# Regression Models on Rand Health Insurance data

Djawed Mancer, Lefafta Rémi

## Contents

# 1 Introduction

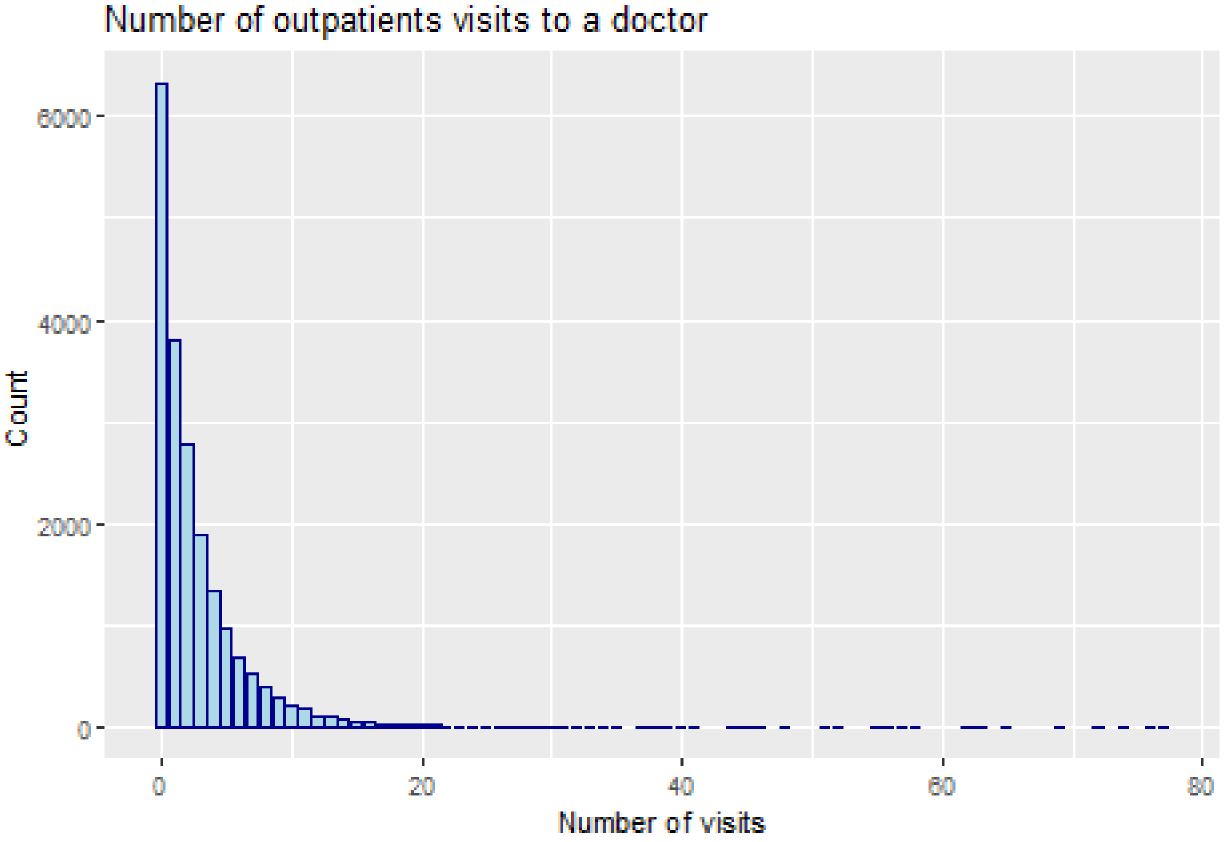
The dataset come from the Rand Health Insurance Experiment, the goal of the experiment was to assess how the patient's use of health services depend on health insurance types. In the following work, we will use the variable `mdu` which is the number of outpatient visits to a medical doctor as our dependent variable. The datasets includesmany variables, but we will use only a subset of them. Moreover, we work on 20186 observations.

Table 1 shows a short summary of our quantitative variables.

Table 1: Quantitative variable summary

| | mdu | lcoins | linc | educdec | age |
|---|---|---|---|---|---|
| Max | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Min | 77.00 | 4.56 | 10.28 | 25.00 | 64.28 |
| Mean | 2.86 | 2.38 | 8.71 | 11.97 | 25.72 |
| Median | 1.00 | 3.26 | 8.98 | 12.00 | 24.20 |
| Sd | 4.50 | 2.04 | 1.23 | 2.81 | 16.77 |
| Var | 20.29 | 4.17 | 1.51 | 7.88 | 281.15 |

The following figure shows how our dependent variable is distributed. As we see, there is an important amount of zeroes. Indeed, for more than 6000 observations `mdu` take zero as value. This excess of zeroes will be taken into account in the last part.



# 2 Poisson regression model

We estimate a Poisson regression model as follow :

$$mdu = \beta_0 + lcoins\beta_1 + idp\beta_2 + linc\beta_3 + female\beta_4 + edycdec\beta_5 + age\beta_6 + bkack\beta_7 + hlthg\beta_8 + hlthf\beta_9 + hlthp\beta_{10}$$

The results are shown in table 6. All of our estimated coefficients are statistically significant at 1%, except for the intercept one. Most of them are positive, excluding `lcoins`, `idp` and `black`. Even if the intercept coefficient is not significant, it represents the predicted number of visits to a doctor for a white male whom self-rated health is excellent and without an individual deductible plan who had average levels of income and coinsurance. In order to interpret this value, we must use the exponentiate, which gives us $exp(0.079) \approx 1.08$. Then, for the enunciated characteristics, the average visit to the doctor is 1.08. Regarding the coefficient associated with `black` which is negative, indicate that the average number of visits to the doctor is smaller for black people than others people. For the `family income` variable, the regression coefficient is 0.072 which represents a $exp(0.072) = 1.07$ increases in medical visits when `family income` increases by 1 for the average individual.

We can check if some variables have no effect on our dependent variable. We will perform a likelihood ratio test in order to compare the full model with a reduce model which does not include `lcoins` and `ldp`.

First, we must estimate a new Poisson regression model without these variables.

Then we compute the likelihood ratio test. The null hypothesis is that `lcoins` and `ldp` have no impact on `mdu`. The Chi-Squared value is 1589.8 and the p-value is inferior to 0.05. We reject the null hypothesis and we keep our first Poisson regression model which is better.

Table 2: Likelihood ratio test

| #Df | LogLik | Df | Chisq | Pr(>Chisq) |
|---|---|---|---|---|
| 11 | -61903.89 | NA | NA | NA |
| 9 | -62698.80 | -2 | 1589.834 | 0 |

# 3 Goodness of fit

## 3.1 Deviance statistic

There are many ways to evaluate the goodness of fit of the Poisson regression model based. First, we will look at deviance statistics. The deviance statistic is defined by :

$$D = 2\sum_{i=1}^{n} y_i \; ln\frac{yi}{\hat{\lambda}_i} - (y_i - \hat{\lambda}_i)$$

with $y_i$ the observed values for `mdu` and $\hat{\lambda}_i$ the fitted values from our model.

We find $D = 82912.06$

In order to calculate the p-value for the deviance goodness of fit test, we calculate the probability to the right deviance value for the chi-squared distribution on 20176 degrees of freedom.
The p-value is 0, then the null hypothesis that our model is correctly specified is rejected at 5% risk. The fitted values significantly differ from the observed values.

### 3.2 Pseudo-$R^2$

Secondly, we will look at pseudo-$R^2$, which is different from the traditional $R^2$. Here, it represents the improvement of our model compared to the null model (intercept only). After calculate, we find a pseudo-$R^2 = 10\%$.

The deviance is reduced by 10% compared to the null model. The improvement of our model is not that good compared to the null model.

# 4 Overdispersion

The overdispersion means that the variance is larger than the mean. One of the main assumption in a Poisson regression model is the equality between the mean and the variance of the dependent variable. However, this assumption is pretty strong and often not respected which lead to the estimation of wrong coefficients. We must do an overdispersion test in order to verify that.

$$H_0 : \alpha = 0 \tag{1}$$
$$H_1 : \alpha > 0 \tag{2}$$

First, thanks to the Poisson model regression, we have $\hat{\lambda_i} = e^{X_i \hat{\beta}}$ which correspond to the fitted values of our model. Secondly, the coefficient $\alpha$ can be estimated by an auxiliary OLS regression. The dependent variable is given by the following expression : $Y = \frac{[(Y_i - \lambda_i)^2 - Y_i]}{\lambda_i}$ with $Y_i$ the observed values for observation $i$. For the predictor variable, we have $X = \alpha \frac{g(\lambda_i)}{\lambda_i}$. Then, we are estimating this OLS : $Y = X + u_i$. We will use different variance formulations, $g(\lambda_i) = \lambda$ and $g(\lambda_i) = \lambda^2$.

Starting with $g(\lambda_i) = \lambda$.

We reject the null hypothesis if t value is superior to 1.96. The t value is calculated as follows : $t = \frac{5.381 - 0}{0.286} = 18.81$.

We reject $H_0$, $\alpha$ is positive, statistically significant at 5% and equal to 5.407 which confirm the presence of overdispersion in our data.

Alternatively, we can use the command dispersiontest from AER for estimating $\alpha$. The dispersion test gives us $\alpha = 5.38$, this confirms our precedent result.

Carry on with $g(\lambda_i) = \lambda^2$

As for the previous case, t value $= 18.21 > 1.96$ we reject $H_0$.

We reject the null hypothesis, $\alpha$ is positive, statistically significant at 5% and equal to 1.676 which confirm the presence of overdispersion in our data. This result is confirmed by the overdispersion test from AER

The results show that using $g(\lambda_i) = \lambda_i^2$ is better. Indeed, $\alpha$ is closer to 0 in our second case, which mean that overdispersion is smaller with this kind of variance formulation. However, Poisson regression model is not the best model due to the presence of overdispersion. Following this, we will estimate different regression model.

# 5 Negative Binomial Model

The negative binomial model relax the assumption made in the Poisson model by introducing a fixed unobserved effect in the conditional mean.

The regression results are in table 6.

### 5.1 Overdispersion

By estimating this model, we have a $\theta$ parameter representing the inverse of the $\alpha$ parameter previously found in the overdispersion test. The estimated $\theta = 0.79$ corresponds to $\alpha = 1.26$. Earlier, we found a $\alpha = 1.67$ with quadratic formulation. Then, the negative binomial model gives some improvement regarding overdispersion.

### 5.2 Poisson model vs Negative Binomial Model

From table 6, we see that sign of each coefficient is identical and coefficient value are really close to each others. One noticeable thing is that standard error is more than twice higher in Negative Binomial Model. In order to compare those two models, we perform a likelihood-ratio test.

Table 3: Likelihood ratio test

| #Df | LogLik | Df | Chisq | Pr(>Chisq) |
|---|---|---|---|---|
| 11 | -61903.89 | NA | NA | NA |
| 12 | -43169.57 | 1 | 37468.63 | 0 |

We conclude that Negative Binomial Model performs better than the Poisson model for our data.

## 6 Marginal effect

The result is shown in table 6. The coefficient associate with the variable `lcoins` has a value of $\approx -0.076$ and is statistically significant at 1%. Then, each one-unit increase in `lcoins`, the expected log count of the number of visits to a doctor decreases by 0.076. We can interpret this result as the elasticity of visits to a doctor, according to the coinsurance rate. Then, the elasticity is equal to $-7.6\%$ for the baseline individual. The coefficient associate to `hlthp` is the expected difference in log count between an individual self rating himself in poor health and the reference is a individual rating himself in excellent health. The expected log count for an individual in poor health is 0.899 higher than the expect log count for an individual in excellent health.

## 7 Hurdle Model

In figure 1, we have seen that our dependent variable was suffering from many zeroes. More exactly, there are 6308 zeroes in our observation. From this statement, we must use a model which takes into account this. The Hurdle model use different distribution for zero and positive counts. Indeed, we will use a logit model to estimate zero counts and a truncated-at-zero Poisson model for positive counts.

### 7.1 Hurdle Model vs Poisson Model

If we compare the Zero Hurdle model with the Poisson model, coefficients regressions from table 6 have the same sign and all statistically significant. Indeed, the intercept is significant conversely to the intercept from Poisson Model. Most of Zero Hurdle coefficients are larger in absolute value than the Poisson Model. Then, our variables may have a more important influence on the decision to not visit a medical doctor than in Poisson Model. Estimated coefficients in Count Hurdle are, in absolute value, lower than in Poisson model. Furthermore, their signs are identical. Our interpretation is that Count Hurdle estimates number of visits for individuals that went at least once to the doctor. Then, our variables have a lesser impact on this case.

## 7.2 Coefficient interpretation

Table 4: Coefficients for Hurdle Model

| | Count Hurdle | Zero Hurdle |
|---|---|---|
| (Intercept) | 2.5547073 | 0.5605055 |
| lcoins | 0.9612015 | 0.8735297 |
| idp1 | 0.9607666 | 0.7612477 |
| linc | 1.0274776 | 1.1181539 |
| female1 | 1.1619832 | 1.6682754 |
| educdec | 1.0062029 | 1.0582746 |
| age | 1.0038459 | 1.0025255 |
| black1 | 0.7181872 | 0.2719257 |
| hlthg | 1.1009635 | 1.1081037 |
| hlthf | 1.4806832 | 1.5563047 |
| hlthp | 1.9885337 | 2.8427786 |

For the Zero Hurdle model, the baseline odds of having a positive count of medical visit against no visit is 0.56. This odds is increased by 1.67 times for female. Age does not have a significant impact. Self-rating himself in good health increases it by 2.84 times. For the Count Hurdle Model, among those who have positive counts, the average visits to a doctor is 2.55. This is increased by 1.98 times for being in poor health whereas good health increases it by 1.10. Being black decreases it by 0.71.

## 7.3 The Vuong Test

In order to choose the best model, we can perform the Vuong Test. The latter is able to compare predicted probabilities of non-nested models, which is our case. Indeed, if we want to compare the Hurdle Model and the Poisson Model, we cannot use the traditional deviance because these two models are not nested within one another.

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##               Vuong z-statistic             H_A    p-value
## Raw                   -34.00373 model2 > model1 < 2.22e-16
## AIC-corrected         -33.95090 model2 > model1 < 2.22e-16
## BIC-corrected         -33.74191 model2 > model1 < 2.22e-16
```

If the two models differ, the p-value would be lower than 0.05. Then, we conclude that Hurdle Model better fits the data than the Poisson Model.

If we compare the Negative binomial model and the Hurdle model we find out that negative binomial is better.

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##               Vuong z-statistic             H_A    p-value
## Raw                    20.78001 model1 > model2 < 2.22e-16
## AIC-corrected          20.79962 model1 > model2 < 2.22e-16
## BIC-corrected          20.87723 model1 > model2 < 2.22e-16
```

If we stop here, we would choose the negative binomial model.

## 7.4 Observed vs Predicted

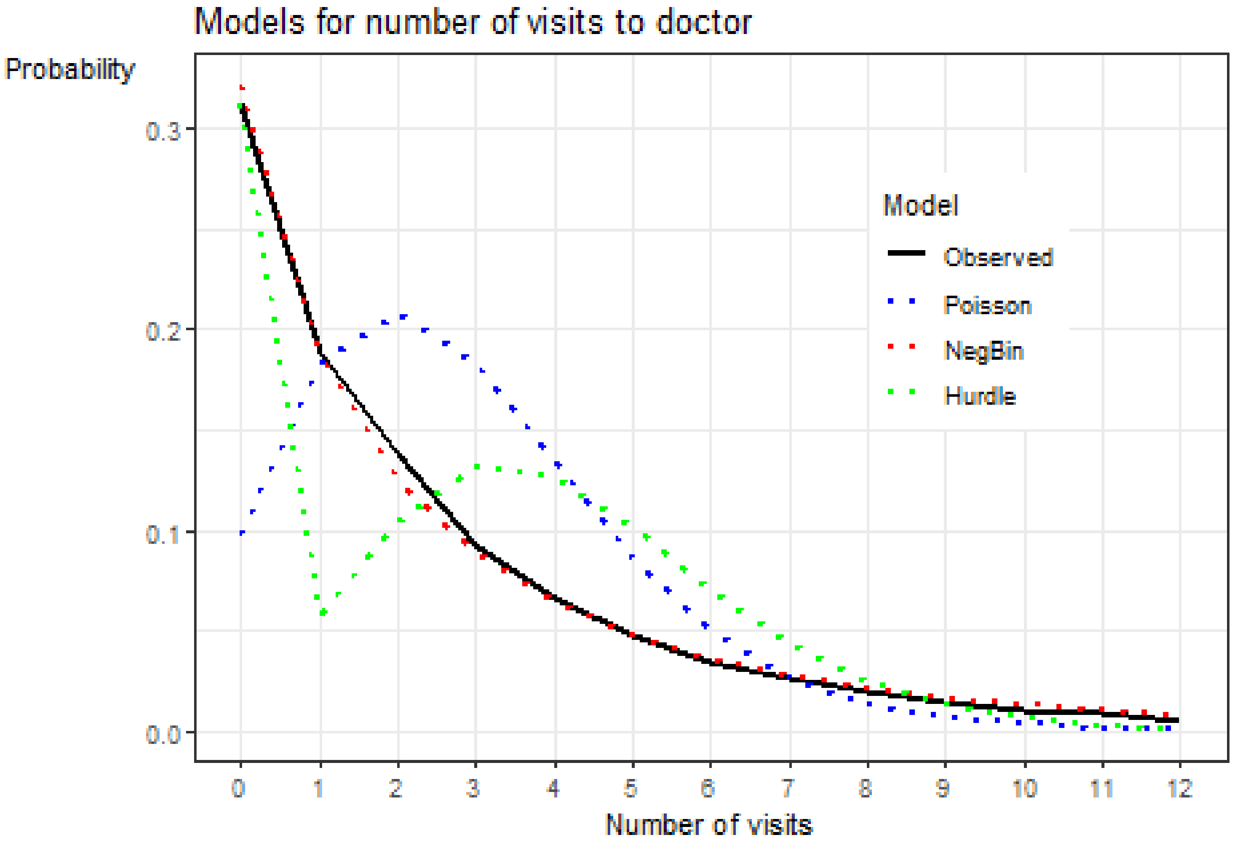
Table 5 shows the observed versus predicted values for each model. Poisson regression predict badly the number of null visits conversely to Hurdle and Negative binomial model. However, it is not easy to tell which model predicts the best. Thereby, we will use MSE as metric. For respectively, Poisson, Negative Binomial and Hurdle model we find 2060589, 522926.5, 870461.8 those values for MSE. The lower one is for Negative Binomial model which confirm the Vuong Test.

Table 5: Observed vs Predicted

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Observed | 6308 | 3815 | 2795 | 1884 | 1345 | 968 | 689 | 531 | 408 | 287 | 206 | 190 | 118 |
| Poisson | 1980 | 3708 | 4222 | 3706 | 2720 | 1744 | 1009 | 541 | 276 | 137 | 68 | 34 | 18 |
| Neg. Binomial | 6476 | 2530 | 3774 | 728 | 1300 | 555 | 428 | 966 | 103 | 1787 | 82 | 24 | 29 |
| Hurdle | 6308 | 1180 | 2124 | 2645 | 2559 | 2052 | 1422 | 878 | 496 | 262 | 132 | 65 | 32 |

## 7.5 Visualisation

We can verify our statement graphically, figure 2 shows that Negative Binomial model overlap the observed observations which is a proof of good approximation. Regarding the other two models, they almost overlap the observed curve if the number of visits is superior to eight. As conclusion, in order to have to best explanation of the data, we choose Negative Binomial Model.


Models for number of visits to doctor
Probability
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
Number of visits
Model
Observed
Poisson
NegBin
Hurdle

# 8 Appendix

Table 6: Models

| | *Dependent variable: mdu* | | | |
|---|---|---|---|---|
| | Poisson | Neg. Binomial | Zero Hurdle | Count Hurdle |
| lcoins | $-0.070^{***}$ | $-0.076^{***}$ | $-0.135^{***}$ | $-0.04^{***}$ |
| | (0.002) | (0.005) | (0.008) | (0.002) |
| idp1 | $-0.129^{***}$ | $-0.107^{***}$ | $-0.273^{***}$ | $-0.04^{***}$ |
| | (0.01) | (0.022) | (0.037) | (0.01) |
| linc | $0.072^{***}$ | $0.077^{***}$ | $0.112^{***}$ | $0.027^{***}$ |
| | (0.005) | (0.009) | (0.014) | (0.005) |
| female1 | $0.290^{***}$ | $0.279^{***}$ | $0.512^{***}$ | $0.15^{***}$ |
| | (0.009) | (0.018) | (0.033) | (0.009) |
| educdec | $0.020^{***}$ | $0.020^{***}$ | $0.057^{***}$ | $0.006^{***}$ |
| | (0.002) | (0.003) | (0.006) | (0.002) |
| age | $0.004^{***}$ | $0.004^{***}$ | $0.003^{***}$ | $0.004^{***}$ |
| | (0.0003) | (0.001) | (0.001) | (0.002) |
| black1 | $-0.781^{***}$ | $-0.806^{***}$ | $-1.302^{***}$ | $-0.331^{***}$ |
| | (0.015) | (0.028) | (0.043) | (0.016) |
| hlthg | $0.116^{***}$ | $0.093^{***}$ | $0.103^{***}$ | $0.096^{***}$ |
| | (0.009) | (0.020) | (0.036) | (0.009) |
| hlthf | $0.493^{***}$ | $0.483^{***}$ | $0.442^{***}$ | $0.392^{***}$ |
| | (0.015) | (0.036) | (0.067) | (0.016) |
| hlthp | $0.899^{***}$ | $0.865^{***}$ | $1.045^{***}$ | $0.688^{***}$ |
| | (0.026) | (0.074) | (0.067) | (0.026) |
| Constant | $0.079^{*}$ | 0.071 | $-0.579^{***}$ | $0.938^{***}$ |
| | (0.045) | (0.087) | (0.142) | (0.046) |
| Observations | 20,186 | 20,186 | 20,186 | 20,186 |
| Log Likelihood | −61,903.890 | −43,170.570 | −54,823.050 | −54,823.050s |
| $\theta$ | | $0.794^{***}$ (0.011) | | |
| Akaike Inf. Crit. | 123,829.800 | 86,363.140 | | |
| *Note:* | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | |