2001

# Heckman's methodology for correcting selectivity bias : An application to road crash costs

Margaret Giles

# Heckman's Methodology for Correcting Selectivity Bias:

# An Application to Road Crash Costs

By

Margaret J. Giles

The Sellenger Centre
Edith Cowan University

School of Finance and Business Economics Working Paper Series
September 2001
Working Paper 01.11
ISSN: 1323-9244

Correspondence author and address:

Margaret J. Giles
The Sellenger Centre
Edith Cowan University
100 Joondalup Drive
Joondalup WA 6027
Phone: 61 (08) 9400 5423
Fax: 61 (08) 9400 5894
Email: m.giles@ecu.edu.au

**Abstract**

Aggregate road crash costs are traditionally determined using average costs applied to incidence figures found in Police-notified crash data. Such data only comprise a non-random sample of the true population of road crashes, the bias being due to the existence of crashes that are not notified to the Police. The traditional approach is to label the Police-notified sample as 'non-random' thereby casting a cloud over data analyses using this sample. Heckman however viewed similar problems as 'omitted variables' problems in that the exclusion of some observations in a systematic manner (so-called selectivity bias) has inadvertently introduced the need for an additional regressor in the least squares procedures. Using Heckman's methodology for correcting for this selectivity bias, Police-notified crash data for Western Australia in 1987/88 is reconciled with total (notified and not notified) crash data in the estimation of the property damage costs of road crashes.

Acknowledgments: The author wishes to thank Professor Paul Miller, The University of Western Australia, for support and valuable comments, as well as session participants at the presentation of an earlier version of this paper at the School of Finance and Business Economics Seminar Series, Edith Cowan University, Western Australia, February 23, 2001.

## 1. Introduction

One of the problems for road crash researchers in the past has been that Police-notified crash data do not include all road crashes. These data not only understate the incidence of road crashes and the damage they cause, by virtue of legislative reporting requirements and the crash characteristics themselves, but they understate this in a non-random way. The implications for statistical inference and prediction from not using randomly sampled data are grave. Berk (1983, p. 386) refers to external validity being threatened and internal validity being made vulnerable in the presence of such biased samples.

In Section 2, examples of the problem of non-random samples are explored. In Section 3, Heckman's solution to this problem is described in relation to the effect of the under-reporting of road crashes on estimates of the property damage costs of road crashes. Section 4 reports on cost models for unreported and reported road crashes, with and without Heckman's correction for selectivity bias. Section 5 discusses the implications of the findings.

## 2. Determination of the General Problem of Selection Bias

A standard approach in applied research attempts to "formulate a (linear regression) model that describes the underlying structure of the behavior of ... variables", Ramanathan (1989, p. 81). The model is specified as $y_i = x_i\beta + u_i$ where $y_i$ is the dependent variable, $\beta$ is the column vector of coefficients, $x_i$ is the row vector of values of the explanatory or independent variables, and $u_i$ is a residual that captures unobserved influences on the dependent variable.[1] Under standard assumptions, the estimates of the vector of coefficients are unbiased.

If a population suffers from selectivity bias then the regression analysis, for example Ordinary Least Squares (OLS), which computes the effects of some characteristics of this population on other characteristics, will be biased. Goldberger (1981) dates the discovery of this bias back to 1903 and Karl Pearson's mathematical examination of the theory of evolution. Pearson had maintained that natural selection modified coefficients of selected body organs when regressed on non-selected organs for the purpose of determining the effect of the formers' size on the latter, Goldberger (1981, p.19-20).

Half a century later, Lush and Shrode (1950) espoused the same conclusions when examining milk production in dairy herds. They found that culling of older, less productive cows was common. Hence they concluded that "if the regression of milk production on age is computed ..... that curve will not show the effects of age alone but will show those effects combined with whatever effects .... culling actually had" (page 338).

In the 1970s, empirical analyses of female labour supply and earnings implicated selectivity bias as having a distorting influence on the estimates and attempted to correct for it, Amemiya (1973), Cain (1973), Heckman (1976, 1979, 1980), Hausman (1977), Crawford (1979). The hypotheses therein focused on the estimation of wage equations based on samples of working women. The results in this instance were biased "because the same sets of variables that determine wages enter in as a criterion for sample eligibility. The estimated wage function confounds the true behavioral wage function with the rules for sample inclusion", Smith (1980, p. 7). For example, the effect of education on female wages is confounded by the existence of a reservation wage at which women would enter the labor market. At low levels of education, only those women at the top of the wage offer distribution will be observed in the sample of working women. At higher levels of education, virtually all women may be observed in the samples. The wage estimates at each level of education are thus biased upwards and the returns to education are under-estimated, Smith (1980, p. 8).

A model that corrects for the sample selectivity outlined above has been developed by Heckman. In the most generally used version of the correction for sample selectivity, the use of non-random samples in the estimation of behavioral relationships is treated as an omitted variables problem. As noted by Berk (1983, p. 388), Heckman has showed that "by excluding some observations in a systematic manner, one has inadvertently introduced the need for an additional regressor that the usual least squares procedures ignore".

## 3. Describing Selection Bias in Relation to the Property Damage Costs of Road Crashes

Heckman (1980) suggested that, when estimating a model based on a sample of data that is generated from a wider population, an equation describing the probability for selection into that database must be developed. This then allows for the construction of a sample

selection term for inclusion in the equation of primary interest, such as the estimated female wage function discussed in Section 2. The inclusion of this sample selection term in the estimating equation solves the omitted variables problem that Heckman has shown to be equivalent to the use of non-random samples. Similarly, in this Section it is argued that generating road crash costs based on reported crashes alone needs to be tempered by the probability of reporting such crashes.

Heckman (1980) considered a two-equation model, which is rewritten here in terms of the probability of reporting a road crash, *PN*, and the property damage costs of a road crash, *PDC*.

$$y_{1i}^{*} = x_{1i}\beta_1 + u_{1i} \tag{1a}$$

$$PN = P(y_{1i}^{*} > 0) = P(y_{1i} = 1) = F(x_{1i}\beta_1) \tag{1b}$$

$$PDC = y_{2i} = x_{2i}\beta_2 + u_{2i} \tag{2}$$

where:

$y_{1i}^{*}$ is the underlying (unobserved) propensities of reporting road crashes to the Police;

$y_{1i}$ is an observed indicator variable with $y_{1i} = 1$ if $y_{1i}^{*} > 0$, that is, if the crash is reported to the Police;

$y_{2i}$ is the natural logarithm of the total costs of road crashes;

$x_{1i}$ is the row vector of variables affecting reporting;

$\beta_1$ is the column vector of coefficients;

$x_{2i}$ is the row vector of variables affecting crash costs;

$\beta_2$ is the column vector of coefficients;

$u_{1i}$ are error terms with $E(u_{1i}) = 0$;

$u_{2i}$ are error terms with $E(u_{2i}) = 0$.

It is likely that some factors affecting the cost of crashes will also affect the reporting of crashes to the Police. That is, elements of $x_{2i}$ will also be found in $x_{1i}$. The specification of these vectors will be discussed later.

Berk (1983, p. 390) refers to (2) as the substantive equation that is of particular interest to the researcher, and (1b) as the selection equation that determines whether particular observations will be in the sample used to estimate equation (2).

The probability of a crash being reported is not observed. Instead, all that is observed is a binary indicator $y_{1i}$, which can be coded to one if a crash is reported to the Police and to zero in cases where crashes are not reported to the Police. As a selection rule, researchers using Police crash data will have access to observations where $y_{1i} = 1$. They will not have observations where $y_{1i} = 0$.

As (2) can only usually be estimated using a non-random sample of crash data, such that $y_{1i}^{*} \geq 0$, the derived parameters, $\beta_2$, may be biased and inconsistent. The substantive equation (2) can be rewritten in terms of expectations as

$$E(y_{2i} / x_{2i}, y_{1i}^{*} \geq 0) = x_{2i}\beta_2 + E(u_{2i} / u_{1i} \geq -x_{1i}\beta_1) \quad (3)$$

There is no guarantee that $E(u_{2i} / u_{1i} \geq -x_{1i}\beta_1)$ equals zero, hence there will be bias in situations where proper account is not taken of the sample selection rule. "Thus the problem of sample selection bias, initially viewed as a missing dependent variable problem, may be reformulated as an ordinary omitted explanatory variable problem", Heckman (1980, p. 210).

Now, the second term on the RHS in (3) can be rewritten as

$$E(u_{2i} / u_{1i} \geq -x_{1i}\beta_1) = [\sigma_{21} / \sqrt{\sigma_{11}}]\lambda_i \quad (4)$$

where:

$\lambda_i = \dfrac{f(Z_i)}{1 - F(Z_i)} \geq 0$, which is the ratio of the height of the density to the right tail area of the standard normal distribution (the inverse Mill's ratio – see Winship (1992, p. 340)). This is the hazard rate "which represents for each observation the instantaneous probability of being excluded from the sample conditional upon being in the pool at risk...The larger the hazard rate, the greater the likelihood that the observation will be discarded", Berk (1983, p. 390). $\lambda_i$ has a number of characteristics including $\delta\lambda_i / \delta Z_i > 0$, $\lim(Z_i \to -\infty)\lambda_i = 0$, $\lim(Z_i \to \infty)\lambda = \infty$, is a monotonic increasing function of $Z_i$, and is a monotonic decreasing function of $\{1 - F(Z_i)\}$;

$\sigma_{21} / \sqrt{\sigma_{11}}$ is the ratio of the correlation between the errors in the selection (probability of reporting) and the substantive (cost) equations and the standard deviation of the reporting error, adapted from Smith (1980, p. 13);

$Z_i = [-x_{1i}\beta_1]/\sigma_{11}$ is the negative of the predicted value from (1b);

$f(Z_i)$ is the density function of $Z_i$;

$F(Z_i)$ is the distribution function of $Z_i$; and

$1 - F(Z_i)$ is the probability that a population observation with characteristics $x_{1i}$ is selected into the observed sample.

If $\lambda_i$ is zero (that is, no observations are omitted from the sample), then $E(y_{2i} / x_{2i}, y_{1i}^* \geq 0) = x_{2i}\beta_2$ and $\hat{\beta}_2$ are unbiased least squares estimators of $\beta_2$. In the case of the property damage costs of that sample of road crashes that are reported to the Police ($y_{1i}^* \geq 0$), it is likely that $\lambda_i \neq 0$. Due to non-random influences on the probability of reporting, there are two possibilities. Firstly, if $\sigma_{21} / \sqrt{\sigma_{11}} > 0$, then it implies that for a given characteristic in equation (1b), large positive errors in (1b) are associated with large positive errors in (2). That is, there is an unmeasured variable (or variables) that results in a crash being both more likely to be reported and relatively more costly. Conservative, risk adverse behavior would most likely reveal itself in this way. Secondly, if $\sigma_{21} / \sqrt{\sigma_{11}} < 0$, then the under-reporting of crashes results in relatively high cost crashes being excluded from the sample. This might occur where, for example, a person with a poor driving record is likely to be averse to reporting a crash to the Police and is relatively more likely to have a high cost crash.

## 4. Selectivity Corrected Estimates of the Property Damage Costs of Road Crashes

### *4.1. The Property Damage Database (PDD)*

In 1989, the Road Accident Prevention Research Unit (Roadwatch) at the University of Western Australia, together with the Australian Road Research Board (ARRB) in Melbourne, funded a project to collect road crash data from 1987/88 insurance claims files held by four major insurance companies in Western Australia (WA). In other Australian States, ARRB was collecting similar information so that, together with the WA data, a nation-wide picture of the cost of vehicle damage for different crash types could be derived. These costs would then be input to aggregate road crash costings based on crash type, as argued by Andreassen

(1991), rather than crash/injury severity as had been the case for previous Australian road crash studies. Atkins (1981, 1982) and Steadman and Bryan (1988)[2].

In WA, the project resulted in a computerized database, the Property Damage Database (PDD) containing 7,630 records (motor vehicle damage claims) pertaining to 125 variables[3]. These variables were based on the information contained on the insurance company motor vehicle claims form required to be completed by the claimant/insured driver. For analysis purposes[4], a subset of 2,168 records was used.

The list of variables and their valid values are given in Table I. More details about the encoding programme and peculiarities and problems with the data collection have been published elsewhere, Giles (1994), Giles, Hendrie and Rosman (1995), Giles, Kroll, Harris and Lam (1991), Harris, Giles, Hendrie and Kroll (1991), Hendrie and Harris (1993).

*4.2. Mean Characteristics of the Samples*

The property damage cost of crashes varies over a wide range of variables and their values[5]. Table II gives the mean values for the variables included in the estimated cost equation (2) for the aggregate sample (Column 2; n=2,168), and the subsets of Police reported crashes (Column 3; n=1,151) and unreported crashes (Column 4; n=1,017).

A number of comparisons from Table II can be highlighted. Firstly, all variables except age are dichotomous. Age is a continuous variable and the mean listed for this variable is therefore the sample mean cost. The mean cost for the aggregate sample is $2,217 compared with $3,259 and $1,038 for the subsets of reported and unreported crashes respectively. These differences demonstrate the likelihood that unreported crashes are less costly and injurious.

Secondly, for all other variables (all of which are categorical), the cost of Police notified crashes is higher than the cost of unreported crashes. For example, the mean cost for the variable 'Crashes at intersections' is $3,486. For reported and unreported crashes for this variable, the mean costs are $3,820 and $1,366 respectively.

Thirdly, some variables are associated with relatively less costly crashes (compared with the overall mean cost) in each of the three samples analysed. Included are the variables

for 'Close to home', 'Large country town', and 'Off road crashes'. Variables associated with relatively high cost crashes are 'Night time', 'Rural road', 'Small country town', 'Insurance Company 4', 'Insurance Company 2', 'Weekend', 'Vans and 4WDs', ‘Crashes at intersections’ and 'Crashes between vehicles travelling in opposing directions'.

Fourthly, some categorical variables are associated with relatively less costly crashes in the subset of reported crashes and with more costly crashes in the corresponding subset of unreported crashes. Variables shown in Table II to have these outcomes are labeled 'Australian-made vehicles', 'Crashes between vehicles traveling in the same direction', ‘Crashes between overtaking vehicles’ and 'Other two-vehicle crashes'.

Fifthly, one categorical variable, 'Insurance Company 3', is associated with relatively more costly crashes in the subset of reported crashes and less costly crashes in the corresponding subset of unreported crashes.

These results support the earlier assertion that the characteristics of reported and unreported crashes differ to the extent that cost estimations based on reported crashes only are likely to be biased.

### *4.3 Benchmark Results for Police Notification and Cost Models*

Table III gives the coefficients in the models of Police notification (Column 2) and cost (Column 3). These are estimates of equations 1(b) and (2) respectively.

The Police notification model is estimated as a logit model[6]. The coefficients give the partial effect on the log odds of reporting a crash to the Police, holding constant all other factors. A positive coefficient will increase the log odds ratio and therefore also increase the probability of the crash being reported. A negative coefficient will reduce the log odds ratio thereby also reducing the probability of the crash being reported. The following discussion pertains to the interpretation of those coefficients shown in Table III Column 2 to be significant at the 5% significance level.

The effects of an explanatory variable on the probability of a crash being reported to the Police in the logit model is given as $\frac{\partial PN}{\partial X_i} = \hat{\beta}(1-PN)PN$ where $PN$ is the probability of Police notification (from equation 1(b)). This partial probability effect is often computed at the mean probability of Police notification. In this data set, the mean probability of Police notification is 0.531, which gives a value of $(1-PN)PN$ of 0.249. Accordingly, partial effects can be obtained by multiplying the coefficients listed in Table III Column 2 by this value and multiplying by 100.

For single vehicle crashes, the estimated coefficient of -1.4040 indicates that Police notification for a single vehicle crash is 24.56% ($e^{-1.4040}$ x100) of the odds of Police notification for two vehicle crashes. In terms of the previously defined partial effects, $\partial PN / \partial X_i$, the partial effect of single vehicle crashes on the probability of a crash being reported is -0.3496 [$= -1.4040(0.249)$] or -34.96%. This is quite a substantial impact, and conforms with the literature on the link between the likelihood of a crash being reported to the Police and the number of vehicles involved in the crash.

Police notification for crashes in which Vehicle 1 is insured with Insurance Company 3 is 154.93% of the odds of Police notification for crashes in which Vehicle 1 is not insured with Insurance Company 3. Moreover, the partial effect of being insured with Insurance Company 3 on the probability of a crash being reported is 0.1090 [$= 0.4378(0.249)$] or 10.90%. This partial effect is much smaller that that calculated for single vehicle crashes, showing that whilst institutional considerations (e.g. the Insurance Company) matter, the actual crash environment is much more important. Police notification of crashes with property damage under $300 is 54.38% of the odds of Police notification for crashes with property damage greater than or equal to $300. The partial effect of property damage under $300 on the probability of a crash being reported is -0.1517 [$= -0.6092(0.249)$] or –15.17%.

Police notification of crashes occurring in large country towns or off-road is 172.72% and 21.40% respectively of the odds of Police notification of crashes not in large country towns and on-roads. The partial effect of crashes in large country towns on the probability of a crash being reported is 0.1361 [$= 0.5465(0.249)$] or 13.61%. The partial effect of off-road crashes on the probability of a crash being reported is -0.3840 [$= -1.5420(0.249)$] or -

38.40%. Furthermore, the odds of Police notification increase by a factor of 351.8834/10,000 for each unit increase in cost and by a factor of 0.9647 for each extra year of age of Vehicle 1.

The partial effect of age on the probability of a crash being reported is given by $\frac{\partial PN}{\partial Age} = \{\beta_{Age} + 2\beta_{Age^2} Age\}\{PN(1-PN)\}$. Evaluated at the mean age of 31.205 years, this equals -0.0213 or about 2 percentage points per year of age. It is noted that this partial effect will be negative up to around 54 years, and it will be positive for ages over 54 years.

The cost model is estimated as a log-linear model where the dependent variable is the natural logarithm of the crash cost per vehicle involved. The coefficients give the partial effect on the log cost per vehicle of a crash, holding constant all other factors. A positive coefficient will increase the log cost per vehicle of a crash. A negative coefficient will reduce the log cost per vehicle of a crash. The following discussion pertains to the interpretation of the coefficients shown in Table III Column 3 to be significant at the 5% level.

It appears that crashes that occur at night time, on rural roads or on roads through small country towns, or involve vehicles insured with either Insurance Company 2 or Insurance Company 3 or vehicles with body types 'Vans and 4WDs', tend to have higher costs per vehicle ($\beta_2 > 0$) than crashes that occur during the day time or on roads in the metropolitan area, or involve vehicles insured with Insurance Company 1 or only one vehicle. It also appears that the property damage cost per vehicle decreases as driver's age increases, and is less for all two-vehicle crash types except vehicles traveling in opposite directions compared with single vehicle crashes ($\beta_2 < 0$).

With a coefficient of determination of around 0.32, and a relatively high number of significant explanatory variables that have coefficients consistent with expectations, this cost model appears to provide an appropriate foundation for the study of the effects of Police notification behavior on the estimated cost of road crashes[7]. Specifically, as Gujarati (1995) states, the included variables may be getting the "credit for the influence that is rightly attributable to (the omitted variable), the latter being prevented from showing its effect explicitly because it is not 'allowed' to enter the model" (page 457). This omitted variable

relates to the notification of the crash to the Police. The next Section explores the role of sample selectivity in this regard.

### *4.4 Selectivity Corrected Estimates*

As discussed earlier, there is concern that the 'police notification' variable is endogenous. This is now examined through an application of Heckman's (1979) model. That is, equation (2) is estimated with the correction for selectivity bias - the adjustment for the probability of being included in the sample. This is a two-stage process. Firstly, the probability of a crash being notified to the Police is modeled, in reduced form, as a function of variables for age (entered as a quadratic), Insurance company, year of manufacture of Vehicle 1, time and location of crash, gender and the number of vehicles involved in the crash. Estimates of this model are obtained using a logit model, and the selectivity correction term, $\lambda$, computed using Lee's (1980) generalization of Heckman's (1979) model. Secondly, $\lambda$ is included as a regressor in the re-estimated log-linear cost equation.

The estimates of the cost model corrected for selectivity bias are shown in Column 4 (n=1,151) of Table IV. Columns 2 and 3 are the results for the uncorrected models of all crashes in the sample (n=2,168) and the subset of reported crashes (n=1,151) respectively. The first point to make concerning the Table IV results is that the selectivity bias correction factor, $\lambda$, is statistically significant and positive. The implication of this finding is that unobserved factors that influence whether or not a crash is reported to the Police also influence the extent/cost of the crash. The way in which these influences are teased out of the data can be explained with reference to equations (3) and (4). The influences of the unobserved[8] characteristics are included in the residuals of both equations such that the covariance between these residuals ($\sigma_{21} / \sqrt{\sigma_{11}}$) is positive. Hence when equation (4) is substituted for the second term on the RHS of equation (3), the coefficient of the previously omitted explanatory variable, now known as $\lambda$, is also positive. Specifically, these results mean that the unobserved characteristics that result in some crashes being less likely to be reported are also characteristics that result in the crashes being less costly. For example, the license status of drivers is not routinely recorded in Insurance or Police databases, Giles (2001). Unlicensed drivers or drivers with suspended licences may not report their crashes in case they incur traffic convictions related to their lack of a licence. Such drivers might also

take more care with their driving to avoid detection by road traffic authorities. Hence, if they are involved in a road crash, the resultant property damage may be minor. Alternatively, some crashes that are more likely to be reported are also more costly. For example, the blood alcohol levels of drivers are not routinely recorded for all drivers. Research shows that fatal and other serious crashes often involve drunk drivers, Smith (1988). Such crashes tend to be more costly in terms of both property damage and injury. They are also more likely to be attended by the Police and hence recorded in the Police road crash database. These results are supported in the literature.

The following comments relate to the variables listed in Table IV. Two main comparisons can be made from Table IV - between Columns 3 and 4, and between Columns 2 and 4. The first set of comparisons illustrates the impact that correction for sample selection can make when dealing with a non-random sample. The second set will demonstrate whether the correction made permits a set of estimates to be obtained that are closer to those obtained when the entire sample is available.

For the first of the comparisons from Table IV, between Columns 3 and 4, four observations can be made. Firstly, many of the regressors have the same sign, magnitude and level of significance in both the uncorrected cost model for reported crashes (Column 3) and the corrected model (Column 4). These include the variables for all five vehicle characteristics and those variables for the crash environment characteristics related to crashes occurring close to home, at night-time, off public thoroughfares, on rural roads or on roads in small country towns, or involving two vehicles travelling in the same direction or overtaking or other two-vehicle crashes. This suggests that, for these characteristics, there appear to be no unobserved factors that might affect both the probability of reporting a crash and the cost of a crash relative to crashes without these characteristics.

Secondly, some of the regressors have different magnitudes and levels of significance. These are the variable for the driver characteristic, age, and variables related to crashes occurring on the weekend, or involving two vehicles travelling in opposing directions. In the case of the variable 'Weekend', the (positive) coefficient in the corrected model is smaller and not significant. This suggests that unobserved factors influencing reporting are increasing the average cost of these crashes. In the case of the variable 'Crashes involving vehicles travelling in opposing directions', the (positive) coefficient in the corrected model is larger and

significant. In this case, the unobserved factors influencing crash reporting are reducing the average cost of these crashes.

Thirdly, the regressor for the crash environment variable 'Intersection' is negative, large and significant in the uncorrected model (Column 3) and positive, small and insignificant in the corrected model (Column 4). This suggests that unobserved factors influencing reporting of crashes to the Police are both reducing the average cost of two-vehicle crashes at intersections, and increasing the statistical importance of this type of crash as a determinant of crash costs.

Finally, the regressor for 'Large country town' is larger in the corrected model. Thus unobserved factors influencing the reporting of crashes to the Police are increasing the average cost of crashes occurring on roads in large country towns.

The second comparison from Table IV is between the regressors shown in Columns 2 and 4. Recall that the estimates in Column 2 are for the total sample. Those listed in Column 4 are for the sub-sample of crashes reported to the Police after a correction has been made for sample selectivity. The important issue that needs to be addressed here is whether the estimates following the correction for sample selectivity (Column 4) are closer to those for the total sample (Column 2) than those obtained without the adjustment for sample selectivity (Column 3). There are two broad observations that can be made.

Firstly, there is a number of variables where the correction for sample selectivity results in a coefficient (Column 4) that is closer in magnitude to that in the total cost model (Column 2). Focusing only on significant coefficients in the sub-sample of Police notified crashes (Column 4), there are 8 such variables. These are both age variables and the variables 'Insurance company 2', 'Vehicles in the same direction', 'Vans and 4WDs', 'Rural road'. 'Vehicles in opposing directions', 'Time' and 'Vehicles overtaking'.

Secondly, there is a number of variables where the correction for sample selectivity results in a coefficient that is further away from that reported for the total cost model. Again, attention is restricted to only the significant coefficients. In this case there are three variables where the coefficient is thus affected. These are the variables 'Insurance company 3', 'Other two-vehicle crash types' and 'Off road'.

Before proceeding to use these results, some comments on the robustness of the estimates need to be made. The sensitivity of the results, to a number of changes in the specifications of the selection (Police notification) and substantive (cost) equations, was examined. Firstly, it was found that the results are not affected greatly by changes to the specifications that involve any variables other than crash type. If these variables are omitted from consideration then the estimated coefficient on the selectivity correction term, $\lambda$, is approximately doubled compared with the findings reported in Table IV (Column 4).

Secondly, if a variable for the number of vehicles involved in the crash is added to the cost model, then $\lambda$ becomes insignificant. It is noted that the variable for number of vehicles in the crash and the crash type variables contain similar information. For example, the variable 'Vehicles in opposing directions' pertains only to two-vehicle crashes. Moreover, the benchmark category for the crash type variables is single vehicle crashes. Thus, including both the crash type variables and a variable for the number of vehicles involved in the crash in the model needs to be done with caution. The results of the logit model for crash reporting are also affected by the inclusion/exclusion of the variable for the number of vehicles involved in the crash. However, the variables that are highly significant when this variable is excluded retain their significance when this variable is added to the model.

Finally, if cost instead of cost per vehicle is used as the dependent variable in the cost model, $\lambda$ is negative. This would suggest that the cost of unreported crashes is higher that the cost of reported crashes. This result is not supported in the literature.

## 5. Discussion

Two questions can now be asked. Firstly, what is influencing these variables for which the regressors in Column 4 of Table IV differ from those in Columns 2 and 3? That is, what driver/vehicle/crash environment characteristics excluded from the general models of police notification and property damage costs contribute to crashes that are both, on average, less costly and less likely to be reported? Mention was made earlier of driver's license status as one such important unobserved characteristic.

A further question is whether correction for sample selectivity permits a better estimate of the cost of a particular set of road crashes. Some examples will show how this might be. Firstly, the mean value of the $\lambda$ variable is 0.5293 and the estimated coefficient is 0.7138. Multiplying these together gives 0.3778. This is an estimate of the difference between the mean (logarithmic) cost per vehicle of reported crashes and the mean for the total sample of reported and unreported crashes. From the data, the actual mean values for these samples are 7.0668 (n=1151) and 6.6725 (n=2168) respectively. This gives a difference of 0.3943. Hence the difference in measured cost for the reported sample and that for the underlying distribution of all crashes is captured reasonably accurately by the selectivity bias correction procedure.

Secondly, for most characteristics, the discussion earlier reveals that the use of the selectivity correction technique will result in more accurate estimates of the impact of variables on the total cost of road crashes. Consider, for example, the age variable. Table V compares impacts of a range of values for age on the cost per vehicle of crashes in three samples – the total sample, the sample of reported crashes without correction for sample selectivity, and the sample of reported crashes with correction for sample selectivity.

It is apparent from Table V that among younger age groups the selectivity bias corrected estimates give a better depiction of the age effects in the total sample than the uncorrected estimates. Among older age groups, there is little basis for choice between the two sets of estimates.

Obviously, for some other variables (for example, ‘Off-road’ and ‘Other two-vehicle crash types’), as discussed above, the correction for sample selectivity does not result in a better foundation for revealing the true picture of road crash costs.

In summary, the inclusion of the selectivity bias correction factor, based on a model of the probability of reporting a road crash, in the cost model for Police reported crashes confirms that unreported crashes are likely to be less costly. There are a number of characteristics of crashes that are not routinely recorded in road crash databases and it is the absence of these from the cost model that is leading to biased estimates of crash costs. Heckman’s sample selectivity correction methodology offers a way of improving on the estimates obtained with non-random samples. While there is improvement in general in this

regard, there are situations in which the correction for sample selectivity actually aggravates the problem. Further research in this area is needed. In this regard, the conclusions of this article are in line with the findings reported by Puhani (2000). These findings are threefold as follows.

Firstly, Puhani concurred with Heckman's own admission that the correction procedure provides "good starting values for … exploratory empirical work", Heckman (1979, p. 160). Secondly, Puhani highlights the problem that a correlation between the exogenous variables in the selection and substantive equations undermines the robustness of Heckman's results. Puhani's final conclusion was that judicious use of Heckman's methodology "should be decided on a case by case basis", Puhani (2000, p. 65). The analysis in this article testifies to such conservatism.

## Notes

1. One problem arising from this specification is that the vector $x$ may not contain all the relevant explanatory variables. This will result in the included regression coefficients (elements of $\beta$) being biased "unless the excluded variable is uncorrelated with every included variable", Ramanathan (1989, p. 185). A second and complementary problem, which results in unbiased and consistent but inefficient estimates, is the inclusion of irrelevant independent variables. Researchers often need to trade-off these problems when choosing which explanatory variables to retain in or discard from their models. Both of these problems differ from the difficulties resulting from missing $y$ values, otherwise known as selection or selectivity bias.
2. Despite this work, two further Australian studies continued to use crash/injury severity to differentiate crash costs, Bureau of Transport and Communications Economics (BTCE) (1992) and Bureau of Transport Economics (2000).
3. Not all of the 125 variables were accessed for each insurance company. For example, only one insurance company collected information on the colour of the insured vehicle on their motor vehicle accident claim form. In some companies, the date of birth of the driver was asked; in others, only the age of the driver was required.
4. Excluded records had missing values on at least one variable, related to non-crash claims (windscreen, fire or theft) or were not randomly sampled (1,607 of the records were non-randomly sampled due to a change to the data collection procedure).
5. The variables included in the multivariate analysis have been recoded from the original data as described in the footnotes to Table II.
6. This specification follows recommendations by Barnard (1989).
7. A Chow test reveals that the determinants of the cost of crashes for the Police notified sample are statistically different from the determinants in the sample of crashes that were not notified to the Police, Chow (1960).
8. Characteristics may be unobserved because they were excluded from the encoding of the crash information into the Property Damage Database, or because the motor vehicle claims files from which the Database was constructed did not contain that information.

## References

Amemiya, T. (1973): "Regression Analysis When the Independent Variable Is Truncated Normal," *Econometrica*, 41, 997-1016.

Andreassen, D. C. (1991): "Model Guidelines for Road Accident Data and Accident Types: Version 1.1," Vermont South, Victoria: Australian Road Research Board.

Atkins, A. S. (1981): "The Economic and Social Costs of Road Accidents in Australia: With Preliminary Cost Estimates for Australia 1978," Melbourne: Centre for Environmental Studies, University of Melbourne.

Atkins, A. S. (1982): "The Economic Costs of Road Accidents in Australia: Some Issues in Estimation, Concept and Application," Vermont South, Victoria: Australian Road Research Board.

Barnard, P. O. (1989): "How Some Advances in Econometrics Might Be Used in Road Crash Analysis," Vermont South, Victoria: Australian Road Research Board.

Berk, R. A. (1983): "An Introduction to Sample Selection Bias in Sociological Data," *American Sociological Review*, 48, 386-398.

Bureau of Transport and Communications Economics (BTCE) (1992): "Social Cost of Transport Accidents in Australia," Canberra: Bureau of Transport and Communications Economics.

Bureau of Transport Economics (2000): "Road Crash Costs in Australia," Canberra: Bureau of Transport Economics.

Cain, G. G., and H. W. Watts (1973): "Toward a Summary and Synthesis of the Evidence," in *Income Maintenance and Labour Supply: Econometric Studies*, ed. by G. G. Cain, and H. W. Watts. Chicago: Markham, 328-367.

Chow, G. C. (1960): "Tests of Equality between Sets of Coefficients in Two Linear Regressions," *Econometrica*, 28, 591-605.

Crawford, D. L. (1979): "Estimating Models of Earnings from Truncated Samples," Madison, WI: Department of Economics, University of Wisconsin.

Giles, M. J. (1994): "Lies, Damned Lies and Road Crash Statistics," Gold Coast International Hotel, Surfers' Paradise, Queensland: 23rd Conference of Economists.

Giles, M. J. (2001): "Data for the Study of Road Crashes in Australia," *The Australian Economic Review*, 34, 222-30.

Giles, M. J., D. Hendrie, and D. L. Rosman (1995): "Characteristics of Reported and Unreported Crash Data in the Property Damage and Road Injury Databases," Perth: School of Finance and Economics, Edith Cowan University.

Giles, M. J., L. Kroll, A. Harris, and K. Lam (1991): "The Development of a Property Damage Cost Database: Preliminary Analyses and First Report to Australian Road Research Board," Perth: Road Accident Prevention Research Board, The University of Western Australia.

Goldberger, A. (1981): "Linear Regression after Selection," *Journal of Econometrics*, 15, 357-366.

Gujarati, D. N. (1995): *Basic Econometrics*. New York: McGraw-Hill.

Harris, A., M. J. Giles, D. Hendrie, and L. Kroll (1991): "The Property Damage Costs of Road Accidents in Western Australia: Final Report to ARRB on the Development of a Property Damage Database," Perth: Road Accident Prevention Research Unit, Department of Public Health, The University of Western Australia.

Hausman, J. A., and D. A. Wise (1977): "Social Experimentation, Truncated Distributions and Efficient Estimation," *Econometrica*, 45, 919-928.

Heckman, J. J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *The Annals of Economic and Social Measurement*, 5, 475-492.

Heckman, J. J. (1979): "Sample Bias as a Specification Error," *Econometrica*, 47, 153-161.

Heckman, J. J. (1980): "Sample Selection Bias as a Specification Error with an Application to the Estimation of Labor Supply Functions," in *Female Labor Supply: Theory and Estimation*, ed. by J. P. Smith. Princeton, New Jersey: Princeton University Press, 206-247.

Hendrie, D., and A. H. Harris (1993): "Vehicle Repair Costs Resulting from Road Accidents in Western Australia," Perth: Road Accident Prevention Research Unit, Department of Public Health, The University of Western Australia.

Lee, L. F., G. S. Maddala, and R. P. Trost (1980): "Asymptotic Covariance Matrices of Two Stage Probit and Two Stage Tobit Methods for Simultaneous Models with Selectivity," *Econometrica*, 491-504.

Lush, J. L., and R. R. Shrode (1950): "Changes in Milk Production with Age and Milking Frequency," *Journal of Dairy Science*, 33, 338-357.

Puhani, P. A. (2000): "The Heckman Correction for Sample Selection and Its Critique," *Journal of Economic Surveys*, 14, 53-68.

Ramanathan, R. (1989): *Introductory Econometrics with Applications*. San Diego: Harcourt Brace Jovanovich.

Smith, D. I. (1988): "Effect on Traffic Safety of Introducing a 0.05% Blood Alcohol Level in Queensland, Australia," *Medicine, Science, and the Law*, 28, 165-170.

Smith, J. P. (1980): *Female Labor Supply: Theory and Estimation*, Princeton, New Jersey: Princeton University Press.

Smith, J. P. (1980): "Introduction," in *Female Labor Supply: Theory and Estimation*, ed. by J. P. Smith. Princeton, New Jersey: Princeton University Press, 3-23.

Steadman, L. A., and R. J. Bryan (1988): "Cost of Road Accidents in Australia," Canberra: Bureau of Transport Economics.

Winship, C., and R. Mare (1992): "Models for Sample Selection Bias," *Annual Review of Sociology*, 18, 327-350.

Table I: Variables and Valid Values

| Variables | Claim Information: |
|---|---|
| 1. | Unique identifier/counter – input by the encoding programme – values 1 to 8,057 (7,630 valid records, 427 blank records). |
| 2. | Insurance company with which Vehicle 1 is insured – values 1 to 4. |
| 3. | Claim number from Insurance company records. |
| 4. | Claim type – values 1 to 5, 9 (not stated). |
| 5. | Claim status – values 1 to 4, 9 (not stated). |
| | Crash Information: |
| 6. | Police notified – values 1 to 2, 9 (not stated). |
| 7. | Number of vehicles involved in the crash – values 1 to 12, 99 (not stated). |
| 8. | Accident hour – values 1 to 24, 88 (not asked), 99 (not stated). |
| 9. | Accident day of week – values 1 to 7, 0 (not stated). |
| 10. | Postcode of crash location – values 800 to 7316, 9999 (not known). |
| 11. | Crash location area – values 1 to 4, 9 (not known). |
| 12. | Prior Road User Movement (PRUM) – events prior to the crash – values 11 to 96, 88 (not a crash), 99 (not stated). See Andreassen (1991). |
| 13. | Chosen Road User Movement (crash type) (CRUM) – values 00 to 97, 88 (not a crash). See Andreassen (1991). |
| 14. | Subsequent Road User Movements (SRUM) I, II and III – values 00 to 97, 88 (not a crash). See Andreassen (1991). |
| 15. | Supplementary Road User Movement (Supp)– values 19 to 97, 88 (not a crash). See Andreassen (1991). |
| | Vehicle and Driver Information: |
| 16. | Vehicle 1 Make. |
| 17. | Vehicle 1 Model. |
| 18. | Vehicle 1 Body Type. |
| 19. | Vehicle 1 Year of Manufacture – values 1908 to 1988, 9999 (not stated). |
| 20. | Vehicle 1 Color – values 00 (not stated), 1 to 36, 88 (not asked). |
| 21. | Vehicle 1 Lamps Lit – values 1 to 2, 8 (not asked), 9 (not stated). |
| 22. | Vehicle 1 Towed – values 1 to 2, 8 (not asked), 9 (not stated). |
| 23. | Vehicle 1 Trip Purpose – values 1 to 2, 8 (not asked), 9 (not stated). |
| 24. | Vehicle 1 Driver's Year of Birth – values 01 to 71, 87 (wrong code), 88 (not asked), 99 (not stated). |
| 25. | Vehicle 1 Driver's Age – values 00 (wrong code), 16 to 89, 99 (not stated). |
| 26. | Vehicle 1 Driver's Gender – values 1 to 2, 9 (not stated). |
| 27. | Vehicle 1 Driver's Home Address Postcode – values 870 to 6962, 9999 (not stated). |
| 28. | Number of injured persons in Vehicle 1 – values 1 to 2, 8 (not asked), 9 (not stated). |
| 29. | Number of passengers in Vehicle 1 – values 0 to 5, 8 (not asked), 9 (not stated). |
| 30. | Vehicle 2 Insurance Company – values 1 to 85, 88 (not asked), 98 (wrong code), 99 (not stated). |
| 31. | Vehicle 2 Make. See Andreassen (1991). |
| 32. | Vehicle 2 Model. See Andreassen (1991). |

Table I: Variables and Valid Values (cont'd)

| Variables | Claim Information: |
|---|---|
| 33. | Vehicle 2 Body Type. See Andreassen (1991). |
| 34. | Vehicle 2 Lamps Lit – values 1 to 2, 8 (not asked), 9 (not stated). |
| 35. | Vehicle 2 Towed – values 1 to 2, 8 (not asked), 9 (not stated). |
| 36. | Vehicle 2 Driver's Age – values 8 to 87, 88 (not asked), 99 (not stated). |
| 37. | Vehicle 2 Driver's Home Postcode – values 820 to 7255, 8888 (not asked), 9999 (not stated). |
| 38. | Vehicle 2 Passengers – values 1 to 2, 8 (not asked), 9 (not stated). |
| 39. | Vehicle 2 Injuries – values 1 to 2, 8 (not asked), 9 (not stated). |
| 40. to 49. | Vehicle 3 as for Vehicle 2. |
| 50. to 59. | Vehicle 4 as for Vehicle 2. |
| 60. to 69. | Vehicle 5 as for Vehicle 2. |
| | Cost Information: |
| | For each vehicle (Vehicles 1 to 5), the following costs might be zero if not applicable or greater than zero if that cost was incurred: |
| | Policy excess |
| | Towing costs |
| | Car hire costs |
| | Investigation costs |
| | Panel-beating costs |
| | Property damage costs |
| | Other costs |
| | Pay-out to the client, in the event of a vehicle write-off |
| | Salvage recoups to the insurance company |
| | Total net cost (non-salvage costs minus salvage recoups) |

Table II: Variables in the Cost Model – Summary Statistics

| Variable | Total Cost (A$) (n=2,168) | | | Total Cost - Police Notified (A$) (n=1,151) | | | Total Cost - Police Not Notified (A$) (n=1,017) | | |
|---|---|---|---|---|---|---|---|---|---|
| Sample statistics[a] | $\overline{X}$ | s | n | $\overline{X}$ | s | n | $\overline{X}$ | s | n |
| Age[b] | 2217 | 2974 | 2168 | 3259 | 3632 | 1151 | 1038 | 1148 | 1017 |
| Australian-made vehicles[c] | 2156 | 2776 | 1104 | 3230 | 3394 | 556 | 1066 | 1218 | 548 |
| Crashes at intersections[d] | 3486 | 3460 | 250 | 3820 | 3588 | 216 | 1366 | 1030 | 34 |
| Crashes between overtaking vehicles[e] | 2524 | 2553 | 26 | 3014 | 2932 | 18 | 1420 | 627 | 8 |
| Crashes between vehicles travelling in opposing directions[f] | 4975 | 4063 | 152 | 5258 | 4074 | 141 | 1347 | 1173 | 11 |
| Crashes between vehicles travelling in the same direction[g] | 2051 | 2206 | 558 | 2426 | 2487 | 382 | 1238 | 1030 | 176 |
| Close to home[h] | 1941 | 2468 | 989 | 3002 | 3049 | 491 | 894 | 858 | 498 |
| Insurance company 2[i] | 4984 | 6309 | 38 | 7415 | 7954 | 19 | 2553 | 2467 | 19 |
| Insurance company 3[j] | 2283 | 3117 | 453 | 3542 | 3792 | 233 | 950 | 1160 | 220 |
| Insurance company 4[k] | 2430 | 3777 | 191 | 3555 | 4884 | 99 | 1219 | 1113 | 92 |
| Large country town[l] | 1985 | 2441 | 106 | 2903 | 2800 | 59 | 832 | 1131 | 47 |
| Night-time[m] | 2457 | 3173 | 797 | 3557 | 3832 | 426 | 1195 | 1341 | 371 |
| Off road crashes[n] | 746 | 838 | 477 | 1273 | 1494 | 70 | 656 | 624 | 407 |
| Other two-vehicle crashes[o] | 1284 | 1310 | 366 | 1674 | 1732 | 136 | 1054 | 908 | 230 |
| Rural road[p] | 3274 | 4497 | 244 | 6198 | 5922 | 96 | 1377 | 1249 | 148 |
| Small country town[q] | 3594 | 6059 | 27 | 5393 | 8722 | 10 | 2536 | 3695 | 17 |
| Vans & 4WDs[r] | 3087 | 4158 | 68 | 5827 | 5411 | 23 | 1686 | 2399 | 45 |
| Weekend[s] | 2453 | 3502 | 596 | 3850 | 4349 | 297 | 1065 | 1350 | 299 |
| All crashes | 2217 | 2974 | 2168 | 3259 | 3632 | 1151 | 1038 | 1148 | 1017 |

Notes:

a $\overline{X}$ is arithmetic mean cost, s is standard deviation, n is sample size.

b Age refers to the age of the driver of Vehicle 1 (the claimant on the motor vehicle insurance claim record from which the records in the Property Damage Database (PDD) are compiled). Some of the insurance companies collected date of birth information. In this case, age was

computed from the difference between the accident date and the date of birth. Other insurance companies collected age and not date of birth information. In two-vehicle accidents (TVAs), the age (or date of birth) of the driver of Vehicle 1 should be reliability recorded, the age of the other driver was less reliably reported.

c During the 1980s in Australia, tariffs on imported cars and spare parts kept the prices on foreign-made cars high relative to Australian-made vehicles and spare parts. In the absence of a specific variable, in the PDD, for country of manufacture of the vehicle, vehicles made by Ford, Toyota, Holden and Chrysler/Mitsubishi were assumed predominantly Australian-made, and all other vehicle makes were assumed to be imported.

d Crashes at intersections include crashes with CRUM (Table I, Variable 13) codes of 10 – 19.

e Crashes between overtaking vehicles include crashes with CRUM (Table I, Variable 13) codes of 50 – 56.

f Crashes between vehicles travelling in opposing directions include crashes with CRUM (Table I, Variable 13) codes of 20 – 27.

g Crashes between vehicles travelling in the same direction include crashes with CRUM (Table I, Variable 13) codes of 30 – 39.

h Home refers to the proximity of the crash to the driver(s) home. In the case of Single Vehicle Accidents (SVAs), the crash is considered close to home if the postcode of the driver's home address is the same as the postcode of the crash location. In the case of Two-Vehicle Accidents (TVAs), the crash is considered close to home if the postcode of either driver's home address is the same as the postcode of the crash location.

i There were four insurance companies with which Vehicle 1 could be insured. These are not identified for ethical reasons and were labelled 1, 2, 3 and 4.

j See i.

k See i.

l The crash location variable (Table I, Variable 11) has four valid values for metropolitan area, large country town, small country town and rural road outside country towns. The metropolitan area was defined in terms of the outskirts shown on the (then) current urban street directory. Large and small country towns were defined in terms of population. Crash location included on and off public thoroughfares.

m Night-time is defined as 5:00 pm (1700 hours) to 5:59 am (0559 hours). Day-time is defined as 6:00 am (0600 hours) to 4:59 pm (1659 hours).

n Off road crashes include crashes with CRUM (Table I, Variable 13) codes of 87.

o Other two-vehicle crashes include crashes with CRUM (Table I, Variable 13) codes of 40 – 45, 47 – 49, 90 – 97.

p See l.

q See l.

r Vans and 4WDs include four-wheel and all-wheel drive vehicles such as land-cruisers, vans such as campervans, minibuses and panel-vans, and buses such as school buses.

s Weekend is defined to include 7 pm (1900 hours) on Friday to midnight (2400 hours) on Sunday.

Table III: Coefficients in the Police Notification and Cost Models

| Crash characteristics | Police Notification: Logit Model[a] (n=2,168) | Cost per Vehicle: Loglinear Model[b] (n=2,168) |
|---|---|---|
| Driver characteristics: | | |
| Age[c] | -0.0870* | -0.0306* |
| Age squared[d] | 0.8079* | 0.00032* |
| Female[e] | 0.1319 | n.a. |
| Vehicle characteristics: | | |
| Insurance company 2[f] | -1.1724 | 0.6745* |
| Insurance company 3[g] | 0.4378* | 0.1563* |
| Insurance company 4[h] | -0.0654 | 0.0831 |
| Cost under $300[i] | -0.6092* | n.a. |
| Cost[j] | 5.8633* | n.a. |
| Australian-made[k] | n.a. | 0.0272 |
| Vans and 4WDs[l] | n.a. | 0.3255* |
| Year of manufacture of Vehicle 1[m] | -0.0359* | n.a. |
| Crash Environment characteristics: | | |
| Large country town[n] | 0.5465* | 0.0547 |
| Small country town[o] | -0.7931 | 0.4950* |
| Rural road[p] | -0.4232 | 0.7022* |
| Intersection[q] | n.a. | -0.0776 |
| Vehicles in opposing directions[r] | n.a. | 0.2566* |
| Vehicles in same direction[s] | n.a. | -0.6037* |
| Vehicles overtaking[t] | n.a. | -0.3895* |
| Other two-vehicle crash types[u] | n.a. | -0.9540* |
| Off-road[v] | -1.5420* | -1.1896* |
| Weekend[w] | -0.0762 | 0.0534 |
| Home[x] | -0.1627 | -0.0669 |
| Time[y] | n.a. | 0.1690* |
| Single vehicle[z] | -1.4040* | n.a. |
| Constant | 72.6272* | 7.6220* |
| Adj $R^2$ | n.a. | 0.3153 |
| Log Likelihood | -1979.990 | n.a. |

Notes:

a Coefficients are estimates of equation 1(b) where the dependent variable, $y_i$, is dichotomous and the model is estimated using logistic regression. The default categories for the categorical variables are 'Male', 'Insurance company 1', 'Cost greater than or equal to $300', 'Metropolitan', 'On-road', Weekday', 'Away from home' and 'Two-vehicle'.

b The natural logarithm of Total (Gross) Cost per Vehicle is the dependent variable. The default categories for the categorical variables are 'Insurance company 1', 'Not Australian-made', 'Not vans and 4WDs', 'Metropolitan', 'Single vehicle', 'Weekday', 'Away from home' and 'Day-time'.

c Mean age is 31.205 years (n=2,168).

d This variable was scaled: age squared = age squared/1,000.

e Only the gender of the driver of Vehicle 1 was included in the PDD.

f 'Insurance Company 2' = 1 for Vehicle 1 insured with Insurance Company 2 (either single or two-vehicle crashes) and 'Insurance Company 2' = 0 for Vehicle 1 insured with other Insurance Companies (either single or two-vehicle crashes).

g 'Insurance Company 3' = 1 for Vehicle 1 insured with Insurance Company 3 (either single or two-vehicle crashes) and 'Insurance Company 3' = 0 for Vehicle 1 insured with other Insurance Companies (either single or two-vehicle crashes).

h 'Insurance Company 4' = 1 for Vehicle 1 insured with Insurance Company 4 (either single or two-vehicle crashes) and 'Insurance Company 4' = 0 for Vehicle 1 insured with other Insurance Companies (either single or two-vehicle crashes).

i Mean cost = $2,216.993 (n=2,168). Cost is included twice. Here it is included as a dichotomous variable with cost either less than $300, or greater than or equal to $300. The cut-off value for reporting a road crash to the Police in 1987/88 in Western Australia was $300.

j Cost here is a continuous variable and is scaled: cost = cost/10,000. for SVAs, aggregate cost is the total cost for Vehicle 1 with any salvage revenue added back in. This cost then represents the total damage (gross cost) to property resulting from the crash and not the net costs of that crash. For TVAs, the total cost of the crash is the sum of the total damage (gross costs) for Vehicles 1 and 2.

k 'Australian-made' = 1 for vehicles made in Australia (Vehicle 1 for single-vehicle crashes; both Vehicle 1 and Vehicle 2 for two-vehicle crashes) and 'Australian-made' = 0 for vehicles not made in Australia (Vehicle 1 for single-vehicle crashes; either Vehicle 1 or Vehicle 2 for two-vehicle crashes).

l 'Vans and 4WDs' = 1 for single-vehicle crashes involving a van or 4WD vehicle and for two-vehicle crashes where both Vehicle 1 and Vehicle 2 were either vans or 4WD vehicles. 'Vans and 4WDs' = 0 for single-vehicle crashes that involved neither vans nor 4WD vehicles or for two-vehicle crashes where either or both Vehicle 1 and Vehicle 2 were not vans or 4WD vehicles.

m The mean year of manufacture of Vehicle 1 is 1980 (n=2,168). This characteristic was only available for Vehicle 1.

n 'Large Town' = 1 for crashes which occur in large towns in rural Western Australia and 'Large Town' = 0 for crashes occurring elsewhere in Western Australia.

o 'Small Town' = 1 for crashes which occur in small towns in rural Western Australia and 'Small Town' = 0 for crashes occurring elsewhere in Western Australia.

p 'Rural Road' = 1 for all crashes outside the Perth metropolitan area which did not occur in large or small country towns, and 'Rural Road' = 0 for all crashes which did not occur on rural roads.

q 'Intersection' = 1 for crash types coded 10 to 19 in the Model Guidelines, Andreassen (1991), and 'Intersection' = 0 for all other crash types.

r 'Vehicles from opposing directions' = 1 for crash types coded 20 to 29 in the Model Guidelines, Andreassen (1991), and 'Vehicles from opposing directions' = 0 for all other crash types.

s 'Vehicles from same direction' = 1 for crash types coded 30 to 39 in the Model Guidelines, Andreassen (1991), and 'Vehicles from same direction' = 0 for all other crash types.

t 'Vehicles overtaking' for crash types coded 50 to 56 in the Model Guidelines, Andreassen (1991), and 'Vehicles overtaking' = 0 for all other crash types.

u 'Other crash types' = 1 for crash types coded 40 to 45, 47 to 49, and 90 to 97 in the Model Guidelines, Andreassen (1991), and 'Other crash types' = 0 for all other crash types.

v 'Off road' = 1 for crashes that occur off public thoroughfares and therefore are outside the legislation for reporting crashes to the Police. Many of these crashes occur in parking lots or on private property, including farms. The Model Guidelines, Andreassen (1991), ignored these crashes so an additional code of 87 was allocated to these crash types. 'Off road' = 0 for crashes occurring on public thoroughfares for all codes in the Model Guidelines excluding the additional code of 87.

w 'Weekend' = 1 for crashes occurring on Saturday, Sunday or after 6.59 pm on Fridays and 'Weekend' = 0 for crashes occurring Monday to Thursday and Friday before 7.00 pm.

x 'Home' = 1 when the postcode of the crash site and the postcode for the address of the driver of Vehicle 1 (in the case of single-vehicle crashes) or the postcode for the address of the driver of either Vehicle 1 or Vehicle 2 (in the case of two-vehicle crashes) are identical. 'Home' = 0 for non-identical postcodes.

y 'Time' = 1 for crashes occurring after 4.59 pm and before 6.00 am (night-time) and 'Time' = 0 for crashes occurring outside these times (day-time).

z 'Single vehicle crashes' = 1 for crash types coded 00 – 09, 46, 60 – 67, 70 – 75, and 80 – 86 in the Model Guidelines, Andreassen (1991), and 'Single vehicle crashes' = 0 for all other crash types.

* denotes coefficients that are significant at 5%.

** denotes coefficients that are significant at 10%.

n.a. not applicable or variable not included in the model.

Table IV: Coefficients in the Loglinear Cost Models[b] – With and without the Correction for Selectivity Bias

| Crash characteristics | Cost per Vehicle<br><br>(n=2,168) | Cost per Vehicle – Reported Crashes without Selectivity Bias Correction Factor<br>(n=1,151) | Cost per Vehicle – Reported Crashes with Selectivity Bias Correction Factor<br><br>(n=1,151) |
|---|---|---|---|
| Driver characteristics: | | | |
| Age[c] | -0.0306* | -0.0172** | -0.0351* |
| Age squared[d] | 0.00032* | 0.00018 | 0.00032* |
| Vehicle characteristics: | | | |
| Insurance company 2[f] | 0.6745* | 0.5994* | 0.6270* |
| Insurance company 3[g] | 0.1563* | 0.1589* | 0.2104* |
| Insurance company 4[h] | 0.0831 | 0.0343 | 0.0288 |
| Australian-made[k] | 0.0272 | -0.0719 | -0.0720 |
| Vans and 4WDs[l] | 0.3255* | 0.5457* | 0.4681* |
| Crash Environment characteristics: | | | |
| Large town[n] | 0.0547 | 0.0397 | 0.1481 |
| Small town[o] | 0.4950* | 0.3426 | 0.2606 |
| Rural road[p] | 0.7022* | 0.9132* | 0.8607* |
| Intersection[q] | -0.0776 | -0.2439* | 0.0120 |
| Vehicles in opposing directions[r] | 0.2566* | 0.0482 | 0.3688* |
| Vehicles in same direction[s] | -0.6037* | -0.7034* | -0.6429* |
| Vehicles overtaking[t] | -0.3895* | -0.6553* | -0.5611* |
| Other two-vehicle crash types[u] | -0.9540* | -1.0767* | -1.2558* |
| Off-road[v] | -1.1896* | -1.4211* | -1.9062* |
| Weekend[w] | 0.0534 | 0.1103** | 0.0777 |
| Home[x] | -0.0669 | -0.0734 | -0.0950 |
| Time[y] | 0.1690* | 0.1532* | 0.1490* |
| $\lambda$ [aa] | n.a. | n.a. | 0.7138* |
| Constant | 7.6220* | 7.7169* | 7.6630* |
| Adj $R^2$ | 0.3153 | 0.3140 | 0.3217 |

Notes:

b to z From Notes to Table III.

aa $\lambda$ is the selectivity bias correction factor, Heckman (1976).

* denotes coefficients that are significant at 5%.

** denotes coefficients that are significant at 10%.

TABLE V: Impact[a] of Age on Crash Cost per Vehicle

| Age | Reported and Unreported crashes[b] (n=2,168) | Reported crashes without selectivity bias correction factor[c] (n=1,151) | Reported crashes with selectivity bias correction factor[d] (n=1,151) |
|---|---|---|---|
| 15 | -0.0210 | -0.0118 | -0.0255 |
| 30 | -0.0114 | -0.0064 | -0.0159 |
| 45 | -0.0018 | -0.0010 | -0.0063 |
| 60 | 0.0078 | 0.0044 | 0.0033 |

Notes:
a Calculated as the partial derivative on cost of age.
b The model reported in Table IV column 2.
c The model reported in Table IV column 3.
d The model reported in Table IV column 4.