Homework 7

Camille Okonkwo

2024-03-30

The `hwdata3.csv` contains data on 1959 patients hospitalized subsequent to undergoing either the CABG or PTCA cardiovascular procedure.

The variables included are:

procedure: 1 = CABG, 0 = PTCA

gender: 1=male, 0 = female

age: age of patient

los: hospital length of stay(days)

type: 1 = emergency/urgent, 0 = elective

1. Fit a Poisson regression model with hospital length of stay as the outcome and procedure, gender, age, and type as covariates (model 1). Write down the model. Is overdispersion a potential problem for this Poisson model?

```
# Fit the Poisson regression model
p.model <- glm(los ~ procedure + gender + age + type, data = hwdata3, family = "poisson")

summary(p.model)
```

```
##
## Call:
## glm(formula = los ~ procedure + gender + age + type, family = "poisson",
##     data = hwdata3)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.554710   0.150124   3.695  0.00022 ***
## procedure    1.121862   0.018971  59.137  < 2e-16 ***
## gender      -0.102885   0.018197  -5.654 1.57e-08 ***
## age          0.010251   0.002086   4.913 8.96e-07 ***
## type         0.189919   0.016844  11.275  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 7541.9  on 1958  degrees of freedom
## Residual deviance: 3471.0  on 1954  degrees of freedom
## AIC: 10457
##
## Number of Fisher Scoring iterations: 5
```

```r
# goodness of fit
pchisq(p.model$deviance, p.model$df.residual, lower.tail=F)
```

```
## [1] 3.946248e-88
```

```r
# deviance
p.model$deviance
```

```
## [1] 3470.967
```

```r
p.model$df.residual
```

```
## [1] 1954
```

```r
sum(resid(p.model,type="pearson")^2)
```

```
## [1] 4342.122
```

```r
ratio = (p.model$deviance) / p.model$df.residual
ratio
```

```
## [1] 1.776339
```

The model formula **los** ~ **procedure** + **gender** + **age** + **type**. Since the ratio of deviance to df is 1.776, which is greater than 1, this suggests potential overdispersion in the model.

2. Refit model 1 with the scale parameter being equal to Pearson chi-square divided by residual DF. Estimate the length of stay rate ratio between patients undergoing CABG and PTCA procedures. Provide the 95% confidence interval and interpret.

```r
p.model2 <- glm(los ~ procedure + gender + age + type, data = hwdata3, family = "quasipoisson")

summary(p.model2)
```

```
##
## Call:
## glm(formula = los ~ procedure + gender + age + type, family = "quasipoisson",
##     data = hwdata3)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.55471    0.22379   2.479 0.013270 *
## procedure    1.12186    0.02828  39.671  < 2e-16 ***
## gender      -0.10289    0.02713  -3.793 0.000153 ***
## age          0.01025    0.00311   3.296 0.000999 ***
## type         0.18992    0.02511   7.564 5.99e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 2.222171)
##
##     Null deviance: 7541.9  on 1958  degrees of freedom
## Residual deviance: 3471.0  on 1954  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```r
coef_procedure <- coef(p.model2)["procedure"]
rate_ratio <- exp(coef_procedure)
```

```r
conf_interval <- exp(confint(p.model2)["procedure", ])
```

```
## Waiting for profiling to be done...
```

3. Use the fitted model in part (2), calculate the expected days of hospital stay for male patients aged 68 who underwent CABG procedure and stayed in an elective type.

```r
# Define the coefficients from the fitted model
intercept <- 0.55471
coef_procedure <- 1.12186
coef_gender <- -0.10289
coef_age <- 0.01025
coef_type <- 0.18992

# Define the values for the specific patient
procedure <- 1  # CABG procedure
gender <- 1  # Male
age <- 68
type <- 0  # Elective type

# Calculate the linear predictor
linear_predictor <- intercept + coef_procedure * procedure +
                    coef_gender * gender + coef_age * age + coef_type * type

# Calculate the expected number of days of hospital stay
expected_stay <- exp(linear_predictor)
expected_stay
```

```
## [1] 9.685985
```

4. Refit model 1 using negative binomial regression. Provide a formal test to decide whether a negative binomial model is needed for this data than a Poisson regression model.

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
# Fit the negative binomial regression model
nb.model = glm.nb(los ~ procedure + gender + age + type, data = hwdata3)

summary(nb.model)
```

```
##
## Call:
## glm.nb(formula = los ~ procedure + gender + age + type, data = hwdata3,
##     init.theta = 9.292017288, link = log)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.54270    0.20759   2.614 0.008942 **
## procedure    1.13067    0.02425  46.633  < 2e-16 ***
## gender      -0.10723    0.02528  -4.242 2.22e-05 ***
```

```
## age          0.01020    0.00289   3.529 0.000418 ***
## type         0.21810    0.02338   9.328  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(9.292) family taken to be 1)
##
##     Null deviance: 4231.5  on 1958  degrees of freedom
## Residual deviance: 1886.3  on 1954  degrees of freedom
## AIC: 9927.3
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  9.292
##           Std. Err.:  0.668
##
##  2 x log-likelihood:  -9915.253
```

```
test.stat = nb.model$theta/nb.model$SE.theta

p_value = 1-pnorm(test.stat)
```

At the 5% level of significance, (and a p-value of 0 which is less than 0.05), we reject the null hypothesis that the Poisson model is sufficient, indicating that a negative binomial model is preferred. This result suggests that the negative binomial model provides a significantly better fit for the data compared to the Poisson model.

5. Use the negative binomial model to estimate the length of stay rate ratio between patients undergoing CABG and PTCA procedures and provide 95% confidence interval. Is the conclusion different from the Poisson model in part (2)?

```
# Extract coefficients from the fitted negative binomial model
coef_nb <- coef(nb.model)

# Coefficient for procedure (CABG vs. PTCA)
coef_procedure_nb <- coef_nb["procedure"]

# Standard error of the coefficient for procedure
se_procedure_nb <- summary(nb.model)$coefficients["procedure", "Std. Error"]

# Calculate the rate ratio
rate_ratio_nb <- exp(coef_procedure_nb)

# Calculate the 95% confidence interval
lower_limit_nb <- exp(coef_procedure_nb - 1.96 * se_procedure_nb)
upper_limit_nb <- exp(coef_procedure_nb + 1.96 * se_procedure_nb)

# Print the rate ratio and confidence interval
print(rate_ratio_nb)
```

```
## procedure
##  3.097718
```

```
print(c(lower_limit_nb, upper_limit_nb))
```

```
## procedure procedure
```

```
##  2.953951  3.248482
```

The results from binomial model and the Poisson model are very similar. Since the 95% confidence intervals of the rate ratios from the negative binomial model and the Poisson model overlap significantly, there may not be a big difference in the estimated rate ratios between the two models.