

## Appendix II. Development of a Rollover Risk Model

In its study of our rating system for rollover resistance (Transportation Research Board Special Report 265), the National Academy of Sciences (NAS) recommended that we use logistic regression rather than linear regression for analysis of the relationship between rollover risk and SSF. We had considered a logistic regression model during the development of the rollover resistance rating system used by NCAP for 2001 to 2003 vehicles, but we observed that it predicted rollover rates that were systematically lower than actual rollover rates for vehicles with low SSF. Our first step was to explore the use of transformations of SSF to create a logistic regression model that better matched actual rollover rates while following the recommendation of the NAS.

A satisfactory logistic regression model using SSF only was the starting point for developing a risk model that used both a vehicle's SSF and its performance in dynamic maneuver tests to predict its rollover rate. We used four binary variables to describe whether or not the vehicle tipped up in two dynamic maneuver tests each performed at two different occupant load conditions. The final model required the results of only the Fishhook maneuver test with the heavy five occupant load and the SSF of a vehicle. The predicted rollover rate determines the rollover resistance rating of the vehicle.

### A. Improving the Fit of the Logistic Regression Model With SSF Only

We had considered logistic regression during the development of the SSF based rating system (66 FR 3393, January 12, 2001), but found that it consistently under-predicted the actual rollover rate at the low end of the SSF range where the rollover rates are high. The NAS study acknowledged this situation and gave the example of another analysis technique (non-parametric) that made higher rollover rate predictions at the low end of the SSF scale. In the NPRM, we discussed our plan to first examine ways to improve the fit of the logistic regression model to the actual rollover rates in the simpler model with SSF as the only vehicle attribute before expanding the logistic regression model to predict rollover rates using maneuver test results and SSF as vehicle attributes. In this way, the addition of maneuver test results is more likely to have an effect that reflects the additional information they represent on rollover causation.

A consultant to the Bureau of Transportation Statistics who lectured on logistic regression suggested that we use a transformation of SSF, like  $\text{Log}(\text{SSF})$ , rather than SSF alone to change the shape of the trend line generated by the logistic regression in our range of interest of SSF. This technique is similar to what we used to improve the fit of the linear regression model in the SSF rating system (Figure II.1). Linear regression creates a "best fit" straight line to predict the relationship between the independent variable, SSF in this case, and the dependent variable, rollover rate per single vehicle crash in this case. However,

the observations of rollover rate for groups of vehicles with a known SSF did not appear to lie on a straight line. The relationship appeared to be exponential with a reduction in rollover rate with increase in SSF much greater at low SSFs than at high SSFs. We used the transformation  $\text{Log}(\text{SSF})$  to replace SSF alone in the linear regression model so that it would compute a "best fit" exponential curve instead of a best fit straight line in order better fit the prediction line to the observations. We referred to Figure II.1 in notices 65 FR 34998 and 66 FR 3388 as a linear regression model because of the analysis technique, but the NAS study refers to it as the exponential model because of its curve shape.

Figure II.2 plots the actual rollover rates as a function of SSF observed for 293,000 single vehicle crashes involving 100 vehicle groups in six states from 1994 to 2001 (not all state's data available in every year). The point designated "actual rate" at each value of SSF gives the proportion of single vehicle crashes for vehicles of that SSF that resulted in rollover. For example, the leftmost point shows that for all single vehicle crashes observed for vehicles with an SSF of 1.00, slightly less than 50% resulted in rollover. There are fewer than 100 data points because the data at each SSF often include the crashes of several vehicles with the same SSF.

Figure II.2 also plots the rollover rates predicted for the same 293,000 crashes by a logistic regression model operating on SSF without transformation as the only vehicle variable. The model was developed from a database that contained the driver characteristic and road condition variables in the state crash reports of 293,000 crashes in six states. Data from Maryland, Florida, North Carolina, Missouri, Utah and Pennsylvania were used because these were the only states with electronic records available to NHTSA in which we could identify the make/model of the vehicle and could be sure whether or not a rollover occurred. The driver variables were gender, age [young (less than 25), old (70 or older), neither], and evidence of alcohol or drug use. The road condition variables were weather, speed limit, curve, hill, darkness, wet or icy surface, and potholes or other bad surface conditions. The SAS logistic regression program used these driver and road variables, the vehicle SSF, the State and the outcome (rollover or not) for each of 293,000 single vehicle crashes to compute the risk model. Figure II.2 shows the exercise of inputting the driver, road, state and vehicle SSF circumstances for each individual crash of the 293,000 back into the risk model to test how well the model can predict the actual rollover outcomes.

In similar fashion as the "actual rate" points on Figure II.2, the "predicted rate" points at each value of SSF give the proportion of single vehicle crashes for vehicles of that SSF that resulted in rollover. The number and circumstances (as well as can be described from state crash report variables) of crashes represented by the actual and predicted rate points are identical. However, in one case the rollover outcomes are the actual outcomes reported in the state

data. But in the other case, the rollover outcomes are the predictions of the risk model given the driver and road variables and vehicle SSF for each actual the crash. The predicted rate points do not lie on a continuous curve when plotted against SSF because the distribution of driver and road variables are different for the single vehicle crashes experienced by each group of vehicles represented by its SSF value.

Figure II.2 shows that the risk model obtained using the untransformed SSF computes predictions that match the actual rollover rates well at SSFs higher than 1.3, but its predictions are consistently low at the low end of the SSF range. The predictions also tend to be too high in the 1.15 to 1.25 SSF range. For this reason we described the form of the curve inherent to the logistic regression computation as being too flat or lacking sufficient curvature to represent rollover risk in our past notices.

Figure II.2 also lists an objective measure of the goodness of fit of the predictions to aid in the comparisons of models with and without using transformations of SSF. It is the  $R^2$  value for linear regression between the predicted and actual rollover rates. Figure II.3 is a plot of predicted versus actual rollover rates taken from Figure II.2. It shows how the  $R^2$  value was obtained. A linear regression of the form " $y = mx$ " computes the best fit line that passes through the origin. The  $R^2$  value that describes the goodness of fit of the points to the line " $y = 0.9673x$ " is 0.752. A perfect set of predictions would cause an  $R^2$  value of 1.0 on the line " $y = 1.0x$ ".

Figures II.4, II.5, and II.6 show the predictions of a series of risk models obtained in the same way as that shown in Figure II.2 except that transformations of SSF were used as the vehicle variable instead of just SSF. The first transformation, shown in Figure II.4, was  $\text{Log}(\text{SSF})$ . This is the transformation currently used in the linear regression rollover risk model. It makes a very small improvement both to the under-predictions at the low end of the SSF range and the over-predictions in the 1.15 to 1.25 SSF range. The  $R^2$  goodness of fit indicator increased to 0.7975.

Next we tried the transformation  $\text{Log}(\text{SSF} - \text{margin})$ . Figure II.5 shows the predictions of a logistic regression model with a margin of 0.85. The subtraction of a margin from SSF makes a large improvement in the fit of the predicted rollover rates to the actual rollover rates in the SSF range of 1.0 to 1.25. The  $R^2$  goodness of fit indicator increased to 0.8811 about the line " $y = 1.0011x$ " for the whole SSF range of data base (1.0 to 1.53). This transformation caused a small sacrifice in the fit of the model at the high end of the SSF range. However, a good fit in the 1.0 to 1.25 SSF range is more important to a rating system because most of the consumer requests for rollover information involve vehicles in this range.

Figure II.6 shows the fit of the model with a margin of 0.9. The  $R^2$  goodness of fit indicator increased slightly to 0.8948 about the line " $y = 1.0091x$ ", but the sacrifice of fit at the high SSF end also increased. Figure II.7 is a plot of predicted versus actual rollover rates taken from Figure II.6. The use