Collision Data

Arden Miller

Executive Summary

The probability of a fatal collision was modelled using data collected from simulated automobile collisions. A logistic regression model that relates the probability of a fatal collision to the "age" of the crash-test dummy and velocity of the automobile at impact was identified. This model indicates that the probability of a fatal collision increases substantially as both age increases from 20 to 65 and as velocity increases from 40 to 60.

Crash Test Data

Data was collected for 58 simulated side impact automobile collisions using crash-test dummies. For each crash it was determined whether or not the injuries sustained by the crash-test dummy would have been fatal. Three aspects of the crashes were varied: the "age" of the crash-test dummy (Age), the maximum acceleration on impact measured on the dummy's abdomen (Acl), and the velocity of automobile at impact (Vel). The variable Age is meant to represent the age of a person involved in an accident and varies from 19 to 65 for this data. The variable Vel and Acl varied from 40 to 60 and Acl varied from 58 to 268. The purpose of the investigation was to investigate how the probability of the accident being fatal is related to these three factors.

A Logistic Regression Model

Logistic regression was used to model the probability that the crash is fatal, π, as a function of the explanatory variables. It was found that only one of Acl and Vel is required in the model. It appears that these variables are providing much the same information about π. The logistic model that contains Age and Vel is used since it fits the data slightly better than the model that contains Age and Acl. The fitted model for this data is:

$$\hat{\pi} = \frac{\exp(-16.98 + 0.1625 \times \mathsf{Age} + 0.2339 \times \mathsf{Vel})}{1 + \exp(-16.98 + 0.1625 \times \mathsf{Age} + 0.2339 \times \mathsf{Vel})}.$$

Table 1 contains the predicted probabilities of a fatal accident for different combination of

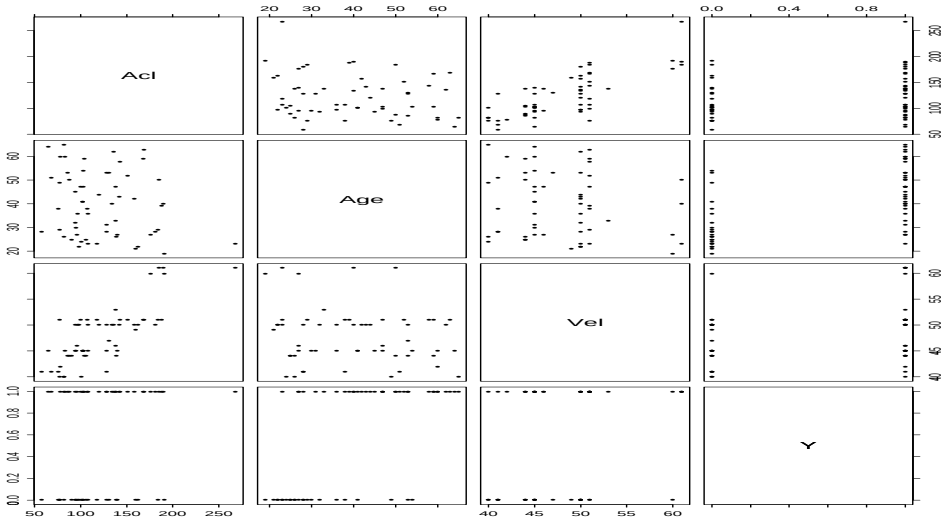

Figure 1: Pairwise Scatterplots of Collision Data.

Table 1: Estimated probabilities of a fatal accident

	Velocity		
Age	40	50	60
20	.012	.115	.574
	(.001, .129)	(.032, .341)	(.179, .893)
35	.126	.598	.939
	(.028, .417)	(.409, .762)	(.667, .992)
50	.623	.945	.994
	(.287, .870)	(.797, .987)	(.914, .9997)
65	.950	.995	.9995
	(.705, .993)	(.937, .9996)	(.978, .99999)

ages. The numbers in brackets represent 95% confidence intervals for the probabilities. Thus for Age = 20 and Vel = 40 the predicted probability is $\hat{\pi} = .012$ and we can be 95% certain that the true probability is between .001 and .129. From this table for any fixed value of velocity it is clear that the probability of a fatal accident increases dramatically as age increases. This is true even taking into account the uncertainty in the estimated For example at a velocity of 40, we can say with 95% confidence that the probability of a fatal accident is between .001 and .129 for age= 20 and is between .705 and .993 for age= 65. The probability of a fatal accident also increases markedly as velocity increases. Suppose age is fixed at 35 then the estimated probability increases from .126 for a velocity of 40 to .939 for a velocity of 60 – the 95% confidence interval goes from (.028, .417) to (.667, .992).

Statistical Appendix

First I tried a logistic model using all three regressors. The output from summary for this model is:

```
                  Value Std. Error   t value
(Intercept) -15.05358809 5.29463455 -2.843178
        Acl   0.01617745 0.01449042  1.116424
        Age   0.17090531 0.04320605  3.955588
        Vel   0.14627932 0.11215342  1.304279

    Null Deviance: 78.67229 on 57 degrees of freedom
Residual Deviance: 43.97593 on 54 degrees of freedom
```

I tested the hypothesis H_o:$\beta_{\text{Acl}} = \beta_{\text{Vel}} = \beta_{\text{Age}} = 0$. The test statistic is $\chi^2_o = 78.67 - 43.98 = 34.69$ which gives a very small p-value and indicates strong evidence that this model has predictive power. The large t value for Age indicates that it provides predictive power in addition to that provided by the other variables. However the "t values" for Acl and Vel are not significant (both are much less than 2) which suggests that we can drop one or both of these variables. I found that if either Acl or Vel is dropped then the remaining one becomes significant. I decided to use the model that contains Age and Vel as it has a smaller residual deviance than the model that contains Age and Acl (45.3 compared to 45.8). The output from summary for this model is:

```
                 Value Std. Error   t value
(Intercept) -16.9844796 5.14715050 -3.299783
        Age   0.1625007 0.04143451  3.921868
        Vel   0.2339056 0.08624799  2.712013

    Null Deviance: 78.67229 on 57 degrees of freedom
Residual Deviance: 45.33153 on 55 degrees of freedom
```

A χ^2 test was used to confirm that dropping Acl was sensible:

```
             Terms Resid. Df Resid. Dev Test Df  Deviance   Pr(Chi)
1 Acl + Age + Vel        54   43.97593
2       Age + Vel        55   45.33153 -Acl -1 -1.355599 0.2443017
```

Another χ^2 test was used to see whether Vel could be dropped as well:

```
      Terms Resid. Df Resid. Dev Test Df  Deviance     Pr(Chi)
1 Age + Vel        55   45.33153
2       Age        56   55.38127 -Vel -1 -10.04974 0.001523691
```

The small p-value indicates that Vel should be retained in the model.

The confidence intervals in Table 1 were found by creating 95% CI's for logit(π) and then using the logistic function to convert these into intervals for π.