Scientific Research Publishing

American Journal of Operations Research, 2018, 8, 17-32
http://www.scirp.org/journal/ajor
ISSN Online: 2160-8849
ISSN Print: 2160-8830

The Use of Item Response Theory in Survey Methodology: Application in Seat Belt Data

Mark K. Ledbetter[1], Norou Diawara[1*], Bryan E. Porter[2]

[1]Mathematics and Statistics Department, Old Dominion University, Norfolk, VA, USA
[2]Department of Psychology, Old Dominion University, Norfolk, VA, USA
Email: mledb001@odu.edu, *ndiawara@odu.edu

How to cite this paper: Ledbetter, M.K., Diawara, N. and Porter, B.E. (2018) The Use of Item Response Theory in Survey Methodology: Application in Seat Belt Data. *American Journal of Operations Research*, **8**, 17-32.
https://doi.org/10.4236/ajor.2018.81002

Received: December 6, 2017
Accepted: January 7, 2018
Published: January 10, 2018

Abstract

Problem: Several approaches to analyze survey data have been proposed in the literature. One method that is not popular in survey research methodology is the use of item response theory (IRT). Since accurate methods to make prediction behaviors are based upon observed data, the design model must overcome computation challenges, but also consideration towards calibration and proficiency estimation. The IRT model deems to be offered those latter options. We review that model and apply it to an observational survey data. We then compare the findings with the more popular weighted logistic regression. *Method*: Apply IRT model to the observed data from 136 sites within the Commonwealth of Virginia over five years collected in a two stage systematic stratified proportional to size sampling plan. *Results*: A relationship within data is found and is confirmed using the weighted logistic regression model selection. *Practical Application*: The IRT method may allow simplicity and better fit in the prediction within complex methodology: the model provides tools for survey analysis.

Keywords

Item Response Theory, Logistic Regression, Sampling Weight

1. Introduction

When sampling methodology is complex, initiatives are employed in statistical analysis to extract the most reliable information from data through the model and its parameters. The goal of this manuscript is to apply the item response theory (IRT) to analyze survey data, and compare the output with one classical test theory (CTT) called logistic regression models as a point of reference.

The sampling methodology used to collect data has a two stage design associated with primary sampling unit (PSU) strata from 15 counties and secondary sampling units (SSU) from 136 road segments within the counties, under National Highway Transportation Safety Authority (NHTSA) guidelines [1]. If sampling weights are ignored, then the model parameter estimates can be biased [2]. In fact, since the sample is collected from a two stage stratified sampling design, standard underlying assumptions of parametric statistical models may be violated, and guidelines based on the statistical design cannot be ignored. [3] [4] and [5] have given suggestions for such complex methodologies. Other authors have applied the methodology to studies. Our intent is to apply the seat belt sampling methodology to predict the seatbelt usage. [6] [7] and [8] have used such methodologies and they concluded that females are more likely to wear seatbelts than males. The relationship between vehicle type and seatbelt use has been explored by [9] [10] and [11] who concluded that seatbelt use in pickup trucks is lower than other passenger vehicles. [12] suggested that passenger and driver use are related. [13] asserts that the seatbelt use is increased in those states within the United States that have primary seatbelt enforcement laws and actively enforce seatbelt use. Studies have also explored relationships between race, socio-economic status, age, rural/urban environments, law enforcement type (primary, secondary), the amount of fines, and the type of road traveled (primary, secondary, tertiary). [14] employed a multivariate approach using the aforementioned factors along with cultural variables to explain the differences in seatbelt use between states using self-reported information, direct observation, and crash reports. However, the validity of self-reported seatbelt use in surveys is questionable compared to observed seatbelt usage [15]. While the methodology is simple to describe, the challenge is found in the statistical analysis tool used to make prediction, especially in the presence of behavioral variables, such as driver gender, vehicle type, traffic volume, road segment length, weather conditions, driver cellphone use, passenger presence, lane, and passenger seatbelt use. The goal is to get meaningful information that can be translated into quantitative measures. [16] and [17] propose the addition of a score variable due to the measurement of concern. Those researchers have incorporated latent traits of data in a score function.

The manuscript presents a comparison of the popular logistic regression presented here along suggestion of the Item Response Theory (IRT) model, and its simple version called the Rasch model [18].

Moreover, ignoring weights may lead to imperfection in the sample (as departing from the reference population) and serious bias in latent variable models [19]. To avoid that problem, we apply a weight function. [14] cautioned about the use of other factors to develop more effective countermeasures for increasing seatbelt use. We propose the weighted logistic and IRT models after variable selections and compare the findings. The manuscript is organized as follows. In Section 2, we present background of data, then build the reference model in Section 3.

In Section 4, the weighting scales are built into the models. The IRT model is presented. We end with a conclusion in Section 5.

2. Overview of Data

Data collected in the summers of 2012, 2013, 2014, 2015, and 2016 for Virginia seat belt use is used as evidence. As mentioned in the previous Section, the data is collected under a two stage design. Primary sampling units (PSU) are county aggregates and were stratified using the five-year average annual VMT (vehicle miles traveled) in millions. Out of 97 total county aggregates, 57 account for 87.2 percent of passenger vehicle crash related fatalities. The 57 eligible county aggregates were grouped by VMT into three strata: low, medium, and high. Within each stratum, five PSU's were selected with PPS where the measure of size (MOS) was the five-year average annual VMT. The PSU sampling weights are calculated by taking the inverse of the five year average annual VMT, and varied from approximately 0.089 to approximately 0.967. Secondary sampling units (SSU) are road segments. Road segments were stratified by type (primary, secondary, and local) and by segment length (short, medium and long) within each county. The eligible SSU were then selected by PPS with segment length as the MOS resulting in 136 selected road sites for observation. The SSU weights are calculated by taking the inverse of the segment length and varied from approximately 0.0001 to approximately 0.1657.

The weighting was added so that information from the whole population would be captured. If the selection mechanism is not informative, the parameter estimates will remain consistent regardless of the weights, and weights should be excluded from the model [20]. Moreover, if the strata sample sizes are large enough, the parameter estimates are unbiased. In sampling surveys, it is not always possible to determine whether the weights are informative. However, the observations should reflect the sampling weights to avoid biased sampling.

The data collected includes the following observed binary data: driver seat belt use (yes, no), driver gender (female, male), passenger present (yes, no), passenger seatbelt use (yes, no), and visible driver cellphone use (yes, no). The other observed data is categorical: vehicle type (car, truck, SUV, van, or minivan), lane of the road (1 - 5, where lane 1 represents the lane furthest to the right and lane 5 denotes the fifth lane from the right in the direction of travel), and weather (sunny/clear, light rain, cloudy, fog, or clear but wet conditions). The VMT for each site observed is classified (Road Class) within each county aggregate as lower, average, and upper. Vehicle type was assigned in no particular order, and later we reclassified it to describe the size of the vehicle which crudely correlates to seatbelt use. Weather is also not ordered in its assignment, and we reclassify it based on severity and impediment of driving ability. The data set also includes the following continuous variables: VMT, road segment length, and selection probabilities determined in the sampling design stage.

3. Unweighted Analysis and Results

Generalized linear models are usually considered in the investigation of the data. First, a classic linear model was suggested to obtain a general relationship between the response (driver seatbelt use) and predictive variables. However, use of a linear model on binary responses is not recommended [21], since predicted values may be outside of the domain of the response variable. From this point forward, a classic model also known as classical test theory (CTT) is considered. We consider first fitting a logistic model to the data.

3.1. Logistic Model

In this model, $p = P(Y = 1)$ is the probability that the driver is wearing a seat belt, and $1 - p = P(Y = 0)$ is the probability that the driver is not wearing a seatbelt. The initial model is:

$$\text{Model 1:}\quad \text{Log}\left[\frac{p}{1-p}\right] = \beta_0 + \beta_v X_v + \beta_r X_r + \beta_g X_g + \beta_s X_s + \beta_l X_l + \beta_c X_c + \beta_w X_w + \beta_{pp} X_{pp} + \beta_{ps} X_{ps}$$

where β_0 denotes the intercept of the model, X_v denotes Vehicle Type (car, truck, SUV, van, or mini-van), X_r denotes Road Classification for VMT (low, average, high), X_g denotes Driver Gender (male/female), X_s denotes the road segment length in mile, X_l denotes Lane in which vehicle observed (right to left), X_c denotes Driver Cell Phone Use (yes/no), X_w denotes Weather (clear, light rain, cloudy, foggy, or clear but wet), X_{pp} denotes Passenger Present (yes/no), X_{ps} denotes Passenger Seatbelt Use (yes/no). This notation is used consistently throughout this manuscript. The weights w_{ij} are obtained as $p_{ij} = p_i * p_{j(i)}$ where p_i is the selected probability of the selected county, and $p_{j(i)}$ is the selection probability of the jth road type selected within the ith county; $i = 1, 2\cdots, 15$, and $j = 1, 2, \cdots, n_i$.

The estimated non-weighted seat belt use for each year is $\hat{p} = 0.83$ for 2012, $\hat{p} = 0.81$ for 2013, $\hat{p} = 0.79$ for 2014, $\hat{p} = 0.84$ for 2015, and $\hat{p} = 0.81$ for 2016.

To simplify the model, the logistic fit is processed with stepwise selection at a 0.15 significance level for both entry into the model and retention in the model. The results are verified using forward selection and backward selection options. The three procedures produce the same results.

Analysis of the effects of weather on seatbelt use revealed inconsistent associations between seatbelt use and weather severity for the five years. Further, the selection process does not identify weather as significant for any combined data. Hence, weather has been removed from the model and the analysis repeated. Analysis of the predictor variables reveals a high correlation (Spearman's correlation coefficient, $r_s = 0.94, p - value < 0.0001$) between road segment length and road class which indicates a confounding condition. Other correlations are less than 0.15 and do not indicate the presence of other confounding effects. As a result, road segment length was removed from the model and the analysis performed again.

Table 1 provides the Wald Test for significance in the selected Model with variables as Vehicle type, Road class, driver gender, and so on. For 2012-2013 combined data, all remaining predictors are significant at $p = 0.01$, while passenger presence is removed due to a p-value > 0.15. For 2012-2014, all predictors are significant at $p = 0.05$. For the combined 2012-2015 data, predictor variables have p-values < 0.005. For the combined data for 2012 through 2016, all five of the remaining predictors are significant at $p < 0.005$.

The close agreement between the models may indicate that the aggregate data follows a standard model which also fits the individual data sets. The test of the global hypothesis of null model, shown in Table 2, of $\beta_i = \beta_j = 0$ for $i \neq j$ versus at least one $\beta_i \neq 0$ ($i, j = r, g, l, c, r,$ or pp depending upon the model) indicates significant evidence exists ($p < 0.0001$) to support the claim that the models are not explained solely by the intercept (*i.e.* the response is not a constant) for all four presented models which is consistent with the Wald Test results in Table 1.

Computational efficiency is measured by Akaike Information Criterion (AIC) numbers [22], displayed in Table 3, which assess the goodness of fit of the model: smaller numbers indicate a better fit. AIC is defined as follows:

$$AIC = 2p + n\log\left(\frac{SS_r}{n}\right),$$

where p is the number of parameters in the model, SS_r is the residual sum of squares, and N is the number of observations in the dataset.

Table 1. Type 3 analysis of effects.

		2012-2013		2012-2014		2012-2015		2012-2016	
Effect	**DF**	**Wald ChiSq**	**Pr > ChiSq**	**Wald ChiSq**	**Pr > ChiSq**	**Wald ChiSq**	**Pr > ChiSq**	**Wald ChiSq**	**Pr > ChiSq**
Vehicle Type	4	513.796	<0.0001	773.573	<0.0001	1005.152	<0.0001	1302.209	<0.0001
Road Classification	2	62.387	<0.0001	63.925	<0.0001	58.591	<0.0001	57.832	<0.0001
Driver Gender	1	51.262	<0.0001	58.242	<0.0001	107.301	<0.0001	145.8213	<0.0001
Lane	4	52.370	<0.0001	57.563	<0.0001	95.317	<0.0001	101.7103	<0.0001
Driver Cell Phone Use	1	25.645	<0.0001	49.523	<0.0001	67.574	<0.0001	75.4237	<0.0001
Passenger Present	1	2.809	0.0937	5.360	0.0206	8.138	0.0043	9.2257	0.0024

Table 2. Testing global null hypothesis: $\beta = 0$.

	2012-2013		2013-2014		2012-2015		2012-2016	
Test	**Chi-Square**	**Pr > ChiSq**	**Chi-Square**	**Pr > ChiSq**	**Chi-Square**	**Pr > ChiSq**	**Chi-Square**	**Pr > ChiSq**
Likelihood Ratio	917.515	<0.0001	1299.731	<0.0001	1758.810	<0.0001	2225.1421	<0.0001
Score	972.383	<0.0001	1377.571	<0.0001	1872.449	<0.0001	2380.1704	<0.0001
Wald	918.137	<0.0001	1305.655	<0.0001	1771.882	<0.0001	2250.9919	<0.0001
DF	13		13		13		13	

The results of the AIC for logistic regression performed on the significant variables identified during the selection process are in the 10 thousands. Since the intercept alone is not a sufficient explanation of the model, we use the values for intercept and covariance. The AIC numbers obtained for individual years are approximately 30% lower than those obtained by [14]; however, the combined data is significantly higher. The significantly higher numbers for the combined data indicate a significant amount of variation in the model, or a less than optimum fit.

3.2. Variable Standardization and Reclassification

Since vehicle types are listed in no particular order, vehicle type is reclassified to indicate size of the vehicle which negatively correlates to driver seatbelt use: *i.e.* in general, the drivers of larger vehicles tend to wear seatbelts less often than drivers of smaller vehicles as suggested in [9]. Preliminary analysis of the data appears to support this hypothesis, so smaller vehicle types are given a larger value to indicate that the driver is more likely to wear a seatbelt. **Table 4** contains the reclassifications of vehicle type. The remaining five predictor variables have positive correlations to driver seatbelt use and reclassification is not necessary. It is known that the variance is larger for population parameters with large values than for population parameters with smaller values. In order to make the variance between variables more homogenous and reduce the overall model variance, each variable of interest was standardized by dividing its value by its third quartile (Q3) in an approach similar to [23]. Standardizing the variables may affect whether they are selected in the model, so all six of the potential predictors are standardized. The Q3 values of the variables after reclassification are listed in **Table 5**. Note that the Q3 values are the same for all five years, and

Table 3. Model fit statistics.

	2012-2013	2012-2014	2012-2015	2012-2016
Criterion	**Intercept and Covariates**	**Intercept and Covariates**	**Intercept and Covariates**	**Intercept and Covariates**
AIC	23015.856	35333.162	48803.129	58559.330
SC	23129.764	35452.647	46926.938	58686.246
−2 Log L	22987.856	35305.162	46775.129	58531.330

Table 4. Reclassification of variables.

Vehicle Type	Original Value	New Value for Size
Car	1	3
Truck	2	1
SUV	3	1
Van	4	1
Mini-Van	5	2

thus the combined Q3 values are constant across time.

3.3. Model Fitting after Standardized and Reclassified Variables

The logistic selection process with $p = 0.15$ for entry and retention in the model is performed on the reclassified and standardized variables. The significant variables indicated prior to standardization in 3.2 above remain significant (Table 6). The model fit statistics are comparable to the previous analysis (Table 7). The global null hypothesis test indicates that the model is not sufficiently described solely by the intercept (Table 8). All variables selected are significant (p-value < 0.0001) for all datasets analyzed. In this analysis, it is reasonable to select the model fit by the combined 2012-2016 data:

Table 5. Third quartiles after reclassification (No weight).

Variable	2012-2013: 75th Percentile (Q3)	2012-2014: 75th Percentile (Q3)	2012-2015: 75th Percentile (Q3)	2012-2016: 75th Percentile (Q3)
Vehicle Type	3	3	3	3
Gender	1	1	1	1
Lane	2	2	2	2
Road Class	3	3	3	3
Cell Phone	1	1	1	1
Passenger Present	1	1	1	1

Table 6. Type 3 analysis of effects for standardized and reclassified variables.

		2012-2013		2012-2014		2012-2015		2012-2016	
Effect	**DF**	**Wald ChiSq**	**Pr > ChiSq**	**Wald ChiSq**	**Pr > ChiSq**	**Wald ChiSq**	**Pr > ChiSq**	**Wald ChiSq**	**Pr > ChiSq**
Vehicle Type	2	158.944	<0.0001	198.594	<0.0001	244.374	<0.0001	303.008	<0.0001
Road Classification	2	63.0613	<0.0001	62.6709	<0.0001	59.2485	<0.0001	59.5541	<0.0001
Driver Gender	1	167.328	<0.0001	227.771	<0.0001	361.160	<0.0001	482.711	<0.0001
Lane	4	67.3511	<0.0001	76.9267	<0.0001	125.775	<0.0001	140.465	<0.0001
Driver Cell Phone Use	1	25.9062	<0.0001	48.904	<0.0001	64.3876	<0.0001	72.7005	<0.0001
Passenger Present	1	7.5306	0.0061	14.047	0.0002	20.3291	<0.0001	22.2701	<0.0001

Table 7. Model fit statistics for standardized and reclassified variables.

	2012-2013		2012-2014		2012-2015		2012-2016	
Test	**Chi-Square**	**Pr > ChiSq**	**Chi-Square**	**Pr > ChiSq**	**Chi-Square**	**Pr > ChiSq**	**Chi-Square**	**Pr > ChiSq**
Likelihood Ratio	575.7315	<0.0001	741.4629	<0.0001	1022.7639	<0.0001	1258.4637	<0.0001
Score	560.1672	<0.0001	722.7031	<0.0001	997.0795	<0.0001	1227.8707	<0.0001
Wald	544.4533	<0.0001	704.7331	<0.0001	972.4031	<0.0001	1198.5248	<0.0001
DF	11		11		11		11	

Table 8. Global null hypothesis: $\beta = 0$ for Standardized and reclassified variables.

	2012-2013	2012-2014	2012-2015	2012-2016
Criterion	**Intercept and Covariates**	**Intercept and Covariates**	**Intercept and Covariates**	**Intercept and Covariates**
AIC	23353.640	35887.431	47535.175	59522.009
SC	23451.276	35989.846	47641.298	59630.794
−2 Log L	23329.640	35863.431	47511.175	59498.009

$$\text{Model 2: } \text{Log}\left[\frac{p}{1-p}\right] = \beta_0 + \beta_v X_v + \beta_r X_r + \beta_g X_g + \beta_l X_l + \beta_c X_c + \beta_{pp} X_{pp}.$$

The variable significance is displayed in Table 6, and the fit estimates are shown in Table 7. The AIC and SC numbers remain undesirably large (Table 8) and indicate that reclassification and standardization are not sufficient actions to improve model fit. Therefore, we investigate the cause for the poor model fit.

In all the previous sections, the AIC, BIC and log likelihood have been used as best measures of goodness fit for the most parsimonious models. They turn out to be high, which is an evidence of over-dispersion, which could be an indication there is more variability in the data than expected from the fitted model, which is an indication of a poor fit. Since the sample size is large, the corrected AIC does not lead us to better improvements. Variables have been selected for each dataset and the selection process results in similar models. We will use these criteria as comparisons when adding the weights to the models considered in the next section.

4. Weighted Statistical Models

4.1. Weights

In all of the above analyses, the weights associated with the data were ignored. However, driver seat belt behavior is intricate and quite certainly involves non-collected data. Ignoring sample weights leads to inflated standard errors and biased estimates [2]. [3] provide guidelines for data analysis under weighted and designed data which reduces bias that would result in over sampled strata. The weights are stratum size and length of road segments. The inclusion of weights results in a significantly different model than selected in Section 3 above as inferred by [5]. Additionally, the goodness of fit criteria is significantly reduced (improved). The sampling plan for the data in this manuscript was developed as a joint effort between two of the authors (N. Diawara and B.E. Porter) and NHTSA. Therefore, in order to correct for bias due to stratum size and length of road segment, we included the weight designed for this analysis in our model, in accordance with NHTSA requirements [1] as:

$$\text{Weight} = \left(\text{Road Segment Length}\right) \times \left(\text{County Selection Probability}\right).$$

In this section, we will compare the results of the analysis based on the sampling weights and validate the appropriateness of the use of the weights.

4.2. Weighted Logistic Models

4.2.1. Model Fitting: Weighted Logistic Regression

Prior to performing analysis on the reclassified and standardized variables, the 75th percentiles for the weighted reclassified variables is determined. The weighted third quartile values are the same as the unweighted values listed in Table 5.

The selection process using the weighted logistic regression model and the SAS® logistic procedure resulted in three significant predictors at $p = 0.15$: driver gender, passenger presence, and vehicle type for 2012-2013 data. The selection process for both the 2012-2014 data and the 2012-2015 data additionally indicates that cell phone use is significant at $p = 0.10$. In the aggregate data for 2012-2016, the selection process results in three significant variables at $p = 0.05$ (see Table 9). There appears to be an increasing significance in the prediction of driver seat belt use by cell phone use ($p > 0.15$ to $p \approx 0.05$) over time. The model is significant as indicated by the global null hypothesis test in Table 10.

There is significant decrease in the AIC when the weights are added to the model, matching in [24] that, in the context of behavioral ecology, a simple controlled model does not show all the complexity of the data. Table 11 contains the AIC and SC values, which are lower than the corresponding unweighted models by a factor of approximately 20. The weights have improved the accuracy of model as it helps reduce the residual variance.

Figure 1 displays the predicted probability of seat belt use (for drivers using a cellphone with a passenger present) versus the vehicle type for each gender. The same general upward trend exists in the weighted model and the unweighted model but using less predictors. Please note that the authors have only included

Table 9. Type 3 analysis of effects for weighted, standardized and reclassified variables.

		2012-2013		2012-2014		2012-2015		2012-2016	
Effect	**DF**	**Wald ChiSq**	**Pr > ChiSq**	**Wald ChiSq**	**Pr > ChiSq**	**Wald ChiSq**	**Pr > ChiSq**	**Wald ChiSq**	**Pr > ChiSq**
Vehicle Type	2	9.3692	0.0092	11.2742	0.0036	13.2448	0.0013	16.9144	0.0002
Driver Gender	1	10.3672	0.0013	12.5182	0.0004	19.3154	<0.0001	24.8218	<0.0001
Driver Cell Phone Use	1	-	-	3.1076	0.0779	3.5323	0.0602	3.7706	0.0522
Passenger Present	1	2.1891	0.1390	2.9222	0.0874	4.2446	0.0394	4.4189	0.0355

Table 10. Global null hypothesis: $\beta = 0$ for weighted, standardized, and reclassified variables.

	2012-2013		2012-2014		2012-2015		2012-2016	
Test	**Chi-Square**	**Pr > ChiSq**	**Chi-Square**	**Pr > ChiSq**	**Chi-Square**	**Pr > ChiSq**	**Chi-Square**	**Pr > ChiSq**
Likelihood Ratio	26.3513	<0.0001	35.1654	<0.0001	46.8481	<0.0001	57.6230	<0.0001
Score	25.3421	<0.0001	34.1321	<0.0001	45.6491	<0.0001	56.2349	<0.0001
Wald	25.5806	<0.0001	33.1892	<0.0001	44.4985	<0.0001	54.8916	<0.0001
DF	4		5		5		5	

Table 11. Model fit statistics for weighted, standardized and reclassified variables.

	2012-2013	2012-2014	2012-2015	2012-2016
Criterion	**Intercept and Covariates**	**Intercept and Covariates**	**Intercept and Covariates**	**Intercept and Covariates**
AIC	1201.336	1812.153	2396.299	3000.441
SC	1242.273	1863.365	2449.364	3054.837
−2 Log L	1191.336	1800.153	2384.299	2988.441

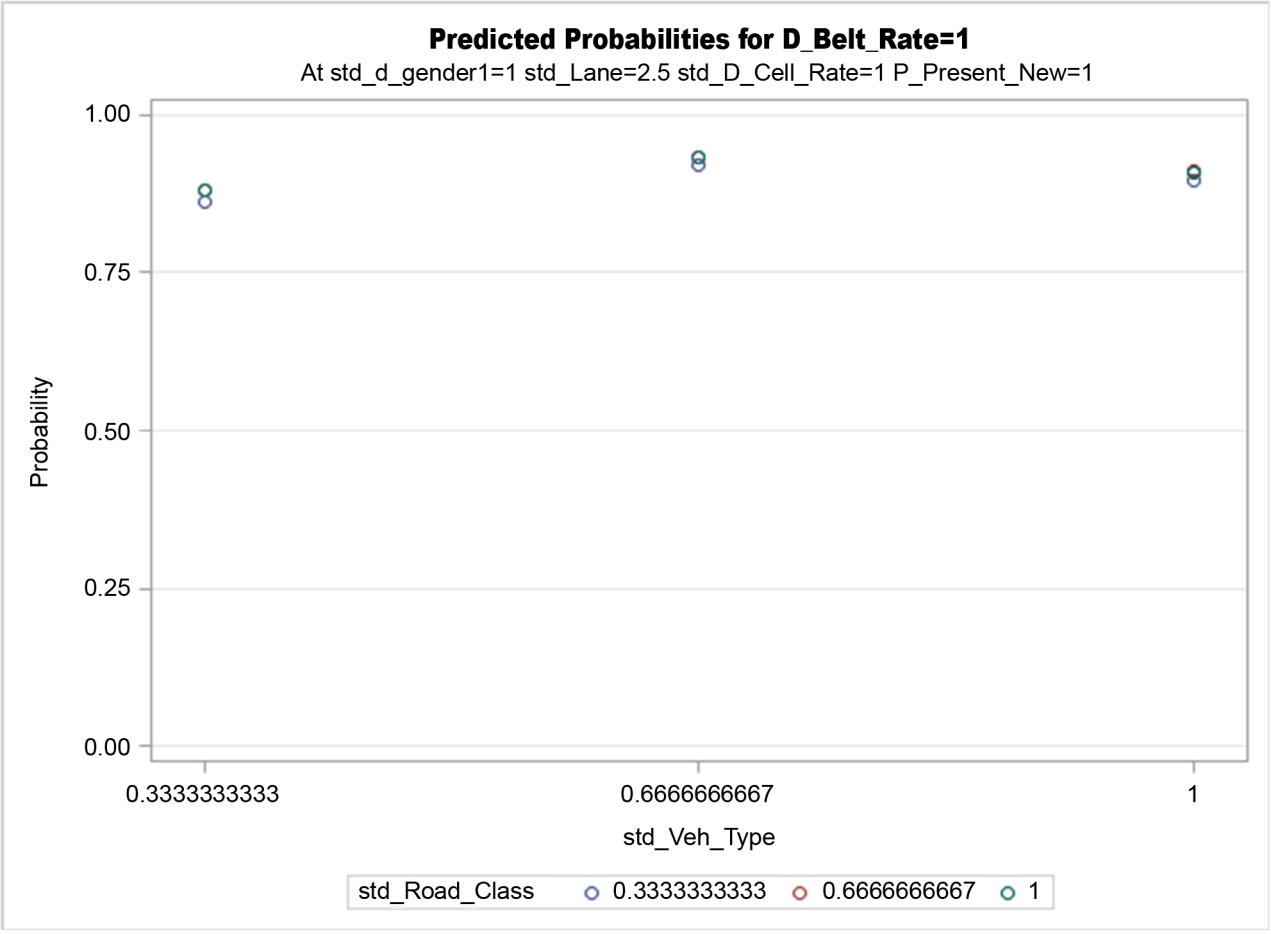

Figure 1. Model 3: Multivariate weighted logistic regression on model with $p = 0.15$ selection (2012-2016 Data).

model but using less predictors. Please note that the authors have only included one chart for this model due to the excessive space required to depict all 24 such combinations.

4.2.2. Model Selection: Weighted Logistic Regression

The final model selected for the 2012-2016 aggregate data is

$$\text{Model 3:}\quad \text{Log}\left[\frac{p}{1-p}\right] = \beta_0 + \beta_v X_v + \beta_g X_g + \beta_c X_c + \beta_{pp} X_{pp}$$

where β_0, β_v, β_g, β_c, and are the estimates calculated using the weights.

As expected, the combination of the data results in an improvement in the

significance of the predictors compared to individual models. However, the models have different selected variables and one of the variables selected for the 2012-2016 combined data has a p-value > 0.05 indicating the necessity for a different analytical method.

One suggestion is to develop an IRT model for prediction of seatbelt use, and it is advisable to include only very significant predictor variables ($p \leq 0.05$). All four selected variables in the aggregate 2012-2016 data model have significance levels less than or very close to 0.05. We explore an IRT model using a selection process with $p = 0.05$ significance on the combined data.

4.3. Weighted Item Response Theory Model

4.3.1. Background

To analyze dichotomous events or polytomous level response data (as usually found in the quality of life field), the item response theory (IRT) model provides a complement to the classical test theory (CTT) as the behavior and characteristic of the driver is not directly understandable. The measurement of driver behavior is not suitable since it is based on qualitative indicators such as the type of vehicle used, and other ad hoc parameters that are not easy to translate into quantitative information to be used in a CTT statistical analysis. Because of that, IRT and its famous Rasch model have also been implemented to measure drivers' behaviors. The IRT model allows the inclusion of the latent factor common to all drivers that can be described by a score function. We applied such a model based on specified traits that reflect the dichotomy of the data such as gender, and made comparisons. We then compare the efficiency and effectiveness of the overall indicators by computing goodness of fit statistics.

4.3.2. Model

Because the model requires consideration of several conditions, the Rasch model is considered, as it provides a tool to analyze characteristics even when they are latent. Such a model can be included in the class IRT in the framework proposed by [17]. Driving habits can be seen as a variable which depends on many factors. Our primary focus is on seat belt use and indicators which give additional information to evaluate seat belt use. We propose to extend the theory of logistic regression to include characteristics associated with driver seatbelt use which is translated into the driver's condition as an associated score. In such a context, the Rasch model ([18] [25]) is an option where we can include each driver's behavior regarding seat belt use. One main concern is the associated measurement of the score. That score is based on the qualitative information to be translated into quantitative measure. Using ideas from [26], we develop a score function that can be used to build the sensitive attributes and behaviors of drivers. As mentioned in [27], the bias reduction is achieved through appropriate weight adjustments.

A score function is built using a linear combination of significant predictor variables. The proposed score attempts to capture the features of vehicle type

driven, driver gender, passenger presence, and driver cellphone use. Those features can alter the probability of seat belt use and they can be seen as sufficient statistics for the response (See [16]). In our case, due to the logistic analysis on driver seat belt use, we propose to use a score function composed of driver gender, vehicle type, passenger presence, and handheld cellphone use as follows:

$$S = X_g + X_v + X_{pp} + X_c$$

where X_g = driver gender (male = 0 and female = 1), X_v = size of vehicle driven standardized by the 3rd quartile (1/3 = SUV/Van/Truck, 2/3 = Minivan, and 1 = car), X_c = passenger presence (present = 1 and not present = 0), and X_c = driver cellphone use (no = 0 and yes = 1).

The final model is

$$\text{Log}\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 S.$$

4.3.3. Results

The logistic regression analysis yields parameter estimates (standard error) $\hat{\beta}_0 = 0.7229$ (0.1384) and $\hat{\beta}_1 = 0.4130$ (0.0609) for the 2012-2016 combined data (Table 12).

The AIC values (Table 13) are comparable to the AIC values in the weighted logistic analysis shown in 4.2.1 indicating a satisfactory fit of the model. The model is significant as indicated by the global null hypothesis test given in Table 14. The odds ratio estimate and its confidence interval are provided in Table 15. Figure 2 shows the regression line and 95% confidence limits for predicted probability of seatbelt use versus the weighted score function. The narrow confidence band and the linear upward trend also indicate a satisfactory fit of the model to the data. All such results conform with the findings by [27] in the bias reductions even in the nonresponse situation, and provide an improvement on their suggested approach.

Table 12. Analysis of maximum likelihood estimates.

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.7229	0.1384	27.2767	<0.0001
Score	1	0.4130	0.0609	46.0207	<0.0001

Table 13. Model fit statistics.

2012-2016	
Criterion	**Intercept and Covariates**
AIC	3002.597
SC	3020.729
−2 Log L	2998.597

Table 14. Testing global null hypothesis: BETA = 0.

Combined 2012, 2013, and 2014			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	47.4665	1	<0.0001
Score	46.7357	1	<0.0001
Wald	46.0207	1	<0.0001

Table 15. Odds ratio estimates.

Effect	Point Estimate	95% Wald Confidence Limits	
Score_Std_Reduced	1.511	1.341	1.703

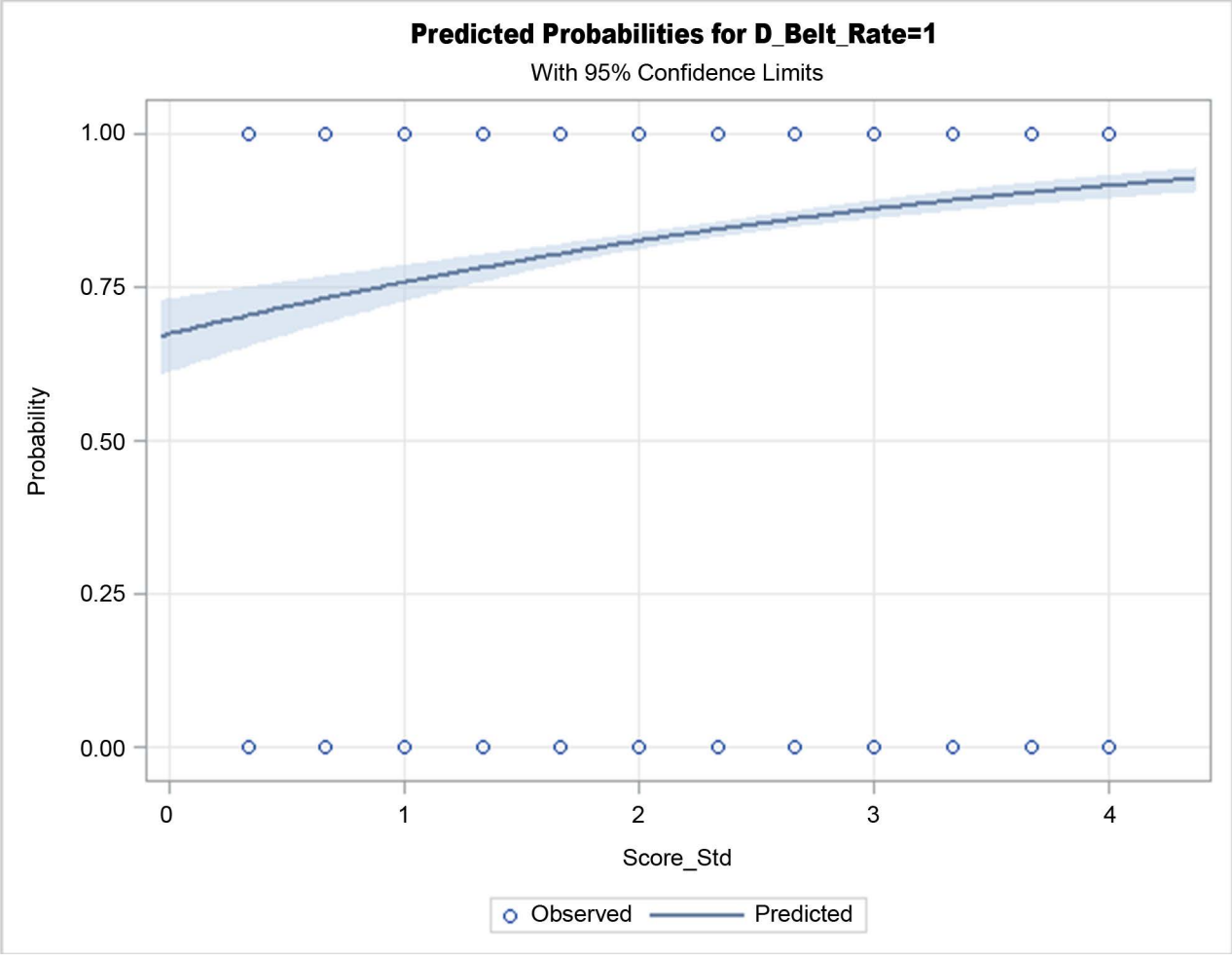

Figure 2. Logistic regression of seatbelt use versus weighted score.

The present IRT model offers many more advantages than the classical test theory (CTT) methods developed in Section 3. The model is parsimonious and allows driver seat belt behavior to be easily estimated from scaled psychometric

item measures under a weighted design model.

5. Conclusions

Driver seatbelt use in the Commonwealth of Virginia may be satisfactorily described using driver gender, vehicle type, passenger presence, and cellphone use in a multivariate logistic model using weights designed specifically for the dataset. However, prediction of seatbelt behavior is more appropriate using item response theory. As such, we have endeavored to build a score function considering driver gender, vehicle type driven, passenger presence, and cellphone usage by applying the IRT model with weights within the model. Fitting a weighted model results in significant improvements in goodness of fit statistics, such as AIC numbers, by factor of approximately 20.

We suggest that a weighted IRT model is more appropriate and it may also potentially include other factors. Such a model could be used to develop programs and more applications of the IRT models.

Acknowledgements

The authors are grateful to the referees and editor for their detailed suggestions, comments and insights, which improved the quality of the paper considerably. The research was made possible by financial support from the Virginia Department of Motor Vehicles via funding from the National Highway Traffic Safety Administration.

References

[1] Uniform Criteria for State Observational Surveys of Seat Belt Use, 23 CFR Part 1340 (2011).

[2] Lohr, S.L. (1999) Sampling: Design and Analysis. Duxbury Press, Pacific Grove, CA.

[3] Thomas, S.L. and Heck, R.H. (2001) Analysis of Large-Scale Secondary Data in Higher Education Research: Potential Perils Associated with Complex Sampling Designs. *Research in Higher Education*, **42**, 517-540. https://doi.org/10.1023/A:1011098109834

[4] Hahs-Vaughn, D.L. (2005) A Primer for Understanding Weights with National Datasets. *The Journal of Experimental Education*, **73**, 221-248. https://doi.org/10.3200/JEXE.73.3.221-248

[5] Korn, E.L. and Graubard, B.I. (1995) Examples of Differing Weighted and Unweighted Estimates from a Sample Survey. *The American Statistician*, **49**, 291-295.

[6] Preusser, D.F., Lund, A.K. and Williams, A.F. (1991) Characteristics of Belted and Unbelted Drivers. *Accident Analysis and Prevention*, **23**, 475-482. https://doi.org/10.1016/0001-4575(91)90013-U

[7] Pickrell, T.M. and Ye, T.J. (2009) Traffic Safety Facts (Research Note): Seat Belt Use in 2008—Demographic Results. National Highway Traffic Safety Administration/ Department of Transportation, Washington DC.

[8] Vivoda, J.M., Eby, D.W. and Kostyniuk, L.P. (2004) Differences in Safety Belt Use by Race. *Accident Analysis and Prevention*, **36**, 1105-1109. https://doi.org/10.1016/j.aap.2003.04.001

[9] Eby, D.W., Fordyce, T.A. and Vivoda, J.M. (2002) A Comparison of Safety Belt Use between Commercial and Noncommercial Light-Vehicle Occupants. *Accident Analysis and Prevention*, **34**, 285-291. https://doi.org/10.1016/S0001-4575(01)00024-0

[10] Glassbrenner, D. and Ye, J. (2006) Seat Belt Use in 2006—Overall Results. Traffic Safety Facts—Research Note (No. DOT HS 810 677). National Center for Statistics and Analysis, Washington DC.

[11] Boyle, J.M. and Vanderwolf, P. (2004) 2003 Motor Vehicle Occupant Safety Survey: Volume 2. Safety Belt Report (DOT-HS-809-789). National Highway Traffic Safety Administration, Washington DC.

[12] Nambisan, S.S. and Vasudevan, V. (2007) Is Seat Belt Usage by Front Seat Passengers Related to Seat Belt Usage by Their Drivers? *Journal of Safety Research*, **38**, 545-555. https://doi.org/10.1016/j.jsr.2007.06.002

[13] Shults, R.A. and Beck, L.F. (2012) Self-Reported Seatbelt Use, United States 2002-2010: Does Prevalence Vary by State and Type of Seatbelt Law? *Journal of Safety Research*, **43**, 417-420. https://doi.org/10.1016/j.jsr.2012.10.010

[14] Molnar, L.J., Eby, D.W., Dasgupta, K., Yang, Y., Nair, V.N. and Pollock, S.M. (2012) Explaining State-to-State Differences in Seat Belt Use: A Multivariate Analysis of Cultural Variables. *Accident Analysis and Prevention*, **47**, 78-86. https://doi.org/10.1016/j.aap.2012.01.006

[15] Özkan, T., Puvanachandra, P., Lajunen, T., Hoe, C. and Hyder, A. (2012) The Validity of Self-Reported Seatbelt Use in a Country Where Levels of Use Are Low. Accident Analysis and Prevention, **47**, 75-77. https://doi.org/10.1016/j.aap.2012.01.015

[16] Hardouin, J.B. and Mesbah, M. (2004) Clustering Binary Variables in Subscales Using an Extended Rasch Model and Akaike Information Criterion. *Communication in Statistics*, *Theory and Methods*, **33**, 1277-1294. https://doi.org/10.1081/STA-120030149

[17] Bartolucci, F. (2007) A Class of Multidimensional IRT Models for Testing Unidimensionality and Clustering Items. *Psychometrika*, **72**, 141-157. https://doi.org/10.1007/s11336-005-1376-9

[18] Rasch, G. (1961) On General Laws and the Meaning of Measurement in Psychology. *Proceedings of the IV Berkeley Symposium on Mathematical Statistics and Probability*, **4**, 321-333.

[19] Kaplan, D. and Ferguson, A.J. (1999) On the Utilization of Sample Weights in Latent Variable Models. *Structural Equation Modeling*: *A Multidisciplinary Journal*, **6**, 305-321. https://doi.org/10.1080/10705519909540138

[20] Asparouhov, T. (2006) General Multi-Level Modeling with Sampling Weights. *Communications in Statistics— Theory and Methods*, **35**, 439-460. https://doi.org/10.1080/03610920500476598

[21] Kleinbaum, D., Kupper, L., Nizam, A. and Rosenberg, E. (2013) Applied Regression Analysis and Other Multivariable Methods. Cengage Learning, Duxbury, Thomson.

[22] Montgomery, D.C., Peck, E.A. and Vining, G.G. (2012) Introduction to Linear Regression Analysis. 5th Edition, Wiley, Hoboken, New Jersey.

[23] Illi, S., Depner, M., Genuneit, J., Horak, E., Loss, G., Strunz-Lehner, C. and von Mutius, E. (2012) Protection from Childhood Asthma and Allergy in Alpine Farm Environments—The GABRIEL Advanced Studies. *Journal of Allergy and Clinical Immunology*, **129**, 1470-1477. https://doi.org/10.1016/j.jaci.2012.03.013

[24] Richards, S.A., Whittingham, M.J. and Stephens, P.A. (2011) Model Selection and Model Averaging in Behavioural Ecology: The Utility of the IT-AIC Framework. *Behavioral Ecology and Sociobiology*, **65**, 77-89. https://doi.org/10.1007/s00265-010-1035-8

[25] Samejima, F. (1996) Evaluation of Mathematical Models for Ordered Polychotomous Responses. *Behaviormetrika*, **23**, 17-35. https://doi.org/10.2333/bhmk.23.17

[26] Mesbah, M. (2010) Statistical Quality of Life. In: Balakrishnan, N., Ed., *Methods and Applications of Statistics in Life and Health Sciences*, Wiley, New York, 839-864. https://doi.org/10.1002/0471667196.ess7129

[27] Beaumont, J.F., Bocci, C. and Haziza, D. (2014) An Adaptive Data Collection Procedure for Call Prioritization. *Journal of Official Statistics*, **30**, 607-621. https://doi.org/10.2478/jos-2014-0040