BLOOD DONATION PROJECT

Hamzah Sami

Stat 410

Background

Blood donations occupy a unique and fundamental aspect of health. They are integral for ensuring the welfare of individuals when their lives are in peril and require a blood transfusion. Since blood has a short shelf life for donations, a number of healthcare organizations organize frequent blood drives in order to increase the quantity of blood they have in supply. Since blood drives are frequent, there are a set of criteria required of blood donors before multiple donations can be made. These criteria range from tattoo policies to not being allowed to donate while ill. For the sake of this project, these policies will not be addressed and instead the focus will be placed on the time duration between consecutive blood donations. For example, in the United States, one must wait a minimum of six weeks between blood donations. Using this value of time durations along with other predictors, we will attempt to predict the number of donations blood donors make.

Dataset

The dataset used for the purpose of this project is managed by the Blood Transfusion Service Center based in Taiwan. This dataset is used by a number of organizations including the UCI Machine Learning and Information Systems Center to solve classifier problems. The dataset consists of a number of predictors: Months between last and current donation, number of donations, volume of blood donated (ml), months since you first donated, donated in March 2007. Of the four predictors, donated in March 2007 is the only categorical variable while the first four listed are numeric. As stated in the Background section, we will be attempting to predict the number of donations made by blood donors and in order to do so, we will use a zero-truncated negative Binomial model. We chose the model because the number off donations is a non-zero quantity but also because it is a predictor with a high level of variance and as such, the data given for the number of donations reflects this observation.

Results

Parameter Estimates for Truncated Negative Binomial Model					
Effect	**marchdonor**	**Estimate**	**Standard Error**	**z Value**	**Pr > |z|**
Intercept		0.5977	0.1102	5.43	<.0001
monthslastdonation		-0.01096	0.006001	-1.83	0.0677
volume_ch		0.3487	0.03630	9.61	<.0001
first		0.01068	0.002051	5.21	<.0001
marchdonor	marchdonor	0.07406	0.08148	0.91	0.3634
marchdonor	ref	0	.	.	.
Scale Parameter		0.06600	0.02066		

Obs	deviance	pvalue
1	277.9	0

Fitted Model for a Zero-Truncated Negative Binomial Model:

λ_hat = exp{0.5977 - 0.01096(monthslastdonation) + 0.3487(volume_ch) + 0.01068(first) + 0.07406(marchdonor)} with the dispersion parameter r_hat = 1/0.066 = 15.2.

Significant predictors at the 5% level. Interpretation of coefficients

The significant predictors at the 5% level are volume_ch and first (months between first and last donation) because the p-values for both predictors are less than 0.05. Since the model fits a zero-truncated negative binomial model, there is no easy interpretation of the estimated regression coefficients. As a result, we omit the interpretation.

Based on the Goodness of Fit test, the fitted model has a better fit of the data because the p-value for the deviance statistic is less than 0.05.

Prediction: The predicted number of blood donations made by a blood donor who waited 11 months between blood donations, has donated 1575 ml, waited 60 months between their first and last donations, and did not donate in March 2007 is about 2.47 donations or approximately 3 donations.

$$y^\circ = \frac{\exp\{0.5977 - 0.01096(11) + 0.3487(1.575) + 0.01068(60) + 0.07406(0)\}}{1- (1+\exp\{0.5977 - 0.01096(11) + 0.3487(1.575) + 0.01068(60) + 0.07406(0)/15.2)\text{^}-15.2\}}$$

Obs	p_donations
101	2.47004

Conclusion:

The model actually predicted my data well which indicated that my assumption that the number of donations did fit a zero-truncated negative binomial model. The deviance statistic had a p-value of 0 which seemed skeptical to me but indicated that the fitted model had a perfect fit of the data since the p-value was less than 0.05. The fact the prediction was somewhat accurate did also help to assuage my concern.

In addition, when I ran the same model in R, I ended up getting a similar answer to the one that I got when using SAS. Furthermore, the regression coefficients of the model end up matching in both SAS and R.

From working with this dataset, I was able to gain more experience with applying the stats concepts I had learned in the class to a new dataset of my choice. Having the opportunity to learn and work with the data ended up helping me understand how to use the regression model more effectively. Furthermore, I was also able to gain a greater appreciation for blood donors as well as organizations such as the Red Cross because the work that they do in blood donations is incredibly meaningful and extremely important for those who are sick.

References

- https://www.medicaldaily.com/10-surprising-facts-about-donating-blood-most-needed-blood-type-time-year-most-408705
- https://www.redcrossblood.org/donate-blood/how-to-donate/how-blood-donations-help/blood-needs-blood-supply.html
- https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center
- https://www.kaggle.com/bonastreyair/predicting-blood-analysis
- https://www.redcross.sg/news-stories/events/642-dropsoflife2018.html

Appendix A

SAS Programming

```
data blooddonors;
input monthslastdonation donations volume first marchdonor$ @@;
cards;

2    12   3000   52   yes
21   7    1750   38   yes
4    1    250    4    yes
11   11   2750   38   yes
4    12   3000   34   no
3    21   5250   42   no
4    2    500    4    yes
14   1    250    14   no
23   2    500    87   yes
14   4    1000   64   no
13   3    750    16   yes
11   7    1750   62   no
5    11   2750   75   yes
4    1    250    4    yes
4    4    1000   26   yes
11   1    250    11   yes
11   6    1500   26   yes
7    14   3500   48   yes
23   14   3500   93   yes
3    4    1000   29   no
2    7    1750   29   yes
4    6    1500   35   yes
5    7    1750   26   no
4    1    250    4    no
2    3    750    38   no
5    14   3500   86   no
2    2    500    11   yes
14   1    250    14   yes
4    3    750    16   no
2    12   3000   52   yes
4    14   3500   86   yes
23   7    1750   88   yes
2    1    250    2    yes
4    7    1750   58   yes
4    2    500    41   no
11   7    1750   29   yes
2    2    500    41   no
14   1    250    14   yes
4    6    1500   28   yes
4    1    250    4    no
2    2    500    4    no
16   6    1500   35   no
16   6    1500   81   no
2    5    1250   26   no
14   3    750    31   yes
2    1    250    2    no
4    4    1000   14   yes
14   3    750    35   yes
14   1    250    14   yes
14   5    1250   28   yes
2    14   3500   57   no
5    24   6000   79   yes
14   4    1000   23   yes
4    6    1500   39   no
23   2    500    38   yes
11   8    2000   52   no
2    7    1750   77   no
4    5    1250   11   no
2    4    1000   35   no
2    15   3750   64   no
4    1    250    4    yes
11   2    500    38   no
12   15   3750   71   yes
2    13   3250   76   yes
11   2    500    38   yes
16   2    500    27   no
9    4    1000   65   no
21   16   4000   64   no
7    10   2500   47   no
4    1    250    4    no
11   5    1250   35   yes
16   4    1000   23   no
38   1    250    38   no
4    1    250    4    no
23   4    1000   52   yes
11   7    1750   64   yes
14   3    750    28   no
4    11   2750   78   no
16   4    1000   33   no
4    5    1250   11   no
38   1    250    38   yes
11   11   2750   38   yes
2    2    500    11   no
2    2    500    4    no
2    14   3500   57   no
21   1    250    21   no
11   1    250    11   no
16   2    500    26   no
21   16   4000   64   no
2    4    1000   26   no
4    1    250    4    yes
4    2    500    52   no
14   3    750    31   no
16   2    500    16   yes
16   11   2750   40   no
11   11   2750   42   no
4    5    1250   23   yes
11   12   3000   58   yes
23   8    2000   64   yes
4    2    500    4    yes
2    1    250    2    yes
11   6    1500   58   no
21   3    750    35   no
7    5    1250   35   yes
11   2    500    16   yes
2    4    1000   11   yes
15   16   4000   82   no
2    10   2500   49   no
4    1    250    4    no
16   3    750    21   no
9    2    500    16   no
4    1    250    4    no
23   7    1750   88   no
4    8    2000   28   no
11   11   2750   38   no
23   3    750    48   no
23   1    250    23   yes
4    7    1750   28   no
```

```
4	16	4000	38	no
2	2	500	23	yes
11	6	1500	58	no
4	9	2250	26	yes
11	4	1000	34	yes
1	26	6500	76	no
2	10	2500	52	no
2	16	4000	81	no
4	6	1500	46	no
11	2	500	11	no
11	4	1000	34	no
14	2	500	14	no
3	5	1250	26	yes
11	1	250	11	yes
11	3	750	15	yes
4	1	250	4	no
4	7	1750	58	no
4	2	500	4	no
11	1	250	11	no
11	3	750	76	no
2	13	3250	32	no
2	12	3000	98	no
2	34	8500	77	no
4	1	250	4	no
2	2	500	11	yes
2	7	1750	77	yes
4	1	250	4	no
23	1	250	23	yes
23	3	750	35	no
4	5	1250	33	no
23	3	750	62	no
2	1	250	2	no
21	2	500	41	no
2	9	2250	22	yes
11	9	2250	33	no
16	6	1500	40	no
16	3	750	19	no
8	15	3750	77	yes
16	1	250	16	yes
2	1	250	2	no
5	24	6000	79	no
11	1	250	11	no
12	15	3750	71	no
14	1	250	14	yes
4	6	1500	23	no
9	9	2250	16	no
2	3	750	52	no
16	3	750	50	no
2	2	500	11	no
4	14	3500	86	no
4	1	250	4	no
3	4	1000	29	no
23	1	250	23	yes
40	1	250	40	no
4	1	250	4	no
2	41	10250	98	no
4	2	500	52	no
21	3	750	38	yes
2	3	750	9	no
2	12	3000	95	no
11	2	500	52	no
4	1	250	4	no
14	1	250	14	no
11	5	1250	33	no
4	9	2250	26	no
16	1	250	16	no
4	16	4000	70	no
14	2	500	14	no
14	3	750	28	no
11	1	250	11	yes
2	1	250	2	no
14	2	500	14	yes
4	23	5750	58	yes
6	3	750	26	no
16	2	500	16	no
14	2	500	29	no
11	7	1750	64	no
16	1	250	16	no
21	2	500	23	no
23	8	2000	46	yes
23	2	500	28	yes
4	11	2750	64	no
;

/*fitting truncated negative binomial model*/
proc format;
value $marchdonorfmt 'no'='ref' 'yes'='marchdonor';
run;

proc fmm;
class marchdonor;
model donations = monthslastdonation volume first marchdonor/dist=truncnegbin;
format marchdonor $marchdonorfmt.;
run;

/*checking model fit*/
proc fmm;
model donations=/dist=truncnegbin;
run;

data deviance_test;
deviance= 1069.8-791.9;
pvalue=1-probchi(deviance,4);
run;

proc print;
run;
```

```
/*using fitted model for prediction*/
data prediction;
input monthslastdonation volume first marchdonor$;
cards;
11 1.575 60 no
;

data blooddonors;
set blooddonors prediction;
run;

proc fmm;
class marchdonor;
model donations = monthslastdonation volume first marchdonor/dist=truncnegbin;
output out=outdata pred=p_donations;
run;

proc print data=outdata(firstobs=101 obs=101);
var p_donations;
run;
```

R Programming

```
> blood.data = read.csv(file = "./Documents/blood-test.csv", header = TRUE, sep = ",")
> install.packages("VGAM")
trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.5/VGAM_1.0-6.tgz'
Content type 'application/x-gzip' length 7852157 bytes (7.5 MB)
==================================================
downloaded 7.5 MB


The downloaded binary packages are in
        /var/folders/xd/_8ybfrln43v0tdn_2w2tdch00000gn/T//RtmpC2RzYX/downloaded_packages
> library(VGAM)
Loading required package: stats4
Loading required package: splines
>
> blood.data$volume <- blood.data$volume/1000
>
> #fitting truncated negative binomial model
> summary(fitted.model<- vglm(donations ~  monthslastdonation+volume+monthsfirst+donationmarch,
data=blood.data, family=posnegbinomial()))

Call:
vglm(formula = donations ~ monthslastdonation + volume + monthsfirst +
   donationmarch, family = posnegbinomial(), data = blood.data)


Pearson residuals:
               Min      1Q Median     3Q    Max
loge(munb)  -2.889 -0.8717 -0.120 0.5047 2.0157
loge(size) -11.797 -0.4551  0.396 0.7159 0.8467

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept):1       0.59763    0.10128   5.901 3.62e-09 ***
```

```
(Intercept):2       2.71746    0.38511   7.056 1.71e-12 ***
monthslastdonation -0.01096    0.00602  -1.821   0.0686 .
volume              0.34880    0.02647  13.178  < 2e-16 ***
monthsfirst         0.01068    0.00201   5.310 1.10e-07 ***
donationmarchyes    0.07405    0.08189   0.904   0.3659
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of linear predictors:  2

Names of linear predictors: loge(munb), loge(size)

Log-likelihood: -395.9334 on 394 degrees of freedom

Number of iterations: 30

Warning: Hauck-Donner effect detected in the following estimate(s):
'(Intercept):2'
Warning message:
In vglm.fitter(x = x, y = y, w = w, offset = offset, Xm2 = Xm2,  :
  convergence not obtained in 30 IRLS iterations
>
> #checking model fit
> intercept.only.model<- vglm(donations ~ 1, data=blood.data,family=posnegbinomial())
> print(deviance<- -2*(logLik(intercept.only.model)-logLik(fitted.model)))
[1] 277.9597
> print(p.value<- pchisq(deviance, df=4, lower.tail=FALSE))
[1] 6.13607e-59
>
> #using fitted model for prediction
> print(bloo <- predict(fitted.model, data.frame(monthslastdonation = 11, volume = 1.575, monthsfirst = 60,
donationmarch = "no"),type="response"))
      [,1]
1 2.424053
```

Appendix
B-SAS
output

The FMM Procedure

Model Information	
Data Set	WORK.BLOODDONORS
Response Variable	donations
Type of Model	Homogeneous Regression Mixture
Distribution	Truncated Negative Binomial
Components	1
Link Function	Log
Estimation Method	Maximum Likelihood

Class Level Information		
Class	**Levels**	**Values**
marchdonor	2	marchdonor ref

Number of Observations Read	200
Number of Observations Used	200

Optimization Information	
Optimization Technique	Dual Quasi-Newton
Parameters in Optimization	6
Mean Function Parameters	5
Scale Parameters	1
Lower Boundaries	1
Upper Boundaries	0
Number of Threads	2

Iteration History

Iteration	Evaluations	Objective Function	Change	Max Gradient
0	5	565.45841188	.	49.18234
1	7	486.38304328	79.07536860	12.14411
2	3	484.38383765	1.99920564	7.267725
3	2	483.70428858	0.67954907	9.063922
4	2	482.75256695	0.95172163	3.876234
5	3	482.11835994	0.63420701	3.536299
6	4	480.2892862	1.82907374	11.40184
7	2	477.13480105	3.15448515	25.61363
8	7	474.94317648	2.19162457	45.17784
9	2	472.77364093	2.16953555	70.19499

The FMM Procedure

Iteration History				
Iteration	**Evaluations**	**Objective Function**	**Change**	**Max Gradient**
10	9	446.87914929	25.89449165	79.87429
11	2	429.81432741	17.06482187	118.9
12	5	417.34663742	12.46768999	94.12813
13	2	411.28677155	6.05986586	1260.179
14	4	406.77980105	4.50697050	571.3306
15	2	403.6363867	3.14341436	395.2024
16	2	399.32291519	4.31347151	262.6971
17	3	398.8308071	0.49210809	224.1611
18	2	398.12124019	0.70956691	164.5389
19	2	397.03747741	1.08376278	116.0991
20	3	396.38659803	0.65087938	36.50997
21	3	395.9917301	0.39486793	16.03637
22	3	395.93580972	0.05592039	1.306962
23	3	395.93351967	0.00229004	0.769872
24	3	395.93342369	0.00009598	0.013313
25	3	395.93342356	0.00000013	0.000809

Convergence criterion (GCONV=1E-8) satisfied.

Fit Statistics	
-2 Log Likelihood	791.9
AIC (Smaller is Better)	803.9
AICC (Smaller is Better)	804.3
BIC (Smaller is Better)	823.7
Pearson Statistic	134.1

Parameter Estimates for Truncated Negative Binomial Model					
Effect	**marchdonor**	**Estimate**	**Standard Error**	**z Value**	**Pr > |z|**
Intercept		0.5977	0.1102	5.43	<.0001
monthslastdonation		-0.01096	0.006001	-1.83	0.0677
volume_ch		0.3487	0.03630	9.61	<.0001
first		0.01068	0.002051	5.21	<.0001
marchdonor	marchdonor	0.07406	0.08148	0.91	0.3634
marchdonor	ref	0	.	.	.
Scale Parameter		0.06600	0.02066		

The FMM Procedure

Model Information	
Data Set	WORK.BLOODDONORS
Response Variable	donations
Type of Model	Non-Mixture
Distribution	Truncated Negative Binomial
Components	1
Link Function	Log
Estimation Method	Maximum Likelihood

Number of Observations Read	200
Number of Observations Used	200

Optimization Information	
Optimization Technique	Dual Quasi-Newton
Parameters in Optimization	2
Mean Function Parameters	1
Scale Parameters	1
Lower Boundaries	1
Upper Boundaries	0
Number of Threads	2

Iteration History				
Iteration	**Evaluations**	**Objective Function**	**Change**	**Max Gradient**
0	5	563.70602539	.	39.43064
1	4	541.69203337	22.01399202	21.22587
2	2	539.71299002	1.97904334	10.30861
3	2	538.72003366	0.99295637	1.549716
4	4	538.61643398	0.10359968	2.94643
5	13	535.81468386	2.80175012	12.77112
6	5	535.53460914	0.28007471	7.273163
7	4	534.92239933	0.61220981	1.120118
8	3	534.91403002	0.00836931	0.404488
9	3	534.91328409	0.00074593	0.011157
10	3	534.91326006	0.00002403	0.001939
11	3	534.91326004	0.00000002	9.13E-7

Convergence criterion (GCONV=1E-8) satisfied.

The FMM Procedure

Fit Statistics	
-2 Log Likelihood	1069.8
AIC (Smaller is Better)	1073.8
AICC (Smaller is Better)	1073.9
BIC (Smaller is Better)	1080.4
Pearson Statistic	187.7

Parameter Estimates for Truncated Negative Binomial Model

Effect	Estimate	Standard Error	z Value	Pr > |z|	Inverse Linked Estimate
Intercept	1.4002	0.1504	9.31	<.0001	4.0561
Scale Parameter	1.8687	0.5378			

Obs	deviance	pvalue
1	277.9	0

The FMM Procedure

Model Information	
Data Set	WORK.BLOODDONORS
Response Variable	donations
Type of Model	Homogeneous Regression Mixture
Distribution	Truncated Negative Binomial
Components	1
Link Function	Log
Estimation Method	Maximum Likelihood

Class Level Information		
Class	**Levels**	**Values**
marchdonor	2	no yes

Number of Observations Read	201
Number of Observations Used	200

Optimization Information	
Optimization Technique	Dual Quasi-Newton
Parameters in Optimization	6
Mean Function Parameters	5
Scale Parameters	1
Lower Boundaries	1
Upper Boundaries	0
Number of Threads	2

Iteration History

Iteration	Evaluations	Objective Function	Change	Max Gradient
0	5	565.45841188	.	49.18234
1	7	486.38304328	79.07536860	12.14411
2	3	484.38383765	1.99920564	7.267725
3	2	483.70428858	0.67954907	9.063922
4	2	482.75256695	0.95172163	3.876234
5	3	482.11835994	0.63420701	3.536299
6	4	480.2892862	1.82907374	11.40184
7	2	477.13480105	3.15448515	25.61363
8	7	474.94317648	2.19162457	45.17784
9	2	472.77364093	2.16953555	70.19499

The FMM Procedure

Iteration History				
Iteration	**Evaluations**	**Objective Function**	**Change**	**Max Gradient**
10	9	446.87914929	25.89449165	79.87429
11	2	429.81432741	17.06482187	118.9
12	5	417.34663742	12.46768999	94.12813
13	2	411.28677155	6.05986586	1260.179
14	4	406.77980105	4.50697050	571.3306
15	2	403.6363867	3.14341436	395.2024
16	2	399.32291519	4.31347151	262.6971
17	3	398.8308071	0.49210809	224.1611
18	2	398.12124019	0.70956691	164.5389
19	2	397.03747741	1.08376278	116.0991
20	3	396.38659803	0.65087938	36.50997
21	3	395.9917301	0.39486793	16.03637
22	3	395.93580972	0.05592039	1.306962
23	3	395.93351967	0.00229004	0.769872
24	3	395.93342369	0.00009598	0.013313
25	3	395.93342356	0.00000013	0.000809

Convergence criterion (GCONV=1E-8) satisfied.

Fit Statistics	
-2 Log Likelihood	791.9
AIC (Smaller is Better)	803.9
AICC (Smaller is Better)	804.3
BIC (Smaller is Better)	823.7
Pearson Statistic	134.1

Parameter Estimates for Truncated Negative Binomial Model					
Effect	**marchdonor**	**Estimate**	**Standard Error**	**z Value**	**Pr > |z|**
Intercept		0.6718	0.1199	5.60	<.0001
monthslastdonation		-0.01096	0.006001	-1.83	0.0677
volume_ch		0.3487	0.03630	9.61	<.0001
first		0.01068	0.002051	5.21	<.0001
marchdonor	no	-0.07406	0.08148	-0.91	0.3634
marchdonor	yes	0	.	.	.
Scale Parameter		0.06600	0.02066		

Obs	p_donations
101	2.47004