Safety Performance Function of Pennsylvania Centre County in Bayesian Framework

Lingyu Li

May 1, 2016

1. Background

Safety performance function (SPF) is one of the most important tools in transportation safety study. It is widely used as a statistical method to evaluate the impact of traffic accident influential factors and predict the expected average crash frequency within a roadway segment associated with a particular geometric design, given time period, and specific traffic volume. In the Highway Safety Manual, published by American Association of State Highway and Transportation Officials, the SPFs are specified as negative binomial regression models considering the count outcome variable as total crash frequency (crash number per year), fatal crash frequency, or fatal and injury crash frequency. For the predictors, the length of the roadway segment is usually considered as an offset variable in the model since crash count and length of the road segment is usually considered to have a direct ratio. So the likelihood of a crash is not changing over the roadway length. Traffic volume is another important predictor since it directly indicates the exposure of vehicles on the road. Annual Average Daily Traffic (AADT) is a general measure of traffic volume. In addition, there are other factors affecting safety such as roadside hazard rate, shoulder rumble strips, the presence of passing zone, number of driver ways to the main roadway, horizontal curve density, degree of curvature per mile, etc. The predicted crash frequency of SPFs could be used as a strong reference to improve the transportation facilities, address some serious safety issues on highways or urban streets. It could also be used to evaluate the safety effectiveness of a countermeasure by processing a before and after study.

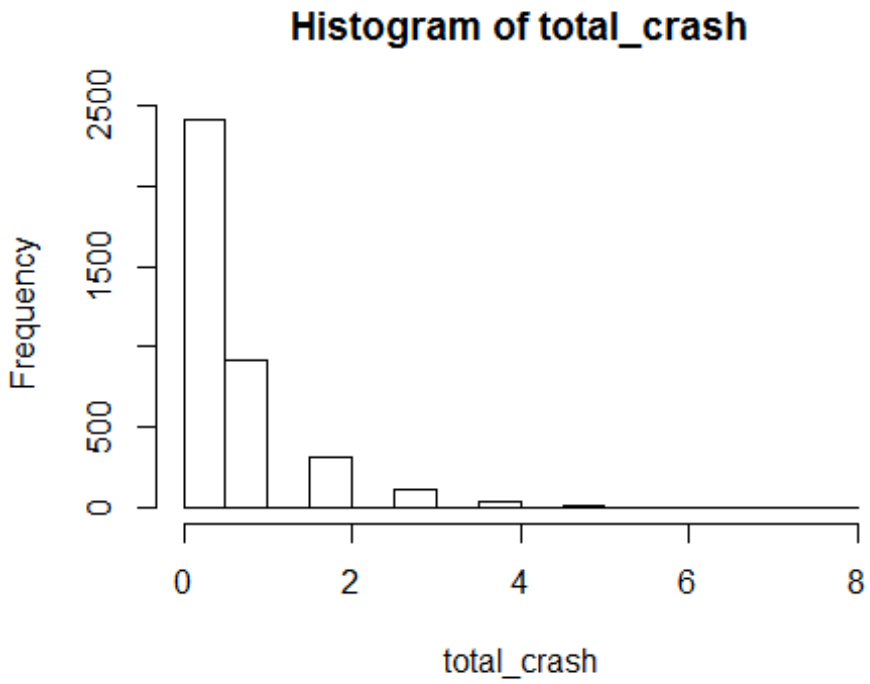
The SPF is usually estimated in the traditional(frequentist) framework using max likelihood estimation (MLE). However, compared with Bayesian estimation, there is some deficiency in the frequentist framework. First, the frequentist framework makes inference about the parameters according to the p-value. The p-value is equal to the probability that the test statistic exceeds the observed value given the null-hypothesis, which represented the extremeness of the observed result under the null-hypothesis. In this framework, there is no inference about the probability of that the null-hypothesis is true and about the alternative. Comparably, in the Bayesian framework, it is accessible to compare the evidence from the dataset and two hypotheses. In addition, the Bayes factor indicates the power of null-hypothesis versus the alternative. It quantifies the power of statistical inference. Besides, Bayesian methods are also advanced in computing methods. Bayesian models have the advantage of being able to handle very complex models, especially for some models that do not have easily calculable likelihood functions. The Markov Chain Monte Carlo (MCMC) sampling estimation methods makes it easier to handle complex function forms. For example, in transportation data analysis, random parameter models are

used to capture the unobserved heterogeneity since it allows parameters to vary across observations (such as roadway segments). Random parameter models are more easily estimated using MCMC method. Thus, it is meaningful to apply Bayesian method on the estimation of safety performance function. In this study, a Pennsylvania centre county two-lane rural highway safety performance function is estimated in Bayesian framework. The dataset used to estimate the model is provided by Pennsylvania Department of Transportation, which include the crash record, traffic information and geometric design features.

2. Data Description

The dataset used in the estimation is complied base on the Pennsylvania center county two-lane rural highway inventory and reported crash inventory from 2005 to 2012. Each crash record in the crash inventory were made by police officers who were assigned to investigate the transportation accidents. The roadway inventory and other traffic and geometric design information were provided by PennDOT. In the data frame, each row is one road segment of one year. There are 477 two-lane rural highway segments in centre county (8 years data, 3816 rows). The crash count is the number of crash occured in a year within the segment. The detailed information of the data is presented as following:

```
crash<-read.csv("centre_tlrh.csv")
attach(crash)

#Data Description

library(pastecs)

## Warning: package 'pastecs' was built under R version 3.2.5

## Loading required package: boot

#responsible varialbe, total crash count of one year one segment

#from the histogram, the distribution of total crash follows poision distribu
tion
#Most of the crash count is zero since the daily traffic volume in rural area
is very low
#Since the variance (0.8168) is larger than the mean (0.5561), we should use
negative binomial distribution to count for overdispersion
stat.desc(total_crash)

##      nbr.val     nbr.null       nbr.na          min          max
## 3.816000e+03 2.417000e+03 0.000000e+00 0.000000e+00 8.000000e+00
##        range          sum       median         mean      SE.mean
## 8.000000e+00 2.122000e+03 0.000000e+00 5.560797e-01 1.463010e-02
## CI.mean.0.95          var      std.dev     coef.var
## 2.868356e-02 8.167756e-01 9.037564e-01 1.625228e+00

hist(total_crash)
```




```
#AADT is the Annual Average Daily Traffic, unit is vehicle/day
stat.desc(aadt_yr)
```

```
##      nbr.val     nbr.null       nbr.na          min          max
## 3.816000e+03 0.000000e+00 0.000000e+00 1.440000e+02 1.404600e+04
##        range          sum       median         mean      SE.mean
## 1.390200e+04 1.301711e+07 2.379000e+03 3.411192e+03 4.773552e+01
## CI.mean.0.95          var      std.dev     coef.var
## 9.358960e+01 8.695443e+06 2.948804e+03 8.644497e-01
```

```
#length of the road segment, unit is mile
stat.desc(length_mi_yr)
```

```
##      nbr.val     nbr.null       nbr.na          min          max
## 3.816000e+03 0.000000e+00 0.000000e+00 1.174242e-01 7.505682e-01
##        range          sum       median         mean      SE.mean
## 6.331440e-01 1.799794e+03 4.751894e-01 4.716441e-01 1.584579e-03
## CI.mean.0.95          var      std.dev     coef.var
## 3.106704e-03 9.581563e-03 9.788546e-02 2.075409e-01
```

```
#roadside hazard rate, there are 7 levels, 1 is the lowest hazard rate and 7
is the highest
#rhr_4 indicator: If the roadside hazard rate is 4 (1) or others (0)
stat.desc(rhr_4)
```

```
##      nbr.val     nbr.null       nbr.na          min          max
## 3.816000e+03 2.912000e+03 0.000000e+00 0.000000e+00 1.000000e+00
##        range          sum       median         mean      SE.mean
```

```
## 1.000000e+00 9.040000e+02 0.000000e+00 2.368973e-01 6.883737e-03
## CI.mean.0.95          var      std.dev     coef.var
## 1.349616e-02 1.808243e-01 4.252345e-01 1.795016e+00

#rhr567 indicator: If the roadside hazard rate is 5 or 6 or 7 (1), or others(
0)
stat.desc(rhr567)

##      nbr.val     nbr.null       nbr.na          min          max
## 3.816000e+03 1.056000e+03 0.000000e+00 0.000000e+00 1.000000e+00
##        range          sum       median         mean      SE.mean
## 1.000000e+00 2.760000e+03 1.000000e+00 7.232704e-01 7.243206e-03
## CI.mean.0.95          var      std.dev     coef.var
## 1.420093e-02 2.002028e-01 4.474402e-01 6.186348e-01

#pass_zone indicator: If there is a passing lane in this segment (1) or not(0
)
stat.desc(pass_zone)

##      nbr.val     nbr.null       nbr.na          min          max
## 3.816000e+03 2.496000e+03 0.000000e+00 0.000000e+00 1.000000e+00
##        range          sum       median         mean      SE.mean
## 1.000000e+00 1.320000e+03 0.000000e+00 3.459119e-01 7.701115e-03
## CI.mean.0.95          var      std.dev     coef.var
## 1.509870e-02 2.263162e-01 4.757270e-01 1.375284e+00

#accessdensity (countinuous): driverway density in the segment (# of driverwa
ys/mile)
stat.desc(accessdensity)

##      nbr.val     nbr.null       nbr.na          min          max
## 3.816000e+03 2.240000e+02 0.000000e+00 0.000000e+00 1.011765e+02
##        range          sum       median         mean      SE.mean
## 1.011765e+02 6.503391e+04 1.248719e+01 1.704243e+01 2.575995e-01
## CI.mean.0.95          var      std.dev     coef.var
## 5.050460e-01 2.532202e+02 1.591289e+01 9.337222e-01

#curve_density (countinuous): horizontal curves density per mile in the segme
nt (# of horizontal curve/mile)
stat.desc(curve_density)

##      nbr.val     nbr.null       nbr.na          min          max
## 3.816000e+03 2.504000e+03 0.000000e+00 0.000000e+00 4.258065e+01
##        range          sum       median         mean      SE.mean
## 4.258065e+01 4.532716e+03 0.000000e+00 1.187819e+00 4.228078e-02
## CI.mean.0.95          var      std.dev     coef.var
## 8.289511e-02 6.821728e+00 2.611844e+00 2.198857e+00

#d_seg_mi (countinuous): degree of curvature per mile in the segment (degress
/100ft/mile), represented the average sharpness of horizontal curve in the se
gments
stat.desc(d_seg_mi)
```

```
##      nbr.val     nbr.null       nbr.na          min          max
## 3.816000e+03 2.504000e+03 0.000000e+00 0.000000e+00 2.754053e+02
##        range          sum       median         mean      SE.mean
## 2.754053e+02 3.540720e+04 0.000000e+00 9.278616e+00 3.959961e-01
## CI.mean.0.95          var      std.dev     coef.var
## 7.763845e-01 5.983982e+02 2.446218e+01 2.636404e+00
```

3. SPF in Frequentist Framework

The presentation of SPF in frequentist framework here is to show the general safety performance function and compare to the estimation in Bayesian framework. The predictors in the model is recommended by the District 2 two-lane rural highway segment SPF according to "Regionalized Safety Performance Functions". In the frequentist framework, the variable rhr_4, rhr567, and curve_density are not statistically significant in the model, which means the probability of observed statistic value exceed test statistic is not high enough, so it fails to reject the null-hyphothesis that these varialbes have no association with the rate.

```
library(MASS)
#negative binomial regression for total crash frequency
model1=glm.nb(total_crash~lnaadt+offset(lnlength)+rhr_4+rhr567+pass_zone+acce
ssdensity+curve_density+d_seg_mi)
summary(model1)
```

```
##
## Call:
## glm.nb(formula = total_crash ~ lnaadt + offset(lnlength) + rhr_4 +
##     rhr567 + pass_zone + accessdensity + curve_density + d_seg_mi,
##     init.theta = 3.346284366, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8296  -0.9497  -0.6232   0.3654   4.1951
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.797614   0.278399 -17.233  < 2e-16 ***
## lnaadt         0.615422   0.029564  20.816  < 2e-16 ***
## rhr_4          0.039661   0.124218   0.319  0.74951
## rhr567        -0.009826   0.121906  -0.081  0.93576
## pass_zone     -0.294933   0.054770  -5.385 7.25e-08 ***
## accessdensity  0.004784   0.001478   3.236  0.00121 **
## curve_density  0.019680   0.016782   1.173  0.24093
## d_seg_mi       0.004563   0.001815   2.514  0.01192 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(3.3463) family taken to be 1)
##
##     Null deviance: 4071.2  on 3815  degrees of freedom
```

```
## Residual deviance: 3390.0  on 3808  degrees of freedom
## AIC: 7106.7
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  3.346
##           Std. Err.:  0.556
##
##  2 x log-likelihood:  -7088.698
```

4. SPF in Bayesian Framework

For Bayesian estimation, firstly, it is necessary to set a prior distribution of the parameters. The prior distribution is set according to the prior knowledge of this problem. According to the property of each parameters, and the prior knowledge from the Distric 2 SPF from "Regionalized Safety Performance Function", the prior of each parameters in negative binomial regression are set. The detailed explanation is presented in the following code. With regard to the selection of predictors,they are the same group of predictors recommended in the SPF report. In order to compare the effects of the predictors, except AADT and Length, which is the exposure and offset variable, all the other varible are input as the standalized form. Then the effect of these geometric design factors can be directly compared (across categorical variables or continuous variables).The data input, model specification, prior setting and MCMC estimation setting is presented as following.

Model specification:

$\text{Log}(\lambda) = \text{lnlength} + \alpha*\text{lnaadt} + \beta_0 + \beta_1*\text{rhr\_4} + \beta_2*\text{rhr567} + \beta_3*\text{pass\_zone} + \beta_4*\text{accessdensity} + \beta_5*\text{curve\_density} + \beta_6*\text{d\_seg\_mi}$

```
#observed counts follows negative binomial distribution

for(i in 1:n){

y[i] ~ dnegbin(p[i],r)
mu[i]<- lnlength[i] + alpha*lnaadt[i] + beta0 + beta1*rhr_4[i] + beta2*rhr567
[i] + beta3*pass_zone[i] +   beta4*accessdensity[i] + beta5*curve_density[i]
beta6*d_seg_mi[i]
lambda[i]<-exp(mu[i])
p[i]<-r/(r+lambda[i])

mu[i]<- lnlength[i] + alpha*lnaadt[i] + beta0 + beta1*rhr_4[i] + beta2*rhr567
[i] + beta3*pass_zone[i] +   beta4*accessdensity[i] + beta5*curve_density[i]
+ beta6*d_seg_mi[i]
```

Prior setting:

```
##r is set to follow a categorical distribution
r ~ dcat(pi[])
for(i in 1:100){pi[i]<-1/100}
##alternative: uniform distribution
# r ~ dunif(0,50)

##coefficient of the explanatory variables
```

```
#prior of alpha on lnaadt (normal distribution)
alpha ~ dnorm(1,1/0.1^2)
#prior of beta0, intercept (normal distribution)
beta0 ~ dnorm(-5,1/1^2)
#prior of coefficients of other geometric design predictors (normal distribut
ion)
beta1 ~ dnorm(0,1/0.5^2)
beta2 ~ dnorm(0,1/0.5^2)
beta3 ~ dnorm(0,1/0.5^2)
beta4 ~ dnorm(0,1/0.5^2)
beta5 ~ dnorm(0,1/0.5^2)
beta6 ~ dnorm(0,1/0.5^2)

#examine the posterior
lambda_mean<-mean(lambda)
p_mean <- mean(p)

#examine the likelihood of data
y_mean <- mean(y)
```

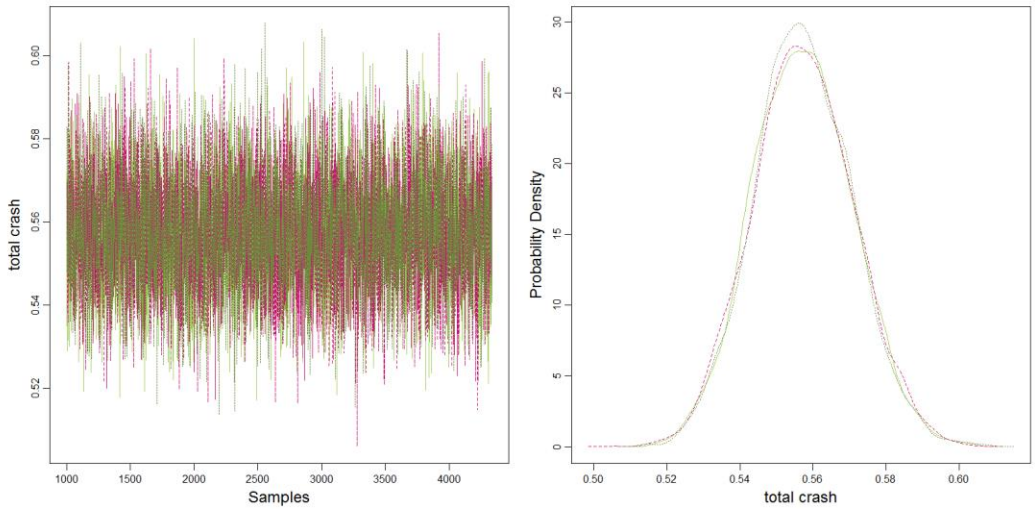
Posterior summary:

Convergence criterion was met for every parameter.

Posterior distribution of the parameters:

	mean	PSD	PCI 2.50%	PCI 97.50%	95% HDI_Low	95% HDI_High	inROPE	stROPE	ltROPE	ESS	RHAT
alpha	0.6467	0.0320	0.5835	0.7078	0.5836	0.7079	0.0000	0.0000	1.0000	36	1.0259
beta0	-5.0687	0.2997	-5.6491	-4.4442	-5.5568	-4.4038	0.0000	1.0000	0.0000	32	1.0434
beta1	0.0459	0.1226	-0.1865	0.2978	-0.1916	0.2921	0.3043	0.2196	0.4761	193	1.0204
beta2	0.0155	0.1207	-0.2091	0.2627	-0.2229	0.2457	0.3140	0.3048	0.3811	164	1.0301
beta3	-0.2980	0.0542	-0.4022	-0.1920	-0.4031	-0.1933	0.0000	1.0000	0.0000	2467	1.0000
beta4	0.0045	0.0015	0.0016	0.0075	0.0015	0.0074	1.0000	0.0000	0.0000	1057	1.0004
beta5	0.0182	0.0175	-0.0161	0.0525	-0.0154	0.0532	0.9664	0.0000	0.0336	865	1.0014
beta6	0.0049	0.0019	0.0011	0.0085	0.0013	0.0087	1.0000	0.0000	0.0000	874	1.0002
lambda_mean	0.5573	0.0135	0.5315	0.5845	0.5314	0.5843	0.0000	0.0000	1.0000	10002	0.9999
p_mean	0.8668	0.0184	0.8468	0.9049	0.8462	0.9041	0.0000	0.0000	1.0000	10002	0.9999
r	3.5024	0.6550	3.0000	5.0000	3.0000	5.0000	0.0000	0.0000	1.0000	10002	0.9999
y_mean	0.5561	0.0000	0.5561	0.5561	0.5561	0.5561	0.0000	0.0000	1.0000	0	1.0000

Markov Chain Monte Carlo (MCMC) sampling:



The estimation result shows that the convergence criterion was met for every parameter. With respect to the effective sample size of MCMC sampling, alpha and beta0 is quite low effective sample size (ESS). Effective sample size measures the amount of independent information (sample size of a completely non-autocorrelated chain). But all the other parameters has a comparably high effective sample size, which indicates a good accuracy.

With regard to the posterior distribution of each parameter:

alpha is the coefficient of lnaadt, the mean of alpha is quite close to the alpha in traditional estimation.

beta0 is the intercept, the exp(beta0) represent the baseline of crash frequency prediction. It's close to the estimate from frequentist framework.

beta1 to beta6 is the posterior distribution of the coefficient of all geometric design factors. The positive coefficient indicates that if the value of the variable increase, there will be more crashes and vice versa. For the dummy variables, beta3 is the only negative coefficient and has the highest absolute value, which means the presence of passing lane affect the most compared with roadside hazard rate 4 and roadside hazard rate 5 or 6 or 7. It reduces total crash frequency. The most interesting finding is that coefficient of rhr567 (beta2) is quite different from the estimates in frequentist framework. In the frequentist framework, rhr567 is not statistically significant due to lack of variation in the dataset. The coefficient distribution of rhr567 in Bayesian estimation should be more convictive since it converges the prior distribution and observed data likelihood of this parameter. For continuous variables, beta6 is the highest, so degree of curvature per mile has the largest impact. If the horizontal in the road segment is sharper, there will be more crashes.

The lambda_mean is the posterior distribution of crash count mean, which is close to the mean of the observed data. It somehow indicates a favorable the model prediction is close to the reality.

5. Conclusion

In summary, the project uses Bayesian method to estimate a negative binomial model to predict the crash frequency in each two-lane rural highway segments in center county Pennsylvania. Roadway length, traffic volume, and geometric design features are considered as the influential factors in the model. According to the MCMC sampling result, the convergence criteria were met for each parameter. The posterior distribution of each parameter is presented. Compared with the estimation in the frequentist framework, most of the coefficient means in Bayesian estimation are close to that in frequentist framework. The difference estimates of rhr567 indicate that Bayesian estimation is more convictive if there is not much variation of an indicator in the dataset.

6. Reference

1) Gigerenzer, Gerd. "Mindless statistics." The Journal of Socio-Economics 33.5 (2004): 587-606.

2) Mannering, F. L., & Bhat, C. R. (2014). Analytic methods in accident research: Methodological frontier and future directions. Analytic Methods in Accident Research, 1, 1–22. http://doi.org/10.1016/j.amar.2013.09.001

3) Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation Research Part A: Policy and Practice, 44(5), 291–305.

4) Donnell, Eric, Vikash Gayah, and Lingyu Li. Regionalized Safety Performance Functions. No. FHWA-PA-2016-001-PSU WO 017. 2016.