A bivariate model for analyzing recurrent multi-type automobile failures

A A Sunethra and M R Sooriyarachchi

Department of Statistics, University of Colombo, Colombo 3, Sri Lanka
E-mail: sunethra@stat.cmb.ac.lk

Abstract. The failure mechanism in an automobile can be defined as a system of multi-type recurrent failures where failures can occur due to various multi-type failure modes and these failures are repetitive such that more than one failure can occur from each failure mode. In analysing such automobile failures, both the time and type of the failure serve as response variables. However, these two response variables are highly correlated with each other since the timing of failures has an association with the mode of the failure. When there are more than one correlated response variables, the fitting of a multivariate model is more preferable than separate univariate models. Therefore, a bivariate model of time and type of failure becomes appealing for such automobile failure data. When there are multiple failure observations pertaining to a single automobile, such data cannot be treated as independent data because failure instances of a single automobile are correlated with each other while failures among different automobiles can be treated as independent. Therefore, this study proposes a bivariate model consisting time and type of failure as responses adjusted for correlated data. The proposed model was formulated following the approaches of shared parameter models and random effects models for joining the responses and for representing the correlated data respectively. The proposed model is applied to a sample of automobile failures with three types of failure modes and up to five failure recurrences. The parametric distributions that were suitable for the two responses of time to failure and type of failure were Weibull distribution and multinomial distribution respectively. The proposed bivariate model was programmed in SAS Procedure Proc NLMIXED by user programming appropriate likelihood functions. The performance of the bivariate model was compared with separate univariate models fitted for the two responses and it was identified that better performance is secured by the bivariate model. The proposed model can be used to determine the time and type of failure that would occur in the automobiles considered here.

1. Introduction
Transportation is a key consideration in any country’s development where transportation by land is more common than other methods of transportation. Automobiles such as lorries, buses, cars are major aspects of land transportation and as of now the automobile industry is highly competitive and diversified. A keen interest is paved both by the manufactures and by the users on the quality and the reliability of automobiles. An unexpected, sudden breakdown in an automobile which can be catastrophic to the extent of fatal road accidents, damages both the customers’ satisfaction and the credibility of the manufacturer. A comprehensive analysis on the failures experienced by an automobile can reveal the most prominent failure types and plausible timings of such failures. When considering the failure mechanism of automobiles, it is of high complexity where it can be a total breakdown or a partial breakdown owing to one or many of the possible failure modes where most common failures are caused...
by defects in engine, battery or wheels, due to a mechanical or an electrical problem, due to a problem in clutch, ignition, cooling, heating, ventilation, and etc. It is noteworthy that these various multiple failure modes are repetitive where once a failure is detected and repaired, it is possible to experience the same failure a multiple times. In a statistical viewpoint, this failure mechanism of an automobile can be defined as a system consisting of recurrent multi-type failures. Therefore, in analyzing the failure data on automobiles this complete process of repeated multi-type failures should be taken into account.

This complex structure in the data possess a great challenge to the analyst mainly due to the non-independence of the data where repetitive failures of an automobile cannot be treated as independent with each other. But, the assumption of independence is a very common assumption in majority of the statistical methods and models. The multiple failure occurrences pertaining to the same vehicle are to be treated as non-independent observations, whereas observations from two different vehicles can be treated as independent observations [1]. Hence, a proper analysis of recurrent failure data should be rich enough to adjust for this correlated structure in the data. Moreover, analyzing only the timing of the failure wouldn’t indicate what type of a failure occurred at the respective timing. Therefore, knowing both the type of the failure and the time of the failure is of higher importance [2]. This postulate the presence of two response variables, namely the type of failure and time of failure as the response variables/dependent variables in fitting models to the data. The nature of these response variables is such that these are highly correlated with each other that is the type of the failure and time of failure are associated with each other. When there is more than one response variable that are associated with each other, fitting a multivariate model/joint model for the two or more responses within a single statistical model is more preferable than fitting separate models for each of the responses. Therefore, this study proposes a bivariate model for time to failure and type of failure by means of analyzing failures experienced by a sample of Sport Utility Vehicles (SUVs) manufactured by a leading automobile manufacturer. The proposed model is well adjusted for the non-independence of the observations in the data.

2. Data

The data for this study was collected from a leading automobile company in Sri Lanka having three types of Sport Utility Vehicles (SUVs) denoted by A, K, and R as original brand names would not be disclosed due to confidentiality of the data. There were 2532 SUVs sold between March 2009 to April 2011 and the failures for these vehicles were reported since March 2009 to May 2011. The mileage on each failure was taken as the time to failure where data consisted of repetitive mileages of failures and the study was limited to a maximum of five failure recurrences and three most common failure modes were considered which are as follows:

- Mode 1: wheels, tyres and vehicle alignment
- Mode 2: brakes-hydraulics, regulator, and servo
- Mode 3: engine-craft shaft, pistons

Thus, SUVs considered in this study constitutes to a maximum of 15 observations per each SUV if five recurrences from all the three failure modes were experienced. The failure instances that were not experienced by an SUV was considered as censored failure times in the last follow up time of the each vehicle. In general, if a vehicle has observed the rth recurrence of a particular failure mode, it will be censored for r+1th recurrence of that failure mode. The SUVs that have not had any failure during the study period will be censored for the first failure instance of all the three failure modes. This data is being previously analyzed by [1] where only a single response of time to failure was considered whereas this study considers both the time and type of failure as response variables in a bivariate model fitted to the data.

3. Descriptive analysis

With respect to the type of the SUV, the majority were of type K (1291, 50.9%) while 750 vehicles were of type R (29.6%) and 491 were of type A (19.4%). Although, a few vehicles have had up to a maximum of 14 recurrent failures, the study was limited to a maximum of 5 recurrences to secure a balance data...
set across each of the recurrences. It is noteworthy that only about 5% of the sold vehicles didn’t have any failure during the reporting period. There were 6013 total failure occurrences out of which 33% were from Failure Mode 1, 32% was from Mode 2 and 35% was from Mode 3.

Figure 1 shows the Kaplan-Meier survival curves drawn against vehicle type and failure mode respectively. Log-rank tests which were carried out to check whether the failure curves differ significantly among the types of SUVs, among failure modes indicated that failure timings are significantly different with respect to the type of SUV and type of failure. A Pearson Chi-squared test carried out to test whether the mode of the failure and type of the vehicle is significantly associated with each other indicated a significant association.

![Figure 1](image-url)  
**Figure 1.** Kaplan Meier curves for time to failure with respect to (a) type of SUV (b) Mode of failure.

In summary, the descriptive analysis revealed that time and type of failure is significantly associated with the type of the vehicle and timing of failure is associated with respective mode of failure. This manifested the need for a bivariate model/joint model where time and type of failure were to be the responses of the bivariate model.

4. Advanced analysis

The advanced analysis of the study establishes the main objective of this study by developing a joint model for time and type of failures which is adjusted for non-independent data. Among several approaches for developing joint models, class of shared parameter models constitute an approach of joining two or more responses by using shared random effects to join the responses [3]. Following this approach, the time and the type of the failure will be joined via a shared random effect assumed for each pair of observation. Similarly, the class of mixed models constitutes models for analyzing non-independent data where correlation in the data is represented by the use of suitable random effects which follows certain probability distributions. In line with the automobile data, observations within a single automobile are correlated and will be represented as a random effect in the joint model fitted to the data.

In summary, this study proposes a joint model for time and the type of failure under the lines of shared parameter models where marginal models for the two responses are of mixed effects models which are adjusted for the non-independence in the data.

4.1. The bivariate model for time to failure and type of failure

Upon identifying the suitable parametric distributions for the response variables of time to failure and type of failure as Weibull distribution and multinomial distribution respectively, the marginal model for the two responses were derived as follows.

Assume that there are $i = 1, 2, \ldots, k$ units/SUVs each having $n_i$ failure recurrences which are correlated within the $i^{th}$ unit. Then, the total number of observations is $N = \sum_{i=1}^{k} n_i$. Let $y_{ij}$ be the mode
of the $j^{th}$ failure of the $i^{th}$ vehicle. Then the mixed effects multinomial logistic regression model for the probability that $y_{ij}=c$ (for $c=1,2,3$) for a given vehicle $i$ conditioned on the random effects $u$ can be written as:

$$
\Pr(y_{ij}=c|u,v) = \frac{\exp(Z_{ijc})}{\sum_{l=1}^{3} \exp(Z_{ijl})} \quad \text{for } l=1,2,3
$$

Where $z_{ijc}=w_{ij} \beta + x_{ij} u + h_i v_i$. Here, $w_{ij}$ is a $p \times 1$ covariate vector, $x_{ij}$ is a design vector for the $N$ random effects used to join the two responses and $h_i$ is the design vector for the $k$ random effect which represents the non-independent data clusters. Similarly, $\beta$ is a $p \times 1$ vector of unknown regression parameters and $u_i$ is a $N \times 1$ vector of unknown random effects which are assumed to follow a normal density with zero mean and variance $\sigma_u^2$. Here, $v_i$ corresponds to a $k \times 1$ vector of unknown random effects which are assumed to follow a normal density to zero mean and variance covariance matrix $\sigma_v^2$ [4].

Correspondingly, the mixed effect Weibull regression model for time to failure can be explained as follows: Let $t_{ij}$ denote the time of the $j^{th}$ failure of the $i^{th}$ vehicle. The probability density function of the mixed effect Weibull regression model is as follows:

$$
f(t_{ij}|u,v) = \frac{1}{\sigma} \exp\left(\frac{t_{ij} - \mu_i}{\sigma}\right) \exp\left(-\exp\left(\frac{t_{ij} - \mu_i}{\sigma}\right)\right)
$$

The corresponding survival function for the Weibull distribution takes the form:

$$
S(t_{ij}|u_i) = \exp\left(-\exp\left(\frac{t_{ij} - \mu_i}{\sigma}\right)\right)
$$

Equation 2 gives the probabilities for uncensored failures and equation 3 gives the probabilities for censored failures. Here, $\mu_i = g_{ij} \alpha + x_{ij} u + h_i v_i$. Where $g_{ij}$ is a $p \times 1$ covariate vector and $x_{ij}$ is a design vector for the $r$ random effects and $\alpha$ is a $p \times 1$ vector of unknown regression parameters. The definition of other terms is similar to the definitions given above for the logistic mixed model.

In line with this automobile dataset, the $u_{ij}$ represents the observational level random effects used for joining the two responses at each failure and $v_i$ corresponds to random effects defined at automobile level to represent the correlated data. For both models, for simplicity’s sake, it is assumed that $u_{ij}$ and $v_i$ are uncorrelated. Therefore, this joint model assumes the two sub models for time and type of failure are independent given the random effects:

$$
a = \begin{pmatrix} u \\ v \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_v^2 \end{pmatrix}\right)
$$

Then, the likelihood of the proposed joint model can be written as:

$$
L(T, Y; \theta) = \prod_{i=1}^{k} \prod_{j=1}^{n_i} f(T_{ij}, Y_{ij}|\theta) = \prod_{i=1}^{k} \prod_{j=1}^{n_i} \left[f(t_{ij}|a)\right]^{\delta_{ij}} \left[S(t_{ij}|a)\right]^{1-\delta_{ij}} f(y_{ij}|a) f(a) \, da
$$

$$
= \int \prod_{i=1}^{k} \prod_{j=1}^{n_i} \left\{ \left[\frac{1}{\sigma} \exp\left(\frac{t_{ij} - \mu_i}{\sigma}\right) \exp\left(-\exp\left(\frac{t_{ij} - \mu_i}{\sigma}\right)\right)\right]^{\delta_{ij}} \right\}
\exp\left(-\exp\left(\frac{t_{ij} - \mu_i}{\sigma}\right)\right) \sum_{\delta=1}^{3} \frac{1}{\sqrt{(2\sigma^2)^{\delta}}} \exp\left(\frac{1}{2} \sum_{\delta=1}^{3} \delta_{ij} \Sigma_{a_i}\right) \, da
$$

4
This model was fitted in SAS procedure Proc NLMIXED by user programming the above joint likelihood as per the two marginal likelihoods. The adaptive Gaussian Quadrature method was used for integral approximation.

4.2. Results
This section presents the results of the statistical models fitted to SUV data. Initially, two separate univariate models for the two responses were fitted, namely a Weibull mixed model and a multinomial logistic mixed model. Both of these models indicated that type of the vehicle is significantly associated with the type and time of failure and the failure recurrence was also used as a covariate in the two models to estimate the time and type of failure at each recurrence were also significant with both of the responses.

Then, the proposed joint model was fitted to the data and the results indicated that type of SUV and recurrence as significant factors to both the responses. It is of great importance to compare the performance of the joint model with the two separate models to gauge the necessity and performance of the joint model over separate models for which the summation of the model fit statistics of the two separate models will be compared with the model fit statistics of the joint model.

| Model fit statistics | Joint model | Separate models | Total |
|----------------------|-------------|-----------------|-------|
| -2Log Likelihood     | 172214      | 142577          | 172225|
| AIC                  | 172274      | 142599          | 172285|
| AICC                 | 1722274     | 142599          | 172285|
| BIC                  | 172449      | 142663          | 172460|

As per the model fit statistics, it can be seen that joint model has achieved lower AIC, AICC and BIC than the sum of those values obtained for the separate models. For example, the sum of AICs of the separate models were 172285(142599+29686) which is higher than the joint model’s AIC of 1722274.

In addition, -2 log L is decreased by 11 for 1 degree of freedom, which is significant 5% level of significance. Upon confirming the superiority of the joint model, table 2 presents the results of the joint model. More importantly, it was identified that the variances of the two random effects were highly significant at 1% level of significance with \( \sigma^2_u = 0.1 \) and \( \sigma^2_v = 0.05 \) which confirms the suitability of the proposed random effects structure.

As per the joint model, the two marginal models for time to failure and type of failure can be written as follows:

\[
\log(t_p) = 11.06 - 0.09(Veh = A) -0.02(Veh = K) + .15(Veh = R) -.15(event = 1) -.003(event = 2) - .44(event = 3) -.061(event t = 4)-.53(event = 5) + 1.2 \log(-\log(1-P))
\]  

(6)

\[
Pr(\text{Mode} = 1) = \exp[1.11 - 0.14(Veh = A) -0.13(Veh = K) + .003(Veh = R) -1.17(event=1) -1.21(event t = 2) -.86(event = 3) -.42(event = 4) -.001(event = 5)]
\]  

(7)

\[
Pr(\text{Mode} = 2) = \exp[.43 -0.01(Veh = A) -.03(Veh = K) + .01(Veh = R) -.64(event = 1) -.41(event = 2) -.24(event = 3) -.14(event = 4) -.00004(event = 5)]
\]  

(8)

From the coefficients of the joint model it can be seen that SUV type R performs significantly better than SUV type A but SUV type K is not significantly different. It is to be noted that mode 3 was taken as the base level in the multinomial model fitted for the type of the failure. Let’s consider the median failure time for the first failure of the SUV type A will be \( t_5 = \exp(11.06 - 0.05 -1.12*(\log(-\log(0.5)))) \) 108440.5 kms. Then, let’s consider the mode of the first failure of the SUV type A. The probability that
the first failure of the Vehicle A will be of Mode 1 is \( \Pr(\text{Mode}=1) = \exp(1.11 - 0.14 - 1.17) = 0.82 \) 82% higher than it will be of mode 3 and the corresponding probability that it will be of Mode 2 is \( \Pr(\text{Mode}=2) = \exp(0.429 - 0.01 - 0.64) = 0.80 \) 80% higher than it will be of Mode 3.

Table 2. Results of the joint model.

| Parameter     | Estimate | Std. Error | Parameter     | Estimate | Std. Error | Parameter     | Estimate | Std. Error |
|---------------|----------|------------|---------------|----------|------------|---------------|----------|------------|
| b0            | 11.06    | 0.01       | m1            | 1.11     | 0.02       | m1            | 0.43     | 0.02       |
| Vehicle A     | -0.09    | 0.02       | Vehicle A     | -0.14    | 0.04       | Vehicle A     | -0.01    | 0.04       |
| Vehicle K     | -0.02    | 0.02       | Vehicle K     | -0.13    | 0.03       | Vehicle K     | -0.03    | 0.03       |
| Vehicle R     | 0.15     | 0.02       | Vehicle R     | 0.003    | 0.03       | Vehicle R     | 0.01     | 0.03       |
| Event 1       | -0.15    | 0.03       | Event 1       | -1.17    | 0.04       | Event 1       | -0.64    | 0.04       |
| Event 2       | -0.003   | 0.02       | Event 2       | -1.21    | 0.04       | Event 2       | -0.41    | 0.04       |
| Event 3       | -0.44    | 0.02       | Event 3       | -0.86    | 0.05       | Event 3       | -0.24    | 0.05       |
| Event 4       | -0.61    | 0.03       | Event 4       | -0.42    | 0.07       | Event 4       | -0.14    | 0.08       |
| Event 5       | -0.53    | 0.04       | Event 5       | 0.001    | 0.10       | Event 5       | 0.00004  | 0.11       |
| gamma         | 1.12     | 0.02       |               |          |            |               |          |            |

5. Discussion and conclusion
The main objective of this study was to propose a suitable joint model for analyzing some automobile failure data where both time and type of failure were the response variables of interest. The proposed model was programmed in SAS procedure Proc NLMIXED. Since the two distributions, namely Weibull and Multinomial were not supported by the Proc NLMIXED procedure, their likelihood functions were programmed to fit the proposed model. A bivariate normal density was used as the distribution of the two random effects which is suitable for the nested random effects found in this data. Since, the covariance between the random effects were constrained to zero for simplicity a bivariate Normal distribution with zero covariance between the random effects were fitted. Comparing the performance of the model specifying two separate random effect distributions and using other parametric distributions in addition to the Normal distribution can be considered as further work. But, it is noteworthy that the joint model fitted with the bivariate normal density with zero covariance outperformed the two separate univariate models. Developing methods for model adequacy testing for the joint models constitutes future work.

References
[1] Amarasinghe N S, Sunethra A A and Sooriyarachchi M R 2013 Jurnal Teknologi 63(2) 65-70
[2] Stamatis D H 2003 Failure mode and effect analysis: FMEA from theory to execution (Milwaukee: ASQ Quality Press)
[3] Rizopoulos D 2012 Joint Models for Longitudinal and Time-to - event data with Applications in R (Boca Raton:CRC Press)
[4] Kuss O and McLerran D 2007 A note on the estimation of the multinomial logistic model with correlated responses in SAS Computer methods and programs in biomedicine 87(3) 262-269