

# STAT 520 Fall 2021 Project

## Replication Study of *Health Care Reform and the Number of Doctor Visits - An Econometric Analysis*

Da Kuang, Zhanqi Wang

### Summary

Our final project based on the paper which evaluates the German health care reform of 1997 (Winkelmann 2004). The individual number of doctor visits is the outcome measure. The data set is from German Socio-Economic Panel for the years 1995-1999, 32837 observations are collected.

As the final project for STAT 520 Applied Econometrics, we decided to reproduce the work in table 1, 2, 3 and part of table 5 as the pedagogical exercises. We will implement five different models including (1) Poisson regression model, (2) negative binomial regression model, (3) Zero-inflated negative binomial regression model, (4) Hurdle negative binomial model, and (5) Hurdle probit poisson log normal model, and evaluate them by log-likelihood, Akaike (AIC) and Bayesian (BIC) Information Criterion. Among all the models, Hurdle negative binomial model demonstrates the best fitness to the data set. Based on the analysis, we found out that the 1997 reform is effective and it affected different groups differently. Our conclusions are consistent with the authors of the original paper.

Besides following the analysis in the original paper, we also generate some visualization to describe the data distributions and demonstrate the fitness of the models. The topics we learned in STAT 520 (Ordinary least square, Heteroscedasticity, Two-stage least squares) give us a solid foundation to understand the advanced topics and models such as Generalized linear models, Zero-inflated models, and Hurdle models in the paper.

### Introduction and Economic Problem of Interest

As one of the Organization for Economic Co-operation and Development (OECD) countries, Germany has a large publicly funded health sector taking a substantial portion of total GDP. A possible explanation for the increase in healthcare spending includes (1) expanding technological and expenditure possibilities in the healthcare services, (2) the surge in demand from the aging population, and incentive structures in the public health sector not promoting the efficient use of resources. One of the significant reforms on the

German healthcare system was the reform of 1997, in which the co-payments for prescription drugs were raised by 200%, and the reimbursements of physicians by state insurance were capped. To evaluate the effect of the healthcare reform of 1997, we want to know (1) whether the reform of 1997 is successful and effective and (2) how the effect of reform differs among different groups of the population.

Some background information needs to be elaborated on before we dig into the analysis. The general aim of this reform is to limit healthcare expenditure and its growth rate. The change in co-payments (for prescription) drugs is the most eminent element of the 1997 reform, with measures like an extended exclusion list of drugs not covered at all by social insurance, price ceilings related to the availability of generics, and a binding overall annual budget for drugs and doctor service.

An increased co-payment has a direct fiscal effect of reducing the proportion of cost covered by the insurer. The increased out-of-pocket expenses would disincentive customers from excessive use of prescription drugs and medical resources as a moral hazard.

The analysis of this paper focuses on the disincentivizing effect on prescription drug use by tracking the number of doctor visits by a person during a given time. In addition to the unavailability of direct information on prescription drug use, a rationale for this approach is that the demand for prescription drugs and the demand for doctor visits are closely related, if not complementary to each other. The implemented policy increased the out-of-pocket expenses for prescription drugs, and the patient might try to persuade the doctor to prescribe in larger size in an appointment or reduce their drug use – in either case we can see a reduction in doctor visits. Though some other effects might affect the number of visits – one might still see a doctor for diagnosis or advice on non-prescription drugs, the number of such visits would not be affected by the increased co-payment as a result of reform. Thus, an analysis of patient visits could help us get some insights into the effectiveness and success of the 1997 Healthcare reform.

In this report, we will first describe the dataset and how to organize them as the systematic components, introduce econometric models like Poisson Regression Model, Negative binomial Regression Model, Zero-Inflation Negative Binomial, Hurdle Negative Binomial Model Regression Model, Two components Negative Binomial Model. After fitting the model to the dataset, we will analyze and discuss the results.

## Data

The GSOEP is an ongoing annual household survey and the author selected five years of data centered around the year of the reform, i.e. 1995-1999. The primary empirical strategy is to pool the data over the five years and estimate the effects of the reforms by comparing the expected number of visits in 1998 and 1996 *ceteris paribus*, i.e., for an individual with given characteristics.

The models that will be estimated in the following sections all include a systematic component (linear predictor) of the type

$$\begin{aligned}
x'_{it}\beta = & \beta_0 + \beta_1 \text{age}_{it} + \beta_2 \text{age}_{it}^2 + \beta_3 \text{years of education}_{it} + \beta_4 \text{married}_{it} + \beta_5 \text{household size}_{it} \\
& + \beta_6 \text{active sport}_{it} + \beta_7 \text{good health}_{it} + \beta_8 \text{bad health}_{it} + \beta_9 \text{self-employed}_{it} \\
& + \beta_{10} \text{full-time employed}_{it} + \beta_{11} \text{part-time employed}_{it} + \beta_{12} \text{unemployed}_{it} \\
& + \beta_{13} \text{equivalent income}_{it} + \beta_{96} (\text{year} = 1996)_{it} + \beta_{97} (\text{year} = 1997)_{it} \\
& + \beta_{98} (\text{year} = 1998)_{it} + \beta_{99} (\text{year} = 1999)_{it}
\end{aligned}$$

The reference year is 1995. In addition, there are three dummies for the quarter in which the interview took place (winter, autumn, spring).

There are three general channels through which these variables can affect the demand for doctor visits. The first is the underlying health status, the second the budget constraint, and the third the preference formation.

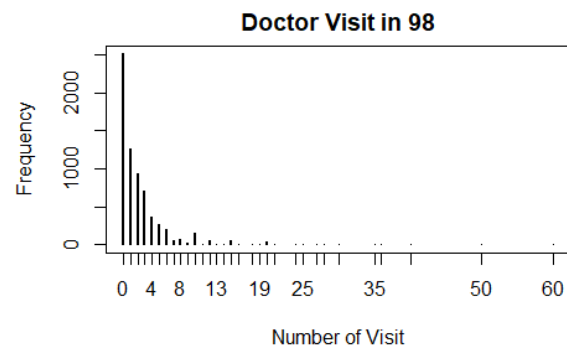
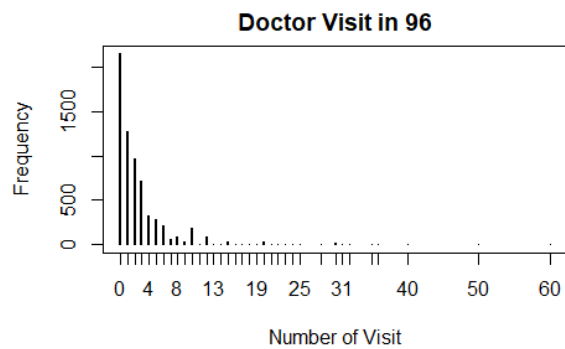
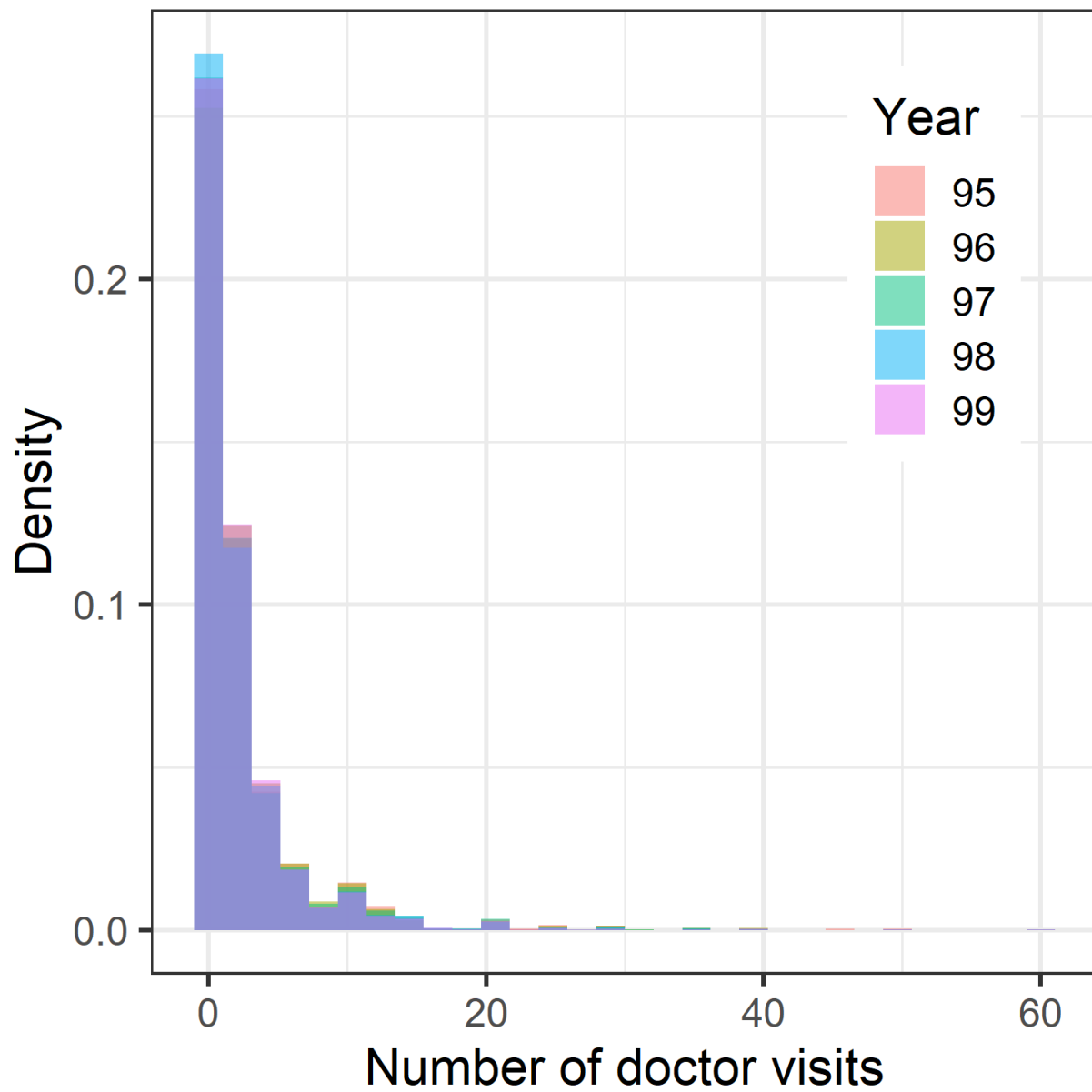
There is no direct health status in the GSOEP. We have the following three sets of proxies instead. Clearly, these are only crude measures of health, and one may want to account for the possibility of additional unobserved heterogeneity to capture any remaining health aspects, as well as other unobserved influences.

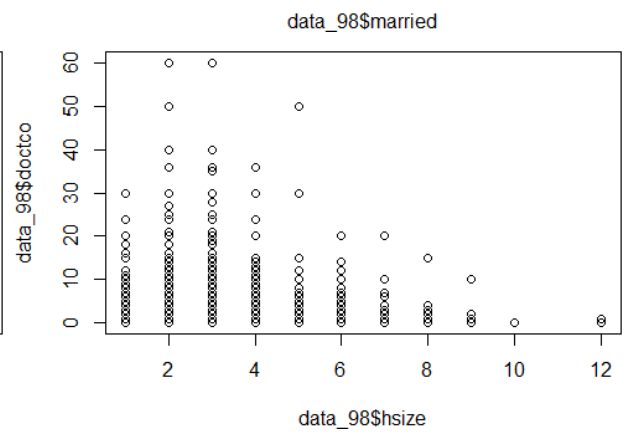
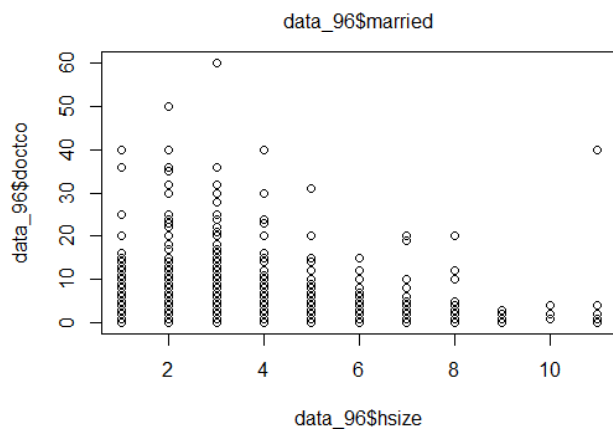
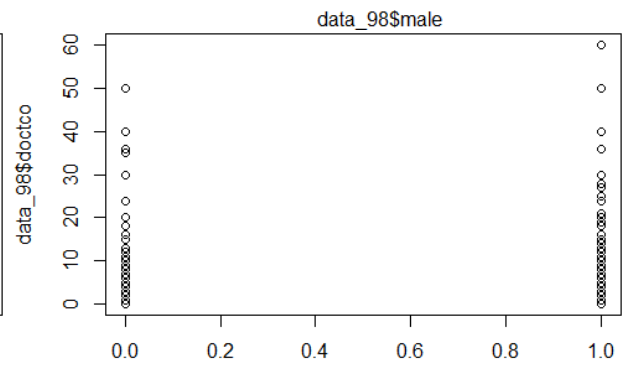
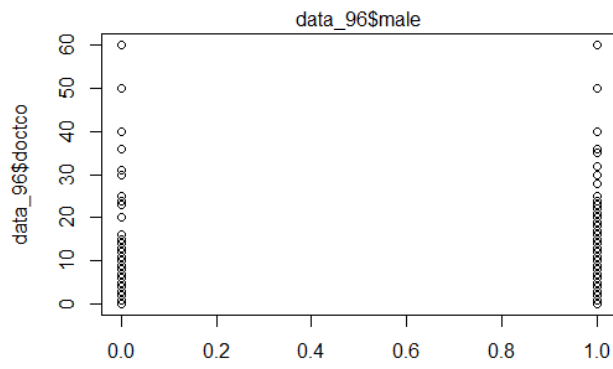
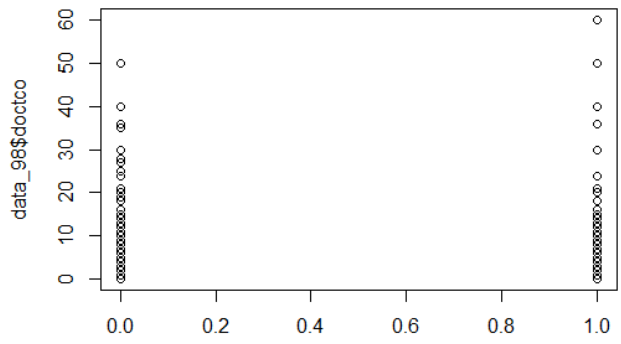
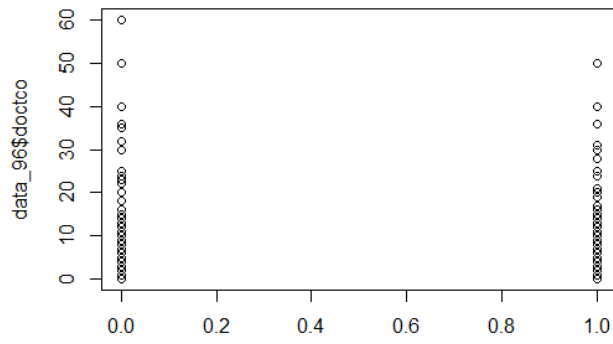
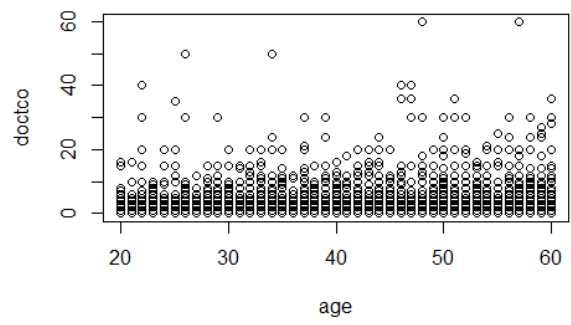
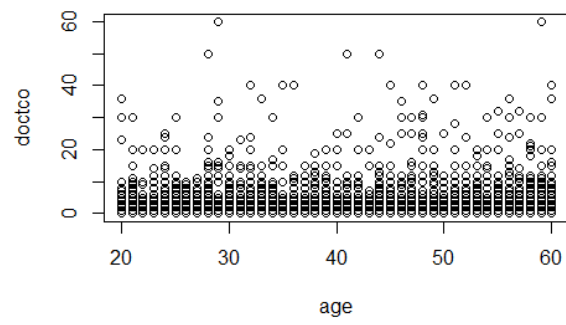
- A time-consistent measure of health over 1995–1999 is provided by a subjective self-assessment in response to the question: “How good do you perceive your own health at current?”, with responses “very good”, “good”, “fair”, “poor” and “very poor”. The two best responses are classified as “good health”, the two worst responses as “bad health”, with fair health being the reference group.
- Another reasonable proxy for health is the age polynomial.
- Finally, engaging in “active sports” (defined as once a week or higher frequencies) acts as a further proxy for good health, although it might have an additional direct effect on the demand for health services as well.

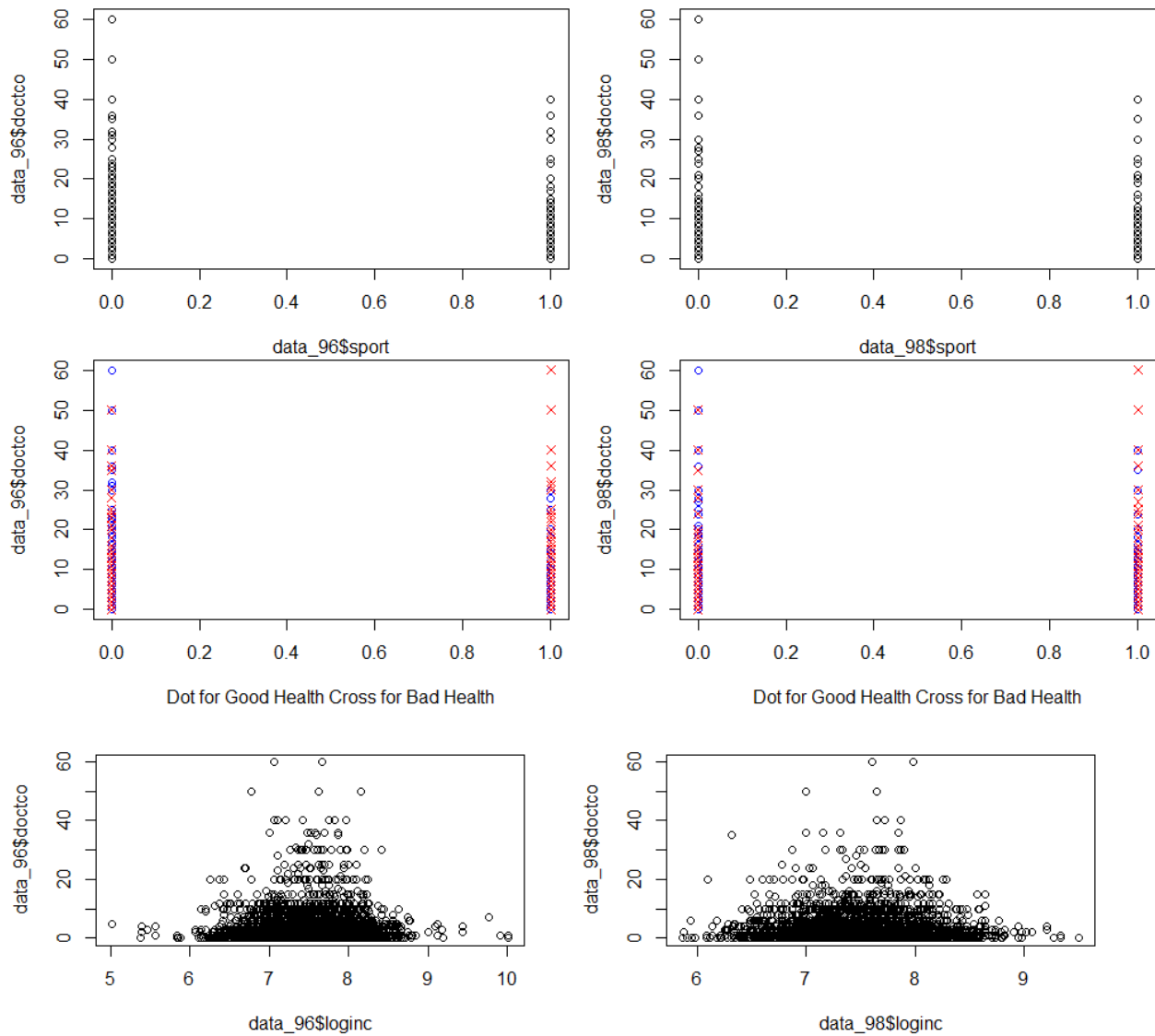
The budget constraint is determined by income and prices. The main price variables are the opportunity costs of a visit to a doctor which, in turn, depend on education level and employment status.

Several of the variables affect more than one aspect at a time. Age, for instance, matters for health, opportunity cost (through the effect of experience on earnings) as well as potentially preferences. Similarly, education is an important factor in determining the optimal investment in health capital. While all those interactions are worth further investigations, it is not the goal of this paper to analyze these various transmission channels – the focus lies on the year dummies, whereas the other variables serve as controls for their potential effects on the changes in visits over time.

## Explorative Analysis of Data







From the histograms above, we can observe that there are a large amount of 0 in the data, and overall the number of 0 increases from 1996 to 1998. Most notably, we can see a decrease in frequency for number of visits between 1 to 13. Such decrease was across groups with different socioeconomical conditions, and it was more apparent on those who were healthy (based on self-reported health conditions and sport active level)

## Econometric Models for consideration

Since the response variables, the number of the doctor visits, are count data, we first consider generalized linear model (GLM) with Poisson distribution.

## Generalized linear model

Generalized linear model (GLM) allows the response variable  $y$  to have an error distribution other than a normal distribution. It has three main components: (1) Systematic component (2) Link function (3) Probability distributions. We define  $\{(y_i, \mathbf{x}_{i*})\}_{i=1}^n$  as following a generalized linear model based on the exponential family  $f_\theta$  if

$$y_i \sim f_{\theta_i}, \quad \theta_i = \mathbf{x}_{i*}^T \boldsymbol{\beta}$$

The exponential family is defined as

$$f_\theta(y) = \exp(\theta y - \psi(\theta)) h(y)$$

Here  $\theta$  is called the natural parameter,  $\psi$  is called the log-partition function, and  $h$  is called the base measure. The distribution with density  $f_\theta$  is called a one-parameter natural exponential family.

## Poisson Regression Model

We take the Poisson distribution as the standard probability distribution for count data

$$P(y_i | \lambda_i) = \frac{\exp(-\lambda_i) \cdot \lambda_i^{y_i}}{y_i!}$$

$$\text{s.t } E(y_i | \lambda_i) = \text{Var}(y_i | \lambda_i) = \lambda_i$$

We assume heterogeneous population with covariance  $x_i$  in regression,  $\lambda_i = \exp(x_i^T \boldsymbol{\beta})$ ,  $y = (y_1, \dots, y_N)^T$ ,  $x = (x_1, \dots, x_N)^T$ . The data was randomly sampled by

$$P(y | x) = \exp \left[ - \sum_{i=1}^N \exp(x_i^T \boldsymbol{\beta}) \right] \prod_{i=1}^N \frac{[\exp(x_i^T \boldsymbol{\beta})]^{y_i}}{y_i!}$$

, and we look for the maximum likelihood (MLE).

The relative change in expected doctor visits as reform effects can be calculated with

$$\begin{aligned} \Delta\%_{(98,96)} &= 100 \times \left[ \frac{E(y_{i,98}|x)}{E(y_{i,96}|x)} - 1 \right] \\ &= 100 \times [\exp(\beta_{98} - \beta_{96}) - 1] \end{aligned}$$

It should be noted that the Poisson Model has a few shortcomings:

1. The assumption exclude the case of unobserved heterogeneity

2. It ignore the Panel Structure of the data and the Standard Errors need to incorporate possible serial correlation (unless we can assume constant individual effect over time)
3. The single index structure of the Poisson Regression model implies that the distribution is determined w.r.t the given mean, and we cannot evaluate the effect of the reform over the distribution (i.e cannot how it affect different sections of the population)

**The random effect and fixed effect models** In statistics, a random effects model, also called a variance components model, is a statistical model where the model parameters are random variables. It assumes that the data from a hierarchy of different populations whose differences relate to that hierarchy and the heterogeneity is uncorrelated with the independent variables. Random effect models assist in controlling for unobserved heterogeneity when the heterogeneity is constant over time and not correlated with independent variables. This constant can be removed from longitudinal data through differencing, since taking a first difference will remove any time invariant components of the model. (Wooldridge 2011)

On the other side, a fixed effects model is a statistical model in which the model parameters are fixed or non-random quantities. It assumes that the individual-specific effects are correlated with the independent variables. It assists in controlling for omitted variable bias due to unobserved heterogeneity when this heterogeneity is constant over time. This heterogeneity can be removed from the data through differencing, for example by subtracting the group-level average over time, or by taking a first difference which will remove any time invariant components of the model.

## Negative binomial regression Model

To model the count data, we can also use Negative binomial regression. It is a generalization of Poisson regression. Negative binomial distribution can be considered as a generalization of Poisson distribution by including a gamma noise variable which has a mean of 1 and a scale parameter of  $\nu$ . The poisson-gamma mixture distribution that results is

$$\Pr(Y = y_i | \mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1} + 1}\right)^{y_i} \left(\frac{1}{\alpha^{-1} + 1}\right)$$

, where  $\mu_i = t_i \mu$  and  $\alpha = \frac{1}{\nu}$ .

## Zero-inflated negative binomial model

Zero-inflated models have been developed to cope with zero-inflated outcome data with over-dispersion (negative binomial) or without (Poisson distribution). In this project, zero-inflated negative binomial regression is applied on the dataset.



A zero-inflated (ZI) model assumes that the zero observations have two different origins: “structural” and “sampling”. The sampling zeros are due to the usual negative binomial distribution, which assumes that those zero observations happened by chance. Zero-inflated models assume that some zeros are observed due to some specific structure in the data (Hu et al 2011)

Formally speaking, the zero-inflated density function is a mix of a point mass at zero and a count distribution.

$$f_{zi}(y; x, z, \beta, \gamma) = f_0(0; z, \gamma)I_{\{0\}}(y) + (1 - f_0(0; z, \gamma)) f_{\text{count}}(y; x, \beta)$$

with  $I(\cdot)$  as the indicator function, inflated probability of observing a zero as  $\pi = f_0(0; z, \gamma)$ , unobserved probability  $\pi$  of belonging to the point mass component modelled by a binomial GLM  $\pi = g^{-1}(z^T \gamma)$  (Zeileis et al)

For the problem of the original paper, the number of the doctor visits is the outcome, some participants may visit zero in certain year because they do not need to and such decision was not affected by the reform. To this end, the number of doctor visit is assumed to be on a negative binomial distribution that includes both zero (the “sampling zeros”) and non-zero counts. Consider the issues with excesssive zeros displayed in the graph above, Zero-inflated negative binomial model can be good candidate in this case.

## Hurdle Models

In contrast to ZI models, a hurdle model assumes that all zero data are from one “structural” source. The positive (i.e., non-zero) data have “sampling” origin, following either truncated Poisson or truncated negative-binomial distribution (Mullahy 1986). The hurdle model has been popular in the health literature, in part because it can be given a structural interpretation that seems to agree well with the intuition of a dual decision structure of the demand process.

For the study discussed in the original paper. It is reasonable to assume that healthy people might have zero number of doctor visit and people who start to feel ill or have not finish the treatment will have non-zero number of doctor visits in a year. Hence, the zero observations can come from only one structural source – the patients. If a subject is considered a patient, they would not have zero number of doctor visits in a whole year and would always have a positive number of visit in a hurdle model, with either truncated Poisson or truncated negative binomial distributions.

Define  $d_i = 1$  if a person does not see a doctor in the period and  $d_i = 0$  if he/she sees a doctor, s.t.  $d_i = 1 - \min(1, y)$ , the probability function of the hurdle model is

$$f(y_i) = f_{1i}^{d_i} [(1 - f_{1i}) f_T(y_i | y_i > 0)]^{1-d_i}$$

, with  $f_{1i} = P(d_i = 1)$ ,  $f_T(y_i | y_i > 0) = \frac{f_2(y_i)}{1 - f_{2i}(0)}$ .

For  $f_{1i}$ , we have the following hurdle functions as candidates

1. Poisson:  $\exp(-\exp(x_i^T \gamma))$
2. Negative Binomial, Type 1:  $(1 + \theta)^{-\exp(x_i^T \theta)/\theta}$
3. Negative Binomial, Type 2:  $[\alpha/(\exp(x_i^T \alpha) + \alpha)^\alpha]$
4. Logit:  $\frac{\exp(x_i^T \gamma)}{1 + \exp(x_i^T \gamma)}$
5. Probit:  $\Theta(x_i^T \gamma)$

For  $f_2$ , we have

1. Poisson
2. Negative Binomial
3. Poisson-log-normal  $p(x) = (\exp(x\mu + x^2\sigma/2) \cdot (2\pi\sigma)^{-1/2})/x! \cdot g(y)$

## Generalized linear model

## Results

**Table 1**

We first reproduce table 1 to generate the summary statistics for the variables of the analysis.

- In general, the health level of people increased
  - Despite volatile averages, the proportion of people in active sports increased from 24.7% to 30.7%
  - The proportion of people reporting good health increased from 56.2% to 59.5%, while the proportion of people reporting poor health decreased from 13.8% to 12.7%%
- Such trends can be meaningful for further analysis for the following two reasons:
  - improvements in the perceived health might be able to explain part of the reduction in the number of doctor visits, and one should control for that in order to isolate the reform effect.
  - these improvements provide some evidence against the possibility that the reforms, while being successful in containing costs, worsened the general health status, illustrating the positive netcontribution to general welfare.
- The average number of doctor visits per quarter declined from 2.66 to 2.35 between 1996 and 1998.
- The decline of the number of doctor visit between 1995 and 1996 is 1%. The increase between 1998 and 1999 is 2%. Therefore, the large drop (7.8%) between 1997 and 1998 in the number of visits clearly coincides with the timing of the reform.

- The increase of the number of doctor visit between 1998 and 1999 also supports the behavior hypothesis. It is because that the 1997 reform enjoyed only a short lifespan. The partial repeal of the 1997 reform was one of the first items on the political agenda of the new government selected in 1998, and a new law lowered the co-payments by between DM 1 and DM 3, effective January 1, 1999.
- The proportion of people reporting good health from 56 to 60% while the proportion reporting bad health decrease from 14 to 13%. This can be the partial reason supporting the idea that the 1997 reform did not worsen the general health status.

	1995	1996	1997	1998	1999
Total Observation	6790.000	6555.000	6480.000	6781.000	6231.000
No. doctor visits	2.687	2.657	2.553	2.353	2.391
Relative change in %	NA	-0.011	-0.039	-0.078	0.016
No. doctor visits (0/1)	0.348	0.328	0.352	0.372	0.346
Age	38.081	38.203	38.471	38.731	38.919
Unemployed (0/1)	0.085	0.084	0.092	0.085	0.075
Active sport (0/1)	0.295	0.247	0.262	0.307	0.266
Good health (0/1)	0.568	0.562	0.581	0.595	0.580
Bad health (0/1)	0.145	0.138	0.134	0.127	0.129

**Table 2**

To fit and describe the number of doctor visit as the count data, we first applied Poisson regression while considering the random and fixed effects. The log-likelihood of the models are as follows.

- The pooled model: -86566.18.
- The random effect model: -70176.72.
- The fixed effect model: -42738.86.

The estimates of coefficients and standard errors for the basic Poisson model, with and without individual specific effects, are displayed in Table II.

	Pooled Estimate	Pooled Robust SE	Random Effects Estimate	Random Effects Robust SE	Fixed Effects Estimate	Fixed Effects Robust SE
Age	-0.01057	0.00660	-0.00927	0.00575		
Age^2	0.00016	0.00008	0.00019	0.00007		
Male	-0.20852	0.02091	-0.29480	0.02028		
Education	-0.00576	0.00370	-0.00424	0.00408		
Married	0.08078	0.02226	0.07733	0.01809	0.10358	0.02507
Household size	-0.05221	0.00777	-0.05547	0.00596	-0.05991	0.00850
Active Sport	0.04657	0.01935	0.00614	0.01257	-0.01015	0.01417
Good health	-0.61088	0.01901	-0.50135	0.01183	-0.43002	0.01277
Bad health	0.81311	0.02250	0.62252	0.01216	0.56610	0.01288
Social assistance	0.08610	0.04444	0.04053	0.02640	0.01763	0.02891
Log(income)	0.09312	0.02286	0.01737	0.01598	-0.01659	0.01924
Year=1996	0.00122	0.02676	-0.01419	0.01110	-0.01022	0.01122
Year=1997	-0.03024	0.02689	-0.05027	0.01143	-0.04030	0.01153
Year=1998	-0.10481	0.02681	-0.10615	0.01187	-0.08439	0.01198
Year=1999	-0.09868	0.02726	-0.10685	0.01233	-0.07929	0.01246

We can make the following observations from Table II

- Men have less doctor visits than women.
- The expected number of doctor visits is u-shaped in age because age is negative.
- The health indicators (Good health and Bad health) have the largest effect among all variables.
- There is a statistically significant decline in the expected number of doctor visits between 1996 and 1998 in all three model specifications.

### Table 3

Four additional models were applied on the same data set. Likelihood ratio tests clearly reject the Poisson model against the alternative models with unobserved heterogeneity. AIC is used to pick the best model among all five models. We also calculated BIC as the alternative criteria and it supported the same conclusion as AIC. Zero-inflated negative binomial and hurdle negative binomial models are the best two models.

Models	Log.likelihood	AIC	BIC
Poisson	-86566.18	173176.4	173361.1
Negative Binomial	-64611.55	129269.1	129462.3
Zero-inflated Negative Binomial	-64369.56	128829.1	129207.1
Hurdle Negative Binomial Model	-64254.90	128599.8	128977.8
Hurdle Probit Poisson Log Normal Model	-77976.35	156040.7	156410.3

**Table 5**

The relative changes in expected doctor visits from 1996 to 1998 for each models are as follows. The basic models (Poisson and Negative Binomial models) estimate the reform effect as -10% and -8%. However the structural models (Zero-inflated negative binomial and Hurdle negative binomial models) estimate the reform effect as -6% and -3%. The differences between the two kinds of models support the idea that the survey, which was conducted in the hospital, was over-represented by patients with more doctor visits.

Models	Delta % (96, 98)
Poisson	-10.060%
Negative Binomial	-8.890%
Zero-inflated Negative Binomial	-6.321%
Hurdle Negative Binomial Model	-3.482%

We also tried to reproduce the finite mixture model analysis. However, the original paper did not provide clear description about how the author divide the population. Here we developed our own strategy for separation and got the same conclusion as the author. We rank all the ids based on the sum of the number of doctor visits in survey from 1995 to 1999. Two-thirds of the population is the low-user group with a mean number of quarterly visits of 0.997, and the one-third belongs to a high-user group with a mean number of 4.839 visits per quarter. The reform effect of the two groups are as follows.

- Lower-user group ( $p = 0.663$ ,  $\mu = 0.997$ ): -8.573883
- Higher-user group ( $p = 0.337$ ,  $\mu = 4.839$ ): -6.185429

## Discussion

As another expansion and alternative to the Hurdle models in the original paper, we applied a zero-inflated negative binomial model (ZINB) on the data. Based on Table 3, ZINB outperformed basic models (Poisson and Negative binomial models) but fitted slightly

worse than Hurdle Negative binomial model. It supports the idea that Hurdle assumption aligns better to the data. In other words, we tend to believe that all zero data are from one “structural” source.

Overall, from Table 5, we can conclude that the reform of 1997 is effective in terms of reducing number of doctor visit. Evaluating the results from Zero-Inflation Negative Binomial vs. Negative Binomial and the result from two-component negative binomial, we can observe that there are effects are unevenly distributed – the left-margin/less frequent users have stronger response to the reform. A possible explanation of this uneven effects is that the higher-users/more frequent users have a consistent and stronger demand for doctor visits, given their potentially unfavorable health status.

The index models used in this research can be applied to study the effect of other medical healthcare events. From 1990 - 2015, multiple attempts were made to improve the solidarity-based system through competition to further improve economic efficiency of the healthcare system and promote technological advancement. (Busse et al 2017) If policy objective is on the effect of promoting efficient and responsible use of healthcare services, we might apply the same analysis to the Act to Strengthen Competition in Statutory Health Insurance in 2007 and the Statutory Health Insurance Care Structures Act, given that the primary policies of 2007 reform are

- mandatory universal coverage
- introduction of a uniform contribution rate, a central reallocation pool and resource allocation to sickness funds according to a morbidity-based risk structure compensation scheme
- choice of tariffs in statutory health insurance

The 2011 Acts are more focused on supplementary premiums and reimbursement for new pharmaceutical products. It should be noted that to evaluate the effect of both reforms, in addition to variables we used in previous analysis, we might want to incorporate more economic variables related to economic growth, business cycles, considering the 2008 recession and 2010s European Sovereign Debt Crisis.

## Bibliography

Busse, Reinhard, Miriam Blümel, Franz Knieps, and Till Bärnighausen. “Statutory Health Insurance in Germany: A Health System Shaped by 135 Years of Solidarity, Self-Governance, and Competition.” *The Lancet* 390, no. 10097 (2017): 882–97. [https://doi.org/10.1016/s0140-6736\(17\)31280-1](https://doi.org/10.1016/s0140-6736(17)31280-1).

Hu, Mei-Chen, Martina Pavlicova, and Edward V. Nunes. “Zero-Inflated and Hurdle Models of Count Data with Extra Zeros: Examples from an HIV-Risk Reduction Intervention Trial.” *The American Journal of Drug and Alcohol Abuse* 37, no. 5 (2011): 367–75. <https://doi.org/10.3109/00952990.2011.597280>.

Mullahy, John. “Specification and Testing of Some Modified Count Data Models.” *Journal of Econometrics* 33, no. 3 (1986): 341–65. [https://doi.org/10.1016/0304-4076\(86\)90002-3](https://doi.org/10.1016/0304-4076(86)90002-3).

Winkelmann, Rainer. “Health Care Reform and the Number of Doctor Visits—an Econometric Analysis.” *Journal of Applied Econometrics* 19, no. 4 (2004): 455–72. <https://doi.org/10.1002/jae.764>.

Wooldridge, Jeffrey M. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: MIT, 2011.

Zeileis, Achim, Christian Kleiber, and Simon Jackman. “Regression Models for Count Data in R.” Accessed January 10, 2022. <https://cran.r-project.org/web/packages/pscl/vignettes/countreg.pdf>.

# Appendix

## The calculation of table 1

```
knitr::opts_chunk$set(echo = TRUE)
library(dplyr)
library(ggplot2)
library(tidyverse)
library(foreign)
```

```
##                               1995                1996                1997                1998
## Total Observation             6.790000e+03 6555.00000000 6480.00000000 6781.00000000
## No. doctor visits            2.687334e+00   2.65659802   2.55339506   2.35319274
## Relative change in %                NA   -0.01143747   -0.03884779   -0.07840632
## No. doctor visits (0/1) 3.484536e-01   0.32784134   0.35200617   0.37162660
## Age                          3.808071e+01  38.20259344  38.47083333  38.73101312
## Unemployed (0/1)          8.527246e-02   0.08360031   0.09197531   0.08538564
## Active sport (0/1)        2.949926e-01   0.24713959   0.26203704   0.30718183
## Good health (0/1)         5.681885e-01   0.56201373   0.58086420   0.59548739
## Bad health (0/1)          1.447717e-01   0.13806255   0.13441358   0.12726736
##                               1999
## Total Observation             6.231000e+03
## No. doctor visits            2.390788e+00
## Relative change in %        1.597627e-02
## No. doctor visits (0/1) 3.460119e-01
## Age                          3.891944e+01
## Unemployed (0/1)          7.478735e-02
## Active sport (0/1)        2.664099e-01
## Good health (0/1)         5.798427e-01
## Bad health (0/1)          1.291927e-01
```

```
write.csv(dis, "../out/table-1.csv")
```

## The calculation of table 2

```
hcg <- read.table("../data/w.data")
colnames(hcg) = c("id", "year", "doctco", "age",
                  "male", "educ", "married", "hsize",
                  "sport", "goodh", "badh", "sozh",
                  "loginc", "ft", "pt", "unemp",
```



```

      "winter", "spring", "fall")

selected_row_names_vec <- c('Age', 'Age^2', 'Male', 'Education', 'Married',
      'Household size', 'Active Sport',
      'Good health', 'Bad health',
      'Social assistance', 'Log(income)',
      'Year=1996', 'Year=1997', 'Year=1998', 'Year=1999')

```

```

emp_vec <- c()
for (idx in seq(dim(hcg)[1])){
  if(hcg[[idx, 'ft']] == 1)
    emp_vec <- c(emp_vec, 'full-time')
  else if (hcg[[idx, 'pt']] == 1)
    emp_vec <- c(emp_vec, 'part-time')
  else if (hcg[[idx, 'unemp']] == 1)
    emp_vec <- c(emp_vec, 'unemployed')
  else
    emp_vec <- c(emp_vec, 'self-employed')
}

```

```

data_df <- hcg %>%
  tibble() %>%
  mutate(age2=age^2) %>%
  mutate(emp=factor(emp_vec)) %>%
  dplyr::select(id, doctco, age, age2, male, educ, married, hsize, sport,
    goodh, badh, emp, sozh, loginc, year, winter, spring, fall)

```

## Column 1: Pooled Poisson Regression

```

model_fit <- glm(
  doctco~age+age2+factor(male)+educ+factor(married)+hsize+
  factor(sport)+factor(goodh)+factor(badh)+factor(sozh)+
  loginc+factor(year)+
  factor(emp)+factor(winter)+factor(spring)+factor(fall),
  family = "poisson", data = data_df)

cov_mat <- vcovHC(model_fit, type = "HC0")
std_err <- sqrt(diag(cov_mat))
q_val <- qnorm(0.975)

pool_df <- cbind(

```

```

Estimate = coef(model_fit)
, "Robust SE" = std_err
, z = (coef(model_fit)/std_err)
, "Pr(>|z|)" = 2 * pnorm(abs(coef(model_fit)/std_err), lower.tail = FALSE)
, LL = coef(model_fit) - q_val * std_err
, UL = coef(model_fit) + q_val * std_err
)

```

```

res_df_col1 <- pool_df[seq(2, 16), ]
rownames(res_df_col1) <- selected_row_names_vec
res_df_col1

```

##	Estimate	Robust SE	z	Pr(> z )
## Age	-0.0105723260	6.598343e-03	-1.60226980	1.090960e-01
## Age^2	0.0001580161	8.007286e-05	1.97340441	4.844951e-02
## Male	-0.2085200124	2.091179e-02	-9.97140747	2.033268e-23
## Education	-0.0057640252	3.701687e-03	-1.55713475	1.194385e-01
## Married	0.0807752520	2.226030e-02	3.62866821	2.848871e-04
## Househould size	-0.0522052984	7.771470e-03	-6.71755749	1.847959e-11
## Active Sport	0.0465699357	1.934927e-02	2.40680609	1.609271e-02
## Good health	-0.6108842053	1.901187e-02	-32.13172022	1.590276e-226
## Bad health	0.8131080333	2.249984e-02	36.13839549	5.661057e-286
## Social assistance	0.0860960122	4.443502e-02	1.93757132	5.267554e-02
## Log(income)	0.0931194922	2.285981e-02	4.07350230	4.631139e-05
## Year=1996	0.0012175406	2.676300e-02	0.04549343	9.637140e-01
## Year=1997	-0.0302387301	2.688853e-02	-1.12459601	2.607603e-01
## Year=1998	-0.1048116755	2.680600e-02	-3.91000836	9.229293e-05
## Year=1999	-0.0986844438	2.726094e-02	-3.61999452	2.946093e-04
##	LL	UL		
## Age	-2.350484e-02	0.0023601889		
## Age^2	1.076213e-06	0.0003149561		
## Male	-2.495064e-01	-0.1675336506		
## Education	-1.301920e-02	0.0014911477		
## Married	3.714586e-02	0.1244046432		
## Househould size	-6.743710e-02	-0.0369734965		
## Active Sport	8.646068e-03	0.0844938038		
## Good health	-6.481468e-01	-0.5736216177		
## Bad health	7.690092e-01	0.8572069048		
## Social assistance	-9.950186e-04	0.1731870429		
## Log(income)	4.831509e-02	0.1379238982		
## Year=1996	-5.123697e-02	0.0536720492		
## Year=1997	-8.293927e-02	0.0224618142		
## Year=1998	-1.573505e-01	-0.0522728853		
## Year=1999	-1.521149e-01	-0.0452539884		

```
logLik(model_fit)
```

```
## 'log Lik.' -86566.18 (df=22)
```

```
nobs(model_fit)
```

```
## [1] 32837
```

## Column 2: Panel Poisson Regression with Random Effects

```
model_fit <- pglm(  
  doctco~age+age2+factor(male)+educ+factor(married)+hsize+factor(sport)+  
    factor(goodh)+factor(badh)+factor(sozh)+  
    loginc+factor(year)+  
    factor(emp)+factor(winter)+factor(spring)+factor(fall),  
  model = "random",  
  family = "poisson",  
  index=c("id"),  
  data = data_df  
)
```

```
res_df_col2 <- cbind(coefficients(summary(model_fit)), confint(model_fit))[seq(2, 16),  
rownames(res_df_col2) <- selected_row_names_vec  
res_df_col2
```

##	Estimate	Std. error	t value	Pr(> t)
## Age	-0.0092679012	5.747857e-03	-1.6124099	1.068728e-01
## Age^2	0.0001884548	7.001972e-05	2.6914535	7.114142e-03
## Male	-0.2948034124	2.027968e-02	-14.5368891	7.073491e-48
## Education	-0.0042384040	4.083674e-03	-1.0378899	2.993213e-01
## Married	0.0773282851	1.809373e-02	4.2737622	1.922020e-05
## Houseould size	-0.0554746693	5.961705e-03	-9.3051688	1.337796e-20
## Active Sport	0.0061441035	1.256618e-02	0.4889398	6.248843e-01
## Good health	-0.5013500804	1.183233e-02	-42.3711971	0.000000e+00
## Bad health	0.6225151267	1.216031e-02	51.1923547	0.000000e+00
## Social assistance	0.0405303262	2.640419e-02	1.5349958	1.247848e-01
## Log(income)	0.0173742593	1.597776e-02	1.0874029	2.768588e-01
## Year=1996	-0.0141918850	1.109711e-02	-1.2788817	2.009387e-01
## Year=1997	-0.0502742933	1.142705e-02	-4.3995875	1.084568e-05
## Year=1998	-0.1061518772	1.186677e-02	-8.9453049	3.709244e-19

```
## Year=1999      -0.1068475570 1.233431e-02 -8.6626293 4.610057e-18
##              2.5 %      97.5 %
## Age           -2.053349e-02 0.0019976911
## Age^2          5.121869e-05 0.0003256909
## Male           -3.345508e-01 -0.2550559762
## Education      -1.224226e-02 0.0037654499
## Married        4.186523e-02 0.1127913376
## Houseould size -6.715940e-02 -0.0437899427
## Active Sport   -1.848515e-02 0.0307733558
## Good health    -5.245410e-01 -0.4781591354
## Bad health     5.986813e-01 0.6463489049
## Social assistance -1.122094e-02 0.0922815930
## Log(income)    -1.394157e-02 0.0486900866
## Year=1996      -3.594181e-02 0.0075580423
## Year=1997      -7.267089e-02 -0.0278776926
## Year=1998      -1.294103e-01 -0.0828934350
## Year=1999      -1.310224e-01 -0.0826727532
```

```
summary(model_fit)$loglik[1]
```

```
## [1] -70176.72
```

## Column 3: Panel Poisson Regression with Fixed Effects

```
model_fit <- pglm(
  doctco~factor(married)+hsize+factor(sport)+
    factor(goodh)+factor(badh)+factor(sozh)+
    loginc+factor(year)+
    factor(emp)+factor(winter)+factor(spring)+factor(fall),
  model = "within",
  family = "poisson",
  index=c("id"),
  data = data_df
)
```

```
partial_rowname <- selected_row_names_vec[seq(5, 15)]
res_df_col3 <- cbind(coefficients(summary(model_fit)), confint(model_fit))[seq(11), ]
rownames(res_df_col3) <- partial_rowname
res_df_col3
```

```
##              Estimate Std. error    t value    Pr(> t)    2.5 %
```

```
## Married          0.10357566 0.025073136  4.1309417  3.612802e-05  0.05443322
## Househould size -0.05990580 0.008495919 -7.0511263  1.774753e-12 -0.07655749
## Active Sport     -0.01015352 0.014165749 -0.7167656  4.735187e-01 -0.03791788
## Good health      -0.43001654 0.012772856 -33.6664353  1.792296e-248 -0.45505088
## Bad health       0.56610474 0.012875245  43.9684621  0.000000e+00  0.54086972
## Social assistance 0.01762670 0.028905203   0.6098105  5.419873e-01 -0.03902646
## Log(income)      -0.01658794 0.019235453 -0.8623629  3.884879e-01 -0.05428874
## Year=1996        -0.01021562 0.011215851 -0.9108200  3.623902e-01 -0.03219829
## Year=1997        -0.04030285 0.011530234 -3.4954056  4.733419e-04 -0.06290169
## Year=1998        -0.08438629 0.011978925 -7.0445624  1.860452e-12 -0.10786455
## Year=1999        -0.07929067 0.012456586 -6.3653612  1.948307e-10 -0.10370513
##                  97.5 %
## Married          0.15271810
## Househould size -0.04325410
## Active Sport     0.01761084
## Good health      -0.40498220
## Bad health       0.59133976
## Social assistance 0.07427985
## Log(income)      0.02111286
## Year=1996        0.01176704
## Year=1997        -0.01770400
## Year=1998        -0.06090803
## Year=1999        -0.05487621
```

```
summary(model_fit)$loglik[1]
```

```
## [1] -42738.86
```

Export the calculations

```
col_1 <- res_df_col1[,c("Estimate", "Robust SE")]
write.csv(col_1, "../out/table-2-pooled.csv")

col_2 <- res_df_col2[,c("Estimate", "Std. error")]
colnames(col_2) <- c("Estimate", "Robust SE")
write.csv(col_2, "../out/table-2-random.csv")

col_3 <- res_df_col3[,c("Estimate", "Std. error")]
colnames(col_3) <- c("Estimate", "Robust SE")
write.csv(col_3, "../out/table-2-fixed.csv")
```

```
col_1
```

##		Estimate	Robust SE
## Age		-0.0105723260	6.598343e-03
## Age^2		0.0001580161	8.007286e-05
## Male		-0.2085200124	2.091179e-02
## Education		-0.0057640252	3.701687e-03
## Married		0.0807752520	2.226030e-02
## Househould size		-0.0522052984	7.771470e-03
## Active Sport		0.0465699357	1.934927e-02
## Good health		-0.6108842053	1.901187e-02
## Bad health		0.8131080333	2.249984e-02
## Social assistance		0.0860960122	4.443502e-02
## Log(income)		0.0931194922	2.285981e-02
## Year=1996		0.0012175406	2.676300e-02
## Year=1997		-0.0302387301	2.688853e-02
## Year=1998		-0.1048116755	2.680600e-02
## Year=1999		-0.0986844438	2.726094e-02

```
col_2 %>%
data.frame() %>%
tibble::rownames_to_column()
```

##	rowname	Estimate	Robust.SE
## 1	Age	-0.0092679012	5.747857e-03
## 2	Age^2	0.0001884548	7.001972e-05
## 3	Male	-0.2948034124	2.027968e-02
## 4	Education	-0.0042384040	4.083674e-03
## 5	Married	0.0773282851	1.809373e-02
## 6	Househould size	-0.0554746693	5.961705e-03
## 7	Active Sport	0.0061441035	1.256618e-02
## 8	Good health	-0.5013500804	1.183233e-02
## 9	Bad health	0.6225151267	1.216031e-02
## 10	Social assistance	0.0405303262	2.640419e-02
## 11	Log(income)	0.0173742593	1.597776e-02
## 12	Year=1996	-0.0141918850	1.109711e-02
## 13	Year=1997	-0.0502742933	1.142705e-02
## 14	Year=1998	-0.1061518772	1.186677e-02
## 15	Year=1999	-0.1068475570	1.233431e-02

```
col_1 %>%
data.frame() %>%
tibble::rownames_to_column(var = 'index') %>%
rename(`Pooled Estimate`=`Estimate`,
       `Pooled Robust SE`=`Robust.SE`) %>%
left_join(
```

```

col_2 %>%
data.frame() %>%
tibble::rownames_to_column(var = 'index') %>%
rename(`Random Effects Estimate`=`Estimate`,
       `Random Effects Robust SE`=`Robust.SE`),
by='index'
) %>%
left_join(
col_3 %>%
data.frame() %>%
tibble::rownames_to_column(var = 'index') %>%
rename(`Fixed Effects Estimate`=`Estimate`,
       `Fixed Effects Robust SE`=`Robust.SE`),
by='index'
) %>%
column_to_rownames(var = 'index') %>%
write.csv("../out/table-2.csv")

```

## The calculation of table 3

```

model_name_vec <- c("Poisson", "Negative Binomial",
                   "Zero-inflated Negative Binomial",
                   "Hurdle Negative Binomial Model",
                   "Hurdle Probit Poisson Log Normal Model")
loglikelihood_vec <- c()
aic_vec <- c()
bic_vec <- c()

```

## Poisson Regression Model

```

model_fit_pos<- glm(
  doctco~age+age2+factor(male)+educ+factor(married)+hsize+
  factor(sport)+factor(goodh)+factor(badh)+factor(sozh)+
  loginc+factor(year)+
  factor(emp)+factor(winter)+factor(spring)+factor(fall),
  data = data_df, family = poisson)
summary(model_fit_pos)

```

```
##
```

```
## Call:
## glm(formula = doctco ~ age + age2 + factor(male) + educ + factor(married) +
##      hsize + factor(sport) + factor(goodh) + factor(badh) + factor(sozh) +
##      loginc + factor(year) + factor(emp) + factor(winter) + factor(spring) +
##      factor(fall), family = poisson, data = data_df)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -4.4143  -1.6487  -0.6007   0.4985  17.0287
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      6.670e-01  8.812e-02   7.569 3.75e-14 ***
## age             -1.057e-02  2.607e-03  -4.056 4.99e-05 ***
## age2             1.580e-04  3.132e-05   5.045 4.54e-07 ***
## factor(male)1    -2.085e-01  7.957e-03 -26.205 < 2e-16 ***
## educ            -5.764e-03  1.635e-03  -3.526 0.000422 ***
## factor(married)1  8.078e-02  9.024e-03   8.951 < 2e-16 ***
## hsize           -5.221e-02  3.071e-03 -17.001 < 2e-16 ***
## factor(sport)1    4.657e-02  8.508e-03   5.474 4.41e-08 ***
## factor(goodh)1   -6.109e-01  8.919e-03 -68.493 < 2e-16 ***
## factor(badh)1     8.131e-01  8.675e-03  93.732 < 2e-16 ***
## factor(sozh)1     8.610e-02  1.782e-02   4.831 1.36e-06 ***
## loginc           9.312e-02  9.297e-03  10.016 < 2e-16 ***
## factor(year)96    1.218e-03  1.062e-02   0.115 0.908726
## factor(year)97   -3.024e-02  1.077e-02  -2.808 0.004984 **
## factor(year)98   -1.048e-01  1.089e-02  -9.623 < 2e-16 ***
## factor(year)99   -9.868e-02  1.112e-02  -8.878 < 2e-16 ***
## factor(emp)part-time -1.589e-02  1.271e-02  -1.251 0.210965
## factor(emp)self-employed 2.376e-01  9.269e-03  25.628 < 2e-16 ***
## factor(emp)unemployed 7.352e-02  1.364e-02   5.392 6.98e-08 ***
## factor(winter)1   5.587e-03  1.265e-02   0.442 0.658754
## factor(spring)1   2.094e-02  1.240e-02   1.690 0.091101 .
## factor(fall)1     7.560e-03  2.439e-02   0.310 0.756605
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 145325  on 32836  degrees of freedom
## Residual deviance: 111603  on 32815  degrees of freedom
## AIC: 173176
##
## Number of Fisher Scoring iterations: 6
```



```
loglikelihood_vec <- c(loglikelihood_vec, logLik(model_fit_pos))
aic_vec <- c(aic_vec, AIC(model_fit_pos))
bic_vec <- c(bic_vec, BIC(model_fit_pos))
```

```
# Adding BIC as an alternative measurement
print(paste0("AIC:", AIC(model_fit_pos)))
```

```
## [1] "AIC:173176.358642251"
```

```
print(paste0("BIC:", BIC(model_fit_pos)))
```

```
## [1] "BIC:173361.143488814"
```

```
print(paste0("Log Likelihood:", logLik(model_fit_pos)))
```

```
## [1] "Log Likelihood:-86566.1793211256"
```

## Negative binomial Regression Model

```
model_fit_negbin <- MASS::glm.nb(
  doctco ~ age + age2 + factor(male) + educ + factor(married) + hsize +
    factor(sport) + factor(goodh) + factor(badh) + factor(sozh) +
    loginc + factor(year) +
    factor(emp) + factor(winter) + factor(spring) + factor(fall),
  data = data_df)
summary(model_fit_negbin)
```

```
##
## Call:
## MASS::glm.nb(formula = doctco ~ age + age2 + factor(male) + educ +
##   factor(married) + hsize + factor(sport) + factor(goodh) +
##   factor(badh) + factor(sozh) + loginc + factor(year) + factor(emp) +
##   factor(winter) + factor(spring) + factor(fall), data = data_df,
##   init.theta = 0.9077874406, link = log)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.1487  -1.2448  -0.3482   0.2619   6.1578
##
```

```
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)      8.836e-01  1.788e-01   4.941 7.75e-07 ***
## age             -1.937e-02  5.256e-03  -3.685 0.000229 ***
## age2             2.650e-04  6.405e-05   4.137 3.52e-05 ***
## factor(male)1    -2.907e-01  1.607e-02 -18.094 < 2e-16 ***
## educ            -4.154e-03  3.185e-03  -1.304 0.192235
## factor(married)1  1.055e-01  1.800e-02   5.862 4.57e-09 ***
## hsize           -5.192e-02  5.947e-03  -8.731 < 2e-16 ***
## factor(sport)1    7.410e-02  1.659e-02   4.467 7.95e-06 ***
## factor(goodh)1   -6.257e-01  1.647e-02 -37.996 < 2e-16 ***
## factor(badh)1     8.263e-01  2.127e-02  38.857 < 2e-16 ***
## factor(sozh)1     1.283e-01  3.802e-02   3.375 0.000739 ***
## loginc           8.542e-02  1.909e-02   4.476 7.62e-06 ***
## factor(year)96     7.413e-03  2.186e-02   0.339 0.734562
## factor(year)97    -3.504e-02  2.203e-02  -1.591 0.111653
## factor(year)98    -8.569e-02  2.192e-02  -3.909 9.26e-05 ***
## factor(year)99    -7.909e-02  2.242e-02  -3.528 0.000418 ***
## factor(emp)part-time -3.151e-02  2.530e-02  -1.246 0.212907
## factor(emp)self-employed 2.227e-01  1.930e-02  11.540 < 2e-16 ***
## factor(emp)unemployed 7.838e-02  2.746e-02   2.854 0.004316 **
## factor(winter)1    1.218e-02  2.511e-02   0.485 0.627780
## factor(spring)1    1.854e-02  2.456e-02   0.755 0.450444
## factor(fall)1     -3.699e-02  4.898e-02  -0.755 0.450105
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.9078) family taken to be 1)
##
##      Null deviance: 43088  on 32836  degrees of freedom
## Residual deviance: 34806  on 32815  degrees of freedom
## AIC: 129269
##
## Number of Fisher Scoring iterations: 1
##
##
##              Theta: 0.9078
##             Std. Err.: 0.0115
##
## 2 x log-likelihood: -129223.0910
```

```
loglikelihood_vec <- c(loglikelihood_vec, logLik(model_fit_negbin))
aic_vec <- c(aic_vec, AIC(model_fit_negbin))
bic_vec <- c(bic_vec, BIC(model_fit_negbin))
```

```
print(paste0("AIC:", AIC(model_fit_negbin)))
```

```
## [1] "AIC:129269.091288148"
```

```
print(paste0("BIC:", BIC(model_fit_negbin)))
```

```
## [1] "BIC:129462.275445918"
```

```
print(paste0("Log Likelihood:", logLik(model_fit_negbin)))
```

```
## [1] "Log Likelihood:-64611.5456440739"
```

## ZINB Regression Model

```
model_zeroinf_negbin = zeroinfl(formula = doctco ~ age + age2 + factor(male) + educ + factor(marr
```

```
## Warning in value[[3L]](cond): system is computationally singular: reciprocal  
## condition number = 5.31751e-17FALSE
```

```
summary(model_zeroinf_negbin)
```

```
##  
## Call:  
## zeroinfl(formula = doctco ~ age + age2 + factor(male) + educ + factor(married) +  
##   hsize + factor(sport) + factor(goodh) + factor(badh) + factor(sozh) +  
##   loginc + factor(year) + factor(emp) + factor(winter) + factor(spring) +  
##   factor(fall), data = data_df, dist = "negbin")  
##  
## Pearson residuals:  
##      Min      1Q  Median      3Q      Max  
## -0.9701 -0.6719 -0.3270  0.2715 21.8498  
##  
## Count model coefficients (negbin with log link):  
##              Estimate Std. Error z value Pr(>|z|)  
## (Intercept)      1.0407428      NA      NA      NA  
## age             -0.0157518      NA      NA      NA  
## age2             0.0002029      NA      NA      NA  
## factor(male)1    -0.1155955      NA      NA      NA  
## educ            -0.0078530      NA      NA      NA
```

```

## factor(married)1      0.0776323      NA      NA      NA
## hsize                 -0.0527561      NA      NA      NA
## factor(sport)1        0.0435412      NA      NA      NA
## factor(goodh)1        -0.5529706      NA      NA      NA
## factor(badh)1         0.7914992      NA      NA      NA
## factor(sozh)1         0.1309518      NA      NA      NA
## loginc                0.0684252      NA      NA      NA
## factor(year)96        -0.0162533      NA      NA      NA
## factor(year)97        -0.0267077      NA      NA      NA
## factor(year)98        -0.0815519      NA      NA      NA
## factor(year)99        -0.0895798      NA      NA      NA
## factor(emp)part-time  -0.0253989      NA      NA      NA
## factor(emp)self-employed 0.2183215      NA      NA      NA
## factor(emp)unemployed 0.0891762      NA      NA      NA
## factor(winter)1       -0.0239917      NA      NA      NA
## factor(spring)1       -0.0074596      NA      NA      NA
## factor(fall)1         -0.0444948      NA      NA      NA
## Log(theta)            0.0396095      NA      NA      NA
##
## Zero-inflation model coefficients (binomial with logit link):
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.572e+01      NA      NA      NA
## age             1.013e-01      NA      NA      NA
## age2           -1.516e-03      NA      NA      NA
## factor(male)1   1.619e+01      NA      NA      NA
## educ           -4.393e-02      NA      NA      NA
## factor(married)1 -4.777e-01      NA      NA      NA
## hsize           1.218e-02      NA      NA      NA
## factor(sport)1  -5.125e-01      NA      NA      NA
## factor(goodh)1   1.071e+00      NA      NA      NA
## factor(badh)1   -1.713e+00      NA      NA      NA
## factor(sozh)1    1.601e-01      NA      NA      NA
## loginc          -3.895e-01      NA      NA      NA
## factor(year)96  -4.278e-01      NA      NA      NA
## factor(year)97   5.894e-02      NA      NA      NA
## factor(year)98   5.165e-02      NA      NA      NA
## factor(year)99  -1.705e-01      NA      NA      NA
## factor(emp)part-time -1.458e-01      NA      NA      NA
## factor(emp)self-employed -8.546e-02      NA      NA      NA
## factor(emp)unemployed 8.809e-04      NA      NA      NA
## factor(winter)1 -4.843e-01      NA      NA      NA
## factor(spring)1 -3.437e-01      NA      NA      NA
## factor(fall)1   -9.360e-02      NA      NA      NA
##
## Theta = 1.0404

```

```
## Number of iterations in BFGS optimization: 81
## Log-likelihood: -6.437e+04 on 45 Df
```

```
print(paste0("AIC:", AIC(model_zeroinf_negbin)))
```

```
## [1] "AIC:128829.113266798"
```

```
print(paste0("Log Likelihood:", logLik(model_zeroinf_negbin)))
```

```
## [1] "Log Likelihood:-64369.5566333991"
```

```
loglikelihood_vec <- c(loglikelihood_vec, logLik(model_zeroinf_negbin))
aic_vec <- c(aic_vec, AIC(model_zeroinf_negbin))
bic_vec <- c(bic_vec, BIC(model_zeroinf_negbin))
```

## Hurdle Negative Binomial Model

```
hurdle_negbin <- hurdle(
  formula = doctco~age+age2+factor(male)+educ+factor(married)+hsize+ factor(sport)+fact
  summary(hurdle_negbin)
```

```
## Warning in sqrt(diag(object$vcov)): NaNs produced
```

```
##
```

```
## Call:
```

```
## hurdle(formula = doctco ~ age + age2 + factor(male) + educ + factor(married) +
##       hsize + factor(sport) + factor(goodh) + factor(badh) + factor(sozh) +
##       loginc + factor(year) + factor(emp) + factor(winter) + factor(spring) +
##       factor(fall), data = data_df, dist = "negbin")
##
```

```
##
```

```
## Pearson residuals:
```

```
##      Min      1Q  Median      3Q      Max
## -0.9796 -0.6637 -0.3207  0.2744 21.8481
```

```
##
```

```
## Count model coefficients (truncated negbin with log link):
```

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.2998218  0.1758057   7.394 1.43e-13 ***
## age             -0.0114729      NaN      NaN      NaN
## age2             0.0001599      NaN      NaN      NaN
## factor(male)1    -0.1005224  0.0199633  -5.035 4.77e-07 ***
```

```

## educ -0.0193004 0.0039623 -4.871 1.11e-06 ***
## factor(married)1 0.0571943 0.0220257 2.597 0.00941 **
## hsize -0.0511680 0.0071557 -7.151 8.64e-13 ***
## factor(sport)1 -0.0052468 0.0207712 -0.253 0.80058
## factor(goodh)1 -0.5493782 0.0207118 -26.525 < 2e-16 ***
## factor(badh)1 0.8036487 0.0152913 52.556 < 2e-16 ***
## factor(sozh)1 0.1515667 0.0476713 3.179 0.00148 **
## loginc 0.0266507 0.0238258 1.119 0.26332
## factor(year)96 -0.0393003 0.0271734 -1.446 0.14810
## factor(year)97 -0.0444285 0.0275466 -1.613 0.10678
## factor(year)98 -0.0747395 0.0275788 -2.710 0.00673 **
## factor(year)99 -0.1211930 0.0279594 -4.335 1.46e-05 ***
## factor(emp)part-time -0.0564210 0.0308661 -1.828 0.06756 .
## factor(emp)self-employed 0.2132689 0.0214050 9.963 < 2e-16 ***
## factor(emp)unemployed 0.1068685 0.0343057 3.115 0.00184 **
## factor(winter)1 -0.0017984 0.0317646 -0.057 0.95485
## factor(spring)1 -0.0131696 0.0310793 -0.424 0.67175
## factor(fall)1 -0.0800171 0.0615170 -1.301 0.19335
## Log(theta) -0.2306514 0.0289391 -7.970 1.58e-15 ***
## Zero hurdle model coefficients (binomial with logit link):
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.1162606 0.2505844 0.464 0.64268
## age -0.0460546 NaN NaN NaN
## age2 0.0005966 NaN NaN NaN
## factor(male)1 -0.6732591 0.0278037 -24.215 < 2e-16 ***
## educ 0.0263569 0.0053480 4.928 8.29e-07 ***
## factor(married)1 0.2330717 0.0301984 7.718 1.18e-14 ***
## hsize -0.0567317 0.0099162 -5.721 1.06e-08 ***
## factor(sport)1 0.2362838 0.0285506 8.276 < 2e-16 ***
## factor(goodh)1 -0.7542977 0.0286202 -26.355 < 2e-16 ***
## factor(badh)1 1.0179366 0.0476283 21.373 < 2e-16 ***
## factor(sozh)1 0.0738980 0.0683965 1.080 0.27995
## loginc 0.2094320 0.0335135 6.249 4.13e-10 ***
## factor(year)96 0.1132972 0.0385986 2.935 0.00333 **
## factor(year)97 0.0035324 0.0384344 0.092 0.92677
## factor(year)98 -0.0950786 0.0378807 -2.510 0.01207 *
## factor(year)99 0.0167302 0.0390497 0.428 0.66834
## factor(emp)part-time -0.0091876 0.0451180 -0.204 0.83864
## factor(emp)self-employed 0.2178569 0.0330445 6.593 4.32e-11 ***
## factor(emp)unemployed 0.0138258 0.0476030 0.290 0.77148
## factor(winter)1 0.0354222 0.0429999 0.824 0.41007
## factor(spring)1 0.0834093 0.0420312 1.984 0.04720 *
## factor(fall)1 0.0506215 0.0838215 0.604 0.54590
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```
##
## Theta: count = 0.794
## Number of iterations in BFGS optimization: 29
## Log-likelihood: -6.425e+04 on 45 Df
```

```
loglikelihood_vec <- c(loglikelihood_vec, logLik(hurdle_negbin))
aic_vec <- c(aic_vec, AIC(hurdle_negbin))
bic_vec <- c(bic_vec, BIC(hurdle_negbin))
```

```
print(paste0("AIC:", AIC(hurdle_negbin)))
```

```
## [1] "AIC:128599.794629153"
```

```
print(paste0("Log Likelihood:", logLik(hurdle_negbin)))
```

```
## [1] "Log Likelihood:-64254.8973145765"
```

## Hurdle Negative Binomial Model conditioned visiting experience

```
novisit_96 = which(data_df$doctco == 0 & data_df$year ==96)
novisit_98 = which(data_df$doctco == 0 & data_df$year ==98)
id_novisit_96 = data_df[novisit_96,]$id
id_novisit_98 = data_df[novisit_98,]$id
novisit_id = c(id_novisit_96, id_novisit_98)
df_visit = subset(data_df, !(data_df$id %in% novisit_id))
hurdle_negbin_visitonly <- pscl::hurdle(
  formula = doctco~age+age2+factor(male)+educ+factor(married)+hsize+ factor(sport)+fact
summary(hurdle_negbin_visitonly)
```

```
## Warning in sqrt(diag(object$vcov)): NaNs produced
```

```
##
## Call:
## pscl::hurdle(formula = doctco ~ age + age2 + factor(male) + educ + factor(married) +
##   hsize + factor(sport) + factor(goodh) + factor(badh) + factor(sozh) +
##   loginc + factor(year) + factor(emp) + factor(winter) + factor(spring) +
##   factor(fall), data = df_visit, dist = "negbin")
##
## Pearson residuals:
##      Min      1Q  Median      3Q      Max
```

```
## -1.0592 -0.7272 -0.3049 0.2704 19.6110
##
## Count model coefficients (truncated negbin with log link):
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.6253552   0.1966004   8.267 < 2e-16 ***
## age             -0.0087108         NaN      NaN      NaN
## age2             0.0001267         NaN      NaN      NaN
## factor(male)1    -0.0719020   0.0219133  -3.281 0.001034 **
## educ            -0.0212191   0.0043462  -4.882 1.05e-06 ***
## factor(married)1  0.0673011   0.0244436   2.753 0.005899 **
## hsize           -0.0437411   0.0080413  -5.440 5.34e-08 ***
## factor(sport)1   -0.0167969   0.0226445  -0.742 0.458231
## factor(goodh)1   -0.5290702   0.0231533 -22.851 < 2e-16 ***
## factor(badh)1     0.7571201   0.0202444  37.399 < 2e-16 ***
## factor(sozh)1     0.1323283   0.0509762   2.596 0.009435 **
## loginc          -0.0051085   0.0258328  -0.198 0.843238
## factor(year)96    -0.0540375   0.0294688  -1.834 0.066696 .
## factor(year)97    -0.0003962   0.0311040  -0.013 0.989837
## factor(year)98    -0.1165372   0.0296932  -3.925 8.68e-05 ***
## factor(year)99    -0.1149076   0.0318083  -3.612 0.000303 ***
## factor(emp)part-time -0.0912362   0.0337438  -2.704 0.006855 **
## factor(emp)self-employed 0.1620038   0.0237142   6.832 8.40e-12 ***
## factor(emp)unemployed 0.0565750   0.0372564   1.519 0.128880
## factor(winter)1   0.0016053   0.0341338   0.047 0.962490
## factor(spring)1   0.0056395   0.0332397   0.170 0.865275
## factor(fall)1     -0.0715543   0.0664898  -1.076 0.281852
## Log(theta)        -0.0202451   0.0274557  -0.737 0.460895
## Zero hurdle model coefficients (binomial with logit link):
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.859e-01  4.441e-01  -0.644 0.51977
## age             -1.538e-02         NaN      NaN      NaN
## age2             3.108e-04         NaN      NaN      NaN
## factor(male)1    -6.312e-01  5.306e-02 -11.895 < 2e-16 ***
## educ            -4.106e-03  1.032e-02  -0.398 0.69063
## factor(married)1  5.762e-02  5.938e-02   0.970 0.33190
## hsize           -5.105e-02  1.924e-02  -2.654 0.00795 **
## factor(sport)1   1.534e-01  5.517e-02   2.780 0.00543 **
## factor(goodh)1   -7.208e-01  5.279e-02 -13.652 < 2e-16 ***
## factor(badh)1     1.048e+00  5.065e-02  20.698 < 2e-16 ***
## factor(sozh)1     1.733e-01  1.389e-01   1.247 0.21225
## loginc          2.751e-01  6.548e-02   4.201 2.66e-05 ***
## factor(year)96    1.841e+01  2.894e+02   0.064 0.94928
## factor(year)97    1.551e-01  5.827e-02   2.662 0.00778 **
## factor(year)98    1.841e+01  2.888e+02   0.064 0.94916
## factor(year)99    1.225e-01  5.898e-02   2.077 0.03780 *
```



```
## factor(emp)part-time      -2.663e-02  8.657e-02  -0.308  0.75838
## factor(emp)self-employed  3.453e-01  5.883e-02   5.869  4.38e-09 ***
## factor(emp)unemployed     7.202e-02  9.263e-02   0.777  0.43690
## factor(winter)1           2.410e-02  8.709e-02   0.277  0.78203
## factor(spring)1           6.375e-02  8.451e-02   0.754  0.45060
## factor(fall)1             1.112e-01  1.890e-01   0.589  0.55610
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta: count = 0.98
## Number of iterations in BFGS optimization: 29
## Log-likelihood: -3.885e+04 on 45 Df
```

## Hurdle Probit Poisson Log Normal Model

```
hurdle_probit_poisson_lognormal = pscl:: hurdle(
  doctco~age+age2+factor(male)+educ+factor(married)+hsize+factor(sport)+factor(goodh)+f
summary(hurdle_probit_poisson_lognormal)
```

```
## Warning in sqrt(diag(object$vcov)): NaNs produced
```

```
##
## Call:
## pscl::hurdle(formula = doctco ~ age + age2 + factor(male) + educ + factor(married) +
##   hsize + factor(sport) + factor(goodh) + factor(badh) + factor(sozh) +
##   loginc + factor(year) + factor(emp) + factor(winter) + factor(spring) +
##   factor(fall), data = data_df, dist = "poisson", zero.dist = "binomial",
##   link = "probit")
##
## Pearson residuals:
##      Min      1Q  Median      3Q      Max
## -2.4682 -0.8749 -0.4462  0.3845 29.0886
##
## Count model coefficients (truncated poisson with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.319e+00  7.786e-02  16.945 < 2e-16 ***
## age           -4.100e-04  1.178e-04  -3.480 0.000501 ***
## age2           2.624e-05      NaN      NaN      NaN
## factor(male)1  -5.158e-02  8.344e-03  -6.182 6.32e-10 ***
## educ          -1.607e-02  1.748e-03  -9.195 < 2e-16 ***
## factor(married)1  3.014e-02  9.466e-03   3.184 0.001454 **
## hsize         -4.078e-02  3.230e-03 -12.625 < 2e-16 ***
```

```

## factor(sport)1      -1.595e-02  9.165e-03  -1.741  0.081766  .
## factor(goodh)1      -4.289e-01  9.780e-03 -43.860  < 2e-16 ***
## factor(badh)1       6.529e-01  8.927e-03  73.134  < 2e-16 ***
## factor(sozh)1       6.656e-02  1.859e-02   3.580  0.000344 ***
## loginc              3.469e-02  9.904e-03   3.503  0.000460 ***
## factor(year)96      -2.824e-02  1.122e-02  -2.517  0.011836 *
## factor(year)97      -3.345e-02  1.138e-02  -2.939  0.003290 **
## factor(year)98      -7.964e-02  1.155e-02  -6.893  5.45e-12 ***
## factor(year)99      -1.129e-01  1.182e-02  -9.556  < 2e-16 ***
## factor(emp)part-time -3.570e-02  1.362e-02  -2.621  0.008755 **
## factor(emp)self-employed 1.764e-01  9.300e-03  18.968  < 2e-16 ***
## factor(emp)unemployed 6.778e-02  1.439e-02   4.711  2.46e-06 ***
## factor(winter)1     -2.488e-04  1.349e-02  -0.018  0.985282
## factor(spring)1     -2.920e-03  1.323e-02  -0.221  0.825270
## factor(fall)1       -2.434e-02  2.600e-02  -0.936  0.349105
## Zero hurdle model coefficients (binomial with probit link):
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)      4.458e-02  2.118e-01   0.211  0.833277
## age              -2.734e-02  7.195e-03  -3.800  0.000145 ***
## age2             3.538e-04  9.308e-05   3.801  0.000144 ***
## factor(male)1     -4.027e-01  1.690e-02 -23.827  < 2e-16 ***
## educ             1.609e-02  3.331e-03   4.831  1.36e-06 ***
## factor(married)1  1.394e-01  1.887e-02   7.384  1.53e-13 ***
## hsize            -3.444e-02  6.147e-03  -5.603  2.11e-08 ***
## factor(sport)1     1.429e-01  1.740e-02   8.214  < 2e-16 ***
## factor(goodh)1     -4.586e-01  1.746e-02 -26.261  < 2e-16 ***
## factor(badh)1      5.720e-01  3.339e-02  17.131  < 2e-16 ***
## factor(sozh)1      4.565e-02  4.126e-02   1.107  0.268504
## loginc            1.293e-01  2.048e-02   6.313  2.74e-10 ***
## factor(year)96     6.711e-02  2.329e-02   2.882  0.003953 **
## factor(year)97     1.684e-03  2.327e-02   0.072  0.942298
## factor(year)98     -5.948e-02  2.298e-02  -2.588  0.009648 **
## factor(year)99     9.178e-03  2.364e-02   0.388  0.697815
## factor(emp)part-time -1.802e-03  2.707e-02  -0.067  0.946939
## factor(emp)self-employed 1.365e-01  2.247e-02   6.074  1.25e-09 ***
## factor(emp)unemployed 9.995e-03  2.898e-02   0.345  0.730136
## factor(winter)1    2.029e-02  2.610e-02   0.778  0.436853
## factor(spring)1    5.097e-02  2.551e-02   1.998  0.045729 *
## factor(fall)1      3.363e-02  5.094e-02   0.660  0.509122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 29
## Log-likelihood: -7.798e+04 on 44 Df

```

```
print(paste0("AIC:", AIC(hurdle_probit_poisson_lognormal)))
```

```
## [1] "AIC:156040.701360903"
```

```
print(paste0("Log Likelihood:", logLik(hurdle_probit_poisson_lognormal)))
```

```
## [1] "Log Likelihood:-77976.3506804513"
```

```
loglikelihood_vec <- c(loglikelihood_vec, logLik(hurdle_probit_poisson_lognormal))
aic_vec <- c(aic_vec, AIC(hurdle_probit_poisson_lognormal))
bic_vec <- c(bic_vec, BIC(hurdle_probit_poisson_lognormal))
```

## Two components Negative Binomial Model

```
doc_vist_df <- data_df %>%
  # filter(year==96 | year==95 ) %>%
  group_by(id) %>%
  summarise(n.visit=sum(doctco), n=n()) %>%
  arrange(n.visit)
n_id <- dim(doc_vist_df)[1]
group1_df <- doc_vist_df %>%
  slice_head(n=ceiling(n_id/3 * 2))
group2_df <- doc_vist_df %>%
  slice_tail(n=ceiling(n_id/3))
print(paste0("mu1:", sum(group1_df$n.visit) / sum(group1_df$n)))
```

```
## [1] "mu1:0.996606048325819"
```

```
print(paste0("mu2:", sum(group2_df$n.visit) / sum(group2_df$n)))
```

```
## [1] "mu2:4.8393893129771"
```

```
is_in_list <- function(x, y){
  x %in% y
}
mask <- sapply(data_df$id, is_in_list, group1_df$id)
group1_data_df <- data_df[mask, ]
mask <- sapply(data_df$id, is_in_list, group2_df$id)
group2_data_df <- data_df[mask, ]
```

```

model_fit_negbin1 <- MASS::glm.nb(
  doctco~age+age2+factor(male)+educ+factor(married)+hsize+
  factor(sport)+factor(goodh)+factor(badh)+factor(sozh)+
  loginc+factor(year)+
  factor(emp)+factor(winter)+factor(spring)+factor(fall),
  data = group1_data_df)
model_fit_negbin2 <- MASS::glm.nb(
  doctco~age+age2+factor(male)+educ+factor(married)+hsize+
  factor(sport)+factor(goodh)+factor(badh)+factor(sozh)+
  loginc+factor(year)+
  factor(emp)+factor(winter)+factor(spring)+factor(fall),
  data = group2_data_df)

```

```
summary(model_fit_negbin1)
```

```

##
## Call:
## MASS::glm.nb(formula = doctco ~ age + age2 + factor(male) + educ +
##   factor(married) + hsize + factor(sport) + factor(goodh) +
##   factor(badh) + factor(sozh) + loginc + factor(year) + factor(emp) +
##   factor(winter) + factor(spring) + factor(fall), data = group1_data_df,
##   init.theta = 1.423811804, link = log)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.8016  -1.1528  -0.6086   0.3645   3.6029
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    9.470e-01  2.331e-01   4.062 4.86e-05 ***
## age           -6.010e-02  6.957e-03  -8.639 < 2e-16 ***
## age2           7.173e-04  8.559e-05   8.381 < 2e-16 ***
## factor(male)1  -2.909e-01  2.160e-02 -13.469 < 2e-16 ***
## educ           2.446e-02  4.103e-03   5.962 2.49e-09 ***
## factor(married)1  9.964e-02  2.434e-02   4.094 4.23e-05 ***
## hsize          -3.107e-02  7.835e-03  -3.965 7.33e-05 ***
## factor(sport)1   1.026e-01  2.171e-02   4.727 2.28e-06 ***
## factor(goodh)1  -4.316e-01  2.195e-02 -19.663 < 2e-16 ***
## factor(badh)1    4.146e-01  3.715e-02  11.158 < 2e-16 ***
## factor(sozh)1    5.419e-02  5.399e-02   1.004 0.315558
## loginc          4.580e-02  2.519e-02   1.818 0.068998 .
## factor(year)96   1.979e-03  2.933e-02   0.067 0.946215
## factor(year)97  -1.093e-01  3.013e-02  -3.627 0.000287 ***
## factor(year)98  -8.766e-02  2.924e-02  -2.998 0.002720 **

```

```
## factor(year)99          -1.677e-02  2.968e-02  -0.565  0.572042
## factor(emp)part-time    -6.405e-03  3.444e-02  -0.186  0.852449
## factor(emp)self-employed 1.065e-01  2.651e-02  4.018  5.86e-05 ***
## factor(emp)unemployed   4.939e-03  3.803e-02  0.130  0.896669
## factor(winter)1         -4.290e-02  3.290e-02  -1.304  0.192295
## factor(spring)1         -4.179e-03  3.185e-02  -0.131  0.895615
## factor(fall)1           4.674e-02  6.228e-02  0.750  0.452979
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.4238) family taken to be 1)
##
##      Null deviance: 21456  on 19740  degrees of freedom
## Residual deviance: 20138  on 19719  degrees of freedom
## AIC: 53413
##
## Number of Fisher Scoring iterations: 1
##
##
##              Theta:  1.4238
##             Std. Err.: 0.0431
##
##  2 x log-likelihood:  -53366.5330
```

```
summary(model_fit_negbin2)
```

```
##
## Call:
## MASS::glm.nb(formula = doctco ~ age + age2 + factor(male) + educ +
##   factor(married) + hsize + factor(sport) + factor(goodh) +
##   factor(badh) + factor(sozh) + loginc + factor(year) + factor(emp) +
##   factor(winter) + factor(spring) + factor(fall), data = group2_data_df,
##   init.theta = 1.522954908, link = log)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.5175  -0.8791  -0.3333   0.2709   5.5421
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.5581610  0.2195036   7.099 1.26e-12 ***
## age           -0.0168959  0.0064472  -2.621 0.008776 **
## age2            0.0002174  0.0000776   2.802 0.005077 **
## factor(male)1    0.0147960  0.0193657   0.764 0.444848
```

```
## educ -0.0088445 0.0039543 -2.237 0.025309 *
## factor(married)1 0.0232419 0.0213646 1.088 0.276654
## hsize -0.0341723 0.0072021 -4.745 2.09e-06 ***
## factor(sport)1 0.0187754 0.0201591 0.931 0.351666
## factor(goodh)1 -0.3723720 0.0202484 -18.390 < 2e-16 ***
## factor(badh)1 0.5874606 0.0212733 27.615 < 2e-16 ***
## factor(sozh)1 0.0957758 0.0428103 2.237 0.025272 *
## loginc 0.0573323 0.0231858 2.473 0.013408 *
## factor(year)96 -0.0311873 0.0256821 -1.214 0.224610
## factor(year)97 -0.0521731 0.0257507 -2.026 0.042756 *
## factor(year)98 -0.0914600 0.0260804 -3.507 0.000453 ***
## factor(year)99 -0.1110335 0.0267087 -4.157 3.22e-05 ***
## factor(emp)part-time -0.0197404 0.0292691 -0.674 0.500027
## factor(emp)self-employed 0.1671559 0.0223114 7.492 6.79e-14 ***
## factor(emp)unemployed 0.0384565 0.0317315 1.212 0.225537
## factor(winter)1 -0.0342259 0.0305298 -1.121 0.262261
## factor(spring)1 -0.0173232 0.0300565 -0.576 0.564375
## factor(fall)1 -0.0661723 0.0601599 -1.100 0.271358
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.523) family taken to be 1)
##
## Null deviance: 17193 on 13099 degrees of freedom
## Residual deviance: 14511 on 13078 degrees of freedom
## AIC: 67560
##
## Number of Fisher Scoring iterations: 1
##
##
## Theta: 1.5230
## Std. Err.: 0.0260
##
## 2 x log-likelihood: -67514.3750
```

Another attempt to divide the population

```
group1_df <- data_df %>%
  arrange(doctco) %>%
  slice_head(n=ceiling(32837/3 * 2))
sum(group1_df$doctco) / ceiling(32837/3 * 2)
```

```
## [1] 0.6669103
```

```
group2_df <- data_df %>%
  arrange(doctco) %>%
  slice_tail(n=ceiling(32837/3))
sum(group2_df$doctco) / ceiling(32837/3)
```

```
## [1] 6.254614
```

## Relative position Negative Binomial Model

*# Try to create subdata based on there relative position in 1996 (mean, or quartiles w*  
*# sep = quantile(data\_df\$doctco[data\_df\$year == 96],probs = 0.66)*

```
sep = quantile(data_df$doctco[data_df$year == 96], probs = 0.67)
```

```
# rownum_sub96 = which(data_df$doctco <= sep & data_df$year ==96)
# rownum_upp96 = which(data_df$doctco > sep & data_df$year ==96)
# length(rownum_sub96) + length (rownum_upp96)
# id_sub96 = data_df[rownum_sub96,]$id
#
# (length(id_sub96) + length(id_upp96))*3
#
# df_sub96 = data_df[data_df$id %in% id_sub96,]
# id_upp96 = data_df[rownum_upp96,]$id
# df_upp96 = data_df[data_df$id %in% id_upp96,]
# dim(data_df)[1]
# mean(df_sub96$doctco)
# mean(df_upp96$doctco)
# dim(df_sub96)[1] + dim(df_upp96)[1]
```

```
sub96 = subset(data_df, (data_df$year == 96 & data_df$doctco < sep))
upp96 = subset(data_df, (data_df$year == 96 & data_df$doctco >= sep))
```

```
(dim(sub96)+dim(upp96)) == dim(data_df)
model_fit_negbin1 <- MASS::glm.nb(
  doctco~age+age2+factor(male)+educ+factor(married)+hsize+
  factor(sport)+factor(goodh)+factor(badh)+factor(sozh)+
  loginc+factor(year)+
  factor(emp)+factor(winter)+factor(spring)+factor(fall),
  data = df_sub96)
summary(model_fit_negbin1)
print(paste0("AIC:", AIC(model_fit_negbin1)))
print(paste0("Log Likelihood:", logLik(model_fit_negbin1)))
```

```
hcg_df2 = subset(data_df, data_df$doctco > sep)
model_fit_negbin2 <- MASS::glm.nb(
  doctco~age+age2+factor(male)+educ+factor(married)+hsize+
  factor(sport)+factor(goodh)+factor(badh)+factor(sozh)+
  loginc+factor(year)+
  factor(emp)+factor(winter)+factor(spring)+factor(fall),
  data = df_upp96)
summary(model_fit_negbin2)
print(paste0("AIC:", AIC(model_fit_negbin2)))
print(paste0("Log Likelihood:", logLik(model_fit_negbin2)))
```

```
table3_df <- data.frame("Models"=model_name_vec,
  "Log-likelihood"=loglikelihood_vec,
  "AIC"=aic_vec,
  "BIC"=bic_vec)
table3_df
```

##	Models	Log.likelihood	AIC	BIC
## 1	Poisson	-86566.18	173176.4	173361.1
## 2	Negative Binomial	-64611.55	129269.1	129462.3
## 3	Zero-inflated Negative Binomial	-64369.56	128829.1	129207.1
## 4	Hurdle Negative Binomial Model	-64254.90	128599.8	128977.8
## 5	Hurdle Probit Poisson Log Normal Model	-77976.35	156040.7	156410.3

```
write_csv(table3_df, "../out/table-3.csv")
```

## The calculation of table 5

From the Poisson Model, the relative change in expected doctor visits from 1996 to 1998 is

```
100*(exp(coef(model_fit_pos)[15]-coef(model_fit_pos)[13])-1)
```

```
## factor(year)98
## -10.06016
```

From the Negative Binomial Model, the relative change in expected doctor visits from 1996 to 1998 is



```
100*(exp(coef(model_fit_negbin)[15]-coef(model_fit_negbin)[13])-1)
```

```
## factor(year)98
##      -8.889712
```

From the Zero Inflation Model, the relative change in expected doctor visits from 1996 to 1998 is

```
100*(exp(coef(model_zeroinf_negbin)[15]-coef(model_zeroinf_negbin)[13])-1)
```

```
## count_factor(year)98
##      -6.321232
```

From the Hurdle Binomial Model, the relative change in expected doctor visits from 1996 to 1998 is

```
100*(exp(coef(hurdle_negbin)[15]-coef(hurdle_negbin)[13])-1)
```

```
## count_factor(year)98
##      -3.481853
```

```
sprintf("From the Two Component Negative Binomial Model, the relative change in expected
sprintf("From the Two Component Negative Binomial Model, the relative change in expected
```

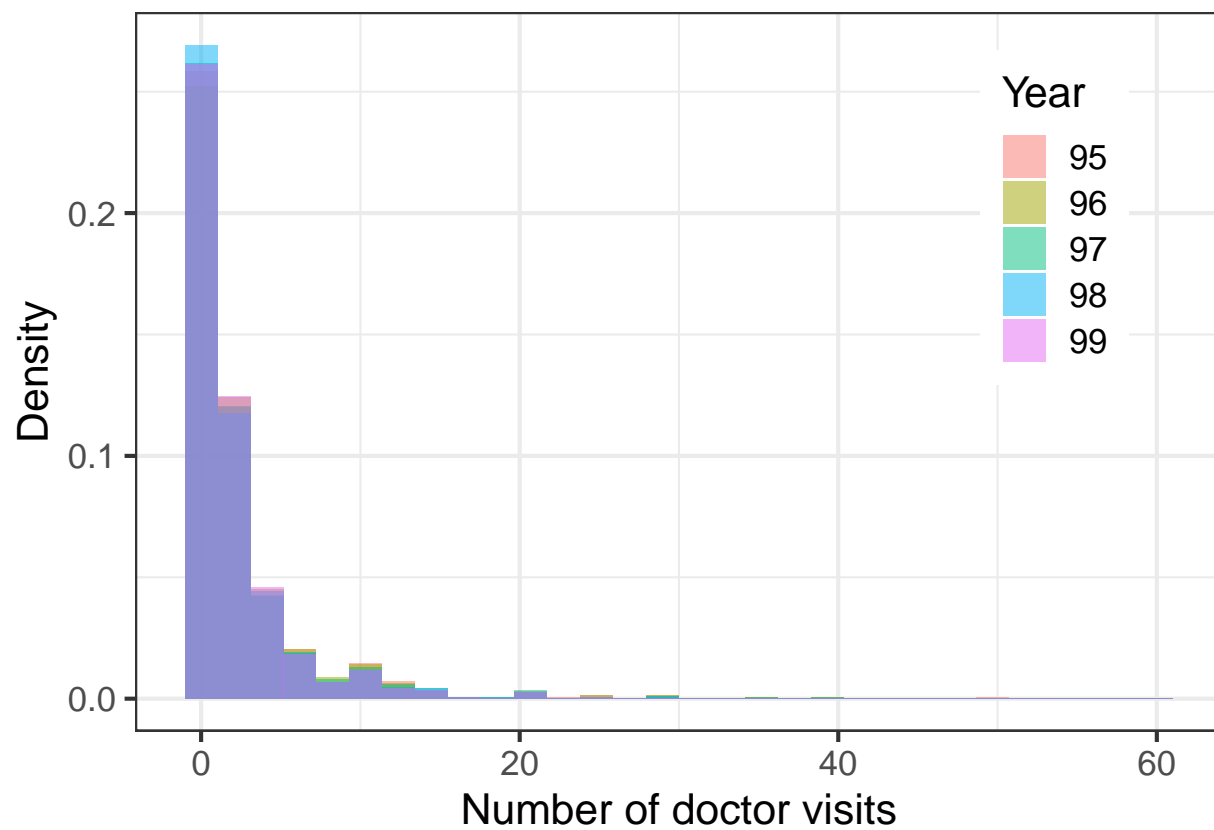
## Box, violin, and histogram plots

```
hist_plt <- hcg %>%
  mutate(Year=factor(year)) %>%
  ggplot() +
  geom_histogram(alpha = 0.5,
                 aes(x = doctco,
                     y = ..density..,
                     fill = Year),
                 position = 'identity',
                 alpha = 0.4) +
  xlab("Number of doctor visits") +
  ylab("Density") +
  theme_bw(base_size=16) +
  theme(legend.justification=c(0.9, 0.9),
        legend.position=c(0.9, 0.9))
```

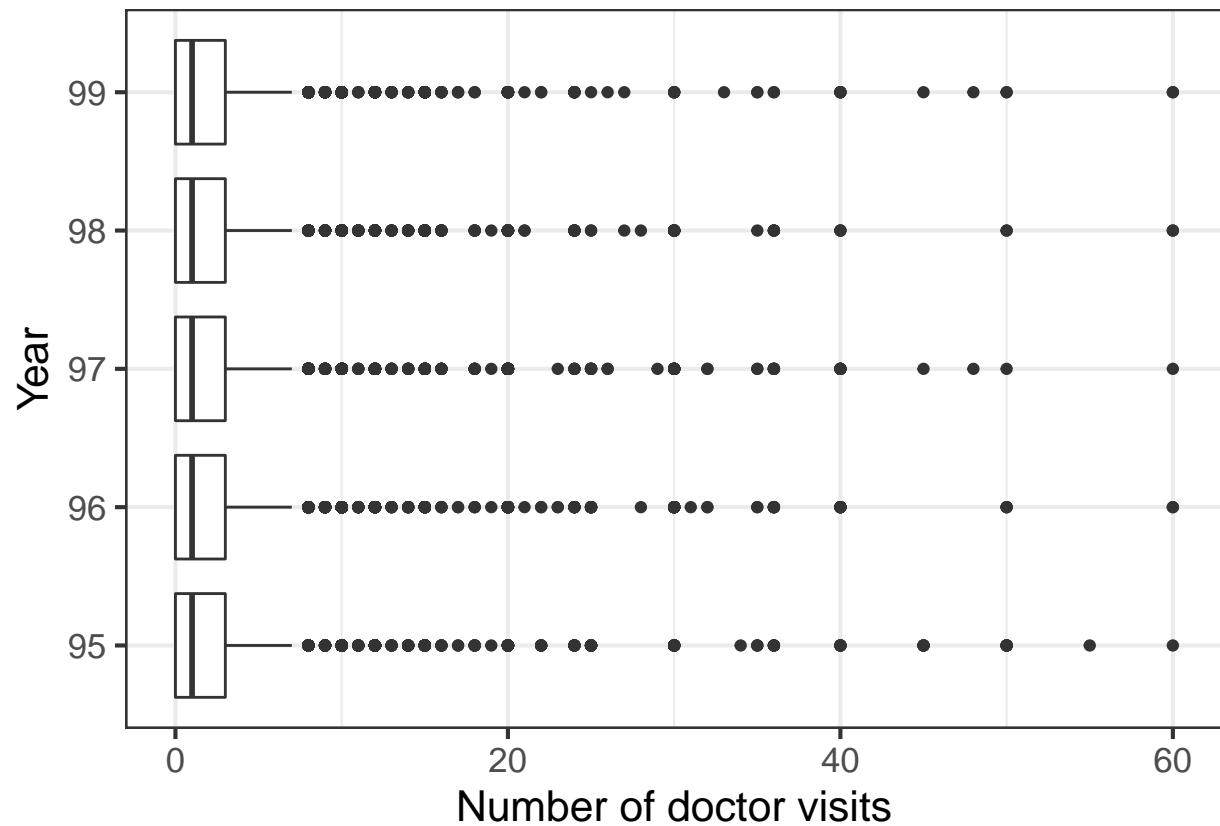
```
## Warning: Duplicated aesthetics after name standardisation: alpha
```

```
hist_plt
```

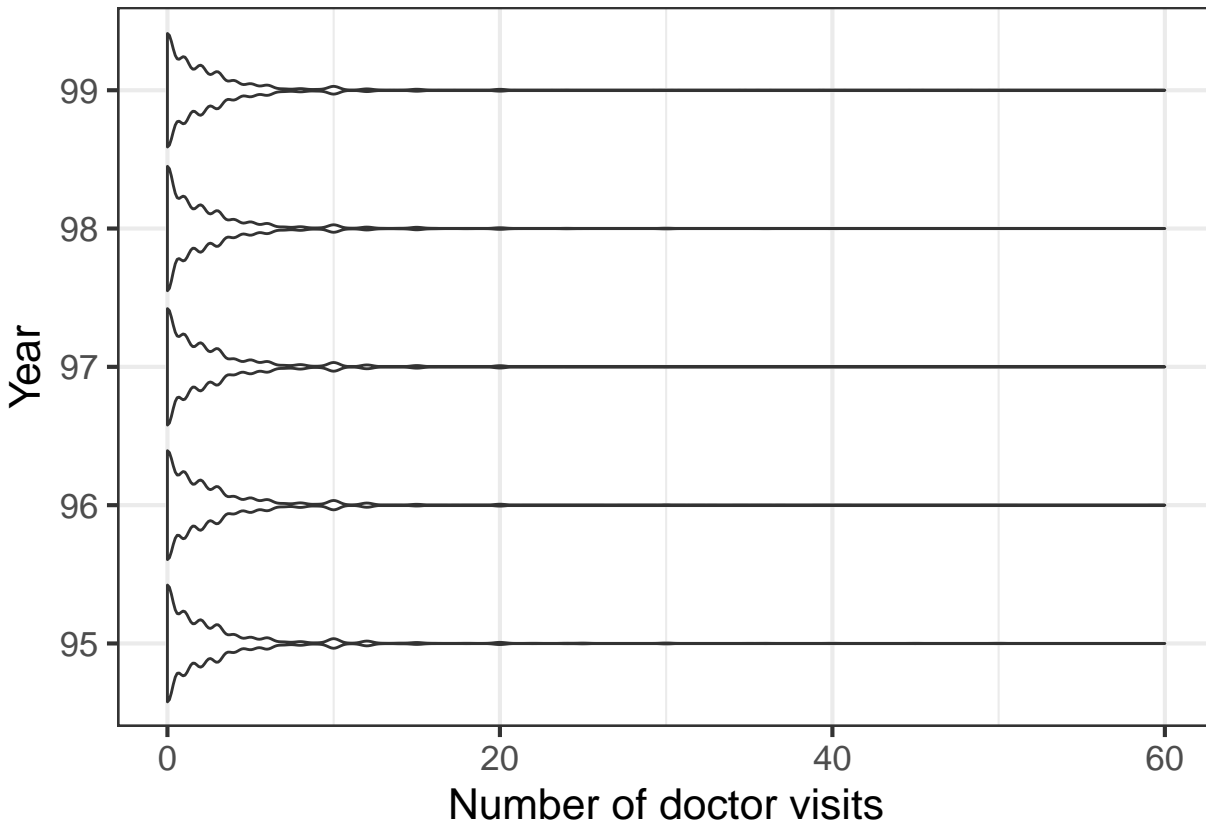
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
box_plt <- hcg %>%  
  mutate(Year=factor(year)) %>%  
  ggplot() +  
  geom_boxplot(aes(x=doctco, y=Year)) +  
  xlab("Number of doctor visits") +  
  ylab("Year") +  
  theme_bw(base_size=16)  
box_plt
```



```
violin_plt <- hcg %>%
  mutate(Year=factor(year)) %>%
  ggplot()+
  geom_violin(aes(x=doctco, y=Year))+
  xlab("Number of doctor visits") +
  ylab("Year") +
  theme_bw(base_size=16)
violin_plt
```



```
ggsave(plot = hist_plt, file = paste0("../out/hist.png"), width = 5, height = 5)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
ggsave(plot = box_plt, file = paste0("../out/box.png"), width = 5, height = 5)
ggsave(plot = violin_plt, file = paste0("../out/violin.png"), width = 5, height = 5)
```

```
ggsave(plot = grid.arrange(box_plt, violin_plt, ncol=2), file = paste0("../out/box_violin.png"), width = 10, height = 10)
```

