(Deaths~Exposure + offset(log(AtRisk)), data=df, family=poisson(link="log"))
mod_b$aic
```

```
## [1] 474.752
```

```
mod_c <- glm(Deaths~Exposure +YearsAfter + offset(log(AtRisk)), data=df, family=poisson(link="log"))
mod_c$aic
```

```
## [1] 215.8553
```

```
mod_d <- glm(Deaths~Exposure + Time + offset(log(AtRisk)), data=df, family=poisson(link="log"))
mod_d$aic
```

```
## [1] 211.7934
```

```
mod_e <- glm(Deaths~Exposure + log(Time) + offset(log(AtRisk)), data=df, family=poisson(link="log"))
mod_e$aic
```

```
## [1] 232.8259
```

```
mod_f <- glm(Deaths~Exposure + log(Time) + log(AtRisk), data=df, family=poisson(link="log"))
mod_f$aic
```

```
## [1] 233.3064
```

```
mod_g <- glm(Deaths~Exposure + log(AtRisk), data=df, family=poisson(link="log"))
mod_g$aic
```

```
## [1] 469.3992
```

```
mod_h <-glm(Deaths~Exposure + log(Time), data=df, family=poisson(link="log"))
mod_h$aic
```

```
## [1] 378.823
```

The best model would be Deaths~Exposure + Time + offset(log(AtRisk)) AIC = 211.7934

followed by

Deaths~Exposure + YearsAfter + offset(log(AtRisk)) AIC = 215.8553

```
mod_d
```

```
##
## Call:  glm(formula = Deaths ~ Exposure + Time + offset(log(AtRisk)),
##     family = poisson(link = "log"), data = df)
##
## Coefficients:
## (Intercept)     Exposure         Time
##   -3.603193     0.001833     0.082960
##
## Degrees of Freedom: 41 Total (i.e. Null);  39 Residual
## Null Deviance:       335.7
## Residual Deviance: 56.04     AIC: 211.8
```

```
beta <- mod_d$coef[3]
V2 <- vcov(mod_d)
CI2 <- c(mod_d$coef[3]-1.96*sqrt(V2[3,3]),
  mod_d$coef[3]+1.96*sqrt(V2[3,3]))
```

A relevant parameter would be the Time (after exposure).

It is estimated to be 0.08296 with a 95% CI of [0.0722058, 0.0937142].

For every 1 year increase after exposure (while keeping the exposure constant), the death rate on the log scale changes by a multiplicative factor of 0.08296.