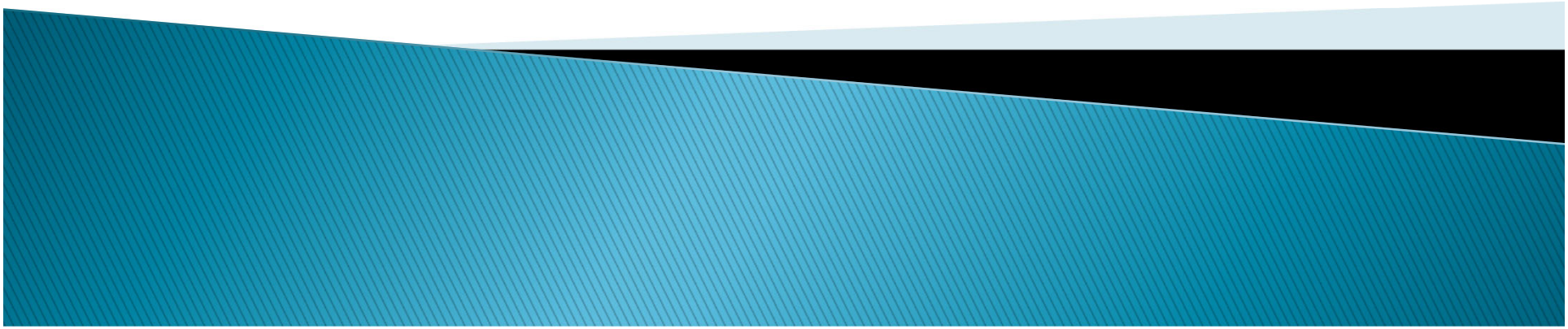
Crash-Severity Modeling

Fall 2021



Recall: Definition (most common)

TABLE 2.3 Comprehensive crash costs (per person) (2018 dollars).

Severity of injuries	Costs
Fatal (K)	$10,855,000
Incapacitating (type A)	$1,187,000
Nonincapacitating (type B)	$327,000
Possible (type C)	$151,000
No injury (property damage only or PDO)	$50,000

Source: NSC.[1]

Values shown above can be used to evaluate highway safety interventions in terms of lives/injuries saved.

Characteristics of Injury Severity Data

- 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 several factors (i.e., driver, vehicle, environment), as discussed in Chapter 2. (see next slide)
- Crash injury severity modeling helps describe, identify, and evaluate the factors contributing to various levels of injury severity.

Characteristics of Injury Severity Data

In crash reports, there are different methods/codes to measure or define injuries (still subjective as it is governed by the opinion of the police officer). In some reports, they classify the injury as the first injury outcome in the sequence of events ("first harmful event"), whereas, in other reports, they defined it as the "most harmful event."

Vehicle swerved to avoid an animal (driver not injured)

↓

Vehicle runs off the traveled-way (driver not injured)

↓

Vehicle hits a break-away pole (speed limit sign) (driver slightly injured by glass, say Type B)

↓

Vehicle goes down the embankment/sideslope and hit the bottom ditch hard (driver is severely injured, say Type A, from the external forces)

↓

Vehicle goes up the backslope and hits a tree (at low speed), the final resting place (driver does not sustain additional injuries)

Characteristics of Injury Severity Data

In crash reports, there are different methods/codes to measure or define injuries (still subjective as it is governed by the opinion of the police officer). In some reports, they classify the injury as the first injury outcome in the sequence of events ("**first harmful event**"), whereas, in other reports, they defined it as the "most harmful event." (**will be classified as 'Type A'**)

Vehicle swerved to avoid an animal (driver not injured)

↓

Vehicle runs off the traveled-way (driver not injured)

↓

Vehicle hits a break-away pole (speed limit sign) (driver slightly injured by glass, say Type B)

↓

Vehicle goes down the embankment/sideslope and hit the bottom ditch hard (driver is severely injured, say Type A, from the external forces)

↓

Vehicle goes up the backslope and hits a tree (at low speed), the final resting place (driver does not sustain additional injuries)

Characteristics of Injury Severity Data

In crash reports, there are different methods/codes to measure or define injuries (still subjective as it is governed by the opinion of the police officer). In some reports, they classify the injury as the first injury outcome in the sequence of events ("first harmful event"), whereas, in other reports, they defined it as the "**most harmful event**." (**will be classified as 'Type A'**)

Vehicle swerved to avoid an animal (driver not injured)

↓

Vehicle runs off the traveled-way (driver not injured)

↓

Vehicle hits a break-away pole (speed limit sign) (driver slightly injured by glass, say Type B)

↓

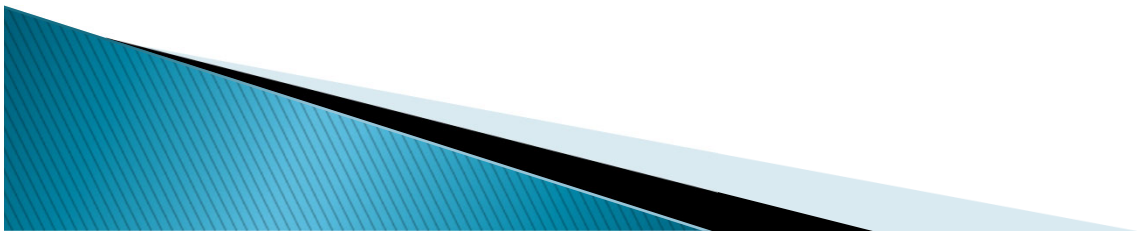
Vehicle goes down the embankment/sideslope and hit the bottom ditch hard (driver is severely injured, say Type A, from the external forces)

↓

Vehicle goes up the backslope and hits a tree (at low speed), the final resting place (driver does not sustain additional injuries)

Characteristics of Injury Severity Data

- 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 relaxed.

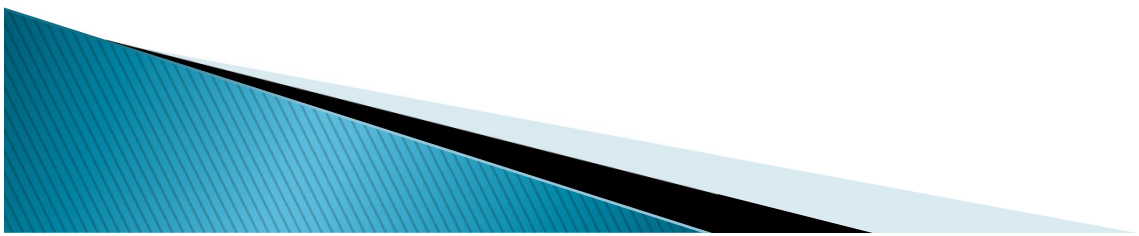


Ordinal nature of crash injury severity data

- 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.
- 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.

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.

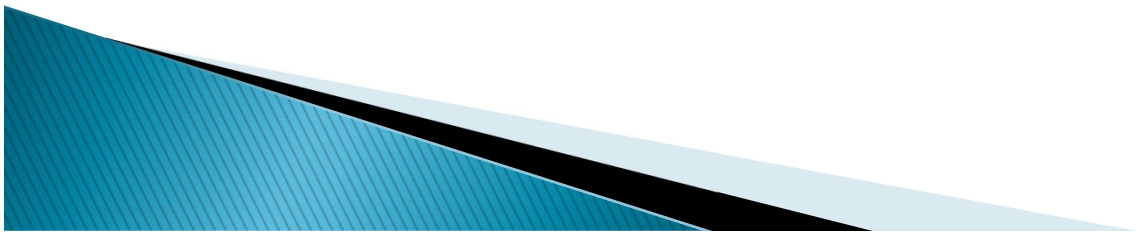


Imbalanced data between injury severity levels

- Crash injury severity data usually are imbalanced on the KABCO scale, where the number of fatal or severe injuries is substantially fewer 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.

Underreporting

- It has been well-documented that crashes with lower severity levels are less likely to be reported to governmental authorities compared to more severe crashes.
- For example, people involved in a reportable property damage only collision (above the minimum reportable threshold) may not be interested in seeing their vehicle insurance premiums go up and would therefore directly pay for the damages themselves or worse, flee from the crash scene (which is more common than we think).
- Furthermore, there is a possibility of inconsistency in the classification of a crash outcome into no injury or possible injury levels; and/or an arbitrary crash threshold for the vehicle or property damages exceeding a certain amount.
- 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.
- There is a lot of variation in the extent of underreporting, which can depend on the study location and severity levels.



Other Data Issues

- Small Sample size
  - Will affect the proportion (see unbalanced data above)
    - See next slide for minimum values.
- Endogeneity
  - An endogenous variable is an explanatory variable whose value is determined or influenced by one or more variables in the model.
  - Carson and Mannering (2001) studied the endogeneity problem by exploring the effectiveness of icewarning signs in reducing the frequency of ice-related crashes.
  - An indicator variable for the presence of an ice warning sign is typically used when developing a crash-frequency model.
  - As ice-warning signs are more likely to be placed at locations with high numbers of ice-related crashes, this indicator variable may be endogenous (the explanatory variable will change as the dependent variable changes).

Minimum Sample Size

Table 4
Three evaluation criteria by sample size for the three models*.

Sample size	Mean of absolute-percentage-bias (APB)			Max of absolute-percentage-bias (APB)			Total root-mean-square-error (RMSE)		
	MNL	ML	OP	MNL	ML	OP	MNL	ML	OP
100	5.50E+13	2.10E+11	143%	9.70E+14	2.90E+12	2.10E+01	7.40E+15	1.60E+13	20.7
500	2.00E+14	1.10E+04	25%	4.50E+15	1.10E+05	94%	1.30E+16	1.20E+06	4.5
2000	16%	26%	11%	45%	167%	40%	12.9	28.7	2.2
5000	9%	13%	5%	27%	52%	20%	7.6	13.7	1.2
10,000	4%	5%	4%	13%	13%	14%	4.7	8.7	0.7
20,000	2%	3%	2%	9%	21%	9%	1.9	3.4	0.4

* MNL: multinomial logit model; OP: ordered probit model; ML: mixed logit model.

In terms of the values of all three criteria, the multinomial logit and mixed logit are more sensitive to small sample sizes than the ordered probit model and this is especially noticeable for the sample sizes equal to 100 and 500. Nonetheless, for a sample size below 500, all models perform poorly.

According to the three criteria, the minimum sample size for the ordered probit, multinomial logit, and mixed logit models should be 2000, 5000 and 10,000, respectively.

Underreporting

- For instance, about three decades ago, Hauer and Hakkert (1988) stated that approximately 20% of severe injuries, 50% of minor injuries, and up to 60% of no-injury crashes were not reported.
- Elvik and Mysen (1999) reported underreporting rates of 30%, 75%, and 90% for serious, slight, and very slight injuries, respectively.
- According to Blincoe et al. (2002), up to 25% of all minor injuries and almost 50% of no-injury crashes were likely to be nonreported.
- The underreporting is a more significant issue in low and middle-income countries than in high-income countries.
- Some studies have proposed methods to minimize this bias even if the underreporting rate is unknown (see Kumara and Chin (2005); Yamamoto et al. (2008); Ma (2009) Ye and Lord (2011); Patil et al. (2012)). (see references in textbook)

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:

Random utility model

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

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.

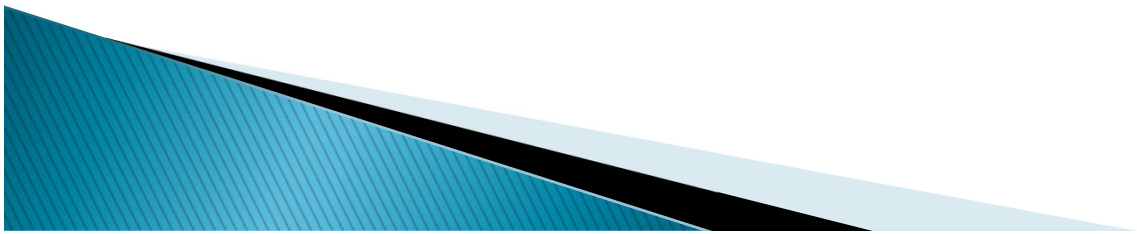
Utility maximization is the process of choosing the alternative with the maximum utility value. In a binary outcome model with injury and no injury, if $U(injury) > U(no\ injury)$, then the probability of injury $P_r(injury) = 1$; and if $U(injury) < U(no\ injury)$, then $P_r(injury) = 0$. This is a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

Random utility model

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:

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

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



Random utility model

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

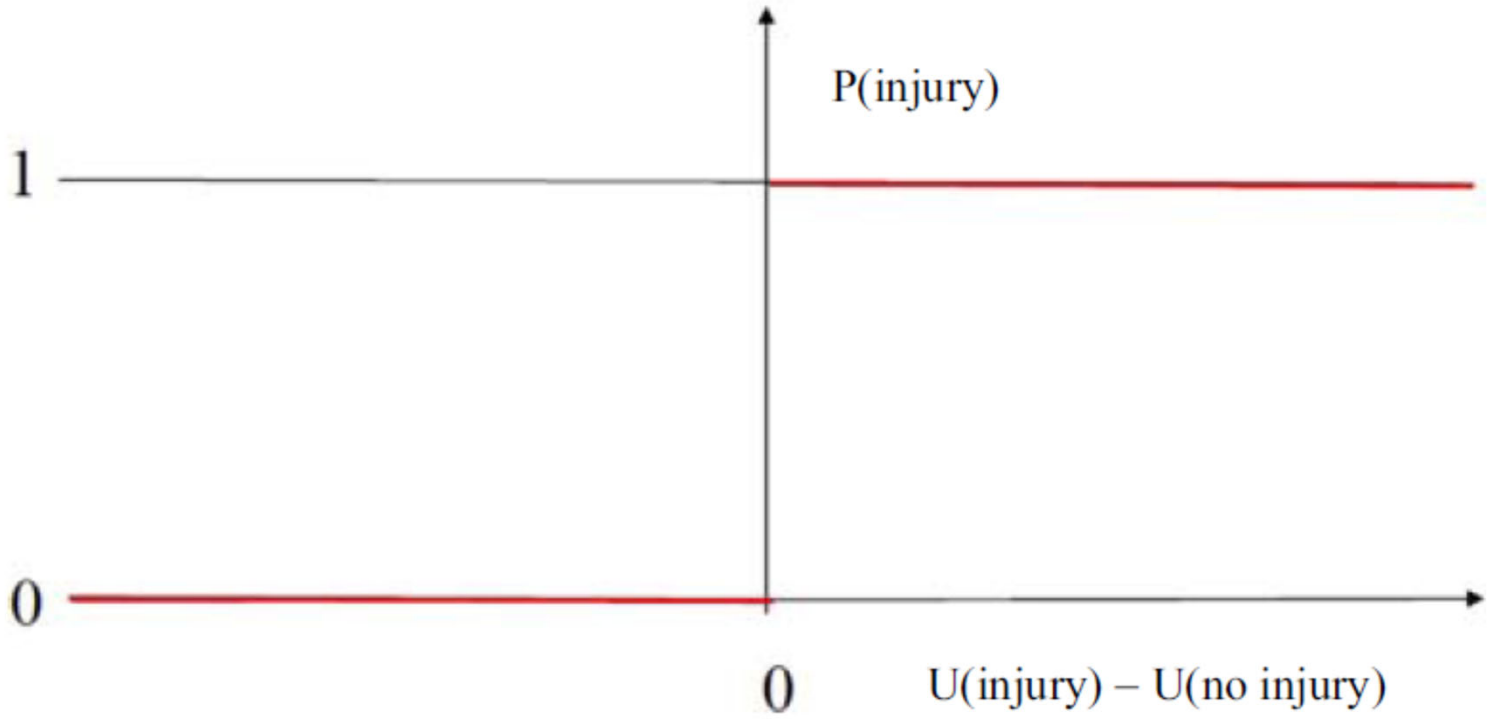


FIGURE 4.1 Deterministic choice of a binary variable.

Random utility model

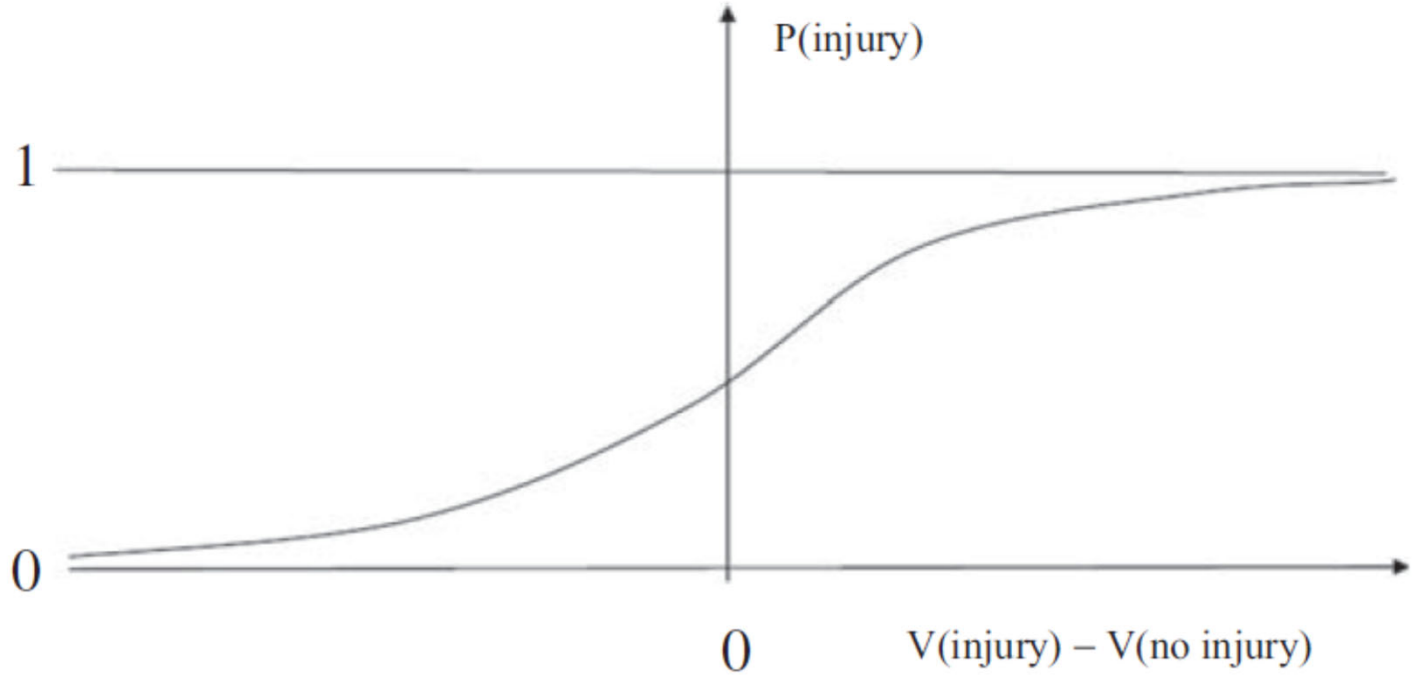


FIGURE 4.2 Stochastic choice of a binary variable.

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).

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 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.

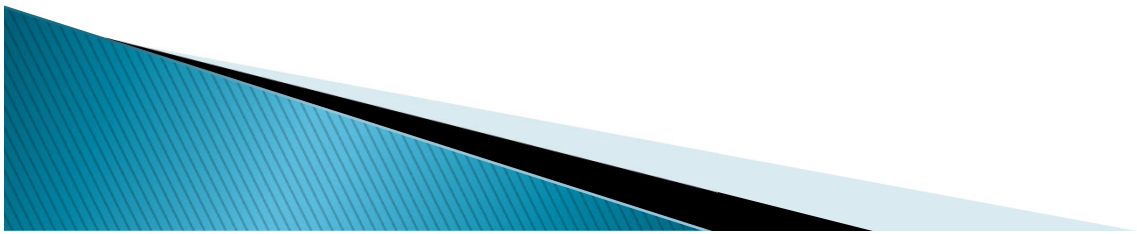
Multinomial logit model

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

If level I is the reference level, the model becomes

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

Note that in crash severity modeling, the lowest injury severity level, i =n 1, (i.e., "no injuries" or "property damage only" (PDO)) is usually set to be the reference level instead of level I.



Non-Ordered – Multinomial Logit

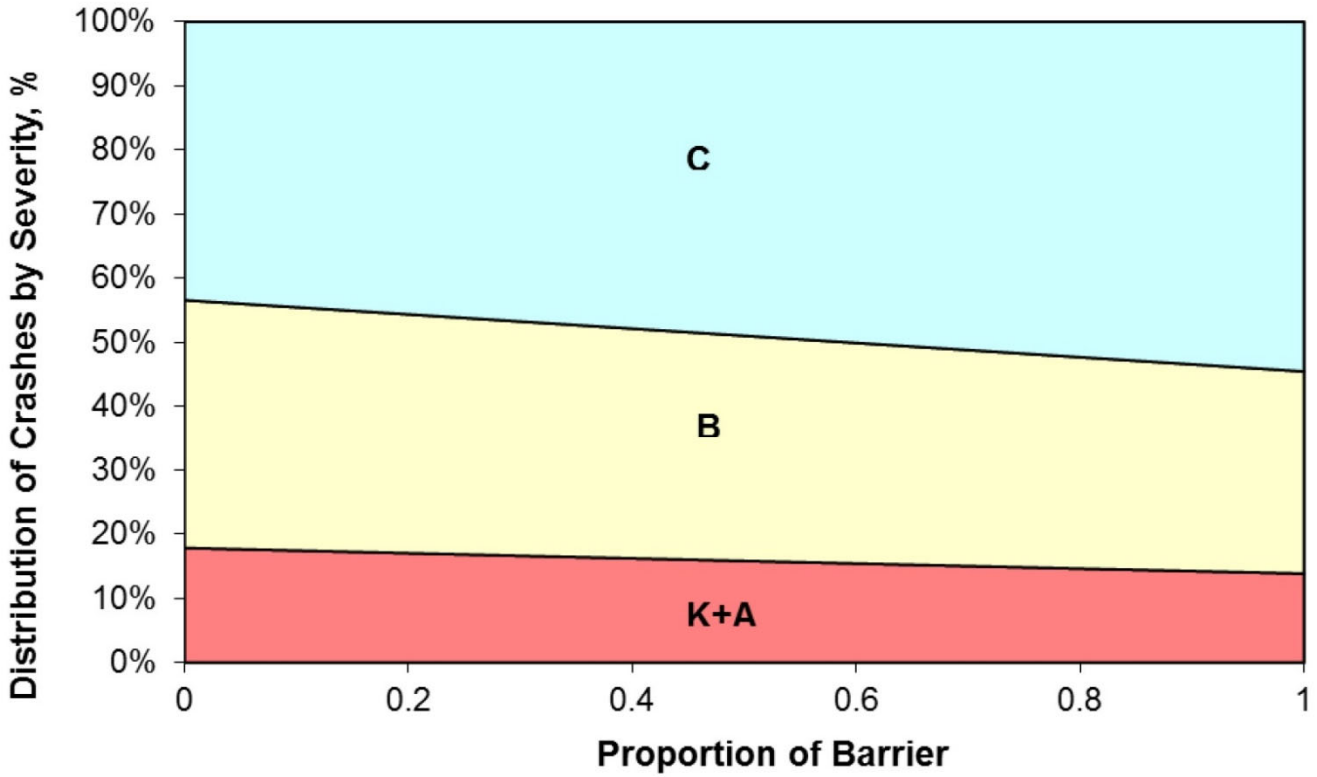


Figure 1. Severity distribution with change in proportion of barrier.

$$V_{K+A} = -1.5373 - 0.4813 \times P_b - 0.228 \times Lanes + 0.6681 \times I_{rural} + 0.426 \times I_{offramp} \quad \textbf{(15)}$$

$$V_B = +0.2355 - 0.4312 \times P_b - 0.435 \times Lanes + 0.6963 \times I_{rural} \quad \textbf{(16)}$$

Nested logit model

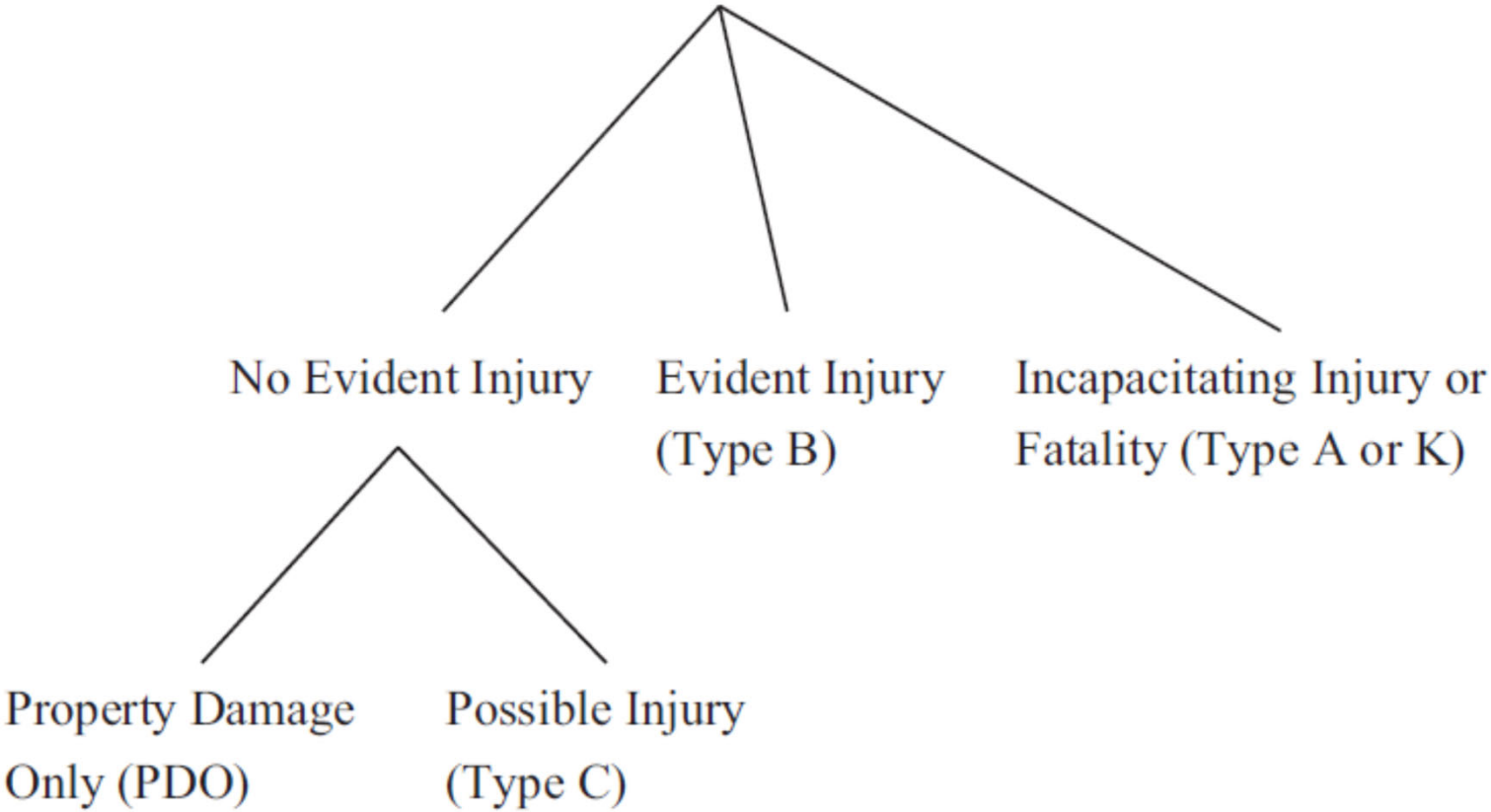


FIGURE 4.3 Nested structure of accident severities.

Mixed Logit Model

This model is similar to the random parameter model for the crash-frequency model. This means that the coefficients are allowed to vary across observations.

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

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

In Milton et al. (2008), they suggested that roadway characteristics better be modeled as fixed parameters, while volume-related variables such as average daily traffic per lane, average daily truck traffic, truck percentage, and weather effects better be modeled as random parameters. They speculated that the random effect of ADT per lane increases injury severity in some cases while decreases it in others may be capturing the response and adaptation of local drivers to various levels of traffic volume.

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).

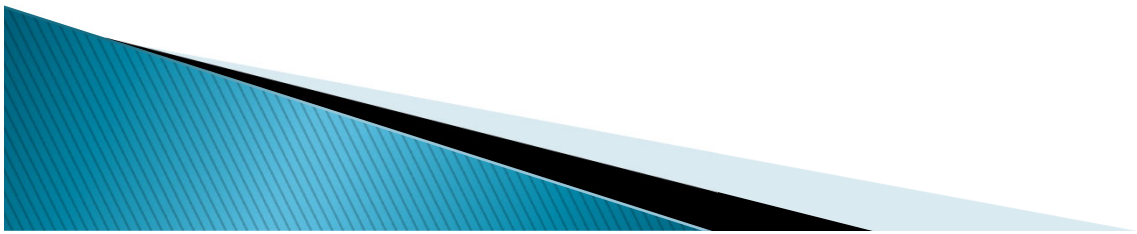
Ordered Logit/Probit

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

$$z_n = \boldsymbol{\beta}' \boldsymbol{X_n} + \varepsilon_n$$

Where $\boldsymbol{X_n}$ is a vector of explanatory variables determining the discrete ordering (i.e., injury severity) for *n* th 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:



Ordered Logit/Probit

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

$$log\left(\frac{P_r(y_n>i)}{1-P_r(y_n>i)}\right) = \alpha_i + \boldsymbol{\beta' X_n} \quad (i = 1, \cdots I-1) \rightarrow \Pr(y_n > i) = \frac{\exp(\alpha_i + \mathbf{x}'_n\boldsymbol{\beta})}{1+\exp(\alpha_i + \mathbf{x}'_n\boldsymbol{\beta})}$$

Ordered Logit/Probit

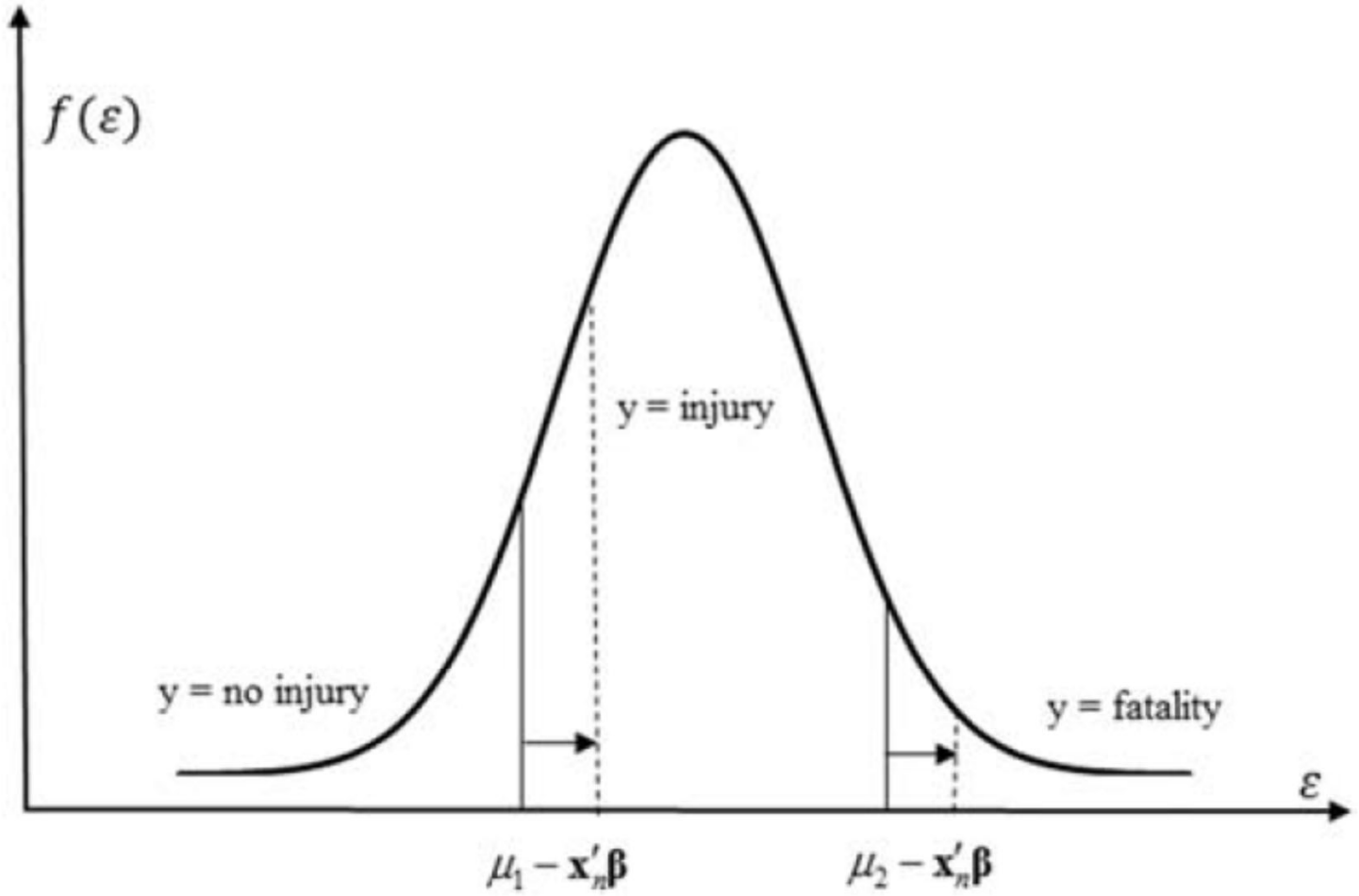
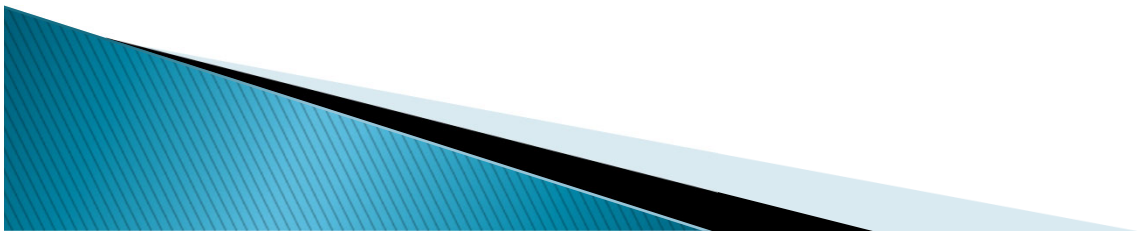


FIGURE 4.4 Illustration of an ordered model with an increase in.$\mathbf{x}'_n\boldsymbol{\beta}$.

Generalized ordered logistic and proportional odds model

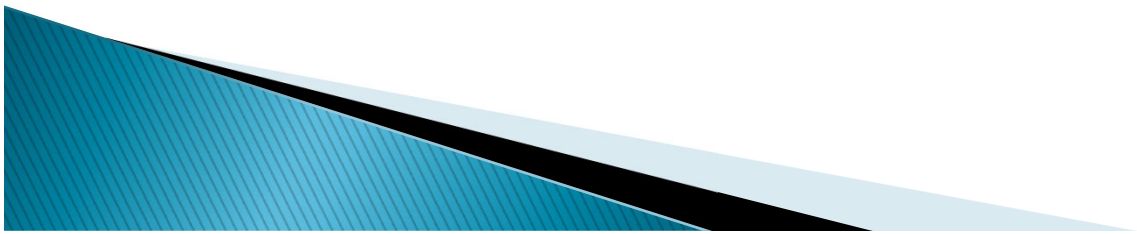
- A generalized ordered logistic model (gologit) provides results similar to those that result from running a series of binary logistic regressions/ cumulative logit models.
- The ordered logit model is a special case of the gologit model where the coefficients β are the same for each category.
- The partial proportional odds model (PPO) is in between, as some of the coefficients β 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.



Generalized ordered logistic and proportional odds model

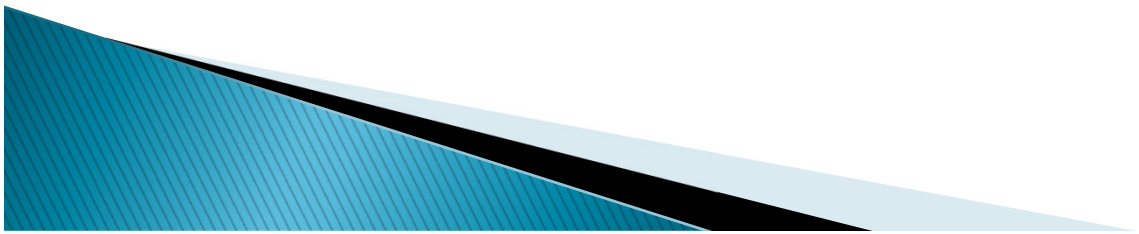
$$\Pr(y_n > i) = \frac{\exp(\mathbf{x}'_n\boldsymbol{\beta}_i - \mu_i)}{1 + \exp(\mathbf{x}'_n\boldsymbol{\beta}_i - \mu_i)}, \quad i = 1, \ldots (I-1)$$

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.



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:



Sequential logistic/probit regression model

$$z_{n1} = \alpha_1 + \mathbf{x}'_n \beta_1 + \varepsilon_{n1}$$
$$z_{n2} = \alpha_2 + \mathbf{x}'_n \beta_2 + \varepsilon_{n2}$$
$$\vdots$$
$$z_{n,I-1} = \alpha_{I-1} + \mathbf{x}'_n \beta_{I-1} + \varepsilon_{n,I-1}$$

where z_{ni} is a continuous latent variable that determines whether the injury severity is observed as i or higher, β_i's are the vectors of estimated parameters, and ε_{ni}'s are error terms that are independent of x_n.

Sequential logistic/probit regression model

- 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."
- It is important to know that the sequential model uses a subpopulation of the data to estimate the variant set of β_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 β_1, but only the crashes with injury type C or higher are used in the second stage to estimate β_2. Crashes with injury type B or higher are used in the second stage to estimate β_3.

Sequential logistic/probit regression model

- Jung et al. (2010) applied the sequential logit model to assess the effects of rainfall on the severity of single-vehicle crashes onWisconsin 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.

Model interpretation

- 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).

Model interpretation

- 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: $\frac{\partial p_i}{\partial x_{ki}} = \beta_{ki} p_i (1 - p_i)$.
- Such marginal effects are called instantaneous rates of change because they are computed for a variable while holding all other variables as constant.
- 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.