MKTG776 HW2

Jordan Farrer

2017-01-31

Contents

1 Question 1

1.1 PMF of Negative Binomial Distribution for Period t

Given the Poisson distribution with rate parameter λt,

$$P(X(t) = t|\lambda) = \frac{(\lambda t)^x e^{-\lambda t}}{x!} \tag{1}$$

and Gamma distribution with shape parameter r and scale parameter α,

$$g(\lambda) = \frac{\alpha^r \lambda^{r-1} e^{-\alpha\lambda}}{\Gamma(r)} \tag{2}$$

where $\Gamma(r)$ is the gamma function. Now define

$$P(X(t) = x) = \int_0^\infty P(X(t) = t|\lambda) g(\lambda) d\lambda \tag{3}$$

$$= \frac{(\lambda t)^x e^{-\lambda t}}{x!} \frac{\alpha^r \lambda^{r-1} e^{-\alpha\lambda}}{\Gamma(r)} d\lambda \tag{4}$$

$$= \frac{\alpha^r t^x}{x!\Gamma(r)} \int_0^\infty \lambda^{r+x-1} e^{-\lambda(\alpha+t)} d\lambda \tag{5}$$

$$\tag{6}$$

We note that this distribution looks similar to the gamma distribution with shape parameter $r + x$ and scale parameter $\alpha + t$,

$$g(\lambda|r + x, \alpha + t) = \frac{(\alpha + t)^{r+x} \lambda^{r+x-1} e^{-\lambda(\alpha+t)}}{\Gamma(r + x)} \tag{7}$$

We multiply the right-hand side of the integral by $\frac{(\alpha+t)^{r+x}}{\Gamma(r+x)}$ and the left-hand side of the integral by the inverse to get

$$= \frac{\alpha^r t^x \Gamma(r+x)}{x!\Gamma(r)(\alpha+t)^{r+x}} \int_0^\infty \frac{(\alpha+t)^{r+x}\lambda^{r+x-1}e^{-\lambda(\alpha+t)}}{\Gamma(r+x)} d\lambda \tag{8}$$

$$\tag{9}$$

We see that the integral over all values of λ means the value of the integral is 1 by the definition of probability distributions. Thus we are left with

$$= \frac{\alpha^r t^x \Gamma(r+x)}{x!\Gamma(r)(\alpha+t)^{r+x}} \tag{10}$$

$$= \frac{\Gamma(r+x)}{x!\Gamma(r)} \frac{\alpha^r t^x}{(\alpha+t)^{r+x}} \tag{11}$$

$$= \frac{\Gamma(r+x)}{\Gamma(r)x!} \frac{\alpha^r t^x}{(\alpha+t)^r(\alpha+t)^x} \tag{12}$$

$$= \left(\frac{\Gamma(r+x)}{\Gamma(r)x!}\right)\left(\frac{\alpha}{\alpha+t}\right)^r\left(\frac{t}{\alpha+t}\right)^x \tag{13}$$

1.2 Forward Recursion Formula

We start by finding $P(X(t)=0)$ using 1.3,

$$P(X(t)=0) = \left(\frac{\Gamma(r+0)}{\Gamma(r)0!}\right)\left(\frac{\alpha}{\alpha+t}\right)^r\left(\frac{t}{\alpha+t}\right)^0 \tag{14}$$

$$= \left(\frac{\Gamma(r)}{\Gamma(r)}\right)\left(\frac{\alpha}{\alpha+t}\right)^r \tag{15}$$

$$= \left(\frac{\alpha}{\alpha+t}\right)^r \tag{16}$$

$$\tag{17}$$

Then we need to simplify

$$\frac{P(X(t)=x)}{P(X(t)=x-1)} = \frac{\left(\frac{\Gamma(r+x)}{\Gamma(r)x!}\right)\left(\frac{\alpha}{\alpha+t}\right)^r\left(\frac{t}{\alpha+t}\right)^x}{\left(\frac{\Gamma(r+x-1)}{\Gamma(r)(x-1)!}\right)\left(\frac{\alpha}{\alpha+t}\right)^r\left(\frac{t}{\alpha+t}\right)^{x-1}} \tag{18}$$

$$= \left(\frac{r+x-1}{x}\right)\left(\frac{t}{\alpha+t}\right) \tag{19}$$

$$= \frac{t(r+x-1)}{x(\alpha+t)} \tag{20}$$

$$\tag{21}$$

Thus the forward-recursion formula is

$$P(X(t)=x)=\begin{cases}\left(\frac{\alpha}{\alpha+t}\right)^r & x=0\\ \frac{t(r+x-1)}{x(\alpha+t)}P(X(t)=x-1) & x=1,2,3,...\end{cases} \tag{22}$$

1.3 Mean of Negative Binomial Distribution for period t

We will use the following relations:

$$\Gamma(x)=(x-1)! \tag{23}$$

$$C_k^n=\binom{n}{k}=\frac{n!}{k!(n-k)!} \tag{24}$$

By the definition of a probability distribution's expectation,

$$E[X(t)]=\sum_{x=0}^{\infty}x\left(\frac{\Gamma(r+x)}{\Gamma(r)x!}\right)\left(\frac{\alpha}{\alpha+t}\right)^r\left(\frac{t}{\alpha+t}\right)^x \tag{25}$$

$$=\sum_{x=0}^{\infty}x\frac{(r+x-1)!}{(r-1)!x!}\left(\frac{\alpha}{\alpha+t}\right)^r\left(\frac{t}{\alpha+t}\right)^x \tag{26}$$

$$=\sum_{x=1}^{\infty}\frac{(r+x-1)!}{(r-1)!(x-1)!}\left(\frac{\alpha}{\alpha+t}\right)^r\left(\frac{t}{\alpha+t}\right)^x \tag{27}$$

We changed the range of the summation in (27) above because x is defined for $x >= 0$ and there cannot be negative factorials. Now, we can multiple by $\frac{r}{r}$,

$$=\sum_{x=1}^{\infty}\frac{r(r+x-1)!}{r!(x-1)!}\left(\frac{\alpha}{\alpha+t}\right)^r\left(\frac{t}{\alpha+t}\right)^x \tag{28}$$

$$=\sum_{x=1}^{\infty}r\binom{r+x-1}{x-1}\left(\frac{\alpha}{\alpha+t}\right)^r\left(\frac{t}{\alpha+t}\right)^x \tag{29}$$

Now, we separate the last two expressions in parentheses and move them outside the summation as they are not expressions of x,

$$=\sum_{x=1}^{\infty}r\binom{r+x-1}{x-1}\left(\frac{\alpha}{\alpha+t}\right)^{r+1}\left(\frac{\alpha}{\alpha+t}\right)^{-1}\left(\frac{t}{\alpha+t}\right)^{x-1}\left(\frac{t}{\alpha+t}\right) \tag{30}$$

$$=r\left(\frac{\alpha}{\alpha+t}\right)^{-1}\left(\frac{t}{\alpha+t}\right)\sum_{x=1}^{\infty}\binom{r+x-1}{x-1}\left(\frac{\alpha}{\alpha+t}\right)^{r+1}\left(\frac{t}{\alpha+t}\right)^{x-1} \tag{31}$$

$$=r\left(\frac{\alpha+t}{\alpha+t}\right)\left(\frac{t}{\alpha+t}\right)\sum_{x=1}^{\infty}\binom{r+x-1}{x-1}\left(\frac{\alpha}{\alpha+t}\right)^{r+1}\left(\frac{t}{\alpha+t}\right)^{x-1} \tag{32}$$

$$=\frac{rt}{\alpha}\sum_{x=1}^{\infty}\binom{r+x-1}{x-1}\left(\frac{\alpha}{\alpha+t}\right)^{r+1}\left(\frac{t}{\alpha+t}\right)^{x-1} \tag{33}$$

Now we can use a change of variable $x = 1 + z$, to get

$$= \frac{rt}{\alpha} \sum_{z=0}^{\infty} \binom{r+1+z-1}{z} \left(\frac{\alpha}{\alpha+t}\right)^{r+1} \left(\frac{t}{\alpha+t}\right)^{z} \tag{34}$$

We can do a change of variable again such that $y = r + 1$, to write (34) as

$$= \frac{rt}{\alpha} \sum_{z=0}^{\infty} \binom{y+z-1}{z} \left(\frac{\alpha}{\alpha+t}\right)^{y} \left(\frac{t}{\alpha+t}\right)^{z} \tag{35}$$

$$= \frac{rt}{\alpha} \sum_{z=0}^{\infty} \frac{(y+z-1)!}{z!(y-1)!} \left(\frac{\alpha}{\alpha+t}\right)^{y} \left(\frac{t}{\alpha+t}\right)^{z} \tag{36}$$

$$= \frac{rt}{\alpha} \sum_{z=0}^{\infty} \frac{\Gamma(y+z)}{\Gamma(y)z!} \left(\frac{\alpha}{\alpha+t}\right)^{y} \left(\frac{t}{\alpha+t}\right)^{z} \tag{37}$$

We see that summation in (37) is simply the negative binomial distribution where x in () is z and the shape parameter r is y. As it is the summation over all possible values of z, this becomes 1. Thus,

$$E[X(t)] = \frac{rt}{\alpha} \tag{38}$$

Resource used in the preceding derivation: http://www.math.ntu.edu.tw/~hchen/teaching/StatInference/notes/lecture16.pdf

2 Question 2

We first load the provided customer data. Below are the first 10 records:

```
pacman::p_load(tidyverse, forcats, pander)
panderOptions('round', 2)
panderOptions('keep.trailing.zeros', TRUE)
options(scipen = 10)

customer_data <-
  readxl::read_excel("khakichinos HW data.xlsx", skip = 1,
            col_names = c("id", "visits", "Empty", "income", "sex", "age","size", "NA1", "NA2")) %>%
  select(id, visits)

customer_data %>%
  head(10)
```

id	visits
1	0
2	5
3	0
4	0
5	0
6	0
7	0

id	visits
8	1
9	0
10	0

Next we define functions that implement the (zero-inflated) Poisson and (zero-inflated) NBD. Using these, we estimate the parameters of each model:

```
# For Zero-inflated Poisson, calculates P(X=x)
fn_zip <- function(x, lambda, pi) {
  p_x <- (lambda^x * exp(-lambda)) / factorial(x)
  if(x == 0) {
    return(pi + (1 - pi) * p_x)
  } else {
    return((1 - pi) * p_x)
  }
}

# For Zero-inflated Negative Binomial Distribution, calculates P(X=x)
fn_zinbd <- function(x, r, alpha, pi) {
  p_x <- (gamma(r + x) / (gamma(r) * factorial(x))) * (alpha / (alpha + 1))^r * (1 / (alpha + 1))^x
  if(x == 0) {
    return(pi + (1 - pi) * p_x)
  } else {
    return((1 - pi) * p_x)
  }
}

# Calculates the log-likelihood of given a type of model
# either poisson or nbd and whether or not it's zero-inflated
# given a vector of number of visits per individual
fn_max_ll <- function(par, model = c('poisson', 'nbd'), zero_inflated = FALSE, visits) {

  if (model == 'poisson') {
    lambda <- par[1]
    if (zero_inflated) {
      pi <- par[2]
    } else {
      pi <- 0
    }
    ll <- sum(log(sapply(visits, fn_zip, lambda, pi)))
  } else {
    r <- par[1]
    alpha <- par[2]
    if (zero_inflated) {
      pi <- par[3]
    } else {
      pi <- 0
    }
    ll <- sum(log(sapply(visits, fn_zinbd, r, alpha, pi)))
  }

  return(-ll)
}

params1 <- nlminb(c(1), fn_max_ll, lower = c(0), upper = c(Inf),
                  model = 'poisson', zero_inflated = FALSE, visits = customer_data$visits)
```

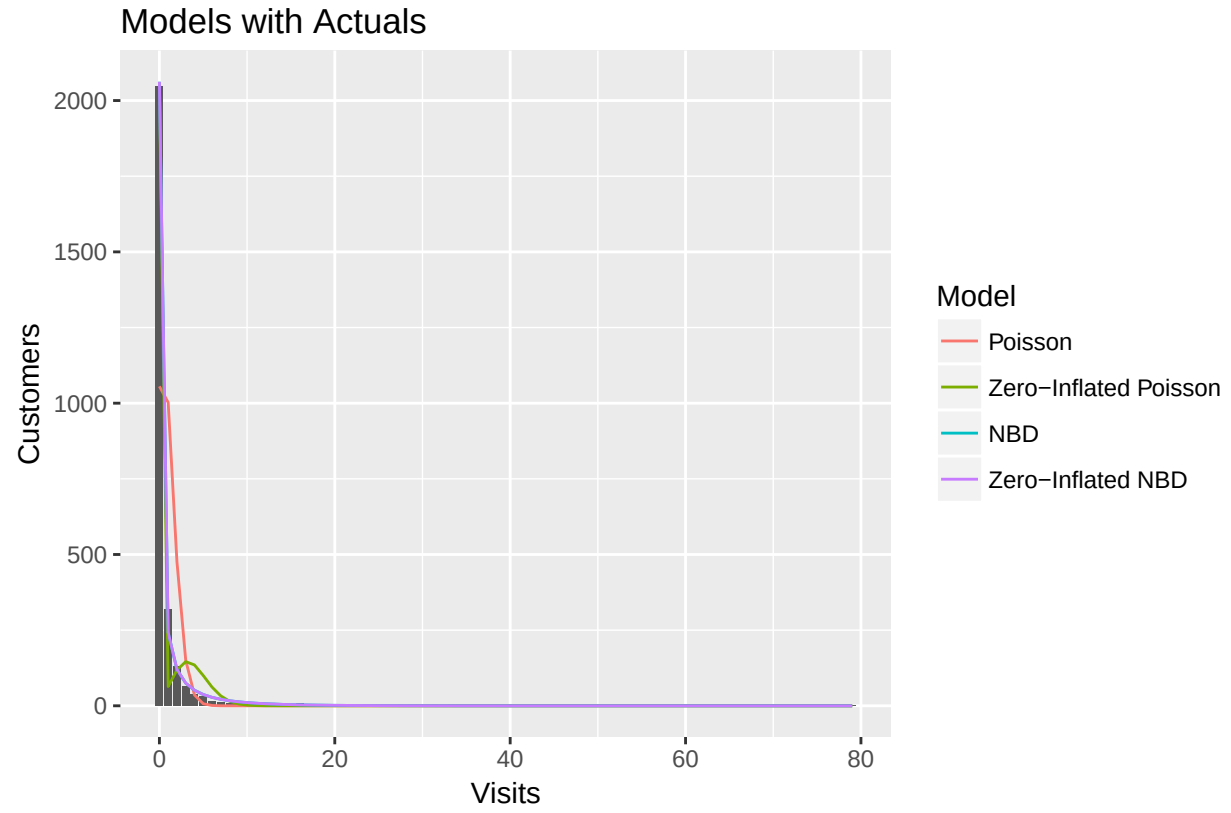
```
params2 <- nlminb(c(1, .5), fn_max_ll, lower = c(0, 0), upper = c(Inf, 1),
                  model = 'poisson', zero_inflated = TRUE, visits = customer_data$visits)

params3 <- nlminb(c(1, 1), fn_max_ll, lower = c(0, 0), upper = c(Inf, Inf),
                  model = 'nbd', zero_inflated = FALSE, visits = customer_data$visits)

params4 <- nlminb(c(1, 1, .5), fn_max_ll, lower = c(0, 0, 0), upper = c(Inf, Inf, 1),
                  model = 'nbd', zero_inflated = TRUE, visits = customer_data$visits)


data_frame(
  model = c('Poisson', "Zero-Inflated Poisson",'NBD',"Zero-Inflated NBD")
  , lamdba = c(params1$par[1], params2$par[1], NA, NA)
  , r = c(NA, NA, params3$par[1], params4$par[1])
  , alpha = c(NA, NA, params3$par[2], params4$par[2])
  , pi = c(NA, params2$par[2], NA, params4$par[3])
  ) %>%
  pander(missing = "")
```

model	lamdba	r	alpha	pi
Poisson	0.95			
Zero-Inflated Poisson	3.70			0.74
NBD		0.13	0.14	
Zero-Inflated NBD		0.13	0.14	0.00

We see that $\pi = 0$ for the Zero-Inflated NBD and thus we are left with just the NBD. We can visualize the results for all number of visits:

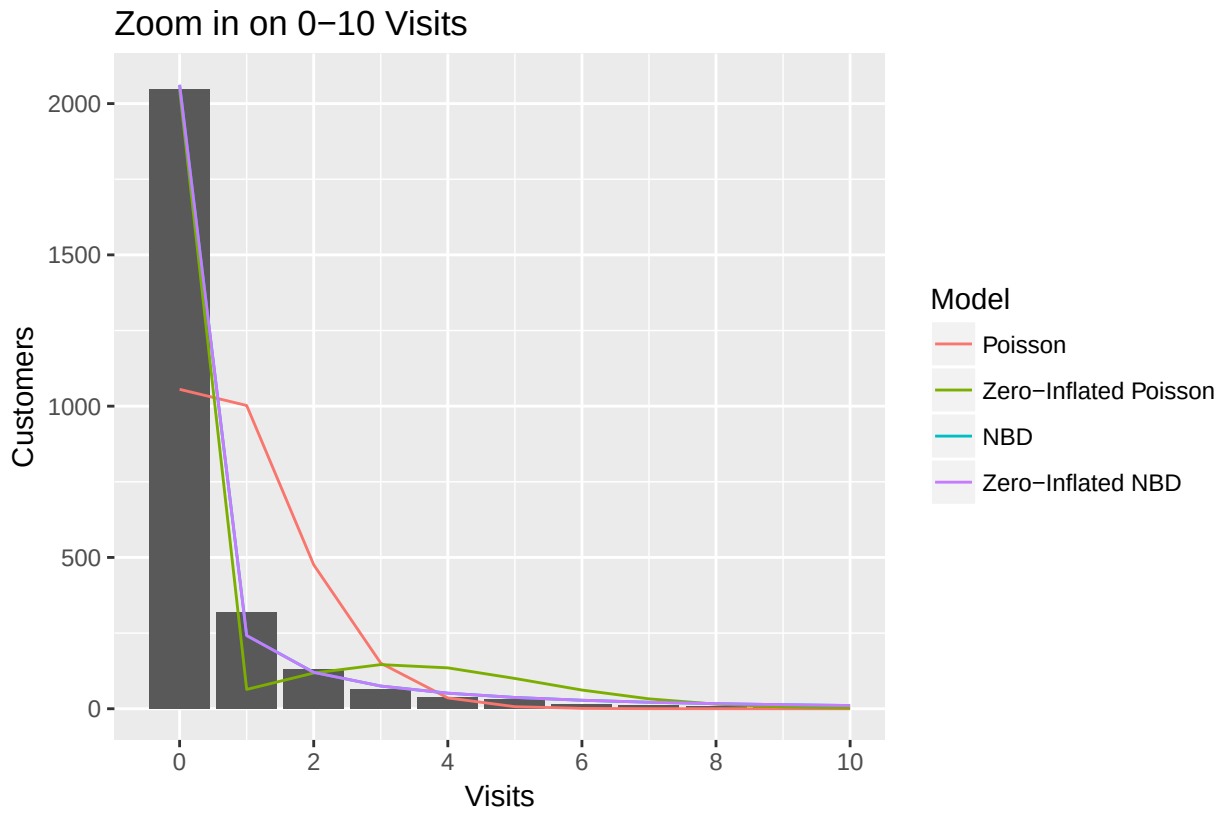
```
results <-
  customer_data %>%
    group_by(visits) %>%
    summarise(
      Actual = n()
    ) %>%
    ungroup() %>%
    mutate(
       Total = sum(Actual)
    ) %>%
    mutate(
      `Poisson` = sapply(visits, fn_zip, params1$par[1], 0) * Total
      , `Zero-Inflated Poisson` = sapply(visits, fn_zip, params2$par[1], params2$par[2]) * Total
      , `NBD` = sapply(visits, fn_zinbd, params3$par[1], params3$par[2], 0) * Total
      , `Zero-Inflated NBD` = sapply(visits, fn_zinbd, params4$par[1],
                                     params4$par[2], params4$par[3]) * Total
    ) %>%
    gather(Type, Customers, -visits, -Total) %>%
    mutate(Type = factor(Type, levels = c('Actual', 'Poisson',
                                          "Zero-Inflated Poisson",'NBD',"Zero-Inflated NBD")))

ggplot() +
  geom_bar(data = results %>% filter(Type == "Actual"),
           aes(x = visits, y = Customers), stat = 'identity') +
  geom_line(data = results %>% filter(Type != "Actual"),
            aes(x = visits, y = Customers, colour = Type)) +
   scale_x_continuous(breaks = scales::pretty_breaks()) +
  labs(colour = "Model", x = "Visits", y = "Customers", title = "Models with Actuals",
       caption = paste0("Note: The NBD and Zero-Inflated NBD are the same ",
```

```
                "model as pi = 0 in the zero-inflated model"))
```



There is a long right tail, so we focus in on just the first 0-10 visits:

```
ggplot() +
  geom_bar(data = results %>% filter(Type == "Actual"), aes(x = visits, y = Customers), stat = 'identity') +
  geom_line(data = results %>% filter(Type != "Actual"), aes(x = visits, y = Customers, colour = Type)) +
  scale_x_continuous(limits = c(NA, 10), breaks = scales::pretty_breaks()) +
  labs(colour = "Model", x = "Visits", y = "Customers", title = "Zoom in on 0-10 Visits",
       caption = "Note: The NBD and Zero-Inflated NBD are the same model as pi = 0 in the zero-inflated model")
```



Zero−Inflated NBD are the same model as pi = 0 in the zero−inflated model

Finally, we compute the χ^2 statistics and then the p-values. As evidenced by the charts above, we have no evidence that the data came from our calculated models.

```
results %>%
  spread(Type, Customers) %>%
  gather(model, expected, -visits, -Total, -Actual) %>%
  mutate(chi.squared = (Actual - expected)^2 / expected) %>%
  group_by(model) %>%
  summarise(chi.squared = sum(chi.squared)) %>%
  inner_join(
    data_frame(
      model = c('Poisson', "Zero-Inflated Poisson",'NBD',"Zero-Inflated NBD")
      , df = c(1,2,2,3)
    )
    , by = c('model')
  ) %>%
  mutate(p.value = pchisq(chi.squared,df=df,lower.tail=FALSE)) %>%
  pander()
```

model	chi.squared	df	p.value
NBD	5.720e+03	2	0
Poisson	5.118e+115	1	0
Zero-Inflated NBD	5.720e+03	3	0
Zero-Inflated Poisson	6.156e+70	2	0

3 Question 3

Based on the examination of the dataset in Bickart & Schmittlein, 1999, we transform the category **3-5** to the integer **4**. We use write a recursive function to calculate NBD and a function to zero-inflate NBD as necessary. We estimate the model parameters and report the results below:

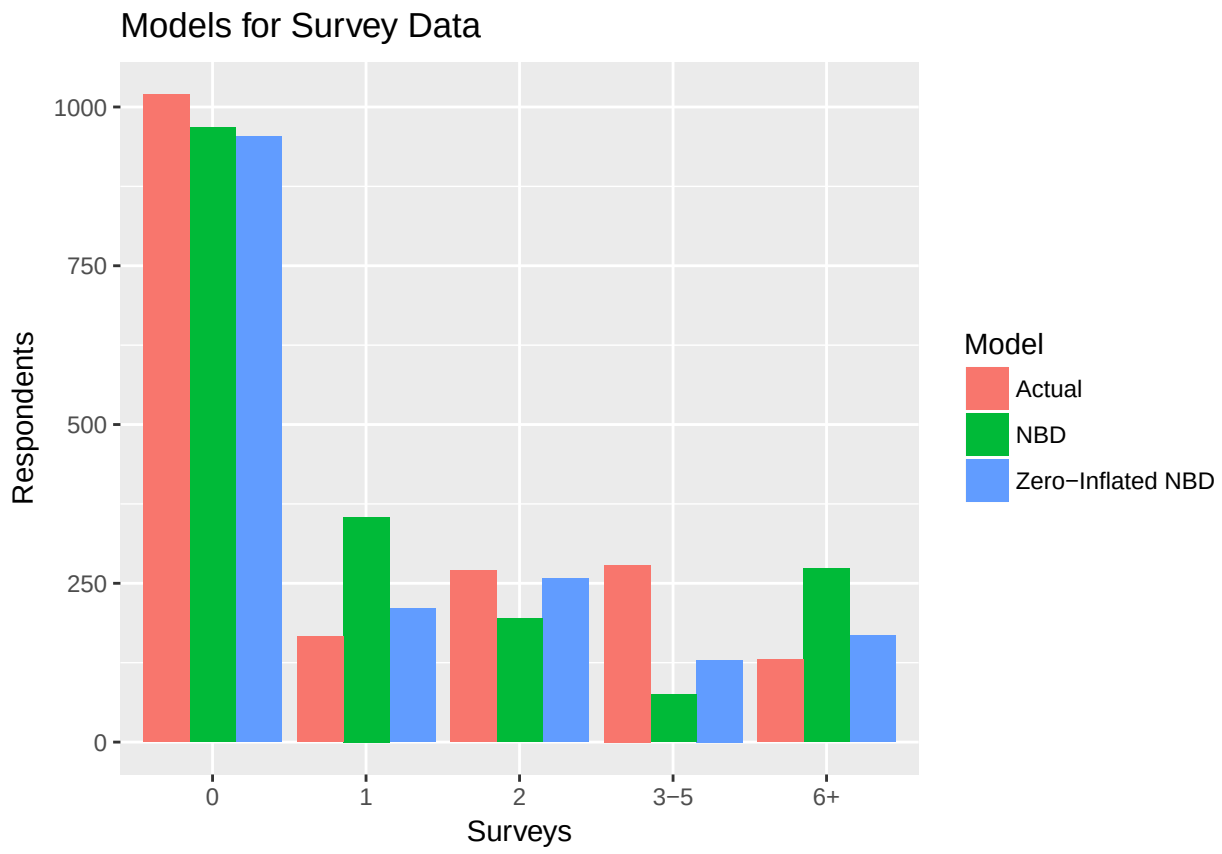
```
survey_data <-
  data_frame(
    surveys = c("0", "1", "2", "3-5", "6+")
    , surveys_int = c(0, 1, 2, 4, 6)
    , respondents = c(1020, 166, 270, 279, 130)
  )

#  A recursive function that calculates NBD model
# for a given x, r, and alpha (aggregated data)
fn_nbd_recursive <- function(x, r, alpha) {
  if(x == 0) {
    p_x <- (alpha / (alpha + 1))^r
    return(p_x)
  } else {
    p_x <- fn_nbd_recursive(x-1, r, alpha) * ((r+x-1)/(x*(alpha+1)))
    return(p_x)
  }
}

# Performs zero-inflation on P(X = x) values provided
# you give x and pi (inflated factor)
fn_zero_inflate <- function(p_x, x, pi) {
    if(x == 0) {
    return(pi + (1 - pi) * p_x)
  } else {
    return((1 - pi) * p_x)
  }
}

# Calculates the MLE for a NBD model with aggregated data
fn_max_ll_nbd <- function(par, zero_inflated = FALSE, surveys, count) {

  r <- par[1]
  alpha <- par[2]
  if (zero_inflated) {
    pi <- par[3]
  } else {
    pi <- 0
  }

  nbd <- map_dbl(surveys, .f = fn_nbd_recursive, r, alpha)
  zi <- map2_dbl(nbd, surveys, .f = fn_zero_inflate, pi)
  all <- c(zi, fn_zero_inflate(1 - sum(zi), -1, pi))
  ll <- sum(log(all) * count)

  return(-ll)
}


# Identify r and alpha for NBD model
survey_params <- nlminb(c(1,1), fn_max_ll_nbd, lower = c(0, 0), upper = c(Inf, Inf),
                 zero_inflated = FALSE, surveys = c(0, 1, 2, 4), count = survey_data$respondents)
```

```r
# Identify r, alpha, and pi for ZI-NBD model
survey_params_zi <- nlminb(c(1,1,.5), fn_max_ll_nbd, lower = c(0,0,0), upper = c(Inf, Inf, 1),
                 zero_inflated = TRUE, surveys = c(0, 1, 2, 4), count = survey_data$respondents)

data_frame(
  model = c("NBD", 'Zero-Inflated NBD')
  , r = c(survey_params$par[1], survey_params_zi$par[1])
  , alpha = c(survey_params$par[2], survey_params_zi$par[2])
  , pi = c(NA, survey_params_zi$par[3])
) %>%
  pander(missing = "")
```

model	r	alpha	pi
NBD	0.5	0.37	
Zero-Inflated NBD	19943.8	8135.12	0.47

Next we plot the results, noting that the zero-inflated NBD does well except for the category **3-5**.

```r
# Calculates the expected number of respondents
# given r, alpha, and pi and the number of
# surveys
fn_calc_expectation <- function(surveys, r, alpha, pi) {
  nbd <- map_dbl(surveys, .f = fn_nbd_recursive, r, alpha)
  zi <- map2_dbl(nbd, surveys, .f = fn_zero_inflate, pi)
  all <- c(zi, fn_zero_inflate(1 - sum(zi), -1, pi))
  return(all)
}

# From the survey data calculate expected number of respondents for each type of
# model
results_q3 <-
  survey_data %>%
    mutate(
      total = sum(respondents)
      , `NBD` = fn_calc_expectation(c(0, 1, 2, 4),
              survey_params$par[1], survey_params$par[2], 0) * total
      , `Zero-Inflated NBD` = fn_calc_expectation(c(0, 1, 2, 4),
              survey_params_zi$par[1], survey_params_zi$par[2], survey_params_zi$par[3]) * total
      )


results_q3 %>%
  select(-surveys_int, -total) %>%
  gather(model, value, -surveys) %>%
  mutate(model = ifelse(model == "respondents", "Actual", model)) %>%
  mutate(model = factor(model, levels = c("Actual","NBD", 'Zero-Inflated NBD'))) %>%
  ggplot(aes(x = surveys, y = value, fill = model)) +
  geom_bar(stat = 'identity', position = 'dodge') +
  labs(x = "Surveys", y = "Respondents", fill = "Model",
       title = "Models for Survey Data")
```



Finally, we see that though there is improvement in the χ^2 statistic for the zero-inflated model, we have no evidence that the data came from either of the models we implemented.

```
results_q3 %>%
  select(-surveys_int, -total) %>%
  gather(model, expected, -surveys, -respondents) %>%
  rename(Actual = respondents) %>%
  mutate(chi.squared = (Actual - expected)^2 / expected) %>%
  group_by(model) %>%
  summarise(chi.squared = sum(chi.squared)) %>%
  inner_join(
    data_frame(
      model = c('NBD',"Zero-Inflated NBD")
      , df = c(2, 3)
    )
    , by = c('model')
  ) %>%
  mutate(p.value = pchisq(chi.squared,df=df,lower.tail=FALSE)) %>%
  pander()
```

model	chi.squared	df	p.value
NBD	757.3	2	0
Zero-Inflated NBD	196.2	3	0