CHAPTER

4

Crash-severity modeling

4.1 Introduction

"Vision Zero," an ambitious, multinational initiative to eliminate traffic fatalities and serious injuries, has made the reduction of fatal and severe injury crashes a top priority for transportation safety stakeholders around the world. To achieve this goal, researchers and safety professionals rely heavily on crash data as they are the most relevant and informative resource for analyzing traffic injuries; however, the causes of an injury are very complicated because they involve a sequence of events and a number of factors (i.e., driver, vehicle, environment), as discussed in Chapter 2—*Fundamentals and Data Collection*. Similar to the methodologies described in Chapter 3—*Crash-Frequency Modeling*, statistical methodologies have been used extensively to explore the intriguing relationships between crash severities and other data elements. In particular, crash injury severity modeling helps describe, identify, and evaluate the factors contributing to various levels of injury severity.

Unlike crash count, which is a nonnegative integer, injury severity has a finite number of outcomes (e.g., killed, injury type A, injury type B, injury type C, no injury) that are categorized on the KABCO scale. Discrete choice and discrete outcome models have been used to handle this type of response variable. Crash severity models are categorized as fixed or random parameter models according to the parameter assumptions. Crash-severity models can also be classified as nonordinal (e.g., multinomial logit (MNL) and multinomial probit) or ordered probabilistic (e.g., ordered probit and order logistic) if an ordinal structure for the response variable is assumed.

Model variations are available if restrictions such as irrelevant and independent alternatives (IIA), proportional odds, or homogeneity are

https://doi.org/10.1016/B978-0-12-816818-9.00005-6

relaxed. Savolainen et al. (2011) performed an extensive review of the methodological alternatives for modeling highway crash injury severity, but the research did not yield an agreement among experts on which model works best with crash severity data. Professionals do agree, however, that both statistical goodness-of-fit and model interpretation should be considered when modeling crash injury severity.

This chapter introduces the methodologies and techniques that have been applied to model crash severity in safety studies. The discussion includes the different forms, constructs, and assumptions of crash severity models due to the prevailing issues of crash data. The theoretical framework and practical techniques for identifying, estimating, evaluating, and interpreting factors contributing to crash injury severities are also explored. In addition, an extensive list of available crash-severity models are described in Appendix B.

4.2 Characteristics of crash injury severity data and methodological challenges

Several prevailing issues related to crash injury data have come to light during model development, including unobserved heterogeneity, omitted variables bias, temporal and spatial correlation, ordinality of injury severity, and imbalanced observations between injury severity levels (Savolainen et al., 2011; Washington et al., 2020; Mujalli, 2016). Some of these issues are discussed in depth in Chapter 3—*Crash-Frequency Modeling* and Chapter 6 - *Cross-sectional and Panel Studies in Safety*, as the roots of the problems are the same. The issues that are specific to crash severity data will be discussed in detail in the following sections.

4.2.1 Ordinal nature of crash injury severity data

An ordinal scale quantitatively categorizes crashes from the highest to lowest levels of injury severity (i.e., KABCO). Recognizing this ordinal structure within data is important because it aids in the selection of an appropriate methodology. Utilizing the intrinsic ordinal information preserved in the data may lead to the estimation of fewer parameters. Additionally, the potential dependency between adjacent categories may share unobserved effects. If such a correlation exists but is not accounted for, it can lead to biased parameter estimates and incorrect inferences (Savolainen and Mannering, 2007). Nevertheless, the ordinality assumption should be exercised with caution, as it can be overly restrictive for models under certain circumstances, such as when lower severity crashes are underreported.

4.2.2 Unobserved heterogeneity

Differences in drivers' risk-taking behaviors, physiological attributes, and other factors lead to unobserved heterogeneity among road users involved in crashes. Data heterogeneity affects the model parameters among injury observations. Large effects, when unaccounted for, could lead to biased parameter estimates and incorrect statistical inferences (McFadden and Train, 2000; Train, 2009).

4.2.3 Omitted variable bias

It is impossible to include all variables relating to injury severity in one model. Some variables, albeit important, may not be available in a crash report (e.g., vehicle mass, speed, collision angles). However, the omission of important explanatory variables can result in inconsistent parameter estimates. This can occur when omitted variables are correlated with variables that are already included in the model, or when omitted variables contribute to different variances among injury severity levels (Washington et al., 2020). If the omission of relevant variables is a critical limitation in a crash prediction model, the model results must be discussed for possible implications on their application.

4.2.4 Imbalanced data between injury severity levels

Crash injury severity data usually are imbalanced on the KABCO scale, where the number of fatal or severe injuries is substantially less than the number of less severe and no injury crashes. This imbalance of data in each injury category presents a challenge for classification algorithms. In predictive modeling, imbalanced data introduce a bias toward the majority that causes less accurate predictions of severe crashes. A common method of treating imbalanced data is to combine similar injury types (i.e., K, A, B, and C) into one category on a new scale (i.e., injury and noninjury) to gain more balanced data. Other methods for handling imbalanced data include resampling techniques that aim to create a balanced injury scale data through oversampling less-representative classes or undersampling overly-representative classes (Mujalli, 2016).

4.3 Random utility model

Crash severity models are driven by the development of econometrics methods. In economics, utility is a measure of relative satisfaction. In the context of safety, we are looking for a combination of factors that lead to the worst injuries. The utility function usually favors the maximum utility

(e.g., high injury severity levels) and is usually a linear form of covariates as follows:

$$U_{ni} = \beta_{0i} + \beta_{1i}x_{n1i} + \beta_{2i}x_{n2i} + \ldots + \beta_{ki}x_{nki} = \mathbf{x}'_{ni}\boldsymbol{\beta}_i \tag{4.1}$$

where U_{ni} is the utility value of crash n with injury severity level i; x_{nki} is the kth variable related to injury level i; β_{0i} is the constant for injury level i; and, β_{ki} is the estimable coefficients for the covariates.

Utility maximization is the process of choosing the alternative with the maximum utility value. In a binary outcome model with injury and no injury, if $U(injury) > U(no\ injury)$, then the probability of injury $P_r(injury) = 1$; and if $U(injury) < U(no\ injury)$, then $P_r(injury) = 0$. This is a deterministic choice that can be depicted in Fig. 4.1.

A random unspecifiable error term, ε_{ni}, is added to the end of Eq. (4.1), as it is difficult to specify each crash observation's utility function with certainty. The utility function becomes a random utility function as follows:

$$U_{ni} = \beta_{0i} + \beta_{1i}x_{n1i} + \beta_{2i}x_{n2i} + \cdots + \beta_{ki}x_{nki} + \varepsilon_{ni} = V_{ni} + \varepsilon_{ni} \tag{4.2}$$

where V_{ni} represents the deterministic portion of U_{ni}.

The addition of a disturbance term ε helps with the previously mentioned issues of variables being omitted from the utility function, an incorrectly specified functional form, use of proxy variables, and unaccounted for variations in β (β may vary across observations). A random utility model leads to an estimable model of discrete outcomes, with I denoting all possible outcomes for observation n, and P_{ni} denoting the probability of observation n with discrete outcome $i(i \in I)$:

$$P_{ni} = \Pr(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj}, \forall j \neq i) = \Pr(\varepsilon_{nj} < V_{ni} + \varepsilon_{ni} - V_{nj}, \forall j \neq i) \tag{4.3}$$

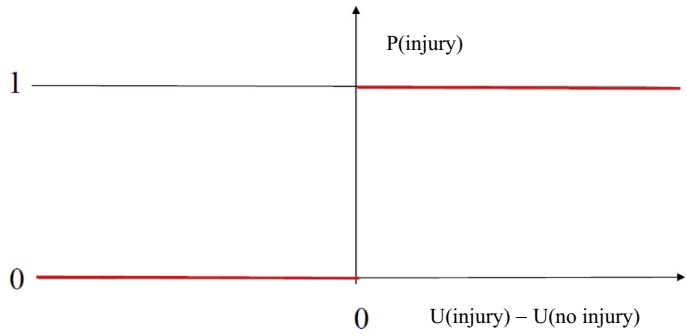


FIGURE 4.1 Deterministic choice of a binary variable.

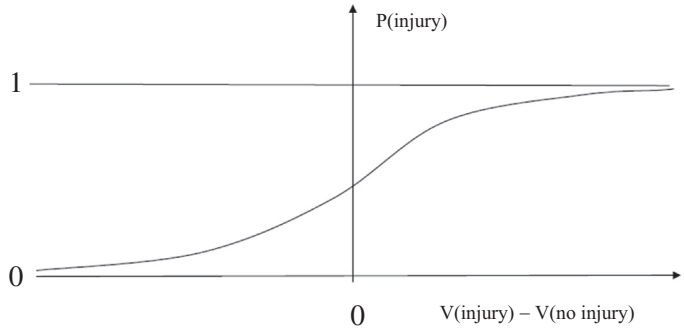


FIGURE 4.2 Stochastic choice of a binary variable.

Models are estimated by assuming a distribution for the random error term, ε's. Now, instead of being a deterministic outcome, the probability of each outcome alternative is determined by the distributional form (Fig. 4.2).

4.4 Modeling crash severity as an unordered discrete outcome

Treating the dependent variable with multiple responses as ordinal or as nominal significantly impacts which methodologies should be considered. From a model estimation perspective, it is desirable for the maximum of a set of randomly drawn values to have the same form of distribution as the one from which they are drawn. An error term (ε) distribution with such a property greatly simplifies model estimation because this property could be applied to the multinomial case by defining the highest utility value of all other options as $x'_{nj}\beta_j(\forall j \neq i)$. The normal distribution does not possess this property because the maximums of randomly drawn samples from the normal distribution are not normally distributed. However, the extreme value distribution is different.

Distributions of the maximums of randomly drawn samples from a distribution are called extreme value distributions (Gumbel, 1958) that can be categorized as Type 1, Type 2, or Type 3. The most common extreme value distribution is Type 1, or the Gumbel distribution. Based on the error distributional assumption of the Gumbel distribution (Type I extreme value), the most known discrete choice model is the MNL model. However, MNL models rely on the independence of irrelevant alternatives (IIA) assumption, which states that the odds of having one outcome category over another do not depend on the presence or absence of other categories. The IIA assumption is violated when there is correlation

among multiple categories, causing the MNL model to generate biased estimates. Nested logit (NL) or mixed logit (ML) models offer a more appropriate methodological approach when IIA cannot be held.

4.4.1 Multinomial logit model

The MNL model has been widely applied in crash severity studies to predict the probability of different crash outcomes. If ε_{ni} is considered known, Eq. (4.3) is the cumulative distribution for ε_{ni} evaluated at $V_{ni}+\varepsilon_{ni}-V_{nj}$. When Gumbel is assumed[1] and ε_{ni}'s are independent, the cumulative distribution over all $j \neq i$ is the product of individual cumulative distributions as (Train, 2009):

$$P_{ni}\Big|\varepsilon_{ni} = \prod_{j\neq i} e^{-\exp[-(V_{ni}+\varepsilon_{ni}-V_{nj})]} \tag{4.4}$$

As ε_{ni} is not given, the choice probability is the integral of $P_{ni}|\varepsilon_{ni}$ over all values of ε_{ni} weighted by its density as follows:

$$P_{ni} = \int \left\{\prod_{j\neq i} e^{-\exp[-(V_{ni}+\varepsilon_{ni}-V_{nj})]}\right\} e^{-\varepsilon_{ni}} e^{-\exp(-\varepsilon_{ni})} d\varepsilon_{ni} \tag{4.5}$$

This results in a closed-form expression known as the MNL model, formulated as follows:

$$P_{ni} = \Pr(y_n = i) = \frac{\exp(\mathbf{x}'_{ni}\boldsymbol{\beta}_i)}{\sum_{i=1}^{I}\exp(\mathbf{x}'_{ni}\boldsymbol{\beta}_i)} \tag{4.6}$$

where $\mathbf{x}_n$ is a vector of explanatory variables that determines the severity of crash observation n, and $\boldsymbol{\beta}_i$ is a vector of estimable coefficients for injury severity level i,$(i \in I)$.

The estimated coefficients $\boldsymbol{\beta}_i$ are usually presented as a log odds ratio between the probability of a given level i and the reference one, resulting in $(I\text{-}1)$ estimates for each independent variable. The odds ratio is defined as the ratio between the probabilities of two specific categories, and it quantifies the propensity of an observation falling into one category compared with another. If level I is the reference level, the model becomes

$$\log\left[\frac{P_n(i)}{P_n(I)}\right] = \mathbf{x}'_{ni}\boldsymbol{\beta}_i \tag{4.7}$$

Note that in crash severity modeling, the lowest injury severity level, $i = 1$, (i.e., "no injuries" or "property damage only" (PDO)) is usually set to be the reference level instead of level I. The latter, however, is a more common choice in commercial statistical software.

[1] Gumbel distribution: $f(x) = e^{-x}e^{-\exp(-x)}$; and $F(x) = e^{-\exp(-x)}$.

Eq. (4.7) shows that the MNL model allows the explanatory variables related to one injury severity, as well as their parameter estimations, to vary. Thus, the MNL model should be an appropriate model when possibilities of different injury severities are related with different contributing factors or are affected differently by the same factor.

Another unique property of the MNL model is the IIA assumption. According to Eq. (4.6), the ratio of any two alternatives A and B is $P(A)/P(B) = \exp(V_A - V_B)$, which is unaffected by another alternative. This property appears to be a major restriction for the use of MNL. Therefore, when the alternatives are distinctly different and independent, the MNL model should work well. Conversely, the MNL cannot be justified when the alternatives share some unobserved effects, and different modeling approaches should be considered.

Maximum likelihood estimate (MLE) is a method for estimating model coefficients. ML estimators are known to have good properties in large samples because they are consistent, asymptotically efficient, and asymptotically normal. In statistical terms, "consistency" means the estimate approaches the true value as the sample size increases indefinitely. Asymptotic efficiency means that in large samples, the estimates have standard errors that are at least as small as those of any other methods. Asymptotically normal means the normal and chi-square distributions can be used to construct confidence intervals and calculate *P*-values for the coefficients.

Exercise 4.1

Estimate a multinomial logit model using the Large Trucks Dataset.

Crashes involving large trucks are generally more severe than those involving other vehicles due to the size, weight, and speed differential between trucks and other vehicles. The exercise is adapted from the large truck safety study published by (Qin et al., 2013a,b). The purpose of the exercise is to identify key contributing factors and their impacts on crash severities involving large trucks using MNL models. The large truck crash dataset includes 10,000 traffic accidents, with 4905 (49%) PDO crashes; 3981 (40%) injury type B (possible injury) and C (nonincapacitating injury); and 1114 (11%) injury type A (incapacitating injury) and K (fatal). The MNL model includes the following explanatory variables: human factors and driver behavior (Young, Old, Female, Alcohol, Drugs, Safety constraints, Speed, Rule violation, Reckless behavior), highway and traffic conditions (Signal, Two-way, None, Total units), environmental factors (Snow, Ice, Wet, and Dark). The following steps are taken to solve the problem.

continued

Exercise 4.1 (*cont'd*)

First, determine the functional form: $P_{ni} = \Pr(y_n = i) = \frac{\exp(\mathbf{x}'_{ni}\boldsymbol{\beta}_i)}{\sum_{i=1}^{I}\exp(\mathbf{x}'_{ni}\boldsymbol{\beta}_i)}$

In this functional form, y_n is the crash injury severity with three levels: PDO ($i = 1$), B or C ($i = 2$); and K or A ($i = 3$). $\mathbf{x}_n$ is a vector of explanatory variables that determines the severity of crash observation n ($n = 1{,}10000$), and β_i is a vector of coefficients for injury severity level i.

Second, estimate the coefficients using the R "mlogit" package:

crash_mnl <- mlogit.data(data_model_ch4, shape = "wide", choice = "INJSVR")

multi_logit <- mlogit(INJSVR ~ 0 | YOUNG + OLD + FEMALE + ALCFLAG + DRUGFLAG + SAFETY + DRVRPC_SPD + DRVRPC_RULEVIO + DRVRPC_RECK + TRFCONT_SIGNAL + TRFCONT_2WAY + TRFCONT_NONE + TOTUNIT + ROADCOND_SNOW + ROADCOND_ICE + ROADCOND_WET + LGTCOND_DARK, data = crash_mnl). The first row of R codes is to shape a data.frame in a suitable form for the mlogit function. The input data file "data_model_ch4" is a wide data.frame (i.e., each row is an observation, as compared to the long form if each row is an alternative). The second row is the model specification. In addition, you can use scripts summary(multi_logit) and AIC(multi_logit) to compute model performance such as AIC, log-likelihood value. It is important to know that in the mlogit package, you need to convert a categorical variable with n values to $(n-1)$ dummy variables. If you choose other R packages such as the multinomial function, then you do not have to transform the data.

Third, present the results of the coefficients.

	B or C			K or A		
Variable	**Estimate**	**Std. Error**	**Pr(>|z|)**	**Estimate**	**Std. Error**	**Pr(>|z|)**
Intercept	−1.1248	0.3605	0.0018	−3.0400	0.5082	0.0000
Young	0.0903	0.0595	0.1289	0.2779	0.0907	0.0022
Old	−0.0429	0.0595	0.4711	0.5054	0.0847	0.0000
Female	0.8380	0.0530	0.0000	0.5557	0.0807	0.0000
Alcohol	0.1466	0.1286	0.2542	0.8121	0.1556	0.0000
Drugs	1.5013	0.4077	0.0002	2.5610	0.4193	0.0000
Safety constraints	−0.7098	0.3319	0.0325	−0.9989	0.4369	0.0222
Speed	0.5290	0.0556	0.0000	0.6722	0.0849	0.0000

Exercise 4.1 (*cont'd*)

	B or C			K or A		
Variable	**Estimate**	**Std. Error**	**Pr(>|z|)**	**Estimate**	**Std. Error**	**Pr(>|z|)**
Rule violation	0.3192	0.0611	0.0000	0.9368	0.0888	0.0000
Reckless behavior	0.2263	0.0507	0.0000	0.3559	0.0767	0.0000
Signal	0.6930	0.1471	0.0000	0.6216	0.2727	0.0227
Two-way	0.7419	0.1555	0.0000	1.3280	0.2709	0.0000
None	0.4295	0.1379	0.0018	0.8710	0.2580	0.0007
Total units	0.3264	0.0269	0.0000	0.3849	0.0358	0.0000
Snow	−0.6935	0.0753	0.0000	−1.0676	0.1305	0.0000
Ice	−0.5375	0.1080	0.0000	−0.7336	0.1836	0.0001
Wet	0.0467	0.0675	0.4891	−0.3037	0.1113	0.0064
Dark	0.0991	0.0613	0.1059	0.3775	0.0901	0.0000

AIC: 17,869.32; Log-Likelihood: −8898.7, McFadden R^2: 0.062873.

Finally, summarize your findings. In the MNL model, the coefficient estimates are explained as the comparison between injury level i and the base level PDO ($i = 1$). As can be seen in the table, a driver usually sustained more severe injuries when alcohol or drugs were involved. If a driver was influenced by drugs, his or her chance of getting injured increases drastically, with respective probabilities of level B or C and level K or A being 4.49 ($e^{1.5013}$) times and 12.95 (or $e^{2.561}$) times that of PDO. The exponentiated value of the logit coefficients is also called the odds ratio. Other factors relating to unsafe driving behavior, such as speeding, violating the traffic rules, and driving recklessly, all suggest an increased probability of serious injuries.

4.4.2 Nested logit model

McFadden (1981) developed the generalized extreme value model (GEV) to overcome the IIA limitation in the MNL model. The NL model is the most well-known GEV model. The NL model generates two kinds of crash outcomes: those that are part of a nest (crash outcomes that are

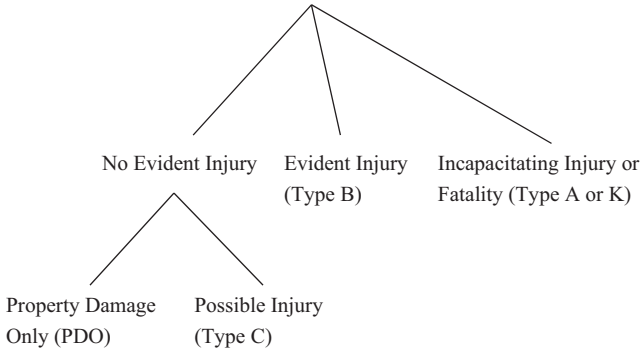


FIGURE 4.3 Nested structure of accident severities.

correlated) and those that are not. By grouping outcomes that share unobserved effects into conditional nests, the shared unobserved effects are canceled out in each nest. Shankar et al. (1996) observed that the "property damage only" and "possible injury" severity levels were correlated due to shared unobserved factors, which is a sign that IIA has been violated. The authors proposed the nested structure shown in Fig. 4.3. The structure combines the two severity levels into one nest named "No Evident Injury," which is independent of the other two nests (evident injury and disabling injury or fatality). The crash severity probabilities for the nested outcome in the NL model consist of the nest probability as well as the outcome probability inside the nest.

Assuming the disturbances are generalized extreme value distributed, the nested logit can be formulated as (see McFadden, 1981):

$$P_{ni} = \frac{\exp[\mathbf{x}'_{ni}\boldsymbol{\beta}_i + \varphi_i L_{ni}]}{\sum_{\forall I}\exp[\mathbf{x}'_{nI}\boldsymbol{\beta}_I + \varphi_I LS_{nI}]} \tag{4.8a}$$

$$P_n(j|i) = \frac{\exp\left[\mathbf{x}'_{nj}\boldsymbol{\beta}_{j|i}\right]}{\sum_{\forall J}\exp\left[\mathbf{x}'_{nJ}\boldsymbol{\beta}_{J|i}\right]} \tag{4.8b}$$

$$LS_{ni} = LN\left[\sum_{\forall J}\exp\left(\mathbf{x}'_{nJ}\boldsymbol{\beta}_{J|i}\right)\right] \tag{4.8c}$$

where P_{ni} is the unconditional probability of crash n resulting in injury outcome i; $\mathbf{x}$ are vectors of characteristics that determine the probability of injury severity, and $\boldsymbol{\beta}$ are vectors of estimable parameters of injury outcome. $P_n(j\,|\,i)$ is the probability of crash observation n having injury outcome j conditioned on the injury outcome being in category i. J is the conditional set of outcomes on i. I is the unconditional set of outcome categories (for example, the upper three branches in Fig. 4.3: "no evident injury," "evident injury," and "disabling injury or fatality"), LS_{ni} is the inclusive value (logsum) that can be considered as the expected

maximum value of the attributes that determine probabilities in severity category i, and φ_i is an estimable parameter.

According to Eq. (4.8b), grouping property damage only and possible injury crashes that share common unobserved effects in "no evident injury" can cancel out the unobserved effects and therefore preserve the independence assumption.

To be consistent with McFadden's generalized extreme value derivation of the model, the parameter estimate for φ_i in the nested logit model must be between zero and one. If φ_i equals to one or is not significantly different from one, there is no correlation between the severity levels in the nest, meaning the model reduces to the multinomial logit model. If φ_i equals to zero, a perfect correlation is implied among the severity levels in the nest, indicating a deterministic process by which crashes result in particular severity levels. The t-test can be used to test if φ_i is significantly different from 1: $t = \frac{\varphi_i - 1}{S.E.(\varphi_i)}$. Because φ_i is less than or equal to one, this is a one-tailed t-test (half of the two-tailed t-test) (more details about the t-test can be found in Chapter 5 - *Exploratory Analyses of Safety Data*). It is important to note that the typical t-test implemented in many commercial software packages are against zero instead of one. Thus, the t value must be calculated manually. The IIA assumption for an MNL model can also be tested with the Hausman-McFadden (1984) test that has been widely implemented in commercial statistical software.

Model estimates can be produced in a sequential fashion (i.e., estimating the conditional model as in Eq. 4.8b) using only the data in the sample that observed the subset of injury outcomes J; then, the logsum in Eq. (4.8c) is calculated using all observations, including those with injury severity J and those without. Lastly, the calculated logsums can be used as an independent variable in Eq. (4.8a). The sequential estimation procedure, however, may generate small variance−covariance matrices that lead to inflated t-statistics. The full information maximum likelihood (FIML) estimation will not have this problem. Its log-likelihood is $\ln L = \sum_i \ln\left[\text{Prob}(\text{twig}|\text{branch})_i \times \text{Prob}(\text{branch})\right]$ where twig is the outcome in the nest and branch is not. FIML is more efficient than two-step estimation and it ensures appropriate estimation of variance-covariance matrices (see Greene, 2000 for additional details).

Exercise 4.2

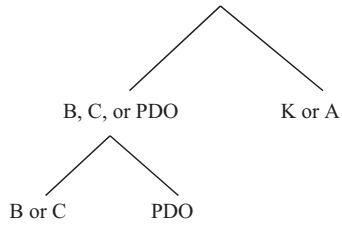
Estimate a nested logit model using the Large Trucks Dataset and compare the model results with Exercise 4.1.

The exercise uses the same dataset as Exercise 4.1. Crash injury severity levels include PDO, injury type B (possible injury), C (noninca-pacitating injury), injury type A (incapacitating injury) and K (fatal). The solution is as follows.

continued

Exercise 4.2 (*cont'd*)

First, establish the nested structure of crash severities.



Second, determine the functional form based on Eq. 4.8 (a–c). For example, $P_n(j|i)$ is the probability of crash n having injury outcome B or C conditioned on the injury outcome being in a category not a K or A injury. I is the unconditional set of outcome categories (for example, the upper three branches in the figure: no K/A injury and K/A injury). LS_{ni} is the inclusive value (logsum).

Third, estimate the coefficients using the R "mlogit" package:

nested_logit <- mlogit(INJSVR ~ 0 | YOUNG + OLD + FEMALE + ALCFLAG + DRUGFLAG + SAFETY + DRVRPC_SPD + DRVRPC_RULEVIO + DRVRPC_RECK + TRFCONT_SIGNAL + TRFCONT_2WAY + TRFCONT_NONE + TOTUNIT + ROADCOND_SNOW + ROADCOND_ICE + ROADCOND_WET + LGTCOND_DARK, data = crash_mnl, nests = list(KA = c("3"), non_KA = c("1", "2")), un.nest.el = TRUE). To obtain the model performance, use summary(nested_logit) and AIC(nested_logit).

Fourth, present the results of the coefficients.

	B or C (lower nest)			K or A		
Variable	**Estimate**	**Std. Error**	**Pr(>|z|)**	**Estimate**	**Std. Error**	**Pr(>|z|)**
Intercept	−0.8087	0.7161	0.2588	−3.0882	0.4798	0.0000
Young	0.0632	0.0677	0.3507	0.2658	0.0920	0.0039
Old	−0.0278	0.0487	0.5680	0.5136	0.0839	0.0000
Female	0.6021	0.5037	0.2320	0.4299	0.2815	0.1267
Alcohol	0.0975	0.1219	0.4239	0.7874	0.1506	0.0000
Drugs	1.0946	0.9586	0.2535	2.3048	0.6835	0.0007
Safety constraints	−0.5025	0.4731	0.2881	−0.8868	0.4541	0.0508

Exercise 4.2 (*cont'd*)

	B or C (lower nest)			K or A		
Variable	**Estimate**	**Std. Error**	**Pr(>|z|)**	**Estimate**	**Std. Error**	**Pr(>|z|)**
Speed	0.3802	0.3187	0.2330	0.5965	0.1864	0.0014
Rule violation	0.2284	0.1952	0.2420	0.8945	0.1265	0.0000
Reckless behavior	0.1621	0.1394	0.2447	0.3264	0.0980	0.0009
Signal	0.4969	0.4280	0.2457	0.5304	0.3382	0.1168
Two-way	0.5331	0.4587	0.2451	1.2299	0.3436	0.0003
None	0.3059	0.2746	0.2652	0.8153	0.2862	0.0044
Total units	0.2328	0.1954	0.2337	0.3304	0.1208	0.0062
Snow	−0.5044	0.4233	0.2334	−0.9792	0.2309	0.0000
Ice	−0.3875	0.3326	0.2439	−0.6646	0.2431	0.0063
Wet	0.0309	0.0550	0.5740	−0.3124	0.1107	0.0048
Dark	0.0756	0.0769	0.3254	0.3650	0.0961	0.0001
iv (inclusive value)	0.7161	0.5989	0.2319			

AIC: 17,871.11; Log-Likelihood: −8898.6; McFadden R^2: 0.062884.

Finally, compare the nested logit model with the multinomial logit model in Exercise 4.1. None of the variables in the lower nest seem to be statistically significant. The AIC value of the NL model (17871.11) is also greater than that of the MNL model (17869.32), indicating inferior performance. The inclusive value is 0.7161, and its t-value is −0.474 (calculated as $\frac{0.7161-1}{0.5989}$). Apparently, the log-sum coefficient is not significantly different from 1. When the inclusive value is equal to one or not significantly different from 1, there is no correlation between the severity levels in the nest, meaning the model reduces to a simple multinomial logit model. We can conclude that for this dataset, the MNL model is more appropriate.

4.4.3 Mixed logit model

The ML model (also known as the random parameters logit model) is highly flexible because it can approximate any random utility model (McFadden and Train, 2000). The mixed logit model addresses the limitations of the multinomial logit by allowing for heterogeneous effects and correlation in unobserved factors (Train, 2009).

The mixed logit is a generalization of the multinomial structure in which the parameter vector $\boldsymbol{\beta}$ can vary across each observation. We can consider the mixed logit probability as a weighted average of the logit function at different values of parameter $\boldsymbol{\beta}$. The weighted average of several functions is called a mixed function, and the density that provides the weights is called the mixing distribution. In crash severity modeling applications, common mixing distributions include normal, lognormal, uniform, and triangular. Thus, the mixed logit is the integral of standard logit probabilities over a density of parameters, specified as (Train, 2009):

$$P_{ni}(i) = \int \frac{\exp(\mathbf{x}'_{ni}\boldsymbol{\beta}_i)}{\sum_J \exp(\mathbf{x}'_{nJ}\boldsymbol{\beta}_J)} f(\boldsymbol{\beta}|\boldsymbol{\phi}) d\boldsymbol{\beta} \tag{4.9}$$

where $f(\boldsymbol{\beta}|\boldsymbol{\phi})$ is a density function of $\boldsymbol{\beta}$, and $\boldsymbol{\phi}$, and is a vector of parameters that specify the density function, with all other terms as previously defined.

The injury severity level probability is a mixture of logits. When all parameters $\boldsymbol{\beta}$ are fixed, the model reduces to the multinomial logit model. When $\boldsymbol{\beta}$ is allowed to vary, the model is not in a closed form, and the probability of crash observation n having a particular injury outcome i can be calculated through integration. Simulation-based maximum likelihood methods such as Halton draws are usually used.

The choice of the density function of $\boldsymbol{\beta}$ depends on the nature of the coefficient and the statistical goodness of fit. The lognormal distribution is useful when the coefficient is known to have the same sign for each observation. Triangular and uniform distributions have the advantage of being bounded on both sides. Furthermore, triangular distribution assumes that the probability increases linearly from the beginning to the midrange and then decreases linearly to the end. A uniform distribution assumes the same probability for any value within the range.

If a coefficient is no longer fixed but random, its interpretation can be tricky because the impact of X on the injury outcome is case-specific. In Milton et al. (2008), the authors suggested that roadway characteristics were better modeled as fixed parameters, while volume-related variables such as average daily traffic per lane, average daily truck traffic, truck percentage, and weather effects were better when modeled as random parameters. The authors also speculated that the random effect of ADT per lane increases injury severity in some cases while it decreases it in others, which captures the response and adaptation of local drivers to various levels of traffic volume. The number of interchanges per mile was

also found to be a random coefficient, suggesting some interchanges may be more complex in design and traffic patterns than others. In Chen and Chen (2011), the authors concurred that weather characteristics such as snowy or slushy surface conditions and a light traffic indictor appeared to be random coefficients. Compared to the fixed parameters, the "randomness" may present new insights for a more comprehensive and better understanding of the complex relationship between observed factors and crash injury outcome.

Exercise 4.3

Estimate a mixed logit model using Large Trucks Dataset.

This exercise uses the same dataset as Exercise 4.1, the presentation of coefficients for fixed parameters in the ML model is the same as the MNL model except when the coefficient is a random variable; in this case, the standard deviation of the coefficient is displayed. For computation efficiency, fewer explanatory variables are tested for the mixed logit model. The explanatory variables are Old, Female, Alcohol, Speed, Snow, and Dark. The dependent variable is the crash injury severity level: PDO ($i = 1$), B or C ($i = 2$); and K or A ($i = 3$). The solution is as follows.

First, determine the density function $f(\boldsymbol{\beta}|\boldsymbol{\phi})$ in the R "mlogit" package, random parameter object "rpar" contains all the relevant information about the distribution of random parameters. Currently, the normal ("n"), log-normal ("ln"), zero-censored normal ("cn"), uniform ("u") and triangular ("t") distributions are available. For illustration, normal distribution is chosen as the density function of random parameter $\boldsymbol{\beta}$.

Second, estimate the coefficients using the R "mlogit" package:

crash_data_mixed <- mlogit.data(data_mixed_ch4, shape = "long", choice = "INJSVR", chid.var = "ID", alt.var = "OUTCOME")

mixed_logit <- mlogit(INJSVR ~ OLD_2 + OLD_3 + FEMALE_2 + FEMALE_3 + ALCFLAG_2 + ALCFLAG_3 + DRVRPC_SPD_2 + DRVRPC_SPD_3 + ROADCOND_SNOW_2 + ROADCOND_SNOW_3 + LGTCOND_DARK_2 + LGTCOND_DARK_3, data = crash_data_mixed, rpar = c(FEMALE_2 = 'n', FEMALE_3 = 'n', ALCFLAG_3 = 'n', DRVRPC_SPD_3 = 'n', ROADCOND_SNOW_2 = 'n', ROADCOND_SNOW_3 = 'n', LGTCOND_DARK_3 = 'n'), panel = FALSE, correlation = FALSE, R = 100, halton = NA). Note that the input data "data_mixed_ch4" is a long data.frame (i.e., each row is an alternative, as compared to the wide data.frame if each row is an observation). choice refers to the injury severity of a crash. chid.var refers to the crash case ID that contains the choice. alt.var refers to the name of the variable that contains the alternative crash outcome. Additionally, the explanatory variables that change by injury severity level should be expanded for each crash outcome except for the base outcome that is PDO in this

continued

Exercise 4.3 (*cont'd*)

exercise. For example, if a crash (ID = 1) involves an old driver with injury severity level 2, the long form should be formatted as follows:

ID	OUTCOME	INJSVR	OLD_2	OLD_3
1	1	0	0	0
1	2	1	1	0
1	3	0	0	1

Third, present the results of the coefficients.

	B or C			K or A		
Variable	**Estimate**	**Std. Error**	**Pr(>|z|)**	**Estimate**	**Std. Error**	**Pr(>|z|)**
Intercept	−0.4882	0.0346	0.0000	−1.9324	0.0637	0.0000
Old	−0.0200	0.0587	0.7340	0.6171	0.0985	0.0000
Female	0.9975	0.0551	0.0000	0.1470	0.6700	0.8264
Alcohol	0.2647	0.1272	0.0375	−3.6985	8.5099	0.6638
Speed	0.3429	0.0513	0.0000	−0.0065	0.4962	0.9896
Snow	−0.8892	0.2158	0.0000	−10.8309	11.2638	0.3363
Dark	−0.0737	0.0616	0.2318	0.2168	0.3050	0.4771
sd. Female_K or A				1.5142	0.8969	0.0914
sd. Alcohol_K or A				−8.7659	12.5946	0.4864
sd. Speed_K/A				−1.0809	0.7895	0.1710
sd. Snow_B or C	1.8159	0.6711	0.0068			
sd. Snow_K or A				8.6369	7.9126	0.2750
sd. Dark_K or A				−0.3950	1.1620	0.7339

AIC: 18,659.75, Log-Likelihood: −9309.9, McFadden R^2: 0.030,777.

Exercise 4.3 (*cont'd*)

Finally, summarize the findings. The ML model can account for the data heterogeneity by treating coefficients as random variables. In this exercise, the coefficients associated with K or A injuries are more of interest and therefore, selected for testing whether or not they are random parameters. The coefficient of Snow is tested for both injury types B or C and injury types K or A to see if snowy pavement has varying effects on injury severity due to driver's risk compensation. According to the model outputs, the snowy surface parameter for truck K or A injuries is fixed (−10.830); and for severity B or C, it is normally distributed with a mean of −0.8892 and a standard deviation of 1.8159, meaning that 31% of truck crashes occurring on snowy pavement have an increased possibility of B or C injuries. It is plausible that people often drive more slowly and cautiously on snowy roads but that the slick conditions still have a tendency to cause crashes.

4.5 Modeling crash severity as an ordered discrete outcome

The primary rationale for using ordered discrete choice models for modeling crash severity is that there is an intrinsic order among injury severities, with fatality being the highest order and property damage being the lowest. Including the ordinal nature of the data in the statistical model defends the data integrity and preserves the information. Second, the consideration of ordered response models avoids the undesirable properties of the multinomial model such as the independence of irrelevant alternatives in the case of a multinomial logit model or a lack of closed-form likelihood in the case of a multinomial probit model. Third, ignoring the ordinality of the variable may cause a lack of efficiency (i.e., more parameters may be estimated than are necessary if the order is ignored). This also increases the risk of obtaining insignificant results.

Although there are many positives to the ordered model, the disadvantage is that imposing restrictions on the data may not be appropriate despite the appearance of a rank. Therefore, it is important to test the validity of the ordered restriction. The rest of the section will introduce three types of ordered choice models: the ordinal probit/logistic model, the generalized ordered and proportion odds model, and the sequential logit/probit model.

4.5.1 Ordinal probit/logistic model

The ordinal logit/probit model applies a latent continuous variable, z_n, as a basis for modeling the ordinal nature of crash severity data. z_n is specified as a linear function of $\mathbf{x}_n$:

$$z_n = \mathbf{x}'_n\boldsymbol{\beta} + \varepsilon_n \tag{4.10}$$

Where $\mathbf{x}_n$ is a vector of explanatory variables determining the discrete ordering (i.e., injury severity) for nth crash observation; $\boldsymbol{\beta}$ is a vector of estimable parameters; and, ε_n is an error term that accounts for unobserved factors influencing injury severity.

A high indexing of z is expected to result in a high level of observed injury y in the case of a crash. The observed discrete injury severity variable y_n is stratified by thresholds as follows:

$$y_n = \begin{cases} 1, & \text{if } z_n \le \mu_1 (\text{PDO or no injury}) \\ 2, & \text{if } \mu_1 < z_n \le \mu_2 (\text{injury C}) \\ 3, & \text{if } \mu_2 < z_n \le \mu_3 (\text{injury B}) \\ 4, & \text{if } \mu_3 < z_n \le \mu_4 (\text{injury A}) \\ 5, & \text{if } \mu_4 < z_n (\text{K or fatal injury}) \end{cases} \tag{4.11}$$

where the μ_s are estimable thresholds, along with the parameter vector $\boldsymbol{\beta}$. The model is estimated using maximum likelihood estimation (Greene, 2000).

If the random error term ε is assumed to follow a standard normal distribution, the model is an ordered probit model. The probabilities associated with the observed responses of an ordered probit model are as follows:

$$\begin{aligned} P_n(1) &= \Pr(y_n = 1) = \Pr(z_n \le \mu_1) = \Pr(\mathbf{x}'_n\boldsymbol{\beta} + \varepsilon_n \le \mu_1) = \Pr(\varepsilon_n \le \mu_1 - \mathbf{x}'_n\boldsymbol{\beta}) \\ &= \Phi(\mu_1 - \mathbf{x}'_n\boldsymbol{\beta}) \\ P_n(2) &= \Pr(y_n = 2) = \Pr(\mu_1 < z_n \le \mu_2) = \Pr(\mu_1 < \mathbf{x}'_n\boldsymbol{\beta} + \varepsilon_n \le \mu_2) \\ &= \Pr(\varepsilon_n \le \mu_2 - \mathbf{x}'_n\boldsymbol{\beta}) - \Pr(\varepsilon_n \le \mu_1 - \mathbf{x}'_n\boldsymbol{\beta}) \\ &= \Phi(\mu_2 - \mathbf{x}'_n\boldsymbol{\beta}) - \Phi(\mu_1 - \mathbf{x}'_n\boldsymbol{\beta}) \\ &\vdots \\ P_n(i+1) &= \Phi(\mu_{i+1} - \mathbf{x}'_n\boldsymbol{\beta}) - \Phi(\mu_i - \mathbf{x}'_n\boldsymbol{\beta}) \\ &\vdots \\ P_n(I) &= \Pr(y_n = I) = \Pr(z_n > \mu_{I-1}) \\ &= 1 - \Phi(\mu_{I-1} - \mathbf{x}'_n\boldsymbol{\beta}) \end{aligned} \tag{4.12}$$

where i is the ith level of injury and I represents the highest injury level (i.e., fatal). $\Phi()$ is the cumulative standard normal distribution.

The ordinal logistic model, also called the cumulative logit model, is formulated when the discrete outcomes are treated as the cumulative distribution of the response. Let $\Pr(y_n > i)$ represent the cumulative probabilities of the observation y_n belonging to categories higher than i, we specify the ordinal logistic model as:

$$\log\left(\frac{\Pr(y_n > i)}{1 - \Pr(y_n > i)}\right) = \alpha_i + \mathbf{x}'_n\boldsymbol{\beta} \quad (i = 1, \ldots, I-1) \tag{4.13}$$

Where α_i is different for each of the equation. $\boldsymbol{\beta}$ is a single set of coefficients.

$$\Pr(y_n > i) = \frac{\exp(\alpha_i + \mathbf{x}'_n\boldsymbol{\beta})}{1 + \exp(\alpha_i + \mathbf{x}'_n\boldsymbol{\beta})} \tag{4.14}$$

We can also derive the formulation based on latent variable Z_n and assume the error term ε_n to be logistically distributed across observations whose CDF is $F(\varepsilon_n) = \frac{\exp(\varepsilon_n)}{1+\exp(\varepsilon_n)}$. The equation is as follows.

$$\begin{aligned}\Pr(y_n > i) &= \Pr(Z_n > \mu_i) = \Pr(\varepsilon_n > \mu_i - \mathbf{x}'_n\boldsymbol{\beta}) \\ &= \frac{1}{1 + \exp(\mu_i - \mathbf{x}'_n\boldsymbol{\beta})} = \frac{\exp(\mathbf{x}'_n\boldsymbol{\beta} - \mu_i)}{1 + \exp(\mathbf{x}'_n\boldsymbol{\beta} - \mu_i)}\end{aligned} \tag{4.15}$$

As can be found, Eqs. (4.14) and (4.15) have the same form except for different symbols for the intercept variable.

It is worth pointing out that in Eq. (4.13), we assume that the regressors $\mathbf{x}_n$ do not include a column of ones because the constant is absorbed in the cutpoints (i.e., thresholds) $-\mu_i$ or α_i. Due to the increasing nature of the thresholds, the positive sign of β indicates higher injury severity when the value of the associated variables increases, while a negative sign suggests the opposite. β does not depend on the placement of the threshold and stays the same across categories. The threshold values affect the intercepts and the relative numbers of crashes that are located in different categories.

McCullagh (1980) refers to ordinal models as proportional odds models because the covariates $\boldsymbol{X}$ increases or decreases the odds of a response in the category higher than i by the factor $\exp(\mathbf{x}'_n\boldsymbol{\beta})$, meaning the effect is a proportionate change for all response categories. In contrast to the MNL coefficient $\boldsymbol{\beta_i}$, which varies by the injury outcome i, one important restriction associated with the $\boldsymbol{\beta}$ of an ordered logit/probit model is the proportional odds assumption (i.e., the parallel regression assumption or the parallel lines assumption). The use of the order probit/logit may be inappropriate if this assumption is violated.

The proportional odds assumption restriction also creates an unintended consequence concerning how the explanatory variables affect the probabilities of the discrete outcome. Consider a model of three injury levels—no injury, injury, and fatality. Suppose that one of the contributing factors in determining the level of the injury is airbag. As shown in

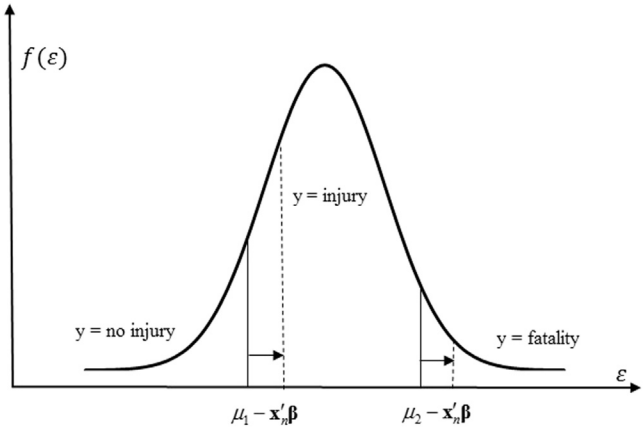


FIGURE 4.4 Illustration of an ordered model with an increase in.$\mathbf{x}'_n\boldsymbol{\beta}$.

Fig. 4.4, a negative parameter of the airbag indicator (1 if it was deployed and zero otherwise) implies that $-\mathbf{x}'_n\boldsymbol{\beta}$ becomes greater and hence, shifts values to the right on the x-axis. Thus, the model constrains the effect of the seatbelt to simultaneously decrease the probability of a fatality and increase the no injury probability. We know for a fact that the activation of an airbag may cause injury and thus, decrease no injury; but unfortunately, ordered models cannot account for this bidirectional possibility because the shift in thresholds is constrained to move in the same direction.

Exercise 4.4

Estimate an ordinal probit model and an ordinal logistic model using the Large Trucks Dataset.

This exercise uses the same dataset as Exercise 4.1. In this exercise, an ordinal probit and an ordinal logistic regression model are respectively applied to recognize the ordinality of injury level, the dependent variable. The problem is solved using the following steps.

First, determine the functional form: Eq. (4.12) for the ordinal probit model and Eq. (4.15) for the ordinal logistic model. In both equations, the μs are estimable thresholds, along with the parameter vector β.

Second, estimate the coefficients using the R "ordinal" package:

crash_data_ordinal <- data_model_ch4

op_model <- clm(as.factor(INJSVR) ~ YOUNG + OLD + FEMALE + ALCFLAG + DRUGFLAG + SAFETY + DRVRPC_SPD + DRVRPC_RULEVIO + DRVRPC_RECK + TRFCONT_SIGNAL + TRFCONT_2WAY + TRFCONT_NONE + TOTUNIT + ROADCOND_SNOW + ROADCOND_ICE + ROADCOND_WET+ LGTCOND_DARK,

Exercise 4.4 (*cont'd*)

data = crash_data_ordinal, link = "probit")

Note that the response (INJSVR) should be a factor, which will be interpreted as an ordinal response with levels ordered as in the factor. Replace "probit" with "logit" if you want to run an ordinal logit model. Other distribution options are as follows: "cloglog", "loglog", "cauchit", "Aranda-Ordaz", "log-gamma".

Third, present the model results of the coefficients.

	Ordinal probit model			Ordinal logit model		
Variable	**Estimate**	**Std. Error**	**Pr(>|z|)**	**Estimate**	**Std. Error**	**Pr(>|z|)**
Young	0.0963	0.0309	0.0019	0.1514	0.0521	0.0037
Old	0.1285	0.0307	0.0000	0.1961	0.0520	0.0002
Female	0.3398	0.0270	0.0000	0.6116	0.0454	0.0000
Alcohol	0.2977	0.0623	0.0000	0.4945	0.1082	0.0000
Drugs	1.0187	0.1459	0.0000	1.7663	0.2455	0.0000
Safety constraints	−0.4321	0.1642	0.0085	−0.7798	0.2881	0.0068
Speed	0.3090	0.0288	0.0000	0.5299	0.0488	0.0000
Rule violation	0.3329	0.0315	0.0000	0.5416	0.0534	0.0000
Reckless behavior	0.1569	0.0264	0.0000	0.2644	0.0446	0.0000
Signal	0.3353	0.0799	0.0000	0.5744	0.1342	0.0000
Two-way	0.5364	0.0830	0.0000	0.9019	0.1403	0.0000
None	0.3190	0.0752	0.0000	0.5144	0.1265	0.0000
Total units	0.1635	0.0120	0.0000	0.2867	0.0210	0.0000
Snow	−0.4450	0.0400	0.0000	−0.7666	0.0682	0.0000
Ice	−0.3358	0.0577	0.0000	−0.5727	0.0974	0.0000
Wet	−0.0615	0.0357	0.0844	−0.0906	0.0596	0.1288
Dark	0.1237	0.0319	0.0001	0.2108	0.0540	0.0001
Threshold coefficients						
1|2	0.5518	0.1811		0.8867	0.3154	
2|3	1.8791	0.1818		3.1692	0.3171	
AIC	18,072.64			18,036.83		

continued

Exercise 4.4 (*cont'd*)

Finally, summarize the findings. The positive coefficients suggest the likelihood of more severe injuries. Thus, all explanatory variables except for the use of safety constraints and adverse weather are associated with more severe injuries. The coefficient estimates of both models are consistent in signs and magnitudes. The threshold coefficients of the ordinal logistic model are greater than these of the ordinal probit model. According to the AIC value, the ordinal logistic model is slightly better than the ordinal logit model.

4.5.2 Generalized ordered logistic and proportional odds model

A generalized ordered logistic model (gologit) provides results similar to those that result from running a series of binary logistic regressions/cumulative logit models. The ordered logit model is a special case of the gologit model where the coefficients $\boldsymbol{\beta}$ are the same for each category. The partial proportional odds model (PPO) is in between, as some of the coefficients $\boldsymbol{\beta}$ are the same for all categories and others may differ. A gologit model and an MNL model, whose variables are freed from the proportional odds constraint, both generate many more parameters than an ordered logit model. A PPO model allows for the parallel lines/proportional odds assumption to be relaxed for those variables that violate the assumption.

In the gologit model, the probability of crash injury for a given crash can be specified as $(I-1)$ set of equations:

$$\Pr(y_n > i) = \frac{\exp(\mathbf{x}'_n\boldsymbol{\beta}_i - \mu_i)}{1+\exp(\mathbf{x}'_n\boldsymbol{\beta}_i - \mu_i)}, \quad i = 1, \ldots (I-1) \tag{4.16}$$

where μ_i is the cut-off point for the ith cumulative logit. Note that Eq. (4.16) is different from Eq. (4.14) in that $\boldsymbol{\beta}_i$ is a single set of coefficients that vary by category i.

In the PPO model formulation, it is assumed that some explanatory variables may satisfy the proportional odds assumption while some may not. The cumulative probabilities in the PPO model are calculated as follows (Peterson and Harrell, 1990):

$$\Pr(y_n > i) = \frac{\exp\left(\mathbf{x}'_n\boldsymbol{\beta} + \mathbf{T}'_n\boldsymbol{\gamma}_i - \mu_i\right)}{1+\exp\left(\mathbf{x}'_n\boldsymbol{\beta} + \mathbf{T}'_n\boldsymbol{\gamma}_i - \mu_i\right)}, \quad i = 1, \ldots (I-1) \tag{4.17}$$

where $\mathbf{x}_n$ is a $(p \times 1)$ vector of independent variables of crash n, $\boldsymbol{\beta}$ is a vector of regression coefficients, and each independent variable has a β coefficient. $\mathbf{T}_n$ is a $(q \times 1)$ vector $(q \leq p)$ containing the values of crash n on the subset of p explanatory variables for which the proportional odds assumption is not assumed, and $\boldsymbol{\gamma}_i$ is a $(q \times 1)$ vector of regression coefficients. So, $\boldsymbol{\gamma}_i$ represents a deviation from the proportionality $\boldsymbol{\beta}$ and $\mathbf{T}_n^{'}\boldsymbol{\gamma}_i$ is an increment associated only with the ith cumulative logit, $i = 1, \ldots, (I-1)$.

An alternative but a simplified way to think about the PPO model is to have two sets of explanatory variables: $\mathbf{x}_1$, the coefficients of which remain the same for all injury severities and $\mathbf{x}_2$, the coefficients of which vary across injury severities. Note that $\mathbf{x}_1$ and $\mathbf{x}_2$ have no common variables. The PPO model is specified in Eq. (4.18):

$$\Pr(y_n > i) = \frac{\exp(\mathbf{x'}_1\boldsymbol{\beta}_1 + \mathbf{x'}_2\boldsymbol{\beta}_{2i} - \mu_i)}{1 + \exp(\mathbf{x'}_1\boldsymbol{\beta}_1 + \mathbf{x'}_2\boldsymbol{\beta}_{2i} - \mu_i)} \tag{4.18}$$

where $\boldsymbol{\beta}_1$ is a vector of parameters to be estimated for $\mathbf{x}_1$ and is the same for all injury severities, and $\boldsymbol{\beta}_{2i}$ is a vector of parameters to be estimated for $\mathbf{x}_2$ and varies across injury severities. The proportion assumption dictates whether a coefficient is the same or different. Parameterization in Eqs. (4.17) or (4.18) depends on the statistical software packages (Williams, 2006).

The gologit/PPO model has been applied in several recent studies as an extension of the ordered logit model, and results show that they consistently outperform conventional ordered response models (Wang and Abdel-Aty, 2008; Qin et al., 2013a,b; Yasmin and Eluru, 2013). According to Williams (2016), the gologit/PPO model usually provides a substantially better fit to the data than the ordered logit model and is also much more parsimonious than other alternatives. However, interpretation and justification are less straightforward for the gologit model than it is for the ordered logit model. A test devised by Brant (1990) is commonly used to test parallel regression assumption. The Brant test is available in most of the statistical software packages.

Exercise 4.5

Estimate a PPO model using the Large Trucks Dataset.

The PPO model in this exercise uses the same crash data as Exercise 4.1. In the ordinal model, the coefficients use cumulative probability to estimate the log odds ratio between all possibilities of injury severities higher than level i and all possibilities of injury severities lower than

continued

Exercise 4.5 (*cont'd*)

and equal to level I (see Eq. 4.13). However, it is important to point out that some statistical software packages use $\log\left(\frac{\Pr(y_n>i)}{1-\Pr(y_n>i)}\right) = \alpha_i + \mathbf{x}'_n\boldsymbol{\beta} \quad (i=1,\ldots,I-1)$, where the log odds ratio is for a lower level. The solution is as follows.

First, determine the functional form: Eq. (4.18) for the PPO model where β_1 is a vector of parameters to be estimated for $\mathbf{x}_1$ and is the same for all injury severities, and β_{2i} is a vector of parameters to be estimated for $\mathbf{x}_2$ and varies across injury severities.

Second, the coefficients are estimated using the R "ordinal" package:

ppo_model <- clm(as.factor(INJSVR) ~ YOUNG + ALCFLAG + DRUGFLAG + SAFETY + DRVRPC_SPD + DRVRPC_RULEVIO+ DRVRPC_RECK +TRFCONT_SIGNAL+ TRFCONT_2WAY + TRFCONT_NONE + TOTUNIT +LGTCOND_DARK,

nominal = ~ OLD + FEMALE + ROADCOND_SNOW + ROADCOND_WET + ROADCOND_ICE,

data = crash_data_ordinal, link = "logit")

As can be seen, nominal effects (e.g., Old, Female) are where the assumption that regression parameters have the same effect across all thresholds is relaxed. Other estimates (e.g., Young, Alcohol, Drug, Safety constraints) are fixed for all severities. Readers can decide which coefficient should be a fixed value or not based on experience or trial and error.

Third, present the results of the coefficients.

Variable	Estimate	Std. Error	Pr(>|z|)
Young	0.1482	0.0529	0.0051
Alcohol	0.4990	0.1087	0.0000
Drug	1.7142	0.2449	0.0000
Safety constraints	−0.7610	0.2856	0.0077
Speed	0.5247	0.0490	0.0000
Rule violation	0.5410	0.0537	0.0000
Reckless behavior	0.2654	0.0448	0.0000
Signal	0.5770	0.1355	0.0000
Two-way	0.9097	0.1416	0.0000
None	0.5176	0.1277	0.0001
Total units	0.2827	0.0210	0.0000
Dark	0.2049	0.0542	0.0002

Exercise 4.5 (*cont'd*)

Variable	Estimate	Std. Error	Pr(>|z|)
Threshold coefficients			
1|2.(Intercept)	0.9253	0.3136	
2|3.(Intercept)	3.0458	0.3161	
1|2.Old	−0.0980	0.0548	
2|3.Old	−0.4868	0.0764	
1|2.Female	−0.7905	0.0506	
2|3.Female	−0.1250	0.0728	
1|2.Snow	0.7558	0.0703	
2|3.Snow	0.8447	0.1196	
1|2.Wet	0.0224	0.0644	
2|3.Wet	0.3475	0.1044	
1|2.Ice	0.5680	0.1009	
2|3.Ice	0.5654	0.1722	
AIC		17,923.93	

Readers can also run an ANOVA test between the PPO model and the proportional odds model such as the ordered probit model (Exercise 4.4) using "anova(op_model, ppo_model)".

	no.par	AIC	logLik	LR.stat	df	Pr(>Chisq)
op_model	19	18,073	−9017.3			
ppo_model	24	17,924	−8938	158.71	5	$<2.2e^{-16}$

Finally, summarize the findings. The PPO model treats crash severities as the ordinal response variable but allows the coefficients to vary across levels if the proportional odds assumption is violated. The table shows that there is a high chance of severe injuries when the driver is under the influence of alcohol. The odds of sustaining a more severe injury under the influence of alcohol is 1.64 ($e^{0.499}$) across all injury levels. However, the sign of the threshold coefficients (nominal effects) seems to be opposite to those in a proportional odds model. For example, the coefficients of variable Old are −0.0980 from PDO to K/A/B/C and −0.4868 from PDO/B/C to K/A in this exercise. They are 0.1285 and 0.1961 in

continued

Exercise 4.5 (*cont'd*)

the ordinal probit and ordinal logistic model, respectively in Exercise 4.4. The plausible reason is that the threshold coefficients may be coded differently in "clm" in the R "ordinal" package (Christensen, 2018). It is a practice in some statistical software packages (e.g., SPSS and Stata) to flip the sign of the coefficients in the proportional odds model so that each odds ratio (e^{β}) represents the factor increase in the odds of moving into a higher ordered category instead of a lower ordered category for each one-unit increase in X.

4.5.3 Sequential logistic/probit regression model

Although the generalized ordered logit model relaxes the proportional odds assumption by allowing some or all of the parameters to vary by severity levels, the set of explanatory variables is invariant over all severity levels. The sequential logit/probit regression model should be considered when the difference in the set of explanatory variables at each severity level is important. Sequential logit/probit regression allows different regression parameters for different severity levels. A sequential logit/probit model supposes $(I-1)$ latent variables given as $(I-1)$ sets of equations:

$$\begin{aligned} z_{n1} &= \boldsymbol{\alpha}_1 + \mathbf{x}'_n\boldsymbol{\beta}_1 + \varepsilon_{n1} \\ z_{n2} &= \boldsymbol{\alpha}_2 + \mathbf{x}'_n\boldsymbol{\beta}_2 + \varepsilon_{n2} \\ &\vdots \\ z_{n,I-1} &= \boldsymbol{\alpha}_{I-1} + \mathbf{x}'_n\boldsymbol{\beta}_{I-1} + \varepsilon_{n,I-1} \end{aligned} \tag{4.19}$$

where z_{ni} is a continuous latent variable that determines whether the injury severity is observed as i or higher, $\boldsymbol{\beta}'_i$s are the vectors of estimated parameters, and ε_{ni}'s are error terms that are independent of $\mathbf{x}_n$.

The sequential model is a type of hierarchical model where lower stages mean lower injury severity. For example, stage 1 of the KABCO scale may be KABC versus O; stage 2 may be KAB versus C and stage 3 may be KA versus B. This change in definition matters when explaining the model results. Moreover, the hierarchical structure can be arranged from low to high or from high to low, which can also be called "forward" or "backward." Use the same example in a "backward" format, stage 1 may be KA versus BCO; stage 2 may be B versus CO; stage 3 may be C versus O. In addition, the choice of contrast will affect the interpretation of

the model results. It is important to know that the sequential model uses a subpopulation of the data to estimate the variant set of $\boldsymbol{\beta}_i$. The subpopulation decreases as the stages progresses forward or backward. In the forward format, all data are used in the first stage to estimate $\boldsymbol{\beta}_1$, but only the crashes with injury type C or higher are used in the second stage to estimate $\boldsymbol{\beta}_2$. Crashes with injury type B or higher are used in the second stage to estimate $\boldsymbol{\beta}_3$.

Observed response variable y_n, is defined as:

$$y_n = \begin{cases} 1, \text{ if } z_{n1} \le 0 \\ 2, \text{ if } z_{n1} < 0 \text{ and } z_{n2} \le 0 \\ 3, \text{ if } z_{n1} < 0, \ z_{n2} < 0 \text{ and } z_{n3} \le 0 \\ \vdots \\ I, \text{ if } z_{n1} < 0, \ z_{n2} < 0, \ldots, \text{ and } z_{n,I-1} \le 0 \end{cases} \tag{4.20}$$

The probability of y_n with different injury severities is written as follows:

$$\begin{aligned} P_n(1) &= \Pr(y_n = 1) = \Lambda(-(\boldsymbol{\alpha}_1 + \mathbf{x}'_n\boldsymbol{\beta}_1)) \\ P_n(2) &= \Pr(y_n = 2) = \Lambda(-(\boldsymbol{\alpha}_1 + \mathbf{x}'_n\boldsymbol{\beta}_1))\Lambda(-(\boldsymbol{\alpha}_2 + \mathbf{x}'_n\boldsymbol{\beta}_2)) \\ P_n(3) &= \Pr(y_n = 3) = \Lambda(-(\boldsymbol{\alpha}_1 + \mathbf{x}'_n\boldsymbol{\beta}_1))\Lambda(-(\boldsymbol{\alpha}_2 + \mathbf{x}'_n\boldsymbol{\beta}_2))\Lambda(-(\boldsymbol{\alpha}_3 + \mathbf{x}'_n\boldsymbol{\beta}_3)) \\ &\vdots \\ P_n(I) &= \Pr(y_n = I) = \prod_{i}^{I-1} \Lambda(-(\boldsymbol{\alpha}_i + \mathbf{x}'_n\boldsymbol{\beta}_i)) \end{aligned} \tag{4.21}$$

where $\Lambda()$ represents the standard logistic CDF for the sequential logit model and the standard normal CDF for the sequential probit model. Take $P_n(1)$ as an example: for the standard logistic CDF, $F(-(\boldsymbol{\alpha}_1 + \mathbf{x}'_n\boldsymbol{\beta}_1)) = \frac{1}{1+\exp(\boldsymbol{\alpha}_1 + \mathbf{x}'_n\boldsymbol{\beta}_1)}$; and for the standard normal CDF, $F(\boldsymbol{\alpha}_1 + \mathbf{x}'_n\boldsymbol{\beta}_1) = \Phi(\boldsymbol{\alpha}_1 + \mathbf{x}'_n\boldsymbol{\beta}_1)$.

The probability of injury level i is the product of individual cumulative functions. This formulation shows one major limitation of the sequential logit/probit model in that the model assumes the independence between error terms. On the other hand, an important practical feature of the hierarchy model is that the multinomial likelihood factors into the product of binomial likelihoods.

Jung et al. (2010) applied the sequential logit model to assess the effects of rainfall on the severity of single-vehicle crashes on Wisconsin interstate highways. The sequential logit regression model outperformed the ordinal logit regression model in predicting crash severity levels in rainy weather when comparing goodness of fit, parameter significance, and

prediction accuracies. The sequential logit model identified that stronger rainfall intensity significantly increases the likelihood of fatal and incapacitating injury crash severity, while this was not captured in the ordered logit model. Yamamoto et al. (2008) also reported superior performance and unbiased parameter estimates with sequential binary models as compared with traditional ordered probit models, even when underreporting was a concern.

4.6 Model interpretation

Statistical modeling is only used as a data-driven tool for measuring the effects of variables on crash injury severity levels. It is the expert domain knowledge that eventually helps to explain what factors cause or contribute to more severe injuries, the safety problem to be solved, and the context within. To properly interpret model results, we need to be wary of the data formats as they can be structured differently because of different methods. The dependent variable can be treated as individual categories, categories higher than level i, or categories lower than level i. Independent variables can be continuous, indicator (1 or 0) or categorical. Categorical variables should be converted to dummy variables, with a dummy variable assigned to each distinct value of the original categories. The coefficient of a dummy variable can be interpreted as the log-odds for that particular value of dummy minus the log-odds for the base value which is 0 (e.g., the odds of being injured when drinking and driving is 10 times of someone who is sober). The percent change in the odds for each 1-unit increase in the continuous independent variable is calculated by subtracting 1 from the odds ratio and then multiplying by 100 or, $100\left(e^{\beta}-1\right)$.

The key concepts of marginal effect and elasticity are fundamental to understanding model estimates. The marginal effect is the unit-level change in y for a single-unit increase in x if x is a continuous variable. In a simple linear regression, the regression coefficient of x is the marginal effect, $\frac{\partial y}{\partial x_k} = \beta_k$. Due to the nonlinear feature of logit models, the marginal effect of any continuous independent variable is: $\frac{\partial p_i}{\partial x_{ki}} = \beta_{ki} p_i (1 - p_i)$. Thus, the marginal effect depends on the logit regression coefficient for x_{ki}, as well as the values of the other variables and their coefficients. If x_{ki} is a discrete variable (indicator or dummy variable), the marginal effect of x_{ki} is $[\Pr(i|\mathbf{x}, x_{ki} = 1) - \Pr(i|\mathbf{x}, x_{ki} = 0)]$. Such marginal effects are called instantaneous rates of change because they are computed for a variable while holding all other variables as constant. The marginal effect at the mean is a popular approach for both continuous and discrete variables in which all $\mathbf{x}$'s are at their mean.

Elasticity can be used to measure the magnitude of the impact of specific variables on the injury-outcome probabilities. For a continuous variable, elasticity is the % change in y given a 1% increase in x. It is computed from the partial derivative with respect to the continuous variable of each observation n. The equation uses the partial derivative of the MNL $P(i)$ to express the elasticity of the continuous variable as (n subscripting omitted):

$$E_{x_{ki}}^{P(i)} = \frac{\partial P(i)}{\partial x_{ki}} \times \frac{x_{ki}}{P(i)} = [1 - P(i)]\beta_{ki} x_{ki} \tag{4.22}$$

where β_{ki} is the estimable coefficient associated with x_{ki}. Elasticity values are the percent effect that a 1% change in x_{ki} has on the injury severity probability $P(i)$.

For indicator or dummy variables (those variables taking on values of 0 or 1), a pseudoelasticity percentage can be written as follows:

$$E_{x_{ki}}^{P(i)} = \left[\frac{\exp[\Delta\beta_i x_i] \sum_{\forall I} exp(\beta_{kI} x_{kI})}{\exp[\Delta\beta_i x_i] \sum_{\forall I_n} exp(\beta_{kI} x_{kI}) + \sum_{\forall I \neq I_n} exp(\beta_{kI} x_{kI})} - 1 \right] \times 100 \tag{4.23}$$

where I_n is the set of injury severity outcomes with x_{ki} in the function determining the outcome, and I is the set of all possible injury severity outcomes. The pseudo-elasticity of an indicator variable with respect to an injury severity category represents the percent change in the probability of that injury severity category when the variable is changed from zero to one.

References

Brant, R., 1990. Assessing proportionality in the proportional odds model for ordinal logistic regression. Biometrics 46, 1171–1178. https://doi.org/10.2307/2532457.

Chen, F., Chen, S., 2011. Injury severities of truck drivers in single- and multi-vehicle accidents on rural highways. Accid. Anal. Prev. 43 (5), 1677–1688.

Christensen, R.H.B., 2018. Cumulative link models for ordinal regression with the R package Ordinal. Submitt. J. Stat. Softw. 1–40.

Greene, W., 2000. Econometric Analysis, fourth ed. Prentice Hall, Upper Saddle River, NJ.

Gumbel, E.J., 1958. Statistics of Extremes. Columbia University Press, New York.

Hausman, J.A., McFadden, D., 1984. A specification test for the multinomial logit model. Econometrica 52, 1219–1240.

Jung, S.Y., Qin, X., Noyce, D.A., 2010. Rainfall effect on single-vehicle crash severities using polychotomous response models. Accid. Anal. Prev. 42 (1), 213–224.

McCullagh, P., 1980. Regression models for ordinal data. J. Roy. Stat. Soc. B 42 (2), 109–127.

McFadden, D., 1981. Econometric models of probabilistic choice. In: Manski, C., McFadden, D. (Eds.), Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge, pp. 198–272.

McFadden, D., Train, K., 2000. Mixed MNL models for discrete response. J. Appl. Econom. 15 (5), 447–470.
Milton, J.C., Shankar, V.N., Mannering, F., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accid. Anal. Prev. 40 (1), 260–266.
Mujalli, R.O., López, G., Garach, L., 2016. Bayes classifiers for imbalanced traffic accidents datasets. Accid. Anal. Prev. 88, 37–51.
Peterson, B., Harrell Jr., F.E., 1990. Partial Proportional Odds Models for Ordinal Response Variables. Appl. Stat. 205–217.
Qin, X., Wang, K., Cutler, C.E., 2013a. Analysis of crash severity based on vehicle damage and occupant injuries. Transport. Res. Rec. 2386 (1), 95–102.
Qin, X., Wang, K., Cutler, C.E., 2013b. Logistic regression models of the safety of large trucks. Transport. Res. Rec. 2392 (1), 1–10.
Savolainen, P.T., Mannering, F., Lord, D., Quddus, M.A., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accid. Anal. Prev. 43 (5), 1666–1676.
Savolainen, P., Mannering, F., 2007. Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes. Accid. Anal. Prev. 39 (5), 955–963.
Shankar, V., Mannering, F., Barfield, W., 1996. Statistical analysis of accident severity on rural freeways. Accid. Anal. Prev. 28 (3), 391–401.
Train, K., 2009. Discrete Choice Methods with Simulation, second ed. Cambridge university press.
Wang, X.S., Abdel-Aty, M., 2008. Analysis of left-turn crash injury severity by conflicting pattern using partial proportional odds models. Accid. Anal. Prev. 40 (5), 1674–1682.
Washington, S.P., Karlaftis, M.G., Mannering, F., 2020. Statistical and Econometric Methods for Transportation Data Analysis, third ed. (Chapman and Hall/CRC).
Williams, 2016. Understanding and interpreting generalized ordered logit models. J. Math. Sociol. 40 (1), 7–20. https://doi.org/10.1080/0022250X.2015.1112384.
Williams, R., 2006. Generalized ordered logit/partial proportional adds models for ordinal dependent variables. Stata J. 6 (1), 58–82.
Yamamoto, T., Hashiji, J., Shankar, V., 2008. Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. Accid. Anal. Prev. 40, 1320–1329.
Yasmin, S., Eluru, N., 2013. Evaluating alternate discrete outcome frameworks for modeling crash injury severity. Accid. Anal. Prev. 59, 506–521.