432 Class 15

Thomas E. Love, Ph.D.

2024-03-05

## This Week’s Main Topic

**Regression Models for Count Outcomes**

* Modeling approaches illustrated in these slides
  + Poisson Regression and Zero-Inflated Poisson (ZIP)
  + Negative Binomial Regression and Zero-Inflated Negative Binomial (ZINB)
  + Two types of Hurdle model (one Poisson, one NB)

Chapters 24-26 of the Course Notes describe this material, as well as tobit regression, and some additional issues with certain types of count models.

## countreg and topmodels packages

To build rootograms to visualize the results of regression models on count outcomes, I have decided for the moment to continue to use the countreg and topmodels packages, which are currently available only on R-Forge. To install, type:

install.packages("countreg", repos="http://R-Forge.R-project.org")  
install.packages("topmodels", repos="http://R-Forge.R-project.org")

into the R Console within R Studio.

## Setup

knitr::opts\_chunk$set(comment=NA)  
options(width = 60)  
  
library(janitor); library(gt); library(broom)   
library(mosaic); library(Hmisc); library(patchwork)  
library(rsample); library(yardstick)  
library(conflicted) ## resolve conflicts  
library(countreg) ## for rootograms  
library(topmodels) ## for rootograms  
library(MASS) ## for glm.nb to fit NB models  
library(pscl) ## for zero-inflated and hurdle fits  
library(lmtest) ## for Vuong test  
library(tidyverse)  
  
conflicts\_prefer(dplyr::select(), dplyr::filter(), base::sum(),  
 pscl::zeroinfl(), pscl::hurdle())  
  
theme\_set(theme\_bw())

# An Overview

## GLMs for Count Outcomes

We want to build a generalized linear model to predict count data using one or more predictors.

Count data are non-negative integers (0, 1, 2, 3, …)

* the number of COVID-19 hospitalizations in Ohio yesterday
* the number of mutations within a particular search grid
* days in the past 30 where your mental health was poor

We’ll use the Poisson and Negative Binomial probability distributions.

## The Poisson Probability Distribution

The Poisson probability model describes the probability of a given number of events occurring in a fixed interval of time or space.

* If events occur with a constant mean rate, and independently of the time since the last event, the Poisson model is appropriate.
  + A Poisson model might fit poorly due to **overdispersion**, where the variance of Y is larger than we’d expect based on the mean of Y.

## Poisson probability mass function

* is the number of times an event occurs in an interval, and can take the values 0, 1, 2, 3, …
* The parameter (lambda) is equal to the expected value (mean) of and is also equal to the variance of .

## Negative Binomial Distribution

The Negative Binomial distribution models the number of failures in a sequence of independent and identically distributed Bernoulli trials before a specified number of successes occurs.

## Probability Mass Function

Negative binomial’s probability mass function is …

* is the number of failures (units of time) before the th event occurs, and can take the values 0, 1, 2, 3, …
* The mean of the random variable Y which follows a negative binomial distribution is and the variance is .

## Poisson regression

* Poisson regression assumes that the outcome Y follows a Poisson distribution, and that the logarithm of the expected value of Y (its mean) can be modeled by a linear combination of a set of predictors.
  + A Poisson regression makes the strong assumption that the variance of Y is equal to its mean.

We will use glm to fit Poisson models, by using family = "Poisson".

## Dealing with Overdispersion

A Poisson model might fit poorly due to **overdispersion**, where the variance of Y is larger than we’d expect based on the mean of Y.

* *Quasipoisson* models are available which estimate an overdispersion parameter, but we’ll skip those for now.

Instead, we’ll look at other ways (especially zero-inflation and the negative binomial models) to address overdispersion.

## Negative Binomial Regression

* Negative binomial regression is a generalization of Poisson regression which loosens the assumption that the variance of Y is equal to its mean, and thus produces models which fit a broader class of data.

We will demonstrate the use of glm.nb() from the MASS package to fit negative binomial regression models.

## Zero-inflated approaches

* Both the Poisson and Negative Binomial regression approaches may under-estimate the number of zeros compared to the data.
* To better match the zero counts, zero-inflated models fit:
  + a logistic regression to predict the extra zeros, along with
  + a Poisson or Negative Binomial model to predict the counts, including some zeros.

We’ll use zeroinfl() from pscl to fit ZIP and ZINB regressions.

## Hurdle models

A hurdle model predicts the count outcome by making an assumption that there are two processes at work:

* a process that determines whether the count is zero or not zero (usually using logistic regression), and
* a process that determines the count when we know the subject has a positive count (usually using a truncated Poisson or NB model where no zeros are predicted)

We use hurdle() from pscl to fit these.

## Comparing Models

1. A key tool will be a graphical representation of the fit of the models to the count outcome, called a **rootogram**. We’ll use the rootograms produced by the countreg and topmodels packages to help us.
2. We’ll also demonstrate a Vuong hypothesis testing approach (from the lmtest package) to help us make decisions between various types of Poisson models or various types of Negative Binomial models on the basis of improvement in fit of things like bias-corrected AIC or BIC.

## Comparing Models

1. We’ll also demonstrate the calculation of pseudo-R square statistics for comparing models, which can be compared in a validation sample as well as in the original modeling sample.

# The medicare data

## The medicare example

Source: NMES1988 data in R’s AER package, cleaned up to medicare.csv.

Essentially the same data are used in from the University of Virginia on hurdle models.

Data are a cross-section US National Medical Expenditure Survey (NMES) conducted in 1987 and 1988. The NMES is based upon a representative, national probability sample of the civilian non-institutionalized population and individuals admitted to long-term care facilities during 1987.

## Ingesting medicare data

The data are a subsample of individuals ages 66 and over all of whom are covered by Medicare (a public insurance program providing substantial protection against health-care costs), and some of whom also have private supplemental insurance.

medicare <- read\_csv("c15/data/medicare.csv", show\_col\_types = FALSE) |>   
 mutate(across(where(is\_character), as\_factor),  
 subject = as.character(subject))

## The medicare code book

| Variable | Description |
| --- | --- |
| subject | subject number (code) |
| visits | outcome: # of physician office visits |
| hospital | # of hospital stays |
| health | self-rated health (poor, average, excellent) |
| chronic | # of chronic conditions |
| sex | male or female |
| school | years of education |
| insurance | subject (also) has private insurance? (yes/no) |

### Today’s Goal

Predict visits using main effects of the 6 predictors (excluding subject)

## The medicare tibble

medicare |> select(-subject)

# A tibble: 4,406 × 7  
 visits hospital health chronic sex school insurance  
 <dbl> <dbl> <fct> <dbl> <fct> <dbl> <fct>   
 1 5 1 average 2 male 6 yes   
 2 1 0 average 2 female 10 yes   
 3 13 3 poor 4 female 10 no   
 4 16 1 poor 2 male 3 yes   
 5 3 0 average 2 female 6 yes   
 6 17 0 poor 5 female 7 no   
 7 9 0 average 0 female 8 yes   
 8 3 0 average 0 female 8 yes   
 9 1 0 average 0 female 8 yes   
10 0 0 average 0 female 8 yes   
# ℹ 4,396 more rows

## Quick Summary of medicare

medicare |> select(-subject) |> summary()

visits hospital health   
 Min. : 0.000 Min. :0.000 average :3509   
 1st Qu.: 1.000 1st Qu.:0.000 poor : 554   
 Median : 4.000 Median :0.000 excellent: 343   
 Mean : 5.774 Mean :0.296   
 3rd Qu.: 8.000 3rd Qu.:0.000   
 Max. :89.000 Max. :8.000   
 chronic sex school insurance   
 Min. :0.000 male :1778 Min. : 0.00 yes:3421   
 1st Qu.:1.000 female:2628 1st Qu.: 8.00 no : 985   
 Median :1.000 Median :11.00   
 Mean :1.542 Mean :10.29   
 3rd Qu.:2.000 3rd Qu.:12.00   
 Max. :8.000 Max. :18.00

### Adjust order of insurance

medicare <- medicare |>  
 mutate(insurance = fct\_relevel(insurance, "no", "yes"))

I want No first, then Yes, when building models.

## Our outcome, visits

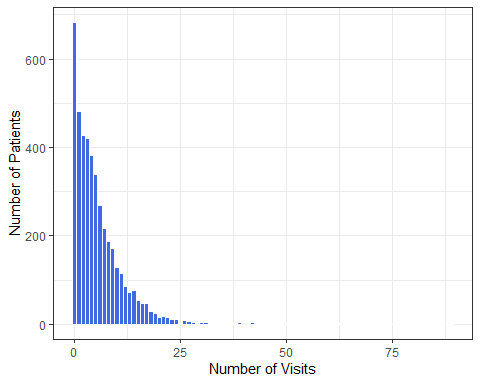
favstats(~ visits, data = medicare)

min Q1 median Q3 max mean sd n missing  
 0 1 4 8 89 5.774399 6.759225 4406 0

describe(medicare$visits) # from Hmisc

medicare$visits   
 n missing distinct Info Mean Gmd   
 4406 0 60 0.992 5.774 6.227   
 .05 .10 .25 .50 .75 .90   
 0 0 1 4 8 13   
 .95   
 17   
  
lowest : 0 1 2 3 4, highest: 63 65 66 68 89

ggplot(medicare, aes(x = visits)) +  
 geom\_histogram(binwidth = 1, fill = "royalblue",   
 col = "white") +  
 labs(y = "Number of Patients", x = "Number of Visits")



## Partitioning the Data

Creating Training and Testing Samples with rsample functions…

set.seed(432)  
med\_split <- initial\_split(medicare, prop = 0.75)  
  
med\_train = training(med\_split)  
med\_test = testing(med\_split)

I’ve held out 25% of the medicare data for the test sample.

dim(med\_train); dim(med\_test)

[1] 3304 8

[1] 1102 8

## Reiterating the Goal

Predict visits using some combination of these 6 predictors…

| Predictor | Description |
| --- | --- |
| hospital | # of hospital stays |
| health | self-rated health (poor, average, excellent) |
| chronic | # of chronic conditions |
| sex | male or female |
| school | years of education |
| insurance | subject (also) has private insurance? (yes/no) |

We’ll build separate training and test samples to help us validate.

# mod\_1: A Poisson Regression

## Poisson Regression

Assume our count data (visits) follows a Poisson distribution with a mean conditional on our predictors.

mod\_1 <- glm(visits ~ hospital + health + chronic +  
 sex + school + insurance,  
 data = med\_train, family = "poisson")

The Poisson model uses a logarithm as its link function, so the model is actually predicting log(visits).

Note that we’re fitting the model here using the training sample alone.

## Complete mod\_1 Summary

summary(mod\_1)

Call:  
glm(formula = visits ~ hospital + health + chronic + sex + school +   
 insurance, family = "poisson", data = med\_train)  
  
Coefficients:  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) 0.886576 0.028813 30.770 < 2e-16 \*\*\*  
hospital 0.163555 0.006710 24.374 < 2e-16 \*\*\*  
healthpoor 0.309610 0.020244 15.294 < 2e-16 \*\*\*  
healthexcellent -0.358758 0.034875 -10.287 < 2e-16 \*\*\*  
chronic 0.137349 0.005266 26.082 < 2e-16 \*\*\*  
sexfemale 0.098325 0.014805 6.641 3.11e-11 \*\*\*  
school 0.031258 0.002111 14.808 < 2e-16 \*\*\*  
insuranceyes 0.200249 0.019484 10.278 < 2e-16 \*\*\*  
---  
Signif. codes:   
0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for poisson family taken to be 1)  
  
 Null deviance: 20618 on 3303 degrees of freedom  
Residual deviance: 17598 on 3296 degrees of freedom  
AIC: 27232  
  
Number of Fisher Scoring iterations: 5

## mod\_1 (Poisson) model coefficients

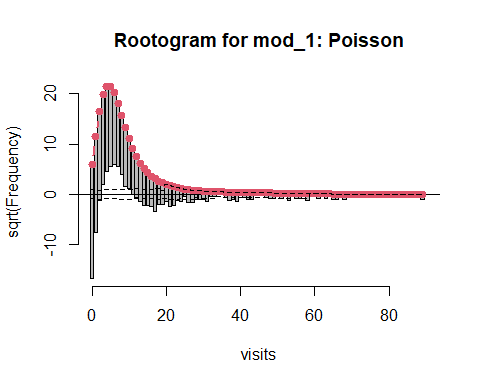
tidy(mod\_1) |> gt() |> fmt\_number(decimals = 3)

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | 0.887 | 0.029 | 30.770 | 0.000 |
| hospital | 0.164 | 0.007 | 24.374 | 0.000 |
| healthpoor | 0.310 | 0.020 | 15.294 | 0.000 |
| healthexcellent | -0.359 | 0.035 | -10.287 | 0.000 |
| chronic | 0.137 | 0.005 | 26.082 | 0.000 |
| sexfemale | 0.098 | 0.015 | 6.641 | 0.000 |
| school | 0.031 | 0.002 | 14.808 | 0.000 |
| insuranceyes | 0.200 | 0.019 | 10.278 | 0.000 |

Harry and Larry have the same values for all other predictors but only Harry has private insurance. mod\_1 estimates Harry’s log(visits) to be 0.2 larger than Larry’s log(visits).

## Visualize fit: (Hanging) Rootogram

plot(rootogram(mod\_1, plot = FALSE), xlim = c(0, 90),   
 main = "Rootogram for mod\_1: Poisson")



See the next slide for details on how to interpret this…

## Interpreting the Rootogram

* The red curved line is the theoretical Poisson fit.
* “Hanging” from each point on the red line is a bar, the height of which represents the observed counts.
  + A bar hanging below 0 indicates that the model under-predicts that value. (Model predicts fewer values than the data show.)
  + A bar hanging above 0 indicates over-prediction of that value. (Model predicts more values than the data show.)
* The counts have been transformed with a square root transformation to prevent smaller counts from getting obscured and overwhelmed by larger counts.
* <https://arxiv.org/pdf/1605.01311> has more on rootograms.
* Our Poisson model (mod\_1) doesn’t fit enough zeros or ones, and fits too many 3-12 values, then not enough of the higher values.

## Store mod\_1 Predictions

We’ll use the augment function to store the predictions within our training sample. Note the use of "response" to predict visits, not log(visits).

mod\_1\_aug <- augment(mod\_1, med\_train,   
 type.predict = "response")  
  
mod\_1\_aug |> select(subject, visits, .fitted) |> head(3)

# A tibble: 3 × 3  
 subject visits .fitted  
 <chr> <dbl> <dbl>  
1 355 19 5.02  
2 2661 3 4.21  
3 2895 0 4.65

## Training Sample mod\_1 Fit

Within our training sample, mod\_1\_aug now contains both the actual counts (visits) and the predicted counts (in .fitted) from mod\_1. We’ll summarize the fit…

mets <- metric\_set(rsq, rmse, mae)  
mod\_1\_summary <-   
 mets(mod\_1\_aug, truth = visits, estimate = .fitted) |>  
 mutate(model = "mod\_1") |> relocate(model)  
mod\_1\_summary |> gt() |> fmt\_number(decimals = 3)

| model | .metric | .estimator | .estimate |
| --- | --- | --- | --- |
| mod\_1 | rsq | standard | 0.100 |
| mod\_1 | rmse | standard | 6.594 |
| mod\_1 | mae | standard | 4.189 |

These will become interesting as we build additional models.

# mod\_2: A Negative Binomial Regression

## Fitting the Negative Binomial Model

The negative binomial model requires the estimation of an additional parameter, called (theta). The default link for this generalized linear model is also a logarithm, like the Poisson.

mod\_2 <- MASS::glm.nb(visits ~ hospital + health + chronic +  
 sex + school + insurance,  
 data = med\_train)

The estimated dispersion parameter value is…

summary(mod\_2)$theta

[1] 1.21109

The Poisson model is essentially the negative binomial model assuming a known .

## Complete mod\_2 summary

summary(mod\_2)

Call:  
MASS::glm.nb(formula = visits ~ hospital + health + chronic +   
 sex + school + insurance, data = med\_train, init.theta = 1.211089878,   
 link = log)  
  
Coefficients:  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) 0.765746 0.065602 11.673 < 2e-16 \*\*\*  
hospital 0.224205 0.022745 9.857 < 2e-16 \*\*\*  
healthpoor 0.360067 0.055608 6.475 9.47e-11 \*\*\*  
healthexcellent -0.335591 0.070353 -4.770 1.84e-06 \*\*\*  
chronic 0.169070 0.013887 12.174 < 2e-16 \*\*\*  
sexfemale 0.109443 0.035920 3.047 0.00231 \*\*   
school 0.030763 0.005037 6.107 1.02e-09 \*\*\*  
insuranceyes 0.237080 0.045780 5.179 2.23e-07 \*\*\*  
---  
Signif. codes:   
0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for Negative Binomial(1.2111) family taken to be 1)  
  
 Null deviance: 4341.7 on 3303 degrees of freedom  
Residual deviance: 3783.1 on 3296 degrees of freedom  
AIC: 18328  
  
Number of Fisher Scoring iterations: 1  
  
 Theta: 1.2111   
 Std. Err.: 0.0388   
  
 2 x log-likelihood: -18309.8280

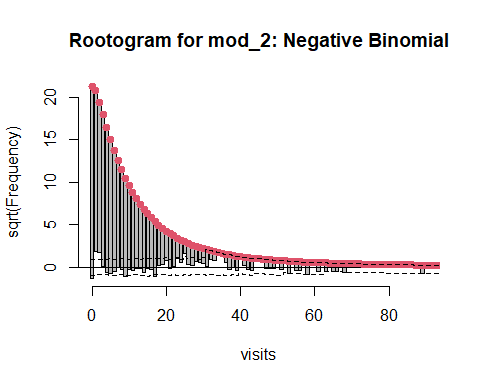
## mod\_2 (NB) coefficients

tidy(mod\_2) |> gt() |> fmt\_number(decimals = 3)

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | 0.766 | 0.066 | 11.673 | 0.000 |
| hospital | 0.224 | 0.023 | 9.857 | 0.000 |
| healthpoor | 0.360 | 0.056 | 6.475 | 0.000 |
| healthexcellent | -0.336 | 0.070 | -4.770 | 0.000 |
| chronic | 0.169 | 0.014 | 12.174 | 0.000 |
| sexfemale | 0.109 | 0.036 | 3.047 | 0.002 |
| school | 0.031 | 0.005 | 6.107 | 0.000 |
| insuranceyes | 0.237 | 0.046 | 5.179 | 0.000 |

## Rootogram for NB Model

plot(rootogram(mod\_2, plot = FALSE), xlim = c(0, 90),   
 main = "Rootogram for mod\_2: Negative Binomial")



Does this look better than the Poisson rootogram?

## Store mod\_2 Predictions

mod\_2\_aug <- augment(mod\_2, med\_train, type.predict = "response")  
  
mod\_2\_aug |> select(subject, visits, .fitted) |> head(3)

# A tibble: 3 × 3  
 subject visits .fitted  
 <chr> <dbl> <dbl>  
1 355 19 5.22  
2 2661 3 4.08  
3 2895 0 4.39

* Note that this *may* throw a warning about who maintains tidiers for negbin models. I’d silence it, as I have here.

## Training Fit for mod\_2

mod\_2\_aug has actual (visits) and predicted counts (in .fitted.)

mod\_2\_summary <-   
 mets(mod\_2\_aug, truth = visits, estimate = .fitted) |>  
 mutate(model = "mod\_2") |> relocate(model)  
mod\_2\_summary |> gt() |> fmt\_number(decimals = 3)

| model | .metric | .estimator | .estimate |
| --- | --- | --- | --- |
| mod\_2 | rsq | standard | 0.078 |
| mod\_2 | rmse | standard | 6.941 |
| mod\_2 | mae | standard | 4.252 |

## Training Sample So Far

The reasonable things to summarize in sample look like the impressions from the rootograms and the summaries we’ve prepared so far.

| Model | Rootogram impressions |
| --- | --- |
| mod\_1 (P) | Many problems. Data appear overdispersed. |
| mod\_2 (NB) | Still not enough zeros; some big predictions. |

## Training Sample Summaries

bind\_rows(mod\_1\_summary, mod\_2\_summary) |>   
 pivot\_wider(names\_from = model,   
 values\_from = .estimate) |>   
 gt() |> fmt\_number(decimals = 3) |>   
 tab\_options(table.font.size = 20)

| .metric | .estimator | mod\_1 | mod\_2 |
| --- | --- | --- | --- |
| rsq | standard | 0.100 | 0.078 |
| rmse | standard | 6.594 | 6.941 |
| mae | standard | 4.189 | 4.252 |

# mod\_3: Zero-Inflated Poisson (ZIP) Model

## Zero-Inflated Poisson (ZIP) model

The zero-inflated Poisson model describes count data with an excess of zero counts.

The model posits that there are two processes involved:

* a logistic regression model is used to predict excess zeros
* while a Poisson model is used to predict the counts

We’ll use the pscl package to fit zero-inflated models.

mod\_3 <- pscl::zeroinfl(visits ~ hospital + health +   
 chronic + sex + school + insurance,  
 data = med\_train)

## mod\_3 ZIP coefficients

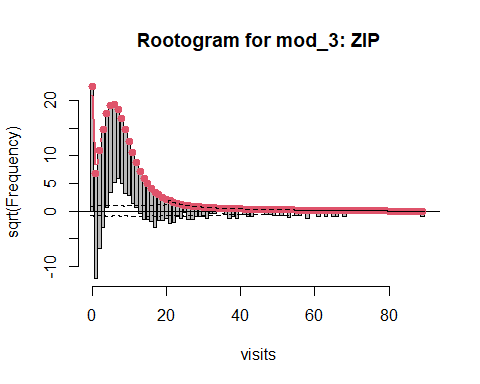
Sadly, there’s no broom tidying functions for these zero-inflated models.

summary(mod\_3)

Call:  
pscl::zeroinfl(formula = visits ~ hospital + health +   
 chronic + sex + school + insurance, data = med\_train)  
  
Pearson residuals:  
 Min 1Q Median 3Q Max   
-5.4401 -1.1618 -0.4769 0.5699 24.3630   
  
Count model coefficients (poisson with log link):  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) 1.298380 0.029608 43.853 < 2e-16 \*\*\*  
hospital 0.160449 0.006780 23.664 < 2e-16 \*\*\*  
healthpoor 0.302326 0.020022 15.099 < 2e-16 \*\*\*  
healthexcellent -0.281826 0.035775 -7.878 3.33e-15 \*\*\*  
chronic 0.097090 0.005420 17.913 < 2e-16 \*\*\*  
sexfemale 0.056219 0.014934 3.765 0.000167 \*\*\*  
school 0.023367 0.002139 10.924 < 2e-16 \*\*\*  
insuranceyes 0.093169 0.019794 4.707 2.52e-06 \*\*\*  
  
Zero-inflation model coefficients (binomial with logit link):  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) 0.25623 0.16807 1.525 0.127381   
hospital -0.29538 0.10252 -2.881 0.003962 \*\*   
healthpoor -0.09642 0.19067 -0.506 0.613093   
healthexcellent 0.32851 0.16898 1.944 0.051891 .   
chronic -0.49292 0.05183 -9.510 < 2e-16 \*\*\*  
sexfemale -0.36437 0.10295 -3.539 0.000401 \*\*\*  
school -0.06165 0.01410 -4.373 1.22e-05 \*\*\*  
insuranceyes -0.66566 0.11996 -5.549 2.87e-08 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1   
  
Number of iterations in BFGS optimization: 20   
Log-likelihood: -1.222e+04 on 16 Df

## Rootogram for ZIP model

plot(rootogram(mod\_3, plot = FALSE), xlim = c(0, 90),   
 main = "Rootogram for mod\_3: ZIP")



## Store mod\_3 Predictions

We have no augment or other broom functions available for zero-inflated models, so …

mod\_3\_aug <- med\_train |>  
 mutate(".fitted" = predict(mod\_3, type = "response"),  
 ".resid" = resid(mod\_3, type = "response"))  
  
mod\_3\_aug |> select(subject, visits, .fitted, .resid) |>  
 head(3)

# A tibble: 3 × 4  
 subject visits .fitted .resid  
 <chr> <dbl> <dbl> <dbl>  
1 355 19 5.31 13.7   
2 2661 3 4.21 -1.21  
3 2895 0 4.59 -4.59

## Training: mod\_3 Fit

mod\_3\_aug now has actual (visits) and predicted counts (in .fitted) from mod\_3, just as we set up for the previous two models.

mod\_3\_summary <-   
 mets(mod\_3\_aug, truth = visits, estimate = .fitted) |>  
 mutate(model = "mod\_3") |> relocate(model)  
mod\_3\_summary |> gt() |> fmt\_number(decimals = 3)

| model | .metric | .estimator | .estimate |
| --- | --- | --- | --- |
| mod\_3 | rsq | standard | 0.108 |
| mod\_3 | rmse | standard | 6.560 |
| mod\_3 | mae | standard | 4.164 |

## Training: Through mod\_3

bind\_rows(mod\_1\_summary, mod\_2\_summary, mod\_3\_summary) |>   
 pivot\_wider(names\_from = model,   
 values\_from = .estimate) |>   
 gt() |> fmt\_number(decimals = 3) |>   
 tab\_options(table.font.size = 20)

| .metric | .estimator | mod\_1 | mod\_2 | mod\_3 |
| --- | --- | --- | --- | --- |
| rsq | standard | 0.100 | 0.078 | 0.108 |
| rmse | standard | 6.594 | 6.941 | 6.560 |
| mae | standard | 4.189 | 4.252 | 4.164 |

Remember we want a larger and smaller values of RMSE and MAE.

## Comparing models with Vuong

Vuong’s test compares predicted probabilities (for each count) in two non-nested models. How about Poisson vs. ZIP?

vuong(mod\_1, mod\_3)

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 -14.59671 model2 > model1 < 2.22e-16  
AIC-corrected -14.51271 model2 > model1 < 2.22e-16  
BIC-corrected -14.25638 model2 > model1 < 2.22e-16

The large negative z-statistic indicates mod\_3 (ZIP) fits better than mod\_1 (Poisson) in our training sample.

# mod\_4: Zero-Inflated Negative Binomial (ZINB) Model

## Zero-Inflated Negative Binomial (ZINB) model

As in the ZIP, we assume there are two processes involved:

* a logistic regression model is used to predict excess zeros
* while a negative binomial model is used to predict the counts

We’ll use the pscl package again and the zeroinfl function.

mod\_4 <- zeroinfl(visits ~ hospital + health + chronic +  
 sex + school + insurance,  
 dist = "negbin", data = med\_train)

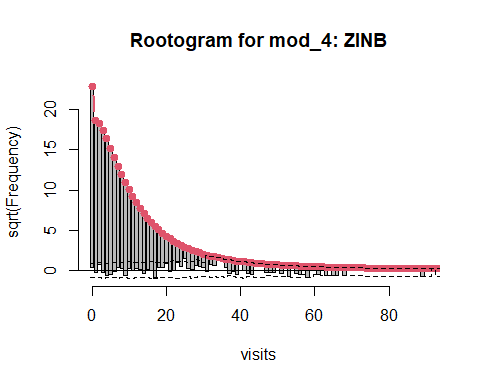
## mod\_4 summary

summary(mod\_4)

Call:  
zeroinfl(formula = visits ~ hospital + health + chronic +   
 sex + school + insurance, data = med\_train, dist = "negbin")  
  
Pearson residuals:  
 Min 1Q Median 3Q Max   
-1.1943 -0.7072 -0.2773 0.3347 17.2775   
  
Count model coefficients (negbin with log link):  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) 1.059477 0.070188 15.095 < 2e-16 \*\*\*  
hospital 0.207087 0.023460 8.827 < 2e-16 \*\*\*  
healthpoor 0.333853 0.052720 6.333 2.41e-10 \*\*\*  
healthexcellent -0.288168 0.073627 -3.914 9.08e-05 \*\*\*  
chronic 0.126107 0.013737 9.180 < 2e-16 \*\*\*  
sexfemale 0.069573 0.035975 1.934 0.05312 .   
school 0.025015 0.004983 5.020 5.18e-07 \*\*\*  
insuranceyes 0.151514 0.048328 3.135 0.00172 \*\*   
Log(theta) 0.389048 0.040783 9.539 < 2e-16 \*\*\*  
  
Zero-inflation model coefficients (binomial with logit link):  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) 0.45740 0.31049 1.473 0.14071   
hospital -0.84886 0.57525 -1.476 0.14004   
healthpoor -0.33730 0.70243 -0.480 0.63109   
healthexcellent 0.27653 0.33753 0.819 0.41262   
chronic -1.16806 0.20536 -5.688 1.29e-08 \*\*\*  
sexfemale -0.56382 0.24204 -2.329 0.01983 \*   
school -0.09233 0.03157 -2.925 0.00345 \*\*   
insuranceyes -1.00819 0.27195 -3.707 0.00021 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1   
  
Theta = 1.4756   
Number of iterations in BFGS optimization: 28   
Log-likelihood: -9102 on 17 Df

## Rootogram for ZINB model

plot(rootogram(mod\_4, plot = FALSE), xlim = c(0, 90),   
 main = "Rootogram for mod\_4: ZINB")



## Store mod\_4 Predictions

Again, there is no augment or other broom functions available for zero-inflated models, so …

mod\_4\_aug <- med\_train |>  
 mutate(".fitted" = predict(mod\_4, type = "response"),  
 ".resid" = resid(mod\_4, type = "response"))  
  
mod\_4\_aug |> select(subject, visits, .fitted, .resid) |>  
 head(3)

# A tibble: 3 × 4  
 subject visits .fitted .resid  
 <chr> <dbl> <dbl> <dbl>  
1 355 19 5.29 13.7   
2 2661 3 4.47 -1.47  
3 2895 0 4.57 -4.57

## Training Sample mod\_4 Fit

mod\_4\_aug now has actual (visits) and predicted counts (in .fitted) from mod\_4.

mod\_4\_summary <-   
 mets(mod\_4\_aug, truth = visits, estimate = .fitted) |>  
 mutate(model = "mod\_4") |> relocate(model)  
mod\_4\_summary |> gt() |> fmt\_number(decimals = 3)

| model | .metric | .estimator | .estimate |
| --- | --- | --- | --- |
| mod\_4 | rsq | standard | 0.094 |
| mod\_4 | rmse | standard | 6.709 |
| mod\_4 | mae | standard | 4.191 |

## Training Sample through mod\_4

bind\_rows(mod\_1\_summary, mod\_2\_summary,   
 mod\_3\_summary, mod\_4\_summary) |>   
 pivot\_wider(names\_from = model,   
 values\_from = .estimate) |>   
 gt() |> fmt\_number(decimals = 3) |>   
 tab\_options(table.font.size = 20)

| .metric | .estimator | mod\_1 | mod\_2 | mod\_3 | mod\_4 |
| --- | --- | --- | --- | --- | --- |
| rsq | standard | 0.100 | 0.078 | 0.108 | 0.094 |
| rmse | standard | 6.594 | 6.941 | 6.560 | 6.709 |
| mae | standard | 4.189 | 4.252 | 4.164 | 4.191 |

What do you think?

## Comparing models with Vuong

How about Negative Binomial vs. ZINB?

vuong(mod\_4, mod\_2)

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 4.808304 model1 > model2 7.6108e-07  
AIC-corrected 4.082004 model1 > model2 2.2324e-05  
BIC-corrected 1.865741 model1 > model2 0.031039

The large positive z-statistics indicate mod\_4 (ZINB) fits better than mod\_2 (Negative Binomial) in our training sample.

# Validation in the Test Sample for our Four Models?

## Validation: Test Sample Predictions

Predict the visit counts for each subject in our test sample.

test\_1 <- predict(mod\_1, newdata = med\_test,  
 type.predict = "response")  
test\_2 <- predict(mod\_2, newdata = med\_test,  
 type.predict = "response")  
test\_3 <- predict(mod\_3, newdata = med\_test,  
 type.predict = "response")  
test\_4 <- predict(mod\_4, newdata = med\_test,  
 type.predict = "response")

## Create a Tibble with Predictions

Combine the various predictions into a tibble with the original data.

test\_res <- bind\_cols(med\_test,   
 pre\_m1 = test\_1, pre\_m2 = test\_2,   
 pre\_m3 = test\_3, pre\_m4 = test\_4)  
  
names(test\_res)

[1] "subject" "visits" "hospital" "health"   
 [5] "chronic" "sex" "school" "insurance"  
 [9] "pre\_m1" "pre\_m2" "pre\_m3" "pre\_m4"

## Summarize fit in test sample for each model

m1\_sum <- mets(test\_res, truth = visits, estimate = pre\_m1) |>  
 mutate(model = "mod\_1")   
m2\_sum <- mets(test\_res, truth = visits, estimate = pre\_m2) |>  
 mutate(model = "mod\_2")   
m3\_sum <- mets(test\_res, truth = visits, estimate = pre\_m3) |>  
 mutate(model = "mod\_3")  
m4\_sum <- mets(test\_res, truth = visits, estimate = pre\_m4) |>  
 mutate(model = "mod\_4")  
  
test\_sum <- bind\_rows(m1\_sum, m2\_sum, m3\_sum, m4\_sum)

## Validation Results: Four Models

test\_sum <- bind\_rows(m1\_sum, m2\_sum, m3\_sum, m4\_sum) |>  
 pivot\_wider(names\_from = model,   
 values\_from = .estimate)  
  
test\_sum |>  
 select(-.estimator) |>   
 gt() |> fmt\_number(decimals = 3) |>   
 tab\_options(table.font.size = 20)

| .metric | mod\_1 | mod\_2 | mod\_3 | mod\_4 |
| --- | --- | --- | --- | --- |
| rsq | 0.103 | 0.108 | 0.099 | 0.097 |
| rmse | 7.212 | 7.205 | 5.907 | 5.967 |
| mae | 4.455 | 4.450 | 3.994 | 4.009 |

* Which model looks best? Is it an obvious choice?

# Hurdle Models

## The Hurdle Model

The hurdle model is a two-part model that specifies one process for zero counts and another process for positive counts. The idea is that positive counts occur once a threshold is crossed, or put another way, a hurdle is cleared. If the hurdle is not cleared, then we have a count of 0.

* The first part of the model is typically a **binary logistic regression** model. This models whether an observation takes a positive count or not.
* The second part of the model is usually a truncated Poisson or Negative Binomial model. Truncated means we’re only fitting positive counts, and not zeros.

# mod\_5: Poisson-Logistic Hurdle Model

## Fitting a Hurdle Model / Poisson-Logistic

In fitting a hurdle model to our medicare training data, the interpretation would be that one process governs whether a patient visits a doctor or not, and another process governs how many visits are made.

## The mod\_5 model

mod\_5 <- hurdle(visits ~ hospital + health + chronic +  
 sex + school + insurance,  
 dist = "poisson", zero.dist = "binomial",   
 data = med\_train)  
mod\_5

Call:  
hurdle(formula = visits ~ hospital + health + chronic +   
 sex + school + insurance, data = med\_train, dist = "poisson",   
 zero.dist = "binomial")  
  
Count model coefficients (truncated poisson with log link):  
 (Intercept) hospital healthpoor   
 1.29892 0.16041 0.30243   
healthexcellent chronic sexfemale   
 -0.28116 0.09697 0.05611   
 school insuranceyes   
 0.02332 0.09351   
  
Zero hurdle model coefficients (binomial with logit link):  
 (Intercept) hospital healthpoor   
 -0.2998 0.3044 0.1114   
healthexcellent chronic sexfemale   
 -0.3705 0.4970 0.3652   
 school insuranceyes   
 0.0637 0.6625

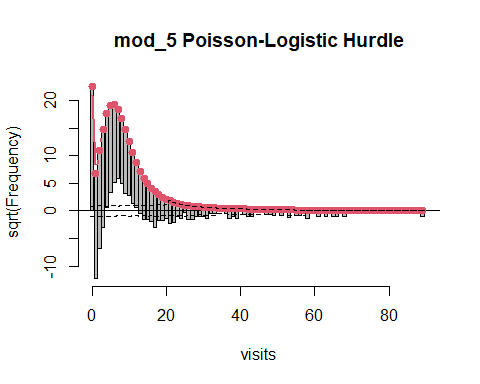
## mod\_5 summary

summary(mod\_5)

Call:  
hurdle(formula = visits ~ hospital + health + chronic +   
 sex + school + insurance, data = med\_train, dist = "poisson",   
 zero.dist = "binomial")  
  
Pearson residuals:  
 Min 1Q Median 3Q Max   
-5.4472 -1.1621 -0.4769 0.5698 24.3403   
  
Count model coefficients (truncated poisson with log link):  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) 1.298920 0.029612 43.865 < 2e-16 \*\*\*  
hospital 0.160410 0.006781 23.656 < 2e-16 \*\*\*  
healthpoor 0.302432 0.020026 15.102 < 2e-16 \*\*\*  
healthexcellent -0.281162 0.035755 -7.864 3.73e-15 \*\*\*  
chronic 0.096971 0.005417 17.901 < 2e-16 \*\*\*  
sexfemale 0.056109 0.014933 3.757 0.000172 \*\*\*  
school 0.023321 0.002138 10.907 < 2e-16 \*\*\*  
insuranceyes 0.093508 0.019789 4.725 2.30e-06 \*\*\*  
Zero hurdle model coefficients (binomial with logit link):  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) -0.29978 0.16500 -1.817 0.069240 .   
hospital 0.30445 0.10235 2.975 0.002934 \*\*   
healthpoor 0.11138 0.18993 0.586 0.557605   
healthexcellent -0.37054 0.16200 -2.287 0.022179 \*   
chronic 0.49699 0.05112 9.722 < 2e-16 \*\*\*  
sexfemale 0.36524 0.10116 3.610 0.000306 \*\*\*  
school 0.06370 0.01382 4.611 4.01e-06 \*\*\*  
insuranceyes 0.66251 0.11789 5.620 1.91e-08 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1   
  
Number of iterations in BFGS optimization: 14   
Log-likelihood: -1.222e+04 on 16 Df

## Rootogram for Poisson-Logistic Hurdle model

plot(rootogram(mod\_5, plot = FALSE), xlim = c(0, 90),   
 main = "mod\_5 Poisson-Logistic Hurdle")



## Store mod\_5 Predictions

No augment or other broom functions for hurdle models, so …

mod\_5\_aug <- med\_train |>  
 mutate(".fitted" = predict(mod\_5, type = "response"),  
 ".resid" = resid(mod\_5, type = "response"))  
  
mod\_5\_aug |> select(subject, visits, .fitted, .resid) |>  
 head(3)

# A tibble: 3 × 4  
 subject visits .fitted .resid  
 <chr> <dbl> <dbl> <dbl>  
1 355 19 5.32 13.7   
2 2661 3 4.20 -1.20  
3 2895 0 4.59 -4.59

## Training Sample mod\_5 Fit

mod\_5\_summary <-   
 mets(mod\_5\_aug, truth = visits, estimate = .fitted) |>  
 mutate(model = "mod\_5") |> relocate(model)  
mod\_5\_summary |> gt() |> fmt\_number(decimals = 3)

| model | .metric | .estimator | .estimate |
| --- | --- | --- | --- |
| mod\_5 | rsq | standard | 0.108 |
| mod\_5 | rmse | standard | 6.560 |
| mod\_5 | mae | standard | 4.164 |

## Training Sample through mod\_5

bind\_rows(mod\_1\_summary, mod\_2\_summary, mod\_3\_summary,   
 mod\_4\_summary, mod\_5\_summary) |>   
 pivot\_wider(names\_from = model,   
 values\_from = .estimate) |>   
 gt() |> fmt\_number(decimals = 3) |>   
 tab\_options(table.font.size = 20)

| .metric | .estimator | mod\_1 | mod\_2 | mod\_3 | mod\_4 | mod\_5 |
| --- | --- | --- | --- | --- | --- | --- |
| rsq | standard | 0.100 | 0.078 | 0.108 | 0.094 | 0.108 |
| rmse | standard | 6.594 | 6.941 | 6.560 | 6.709 | 6.560 |
| mae | standard | 4.189 | 4.252 | 4.164 | 4.191 | 4.164 |

What do you think?

## Are ZIP and Poisson-Logistic Hurdle the Same?

temp\_check <- tibble(  
 subject = mod\_3\_aug$subject,  
 visits = mod\_3\_aug$visits,  
 pred\_zip = mod\_3\_aug$.fitted,  
 pred\_hur = mod\_5\_aug$.fitted,  
 diff = pred\_hur - pred\_zip)  
  
favstats(~ diff, data = temp\_check)

min Q1 median Q3  
 -0.0241247 -0.0004086227 0.0003036167 0.0009285733  
 max mean sd n missing  
 0.03270967 0.0003302558 0.003050025 3304 0

## Vuong test: mod\_3 vs. mod\_5

vuong(mod\_3, mod\_5)

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 1.913241 model1 > model2 0.027859  
AIC-corrected 1.913241 model1 > model2 0.027859  
BIC-corrected 1.913241 model1 > model2 0.027859

There’s some evidence mod\_3 (ZIP) fits a bit better than mod\_5 (Hurdle) in our training sample, though the p value (barely) exceeds 0.05.

# mod\_6: Negative Binomial-Logistic Hurdle Model

## Hurdle Model / NB-Logistic

mod\_6 <- hurdle(visits ~ hospital + health + chronic +  
 sex + school + insurance,  
 dist = "negbin", zero.dist = "binomial",   
 data = med\_train)  
mod\_6

Call:  
hurdle(formula = visits ~ hospital + health + chronic +   
 sex + school + insurance, data = med\_train, dist = "negbin",   
 zero.dist = "binomial")  
  
Count model coefficients (truncated negbin with log link):  
 (Intercept) hospital healthpoor   
 1.07299 0.21828 0.36531   
healthexcellent chronic sexfemale   
 -0.30074 0.12372 0.05795   
 school insuranceyes   
 0.02429 0.13122   
Theta = 1.406   
  
Zero hurdle model coefficients (binomial with logit link):  
 (Intercept) hospital healthpoor   
 -0.2998 0.3044 0.1114   
healthexcellent chronic sexfemale   
 -0.3705 0.4970 0.3652   
 school insuranceyes   
 0.0637 0.6625

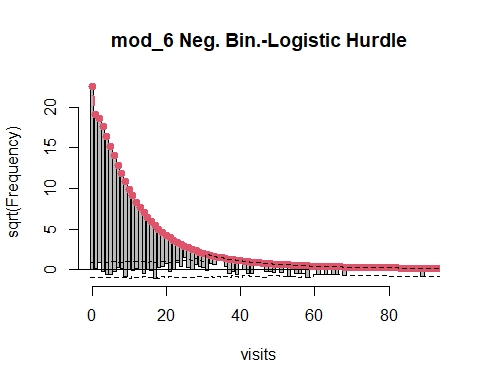
## mod\_6 Summary

summary(mod\_6)

Call:  
hurdle(formula = visits ~ hospital + health + chronic +   
 sex + school + insurance, data = med\_train, dist = "negbin",   
 zero.dist = "binomial")  
  
Pearson residuals:  
 Min 1Q Median 3Q Max   
-1.1756 -0.7080 -0.2771 0.3371 17.4791   
  
Count model coefficients (truncated negbin with log link):  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) 1.072988 0.072095 14.883 < 2e-16 \*\*\*  
hospital 0.218281 0.024288 8.987 < 2e-16 \*\*\*  
healthpoor 0.365310 0.054624 6.688 2.27e-11 \*\*\*  
healthexcellent -0.300745 0.076626 -3.925 8.68e-05 \*\*\*  
chronic 0.123717 0.014253 8.680 < 2e-16 \*\*\*  
sexfemale 0.057948 0.037229 1.557 0.11958   
school 0.024290 0.005156 4.711 2.46e-06 \*\*\*  
insuranceyes 0.131216 0.049295 2.662 0.00777 \*\*   
Log(theta) 0.340775 0.049007 6.954 3.56e-12 \*\*\*  
Zero hurdle model coefficients (binomial with logit link):  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) -0.29978 0.16500 -1.817 0.069240 .   
hospital 0.30445 0.10235 2.975 0.002934 \*\*   
healthpoor 0.11138 0.18993 0.586 0.557605   
healthexcellent -0.37054 0.16200 -2.287 0.022179 \*   
chronic 0.49699 0.05112 9.722 < 2e-16 \*\*\*  
sexfemale 0.36524 0.10116 3.610 0.000306 \*\*\*  
school 0.06370 0.01382 4.611 4.01e-06 \*\*\*  
insuranceyes 0.66251 0.11789 5.620 1.91e-08 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1   
  
Theta: count = 1.406  
Number of iterations in BFGS optimization: 16   
Log-likelihood: -9102 on 17 Df

## Rootogram for NB-Logistic Hurdle model

plot(rootogram(mod\_6, plot = FALSE), xlim = c(0, 90),   
 main = "mod\_6 Neg. Bin.-Logistic Hurdle")



## Store mod\_6 Predictions

mod\_6\_aug <- med\_train |>  
 mutate(".fitted" = predict(mod\_6, type = "response"),  
 ".resid" = resid(mod\_6, type = "response"))  
  
mod\_6\_aug |> select(subject, visits, .fitted, .resid) |>  
 head(3)

# A tibble: 3 × 4  
 subject visits .fitted .resid  
 <chr> <dbl> <dbl> <dbl>  
1 355 19 5.35 13.6   
2 2661 3 4.16 -1.16  
3 2895 0 4.49 -4.49

## Training Sample mod\_6 Fit

mod\_6\_summary <-   
 mets(mod\_6\_aug, truth = visits, estimate = .fitted) |>  
 mutate(model = "mod\_6") |> relocate(model)  
mod\_6\_summary |>   
 gt() |> fmt\_number(decimals = 3)

| model | .metric | .estimator | .estimate |
| --- | --- | --- | --- |
| mod\_6 | rsq | standard | 0.089 |
| mod\_6 | rmse | standard | 6.772 |
| mod\_6 | mae | standard | 4.209 |

## Training Sample through mod\_6

bind\_rows(mod\_1\_summary, mod\_2\_summary, mod\_3\_summary,   
 mod\_4\_summary, mod\_5\_summary, mod\_6\_summary) |>   
 pivot\_wider(names\_from = model, values\_from = .estimate) |>   
 select(-.estimator) |>   
 gt() |> fmt\_number(decimals = 3) |>   
 tab\_options(table.font.size = 20)

| .metric | mod\_1 | mod\_2 | mod\_3 | mod\_4 | mod\_5 | mod\_6 |
| --- | --- | --- | --- | --- | --- | --- |
| rsq | 0.100 | 0.078 | 0.108 | 0.094 | 0.108 | 0.089 |
| rmse | 6.594 | 6.941 | 6.560 | 6.709 | 6.560 | 6.772 |
| mae | 4.189 | 4.252 | 4.164 | 4.191 | 4.164 | 4.209 |

## Vuong test: mod\_4 vs. mod\_6

vuong(mod\_4, mod\_6)

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 0.02994908 model1 > model2 0.48805  
AIC-corrected 0.02994908 model1 > model2 0.48805  
BIC-corrected 0.02994908 model1 > model2 0.48805

There’s some evidence mod\_4 (ZINB) fits better than mod\_6 (NB Hurdle) in our training sample, but not much, based on the large *p* value.

# Validation including Hurdle Models

## Validation: Test Sample Predictions

Predict the visit counts for each subject in our test sample.

test\_5 <- predict(mod\_5, newdata = med\_test,  
 type.predict = "response")  
test\_6 <- predict(mod\_6, newdata = med\_test,  
 type.predict = "response")

## Create a Tibble with Predictions

Combine the various predictions into a tibble with the original data.

test\_res6 <- bind\_cols(med\_test,   
 pre\_m1 = test\_1, pre\_m2 = test\_2,   
 pre\_m3 = test\_3, pre\_m4 = test\_4,   
 pre\_m5 = test\_5, pre\_m6 = test\_6)  
  
names(test\_res6)

[1] "subject" "visits" "hospital" "health"   
 [5] "chronic" "sex" "school" "insurance"  
 [9] "pre\_m1" "pre\_m2" "pre\_m3" "pre\_m4"   
[13] "pre\_m5" "pre\_m6"

## Summarize fit in test sample for each model

m1\_sum <- mets(test\_res6, truth = visits, estimate = pre\_m1) |>  
 mutate(model = "mod\_1")   
m2\_sum <- mets(test\_res6, truth = visits, estimate = pre\_m2) |>  
 mutate(model = "mod\_2")   
m3\_sum <- mets(test\_res6, truth = visits, estimate = pre\_m3) |>  
 mutate(model = "mod\_3")  
m4\_sum <- mets(test\_res6, truth = visits, estimate = pre\_m4) |>  
 mutate(model = "mod\_4")  
m5\_sum <- mets(test\_res6, truth = visits, estimate = pre\_m5) |>  
 mutate(model = "mod\_5")  
m6\_sum <- mets(test\_res6, truth = visits, estimate = pre\_m6) |>  
 mutate(model = "mod\_6")  
  
test\_sum6 <- bind\_rows(m1\_sum, m2\_sum, m3\_sum, m4\_sum,  
 m5\_sum, m6\_sum)

## Validation Results in Test Sample

test\_sum6 <- bind\_rows(m1\_sum, m2\_sum, m3\_sum, m4\_sum,  
 m5\_sum, m6\_sum) |>  
 pivot\_wider(names\_from = model,   
 values\_from = .estimate)  
  
test\_sum6 |>  
 select(-.estimator) |>   
 gt() |> fmt\_number(decimals = 4) |>   
 tab\_options(table.font.size = 20)

| .metric | mod\_1 | mod\_2 | mod\_3 | mod\_4 | mod\_5 | mod\_6 |
| --- | --- | --- | --- | --- | --- | --- |
| rsq | 0.1032 | 0.1082 | 0.0993 | 0.0970 | 0.0992 | 0.0937 |
| rmse | 7.2122 | 7.2051 | 5.9069 | 5.9674 | 5.9072 | 6.0009 |
| mae | 4.4550 | 4.4496 | 3.9943 | 4.0095 | 3.9946 | 4.0265 |

* Now which model would you choose?

## After Spring Break

* Project A due at noon on **Monday 2023-03-18**
* Will get started on Project B once Project A is submitted
* Regression on Multi-Categorical Outcomes will be our first new topic