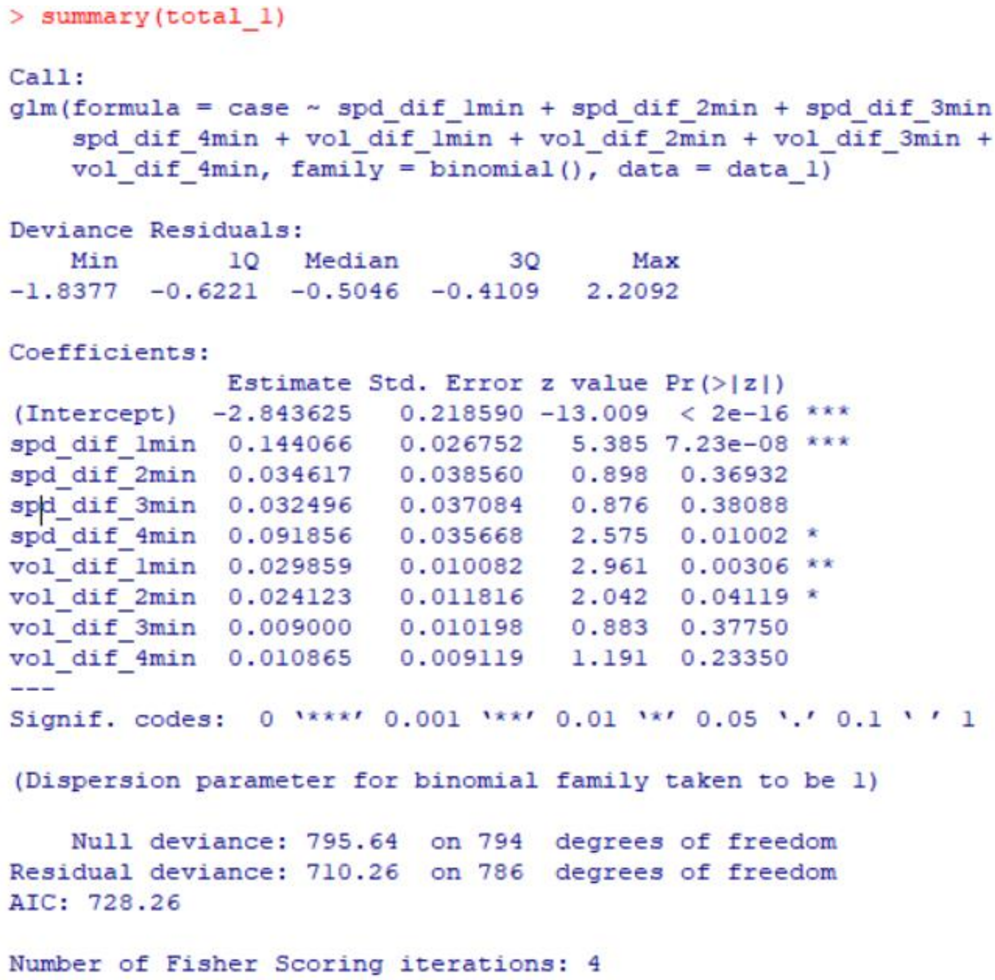
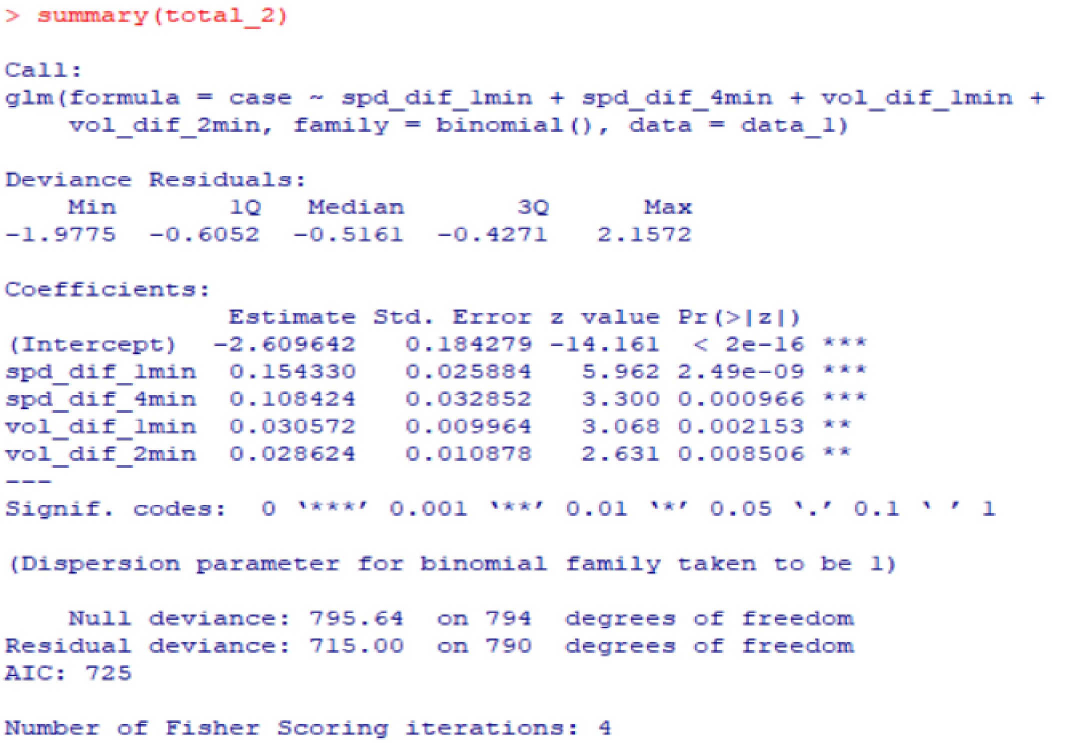
第一步建立事故是否发生与速度和流量之间的模型，由于事故是否发生为二值变量，因此采用逻辑回归模型，得到下图所示结果



```
> summary(total_1)

Call:
glm(formula = case ~ spd_dif_1min + spd_dif_2min + spd_dif_3min +
    spd_dif_4min + vol_dif_1min + vol_dif_2min + vol_dif_3min +
    vol_dif_4min, family = binomial(), data = data_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8377  -0.6221  -0.5046  -0.4109   2.2092

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.843625   0.218590 -13.009  < 2e-16 ***
spd_dif_1min  0.144066   0.026752   5.385 7.23e-08 ***
spd_dif_2min  0.034617   0.038560   0.898  0.36932
spd_dif_3min  0.032496   0.037084   0.876  0.38088
spd_dif_4min  0.091856   0.035668   2.575  0.01002 *
vol_dif_1min  0.029859   0.010082   2.961  0.00306 **
vol_dif_2min  0.024123   0.011816   2.042  0.04119 *
vol_dif_3min  0.009000   0.010198   0.883  0.37750
vol_dif_4min  0.010865   0.009119   1.191  0.23350
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 795.64  on 794  degrees of freedom
Residual deviance: 710.26  on 786  degrees of freedom
AIC: 728.26

Number of Fisher Scoring iterations: 4
```

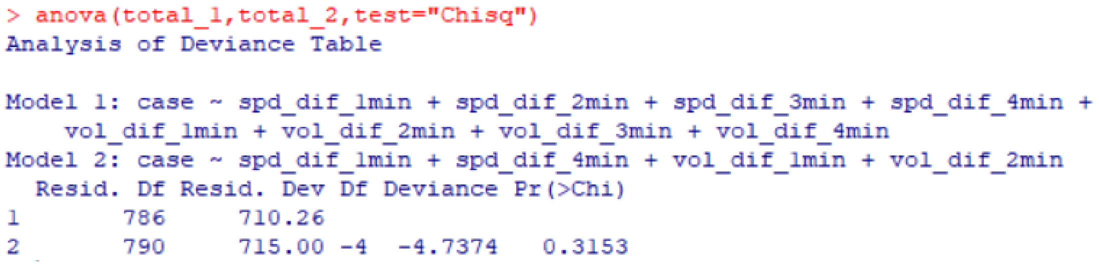
由上可以看出，回归系数比较显著（p<0.05）的是事故发生前 1 分钟内，4 分钟内的速度变化值和事故发生前 1 分钟内，2 分钟内的流量变化值。

因此去除不明显的，用显著变量建立新的逻辑回归模型如下



```
> summary(total_2)

Call:
glm(formula = case ~ spd_dif_1min + spd_dif_4min + vol_dif_1min +
    vol_dif_2min, family = binomial(), data = data_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9775  -0.6052  -0.5161  -0.4271   2.1572

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.609642   0.184279 -14.161  < 2e-16 ***
spd_dif_1min  0.154330   0.025884   5.962 2.49e-09 ***
spd_dif_4min  0.108424   0.032852   3.300 0.000966 ***
vol_dif_1min  0.030572   0.009964   3.068 0.002153 **
vol_dif_2min  0.028624   0.010878   2.631 0.008506 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 795.64  on 794  degrees of freedom
Residual deviance: 715.00  on 790  degrees of freedom
AIC: 725

Number of Fisher Scoring iterations: 4
```

从上面的结果可以看出，4 个参数的回归系数都非常显著，我们再用 anova()函数对两个模型进行比较，用卡方检验，可得



```
> anova(total_1,total_2,test="Chisq")
Analysis of Deviance Table

Model 1: case ~ spd_dif_1min + spd_dif_2min + spd_dif_3min + spd_dif_4min +
    vol_dif_1min + vol_dif_2min + vol_dif_3min + vol_dif_4min
Model 2: case ~ spd_dif_1min + spd_dif_4min + vol_dif_1min + vol_dif_2min
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       786     710.26
2       790     715.00 -4  -4.7374   0.3153
```

可以看出 p=0.3153,并不显著，因此用后面的四个变量模型和前面的拟合的一样好。所以就可以得到最佳模型如下

logit(p) = -2.6096+0.1543spd_dif_1min+0.1084spd_dif_4min+0.0306vol_dif_1min
+0.0286vol_dif_2min;　y~dbern(p).