Statistical modelling and validation using 500 drivers

*Miao Cai**

2019-02-01

1 Logistic regression

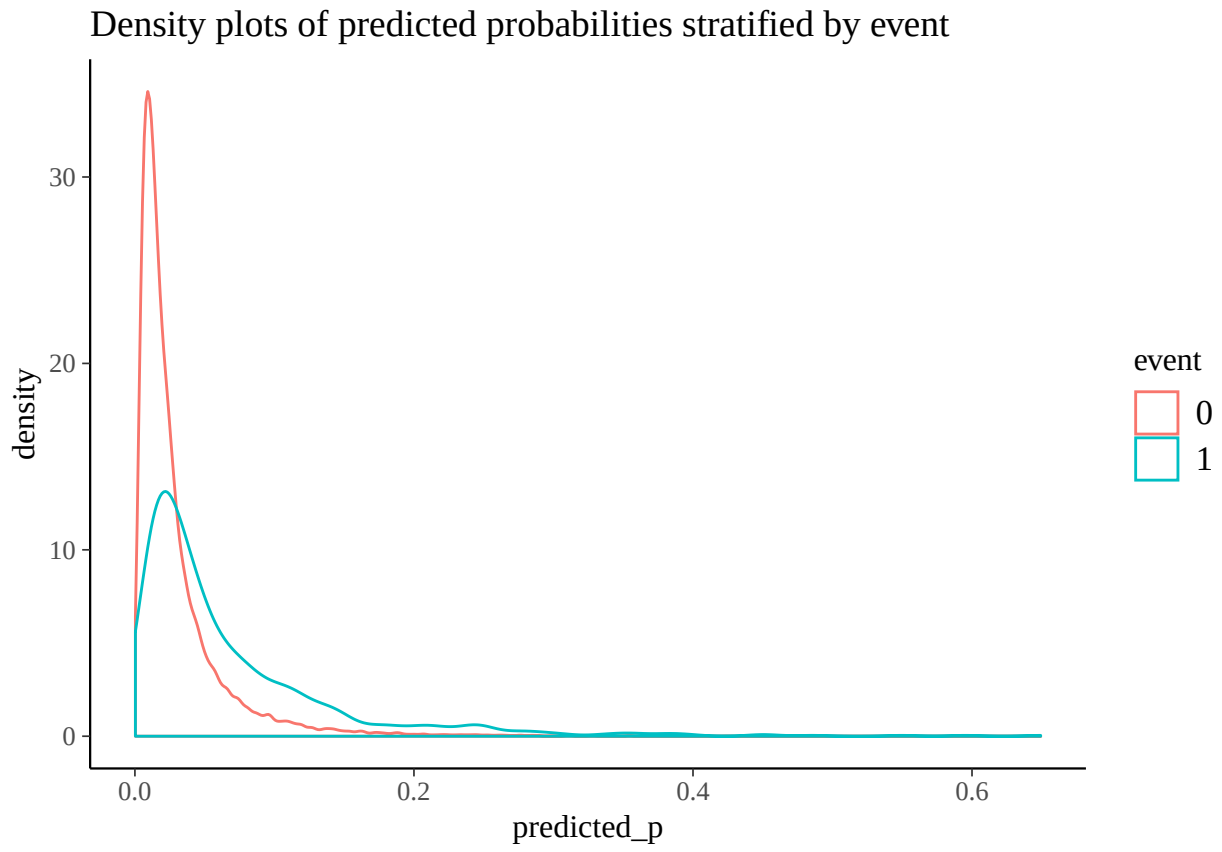
1.1 Logistic regression predicted by cumulative driving time

$$Y \sim Bernoulli(p)$$

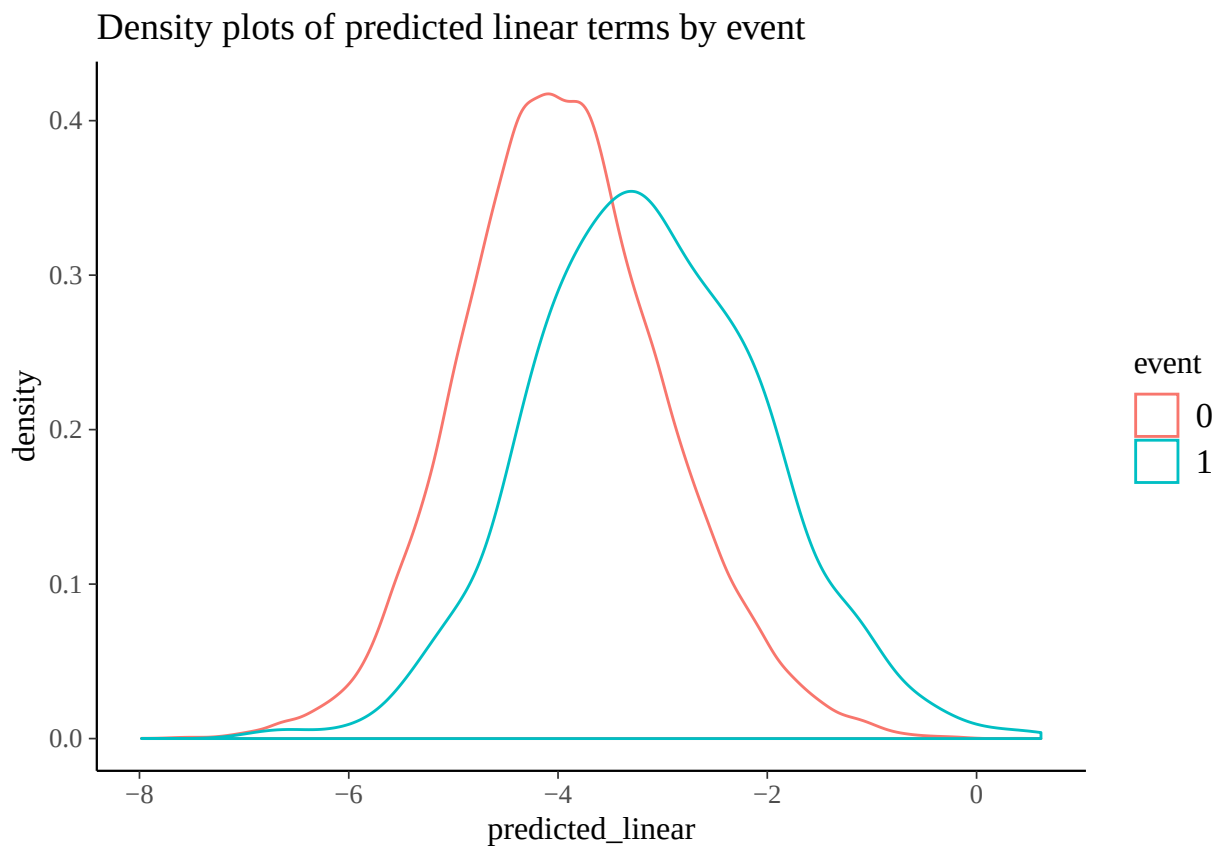
$$\text{Logit}\frac{p}{1-p} = \beta_{1,d(i)} + \beta_{2,d(i)} * CT$$

- 498 drivers
- in total: 283,321 trips
- Train data: 10% in each driver = 28,335 trips
- test data: the rest 90% = 254,733 trips

```
## # A tibble: 3 x 3
##   term        estimate std.error
##   <chr>          <dbl>     <dbl>
## 1 (Intercept)  -4.55      0.117
## 2 cumDrive     -0.0171    0.0180
## 3 travelTime    0.319     0.0191
```



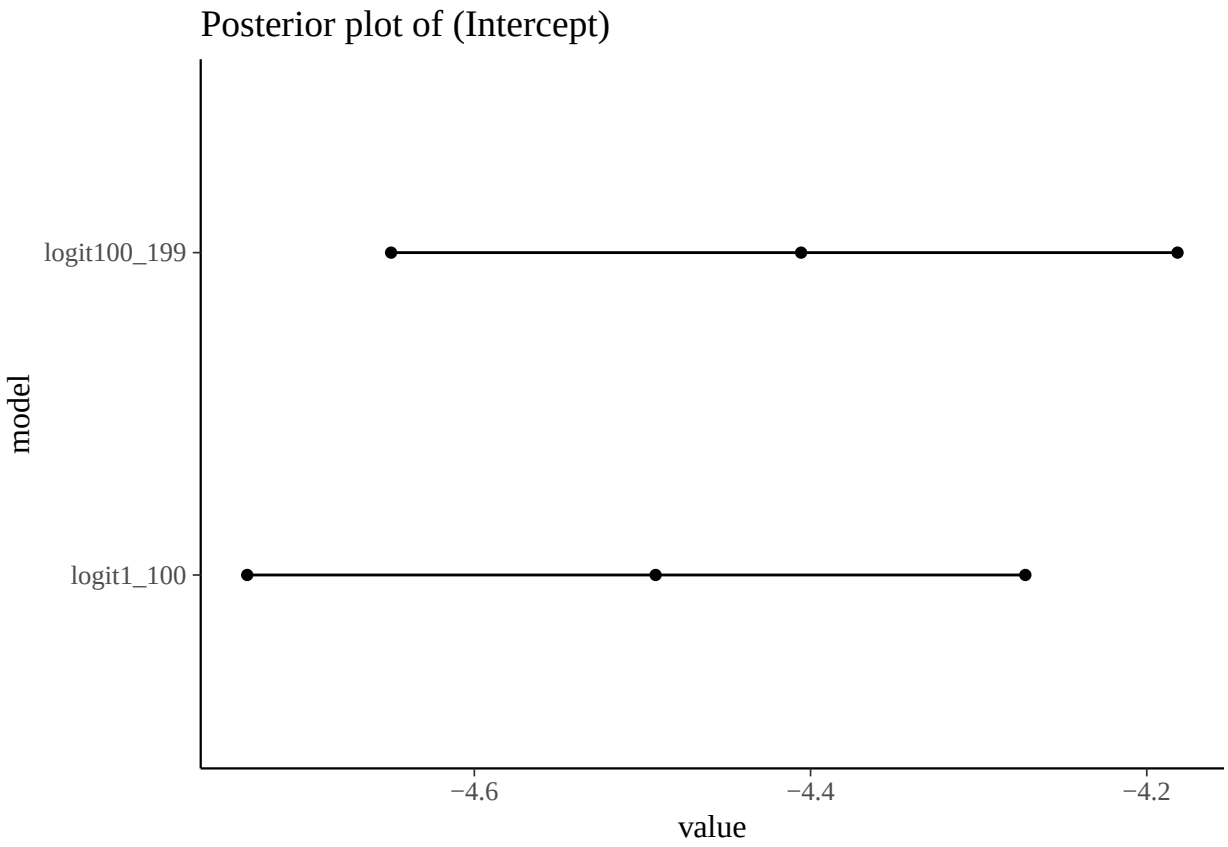
*Department of Epidemiology and Biostatistics, Saint Louis University. Email address miao.cai@slu.edu

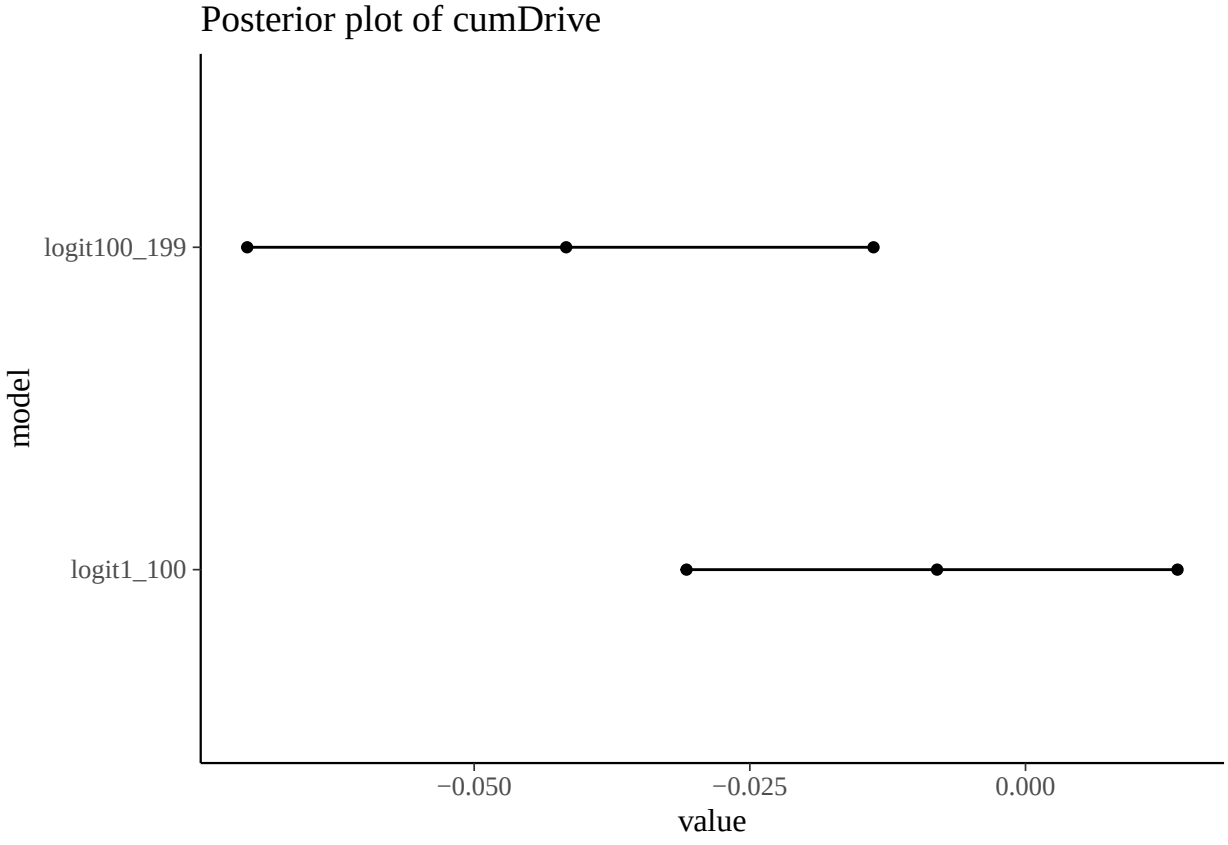


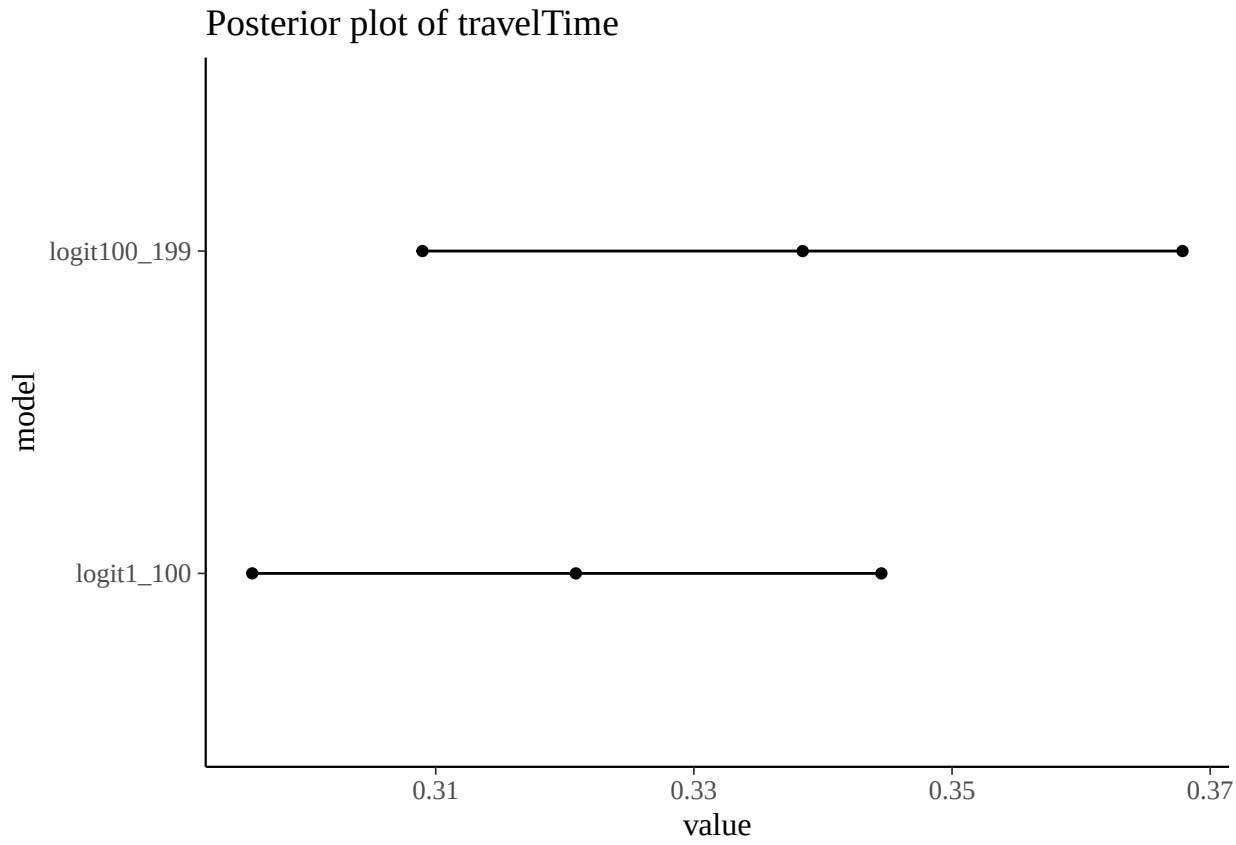

```
## [1] 0.003802281
```

```
## [1] 0.00004051207
```

1.2 Comparing drivers 1-100 and 100-199



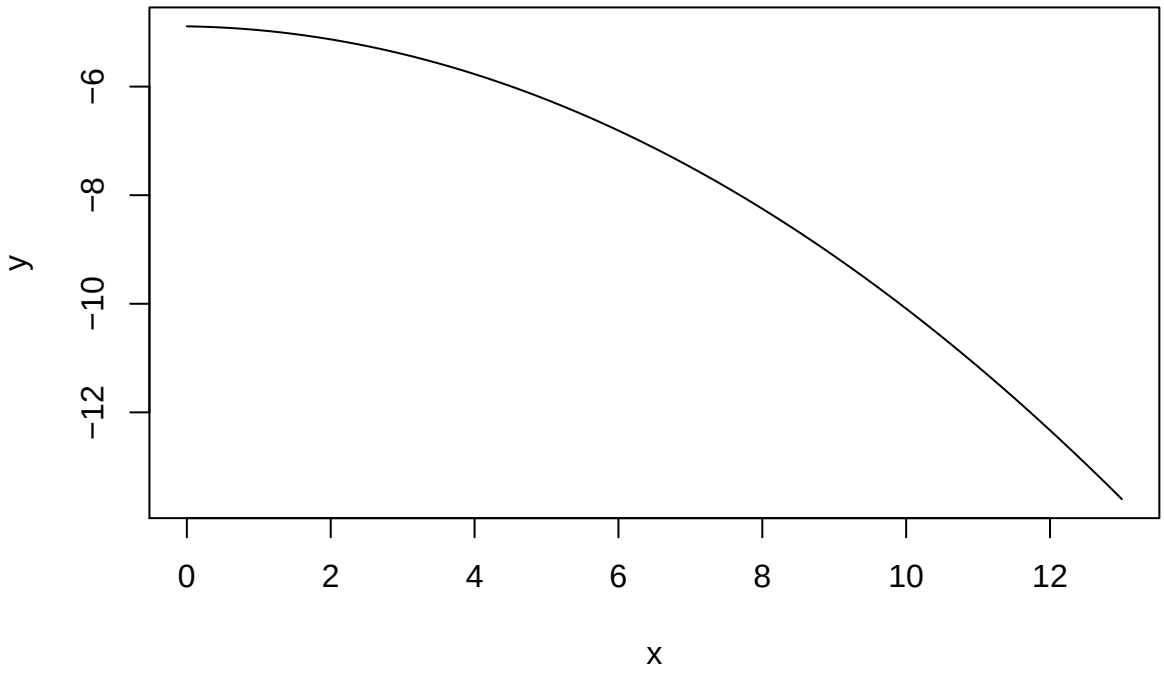


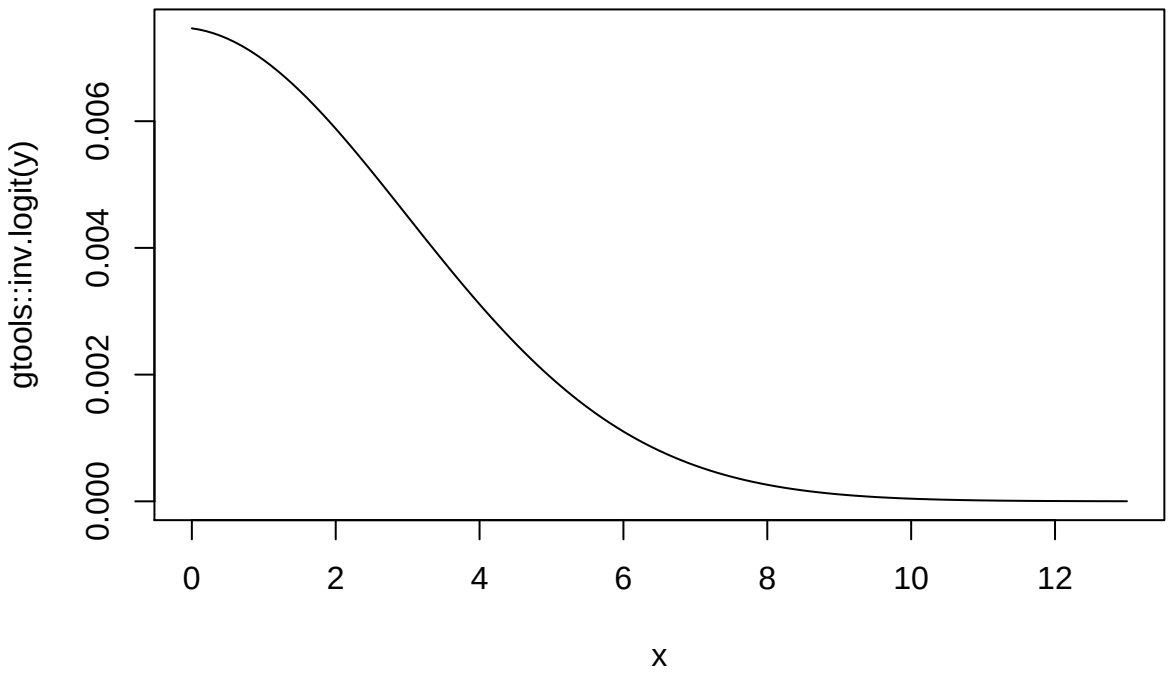


1.3 Adding a quadratic predictor of cumulative driving time square

```
##  Family: bernoulli
##   Links: mu = logit
## Formula: outlogit ~ cumDrive + travelTime + CTsquare + (1 + cumDrive + CTsquare | driverID)
##    Data: t50square (Number of observations: 32807)
## Samples: 4 chains, each with iter = 5000; warmup = 2000; thin = 1;
##          total post-warmup samples = 12000
##
## Group-Level Effects:
## ~driverID (Number of levels: 50)
##                          Estimate Est.Error l-95% CI u-95% CI Eff.Sample
## sd(Intercept)                0.90      0.12     0.68     1.16       3147
## sd(cumDrive)                 0.02      0.01     0.00     0.05       1878
## sd(CTsquare)                 0.01      0.00     0.00     0.01       1981
## cor(Intercept,cumDrive)      0.04      0.45    -0.80     0.86       9806
## cor(Intercept,CTsquare)     -0.02      0.37    -0.73     0.74       8324
## cor(cumDrive,CTsquare)       0.04      0.49    -0.86     0.88       2496
##                          Rhat
## sd(Intercept)            1.00
## sd(cumDrive)             1.00
## sd(CTsquare)             1.00
## cor(Intercept,cumDrive) 1.00
## cor(Intercept,CTsquare) 1.00
## cor(cumDrive,CTsquare)  1.00
##
```

```
## Population-Level Effects:
##            Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
## Intercept     -4.89      0.17    -5.23    -4.55       2472 1.00
## cumDrive      -0.02      0.01    -0.05     0.00       9603 1.00
## travelTime     0.70      0.05     0.60     0.80       5905 1.00
## CTsquare      -0.05      0.01    -0.06    -0.03       4790 1.00
##
## Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
## is a crude measure of effective sample size, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```




gtools::inv.logit(y)
0.000
0.002
0.004
0.006
0
2
4
6
8
10
12
x