

4330 Assignment 4

Ravish Kamath: 213893664

05 December, 2022

Question 1

The file `reform.csv` has a cross-sectional subsample from the German Socio-Economic Panel, which collected data on doctor visits before and after a major health care reform that took place in 1997. The reform increased the copayments for prescription drugs by up to 200% and imposed upper limits on the reimbursement of physicians by the state insurance. The outcome is the number of doctor visits in a three month period. The descriptions of the variables are:

- id**: The patient's ID number
- numvisit**: Number of doctor visits in a 3-month period
- reform**: Before (`reform=0`) or after (`reform=1`) the reform
- badh**: Person is in bad health? (1=yes, 0=no)
- age**: Age in years
- educ**: Education in years
- loginc**: Logarithm of income

- (a) [5 points] Fit a Poisson regression model (with all possible predictors included, other than `id`) to see whether the reform affected the number of doctor visits. Formulate the model so that the estimated rates are per 1-month of follow-up. Report the rate ratio for the reform variable and interpret.
- (b) [2 points] Calculate and interpret the rate ratio for age.
- (c) [2 points] Estimate the 1-month rate of visits for a 30 year old person before the reform, with 12 years education, in bad health, and with an average income (so that `loginc = 7.6989`). Be sure to specify the proper units of the rate.
- (d) [1 point] Estimate the 1-year rate for the person from part (c).
- (e) [2 points] Compute the ratio of 1-month visit rates corresponding to a 10 year increase in age, assuming all other variables are held constant.

Solution

Part A

To fit a Poisson regression model with all the possible predictors included in the question, we will use the `glm` function. However we first need to deal with the offset situation. We will mutate our current reform data to have another column of the number 3 to represent the period of time for the number of doctor visits. This has been previously added however we can show the first 6 data rows look like

```
head(reform)
```

```
##   id numvisit reform badh age educ  loginc period
## 1  3         1     0   0  45 10.5 7.636776      3
## 2  4         9     0   1  53  9.0 7.699212      3
## 3  7        40     0   1  48 10.5 7.057358      3
## 4 12         0     1   0  52 18.0 7.688554      3
## 5 22         1     1   0  42 10.5 7.331879      3
## 6 26         0     0   1  57 10.5 7.428922      3
```

Now we can run our Poisson regression as shown below our code.

```
fit <- glm(numvisit ~ reform + badh + age + educ + loginc + offset(log(period)),
           data= reform, family = poisson)
summary(fit)$coefficients
```

```
##              Estimate Std. Error    z value    Pr(>|z|)
## (Intercept) -1.2728170897 0.316772037 -4.01808538 5.867294e-05
## reform      -0.2273629035 0.031526364 -7.21183392 5.520328e-13
## badh        1.1726351709 0.035255216 33.26132449 1.400250e-242
## age         0.0049815098 0.001473348  3.38108219 7.220094e-04
## educ        -0.0006805681 0.006873469 -0.09901377 9.211273e-01
## loginc      0.1119395623 0.042706815  2.62111711 8.764215e-03
```

The reported rate ratio will be the **exponential** of the reform coefficient in our model named `fit`.

```
RR = exp(fit$coefficients[2])
RR
```

```
##      reform
## 0.7966316
```

The rate of number of doctor visits after the health care reform that took place in 1997 is only **79%** as high as the rate when the reform was not in effect.

Part B

To calculate the rate ratio for age, we will once again take the **exponential** of the age coefficient in our model named `fit`.

```
RR = exp(fit$coefficients[4])
RR
```

```
##      age
## 1.004994
```

To interpret this value, the expected number of doctor visits, as age increases by 1 year, would increase by 0.4%.

Part C

We will now estimate the 1-month rate of visits for a 30 year old individual with 12 years of education, bad health, and the log income of 7.6989.

```
ndat <- data.frame(age = 30, reform = 0, badh = 1, educ = 12, loginc = 7.6989,
                  period = 1)

# Predict with type=response will give the estimated rate.
pred.rate <- predict(fit, newdata = ndat, type="response")
pred.rate
```

```
##          1
## 2.466766
```

To conclude, the estimated number of doctor visits for that specific individual would be between **2-3 visits per month**.

Part D

We will estimate the **1-year rate** for the same individual that was estimated in part (c).

```
ndat <- data.frame(age = 30, reform = 0, badh = 1, educ = 12, loginc = 7.6989,
                  period = 12)
pred.rate1 <- predict(fit, newdata = ndat, type="response")
pred.rate1
```

```
##          1
## 29.60119
```

Hence, the estimated number of doctor visits for that specific individual would be between **29-30 visits per year**.

Part E

```
exp(10*fit$coefficients[4])
```

```
##      age
## 1.051077
```

Question 2

Assume that you have an outcome variable Y_i and predictor variable x_i for $i = 1, \dots, n$, with independent observations. Furthermore, assume that you have the following model:

$$Y_i | x_i \sim \text{Poisson}(\lambda_i)$$

with

$$\log(\lambda_i) = \beta x_i$$

- (a) [3 Points] Find the log-likelihood $\ell(\beta | y_i)$.
- (b) [5 Points] Assume you want to find the maximum likelihood estimate of β . Derive the Newton-Raphson update for $\beta^{(k+1)}$ given you have $\beta^{(k)}$.

Solution

Part A

Let the likelihood function be:

$$L(\beta | y_i) = \prod_{i=1}^n \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$$

Then the log-likelihood will be:

$$\begin{aligned} \ell(\beta | y_i) &= \sum_{i=1}^n \log(e^{-\lambda_i} \lambda_i^{y_i}) - \log(y_i!) \\ &= \sum_{i=1}^n -\lambda_i + \sum_{i=1}^n y_i \log(\lambda_i) - \sum_{i=1}^n \log(y_i!) \end{aligned}$$

Now $\log(\lambda_i) = \beta x_i$ and $\lambda_i = e^{\beta x_i}$

Hence we get the log-likelihood to be:

$$\ell(\beta | y_i) = \sum_{i=1}^n -e^{\beta x_i} + \sum_{i=1}^n y_i \beta x_i - \sum_{i=1}^n \log(y_i!)$$

Part B

$$\begin{aligned}\frac{\partial \ell(\beta|y_i)}{\partial \beta} &= \sum_{i=1}^n -x_i e^{\beta x_i} + \sum_{i=1}^n y_i x_i \\ &= \sum_{i=1}^n x_i (y_i - e^{\beta x_i}) \\ &= \ell'(\beta|y_i)\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ell(\beta|y_i)}{\partial \beta^2} &= - \sum_{i=1}^n x_i^2 e^{\beta x_i} \\ &= \ell''(\beta|y_i)\end{aligned}$$

Hence the Newton-Raphson method will be as follows:

$$\begin{aligned}\beta^{(k+1)} &= \beta_k - \frac{\ell'(\beta|y_i)}{\ell''(\beta|y_i)} \\ &= \beta_k + \frac{\sum_{i=1}^n x_i (y_i - e^{\beta x_i})}{\sum_{i=1}^n x_i^2 e^{\beta x_i}}\end{aligned}$$

Question 3

[5 points] Refer to the model in Question 2. Use the `optim()` function in R to find the maximum-likelihood estimate of β . Test the function using data from `reform.csv` with `numvisit` as Y_i and `age` as x_i . Don't use an offset for this part.

Solution

```
fit = glm(numvisit ~ age, data = reform, family = poisson)
summary(fit)$coefficients

##              Estimate Std. Error   z value    Pr(>|z|)
## (Intercept) 0.40300193 0.055028832  7.323469 2.416420e-13
## age         0.01604085 0.001378441 11.636948 2.674155e-31

yi = reform$numvisit
xi = reform$age

log.lik.pr = function(par){
  b0 = par[1]
  b1 = par[2]

  lam = exp(b0 + b1*xi)

  -sum(dpois(yi, lambda = lam, log = TRUE))
}

opt.pr = optim(par = list(b0 = 1, b1 = 1), fn = log.lik.pr)
opt.pr$par

##          b0          b1
## 1.13125 -0.17500
```

Question 4

The file adult data clean.csv contains education, demographic, and income information from the US census database as of 1994. The three variables of interest are:

- education**: The education level (HS-grad, Bachelors, Masters, Doctorate)
- age**: The age of the individual
- sex**: The sex of the individual

You should run the following code to make sure that HS-grad is the reference category for education, assuming that you read the csv file into a data object called a.data:

```
a.data$education <- factor(a.data$education)
a.data$education <- relevel(a.data$education, ref="HS-grad")
```

- (a) [2 Points] Run a multinomial logistic regression model with education as the outcome and age and sex as predictors. Make sure that HS-grad is the reference category for the outcome.
- (b) [6 Points] Using odds ratios calculated from the model fit, describe the effects of age and sex on the odds of an individual having a Bachelors, Masters, and Doctorate (each compared to the reference category).

Solution

Part A

To run the multinomial regression, we will be using the nnet package. Our reference category will be HS-grad. When running the model we get:

```
fit = multinom(education~ age + sex, data = ad_ed)

## # weights: 16 (9 variable)
## initial value 24942.208145
## iter 10 value 18781.483905
## iter 20 value 17540.685828
## final value 17540.685020
## converged

fit

## Call:
## multinom(formula = education ~ age + sex, data = ad_ed)
##
## Coefficients:
##      (Intercept)          age      sexMale
## Bachelors    -0.7131205 -0.0006913149  0.096781173
## Doctorate    -5.6204287  0.0462546638  0.528634166
## Masters      -3.0003569  0.0286641455  0.007771833
##
## Residual Deviance: 35081.37
## AIC: 35099.37
```

Part B

Let us first make a few tables to indicate each category of education vs. a HS-Grad.

Bachelors vs. HS-Grads

Coefficients	Estimates	StandardErrors
Constant	-0.7131	0.0585
Age	-0.0007	0.0013
Sex (Male)	0.0968	0.0364

Masters vs. HS-Grads

Coefficients	Estimates	StandardErrors
Constant	-3.0004	0.0938
Age	0.0287	0.0019
Sex (Male)	0.0078	0.0565

Doctrate vs. HS-Grads

Coefficients	Estimates	StandardErrors
Constant	-5.6204	0.1988
Age	0.0463	0.0036
Sex (Male)	0.5286	0.1234

Odds of an individual with a bachelors vs. high school graduate:

Age: $e^{-0.007} = 0.9930$

The odds of having a **bachelors** vs. a high school graduate is **0.9930** given that **age differs by 1 year** and their sex is constant.

Sex: $e^{0.09678} = 1.1016$

The odds of having a **bachelors** vs. a high school graduate is **0.9930** given that the **sex of the individual is male** and their age is constant.

Odds of an individual with a masters vs. high school graduate:

Age: $e^{0.0287} = 1.0291$

The odds of having a **masters** vs. a high school graduate is **1.0291** given that **age differs by 1 year** and their sex is constant.

Sex: $e^{0.0078} = 1.0078$

The odds of having a **masters** vs. a high school graduate is **1.0078** given that the **sex of the individual is male** and their age is constant.

Odds of an individual with a doctorate vs. high school graduate:

Age: $e^{0.0463} = 1.0474$

The odds of having a **doctorate** vs. a high school graduate is **1.0474** given that **age differs by 1 year** and their sex is constant.

Sex: $e^{0.5286} = 1.6966$

The odds of having a **doctorate** vs. a high school graduate is **1.6966** given that the **sex of the individual is male** and their age is constant.