Problem 5 - Poisson Regression

Taufiqur Rohman

2022-06-20

Use the data in exercise 5 for this exercise. This problem is about solving the count frequency of people smoking cigarettes in a day using Poisson distribution.

The problems are written in below. For now, let's upload and clean the data set to the R environment.

Let's upload the data set. Looking from the data set, they have 2 data here. The metadata of the data, and the data itself. The metadata are written in the first 10 rows. After that, the data itself is written 3 rows below it. So, I will try to separate both of the data first.

```
path <- '~/Documents/RU/Econometrics II/06-HW-Jun/Exercise_5.xls'

df_exec4_meta <- read_excel(path, skip = 2, n_max = 10, col_names = FALSE)
```

```
## New names:
## * `` -> ...1
```

```
df_exec4 <- read_excel(path, skip = 13)

head(df_exec4)
```

```
## # A tibble: 6 x 10
##    educ cigpric white   age income  cigs restaurn lincome agesq lcigpric
##   <dbl>   <dbl> <dbl> <dbl>  <dbl> <dbl>    <dbl>   <dbl> <dbl>    <dbl>
## 1  16      60.5     1    46  20000     0        0    9.90  2116     4.10
## 2  16      57.9     1    40  30000     0        0   10.3   1600     4.06
## 3  12      57.7     1    58  30000     3        0   10.3   3364     4.05
## 4  13.5    57.9     1    30  20000     0        0    9.90   900     4.06
## 5  10      58.3     1    17  20000     0        0    9.90   289     4.07
## 6   6      59.3     1    86   6500     0        0    8.78  7396     4.08
```

As the data looks clean now, we can move on to the analysis.

Problem 1

The variable cigs is the number of cigarettes smoked per day. How many people in the sample do not smoke at all? What fraction of people claim to smoke 20 cigarettes a day? Why do you think there is a pileup of people at 20 cigarettes?

```
smoke_no <- sqldf("
SELECT
  COUNT(*) AS not_smoke
FROM
  df_exec4
WHERE
  cigs == 0")

smoke_twenty <- sqldf("
SELECT
  SUM(CASE WHEN cigs == 20 THEN 1 ELSE 0 END) AS smoke_twenty,
  ROUND(SUM(CASE WHEN cigs == 20 THEN 1 ELSE 0 END) / CAST(COUNT(*) AS float), 3) AS fraction
FROM
  df_exec4")

smoke_no
```

```
##   not_smoke
## 1       497
```

```
smoke_twenty
```

```
##   smoke_twenty fraction
## 1          101    0.125
```

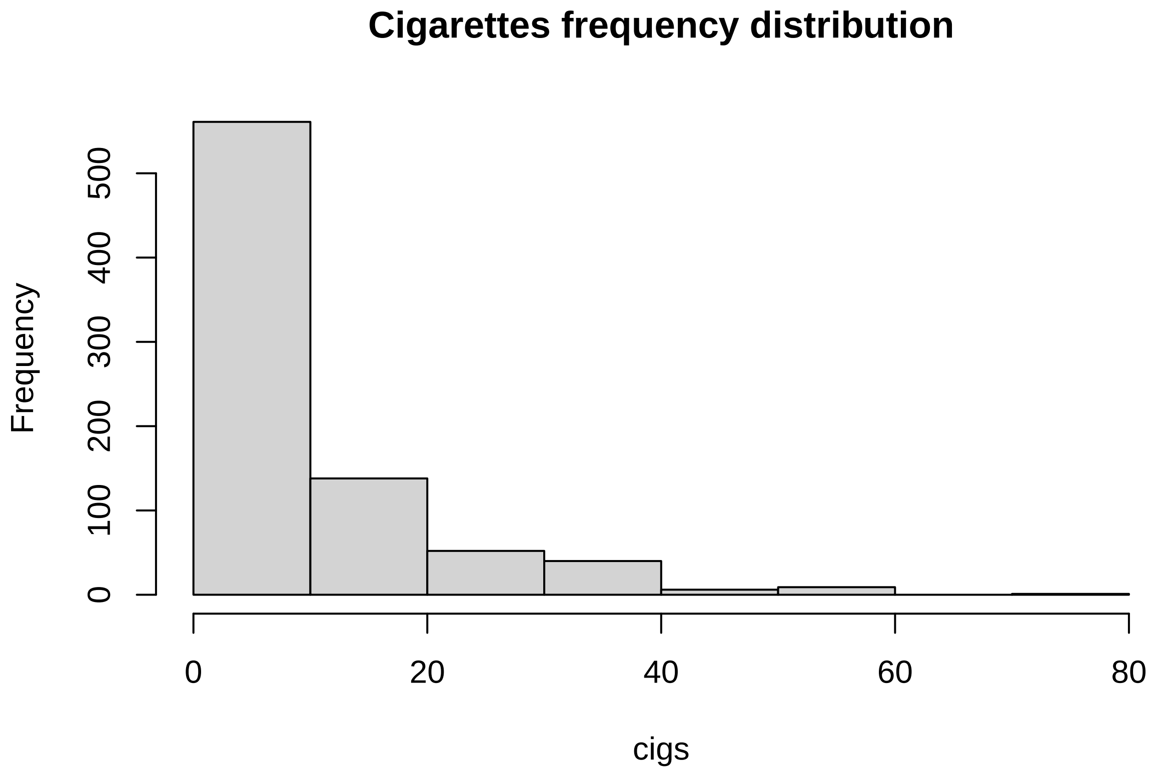
The sample contains a total of 807 people. There are 497 people who do not smoke at all. A total of 101 people consume 20 cigarettes every day. This suggests that 12.52 percent of the population smokes 20 cigarettes each day.

It's unsurprising that the highest percentage of smokers consumes 20 cigarettes each day. This is because one package of cigarettes contains 20 cigarettes, and most smokers will smoke a packet of cigarettes in a single day.

Problem 2

Given your answers to part (1), does cigs seem a good candidate for having a conditional Poisson distribution?

```
cigs <- df_exec4$cigs
hist(cigs, breaks = 10, main = "Cigarettes frequency distribution")
```



Given the lack of a smooth distribution of individuals throughout the categories of number of cigarettes smoked per day, as well as the skewing to the left of the histogram, it does not appear to be a good candidate for Normal distribution. As a result, we can proceed with the Poisson distribution.

Problem 3

Estimate a Poisson regression model for cigs, including log(cigpric), log(income), white, educ, age, and age2 as explanatory variables. What are the estimated price and income elasticities?

Let's create the variable in the data frame first.

```
df_exec4["logcigprc"] <- log(df_exec4$cigpric)
df_exec4["logincome"] <- log(df_exec4$income)
```

Now, run the Poisson regression using the GLM package.

```
formula <- cigs ~ logcigprc + logincome + white + educ + age + agesq
data <- df_exec4

p3_poisson <- glm(formula, data, family = poisson)
summary(p3_poisson)
```

```
##
## Call:
## glm(formula = formula, family = poisson, data = data)
```

```
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -6.138  -4.210  -3.418   2.230  14.412
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.463e+00  6.145e-01   2.381   0.0173 *
## logcigprc   -3.553e-01  1.439e-01  -2.468   0.0136 *
## logincome    8.463e-02  2.011e-02   4.209 2.56e-05 ***
## white       -1.900e-03  3.719e-02  -0.051   0.9593
## educ        -6.010e-02  4.230e-03 -14.209  < 2e-16 ***
## age          1.152e-01  4.961e-03  23.222  < 2e-16 ***
## agesq       -1.379e-03  5.686e-05 -24.249  < 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: 15821  on 806  degrees of freedom
## Residual deviance: 14897  on 800  degrees of freedom
## AIC: 16382
##
## Number of Fisher Scoring iterations: 6
```

The estimated price elasticity is given by the coefficient of, which is -0.355 and the estimated income elasticity is given by the coefficient of, which is 0.085