UNITED STATES MILITARY ACADEMY

HOMEWORK 2

MA478

SECTION H2

COL. CLARK

BY

CADET AIMEE ROHAN'25, CO I3

WEST POINT, NEW YORK

15 FEB 2024

_ MY DOCUMENT IDENTIFIES ALL SOURCES USED AND ASSISTANCE RECEIVED IN COMPLETING THIS ASSIGNMENT.

_AR__ I DID NOT USE ANY SOURCES OR ASSISTANCE REQUIRING DOCUMENTATION IN COMPLETING THIS ASSIGNMENT.

SIGNATURE: 

# MA478 Insurance

Aimee Rohan Ramirez

February 2024

## 1 Introduction

Exploring the impact of demographic and personal factors on the likelihood of car wrecks (TARGET_FLAG) and the associated financial cost (TARGET_AMT), the analysis incorporates variables such as age, sex, education levels, license revoked status, marital status, presence of kids driving, number of kids, and driving purpose (commercial or private). The investigation utilizes logistic regression (Model 3) and multiple linear regression (Model 5) to quantify and interpret these effects statistically. Logistic regression Model 3 demonstrates superior predictive performance with an AUC of 0.7042, while multiple linear regression Model 5, featuring the significant predictor "INCOME," exhibits a lower AIC and improved interpretability. These models offer insurers valuable tools for precise risk assessment and informed decision-making, facilitating tailored adjustments in premiums based on individual driver characteristics. Notably, limitations include the exclusion of SEX and number of kids due to collinearity concerns and potential intertwined effects impacting explanatory power.

## 2 Data Exploration

How do demographic and personal factors, including age, sex, education levels, license revoked status, marital status, presence of kids driving, number of kids, and driving purpose (commercial or private), impact the likelihood of experiencing a car wreck (TARGET_FLAG) and the associated financial cost (TARGET_AMT)? We were given a dataset of 8162 observational units. The median age of our dataset was 45 years with a standard deviation of 8.6 years with sizes of 4376 Females and 3786 Males. Overall, there is missing data in some predictors of interest such as Income(435) and Age (6).

In analyzing the frequency of the TARGET_FLAG variable, it is evident that the occurrence of 0 (no car crash) is more prevalent, with 6007 instances, compared to 1 (car crash), which appears 2148 times. Examining the proportions of TARGET_FLAG by gender indicate similar distributions (Figure 1). Specifically, the proportion of TARGET_FLAG 0 is 0.728 for females and 0.746 for males. For TARGET_FLAG 1, the proportions are 0.272 for females and 0.253 for males. This implies a relatively even distribution of target flags across gender categories despite the males group being smaller.

In the univariate analysis of TARGET_AMT by gender, it is observed that both the mean and median values for the Female (F) and Male (M) groups are closely aligned, especially with a median of 0.000 (Table 1). This alignment suggests a similarity in the average target amounts for both groups. However, it is crucial to note that the standard deviations for both groups are relatively high, indicating substantial variability in the target amounts within each gender category.

Analyzing the association between kids driving and car crashes (TARGET_FLAG), a proportion comparison was conducted. This is due to a size discrepancy between the two groups; the group with no kids driving comprises 7174 units, while the group with kids driving has 981 units. (Figure 2) shows that proportion of accidents was higher in the group with kids driving. This difference in proportions suggests a potential relationship between kids driving and the occurrence of car crashes.

The analysis of crash cost (Target_AMT) in relation to Motor Vehicle Record (MVR) points (Figure 3) suggests a negative association, indicating that having more MVR points may be associated with lower damage costs. It's important to note that this observation is only a possible association and requires further investigation for a comprehensive understanding of this potential relationship.

In identifying associations between predictor variables and two response variables,Table 2 highlights predictors linked to the likelihood of car crashes, including MSTATUS, CAR_USE, EDUCATION, KIDSDRIV, REVOKED. These predictors emerged as factors in predicting the occurrence of car crashes.For the financial aspect of car crashes,TARGET_AMT, Table 3 underscores the influence of predictors such as EDUCATION, CAR_USE, REVOKED, KIDSDRIV, and MSTATUS. Notably, SEX exhibits weaker associationin predicting the financial implications of car crashes. Overall, this provides insight for constructing the models to predict both the likelihood of car crashes and the associated financial impact. The tables referenced encapsulate the specific associations and significances of each predictor in the context of the analysis.

In assessing collinearity, the significance of SEX varied when key predictors (EDUCATION, CAR_USE, MSTATUS) were individually omitted from the models, indicating potential collinearity with these variables.To address multicollinearity concerns, SEX was deliberately excluded from the models to ensure stability and minimize collinear influences. Notably, the predictors related to the number of children at home (HOMEKIDS) and the presence of a driving-age child (KIDSDRIVE) demonstrated a robust association, as evidenced by a Cramér's V test yielding a coefficient of 0.504. However, the influence of HOMEKIDS on the response variables was comparatively weak, leading to its exclusion from the models for enhanced stability. Consequently, it is acknowledged that the models may not fully disentangle the intertwined effects of sex and the number of children at home have on the mentioned predictors.

# 3 Data preparation

Statistically, I transformed the data by first adjusting the levels of categorical variables. I ensured consistent levels for factors such as 'SEX,' 'MSTATUS,' 'TARGET_FLAG,' 'EDUCATION,' 'PARENT1,' 'REVOKED,' 'RED_CAR,' and 'CAR_USE.' This not only improved data consistency but also facilitated proper interpretation.

For the 'KIDSDRIV' variable, I changed it to a binary factor ('Yes' or 'No') based on whether there were kids driving or not. I addressed missing or unexpected values by setting them to 'NO'.

Next, I handled missing values in 'INCOME' by imputing zeros when 'YOJ' (Years on Job) was zero, and for other cases, I filled in missing values with the median income. Additionally, I set the income to zero for cases where 'AGE' was greater than 62, assuming that average people retired at that age.

To address missing values in 'YOJ,' I replaced them with zeros. For 'AGE,' as there were only six missing values, I removed the corresponding rows.

These transformations were performed to ensure uniformity, handle missing data appropriately, and enhance the overall quality of the dataset for subsequent analyses. The use of median imputation and zero filling aligns with common practices to maintain the integrity of the data and minimize the impact of missing information.

# 4 Build Models

**Model 1**: This is a logistic regression model using the logit link as it is very interpretable. The predictors were hand selected from my steps in data exploration where I chose terms that had high associations to TARGET_FLAG. In terms of interpretation, each coefficient represents the change in the log-odds of the event (car crash) happening for a one-unit change in the predictor, holding other variables constant. The signs of the coefficients indicate the direction of the effect on the log-odds.

$$\log\left(\frac{p}{1-p}\right) = -0.62 + \begin{cases} -0.06, \text{High School} \\ -0.61, \text{Bachelors} \\ -0.67, \text{Masters} \\ -0.93, \text{PhD} \\ 0, < \text{High School} \end{cases} + \begin{cases} -0.69 \text{if MSTATUSYes} \\ +0.69 \text{if MSTATUSNo} \end{cases} + \begin{cases} +0.65 \text{CAR\_USECommercial} \\ -0.65 \text{CAR\_USEPrivate} \end{cases} + \begin{cases} 0.70 \text{KIDSDRIVYes} \\ -0.70 \text{KIDSDRIVNo} \end{cases} \tag{1}$$

What predicts to be a safer driver:

- Individuals with higher education (Bachelors, Masters, PhD) are associated with lower odds of a crash. To add, having a Ph.D. is associated with a $-60.6\%$ change in the odds of not crashing in this model.

- Being married is associated with lower odds of a crash. This aligns with common stereotypes of married individuals being perceived as safer drivers.
- Individuals without kids driving are associated with lower odds of a crash compared to those with kids driving.
- Individuals who drive commercially are more at risk for crashing in comparison to those who drive privately.

The model indicates that several predictors are statistically significant, including "EDUCATIONBachelors," "EDUCATIONMasters," "EDUCATIONPhD," "MSTATUSYes," "CAR_USECommercial," and "KIDSDRIVYes". However, "EDUCATIONHigh School" does not appear to be a statistically significant predictor, as its p-value is above the typical significance level of 0.05.

The AUC value for the model is 0.6527. The AUC (Area Under the Curve) is a measure of the model's ability to distinguish between positive and negative cases. Overall, the model demonstrates reasonable predictive ability, as evidenced by the AUC value.

**Model 2** This model had the addition of license revoked status. This model as well was chosen with the logit link due to it being very interpretable and scoring better on the AIC and AUC than the model with the "probit" link.

$$\log\left(\frac{p}{1-p}\right) =$$

$$=\beta_0 + \begin{cases}\beta_1, \text{High School}\\ \beta_2, \text{Bachelors}\\ \beta_3, \text{Masters}\\ \beta_4, \text{PhD}\\ \beta_5, < \text{High School}\end{cases} + \begin{cases}\beta_6, \text{if MSTATUSYes}\\ \beta_7, \text{if MSTATUSNo}\end{cases} + \begin{cases}\beta_8, \text{if CAR\_USECommercial}\\ \beta_9, \text{if CAR\_USEPrivate}\end{cases} + \begin{cases}\beta_{10}, \text{if KIDSDRIVYes}\\ \beta_{11}, \text{if KIDSDRIVNo}\end{cases} + \begin{cases}\beta_{12}, \text{if REVOKEDYes}\\ \beta_{13}, \text{if REVOKEDNo}\end{cases} \tag{2}$$

**Impacts** (reference Table 5 for the coefficients for this model ):

- The coefficients for level of education, married status (MSTATUS), purpose of car (CAR_USE), and whether kids are driving (KIDSDRIV) remain similar to the previous model, indicating their impact on crash risk.
- If a person's license has been revoked (REVOKEDYes), the odds of being involved in a car crash increase by approximately 145.96% (calculated $(e^{.9}-1)*100$) compared to someone with a license that has not been revoked. This suggests that individuals with a revoked license are more likely to be in a car crash, emphasizing the importance of considering the license status when assessing crash risk.

The only predictor in this model that was not significant were the education level being high school(P val = 0.361). The AUC value for the model was 0.666. This indicates moderate predictability,but may have limitations in accurately distinguishing between individuals who experience a car crash and those who do not.

**Model 3**: This model has the addition of the variable of motor vehicle points due to it scoring moderately as a correlation coefficient in Table 2 for the response of someone getting into a car crash (TARGET_FLAG). This model was chosen with the probit link towards the end due to it performing better than the logit link in AUC and considered close to the AIC value of the logit. Overall, this model still had a lower AIC value than the other 2 models. Below is a simplified equation for the model with the coefficients in Table 6.

$$\Phi^{-1}(p) = \beta_0+\beta_1\times\text{EDUCATION}+\beta_2\times\text{MSTATUS}+\beta_3\times\text{CAR\_USE}+\beta_4\times\text{KIDSDRIV}+\beta_5\times\text{REVOKED}+\beta_6\times\text{MVR\_PTS}$$

Impacts:

- MSTATUSYes (-0.3833): The odds decrease for a car crash by approximately 31.5% for being married compared to not being married.

- CAR_USECommercial (0.3563): The odds increase for a car crash by approximately 42.8% for using a car for commercial purposes.
- KIDSDRIVYes (0.3865): The odds increase for a car to be in a crash by approximately 47.4% for having kids who drive.
- REVOKEDYes (0.5204): The odds increase for a car to be in a crash by approximately 67.9% for having a revoked license.
- MVR_PTS (0.1196): The odds increase for a car crash by approximately 12.6% for each additional point in the number of traffic violations.

Examining the model more, the only predictor again that was not significant due to P Value was Education at the high school level (P value = 0.483). The AUC value as well was 0.7042 meaning that the model has acceptable discrimination ability to distinguish between individuals whose car has experienced a car crash (positive class) and those who did not.

**MULTIPLE LINEAR REGRESSION MODELS FOR TARGET_AMT**
**Model 4**

$$\begin{aligned}\text{TARGET\_AMT} =& 1901.5212 - 235.5327 \times \text{EDUCATIONHigh School} + 638.9767 \times \text{EDUCATIONBachelors} \\ & - 626.9749 \times \text{EDUCATIONMasters} - 761.8124 \times \text{EDUCATIONPhD} + 938.1622 \times \text{CAR\_USECommercial} \\ & + 761.0551 \times \text{REVOKEDYes} + 880.1136 \times \text{KIDSDRIVYes} - 846.6190 \times \text{MSTATUSYes}\end{aligned} \tag{3}$$

**IMPACTS**

- High School (-235.5): Holding other factors constant, having an education level of High School is associated with a decrease in the estimated price if there is a car crash by $235.5 compared to the reference level.However, this is not a definite answer as the p-value for this predictor was not smaller than the 0.05 significance level. Overall, looking at the higher educational levels there is a trend in there being a decrease in estimated price of a crash as those values did have significant p-values.
- Commercial car use (938.2): If the car use is commercial, it is associated with an increase in the estimated target amount by $938.2 compared to private car use.
- License Revoked status: If the driver's license is revoked, it is associated with an increase in the estimated target amount by $761.1 compared to not being revoked.
- Kids driving (880.1): If there are kids driving, it is associated with an increase in the estimated target amount by $880.1 compared to no kids driving.
- Marital status(-846.6): Being married is associated with a decrease in the estimated target amount by $846.6 compared to not being married.

In the multiple linear regression model (Model 4), several predictors were identified as statistically significant contributors to the estimated financial impact of car crashes. Education levels played a significant role, with individuals holding Bachelor's, Master's, or Ph.D. degrees associated with decreased crash costs (p-values: 0.000138, 0.000375, 0.000502, respectively). Commercial car use exhibited a substantial positive impact (estimate: 938.2, p-value: ¡ 2e-16), indicating higher estimated costs for crashes involving commercial vehicles. Additionally, having a revoked license (estimate: 761.1, p-value: 1.30e-06), kids driving (estimate: 880.1, p-value: 3.00e-08), and being married (estimate: -846.6, p-value: <2e-16) were significant predictors, highlighting their influence on the financial implications of car crashes. These statistical insights offer a nuanced understanding of the specific factors shaping the economic consequences of automotive incidents in the context of the model.

In summary, while most predictors are significant, the overall explanatory power of the model is limited. This is due to a high AIC (160859.9) and high residual standard error (4643). Consideration of additional predictors or model enhancements may be beneficial for improving model performance.

**Model 5**:

$$\begin{aligned}\text{TARGET\_AMT} =& 2034.843 - 186.6868 \times \text{EDUCATIONHigh School} \\ & - 448.2528 \times \text{EDUCATIONBachelors} - 315.6636 \times \text{EDUCATIONMasters} \\ & - 251.6398 \times \text{EDUCATIONPhD} + 999.8388 \times \text{CAR\_USECommercial} \\ & + 752.7800 \times \text{REVOKEDYes} + 877.3270 \times \text{KIDSDRIVYes} \\ & - 849.1442 \times \text{MSTATUSYes} - 0.005389065 \times \text{INCOME}\end{aligned} \tag{4}$$

**IMPACTS**:

- Car Use: Using the car for commercial purposes is associated with an increase of $999.80 in the car crash price
- Having kids driving is associated with an increase of $877.30 in the car crash price.
- For each additional unit increase in income, the car crash price is expected to decrease by $0.0054. This effect is statistically significant, indicating that higher income is associated with a lower car crash price.
- The effects of education are not statistically significant in this model. Thus, based on their p-values, do not provide enough evidence to reject the null hypothesis that their coefficients are equal to zero. (EDUCATIONHigh School p-value is 0.2662, EDUCATIONMasters p-value is 0.1026, EDUCATIONPhD p-value is 0.3236).

In summary, the model suggests that factors such as car use, license status, presence of kids driving, marital status, and income can impact the car crash price. The AIC did show some improvement, but it still is a relatively high value (160846.7).

# 5 Best Models

The logistic regression model (Model 3) achieved superior predictive performance with an AUC of 0.7042, providing insurers with a reliable tool for assessing car crash risks. In the multiple linear regression framework, Model 5, featuring the significant predictor "INCOME," exhibited a lower AIC and improved interpretability. These findings suggest that insurers can enhance risk assessment and pricing strategies by leveraging these models, contributing to more precise premium adjustments based on individual driver characteristics. The incorporation of income as a predictor enhances the models' ability to explain variations in the target amount, offering insurers a valuable tool for data-driven decision-making. Acknowledging limitations, SEX was excluded due to varying significance when key predictors were omitted, addressing collinearity concerns. The robust association between HOMEKIDS and KIDSDRIVE, evidenced by a Cramér's V coefficient of 0.504, led to the exclusion of HOMEKIDS for enhanced stability. Despite efforts, the models may not fully disentangle the intertwined effects of sex and the number of children at home on the mentioned predictors, impacting explanatory power.

# 6 Appendix

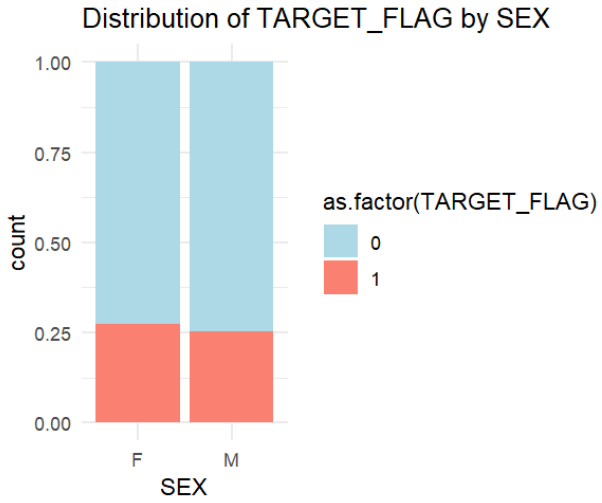


Figure 1: Proportion of Target-Flag by Sex

| | **Females** | **Males** | **Target_AMT** |
|---|---|---|---|
| **Mean** | 1454.516 | 1558.733 | 1503 |
| **Median** | 0.000 | 0.000 | 0.000 |
| **Standard deviation** | 4146.334 | 5277.489 | 4705.129 |
| **Min** | 0.000 | 0.000 | 0.106928 |
| **Max** | 85523.653 | 107586.136 | 107586 |

**Background:** These statistics provide a detailed breakdown of the financial implications of car wrecks (Target_AMT), with gender-specific insights complementing the overall picture.

Table 1: Summary Statistics for Cost of Car Crash (TARGET-AMT)

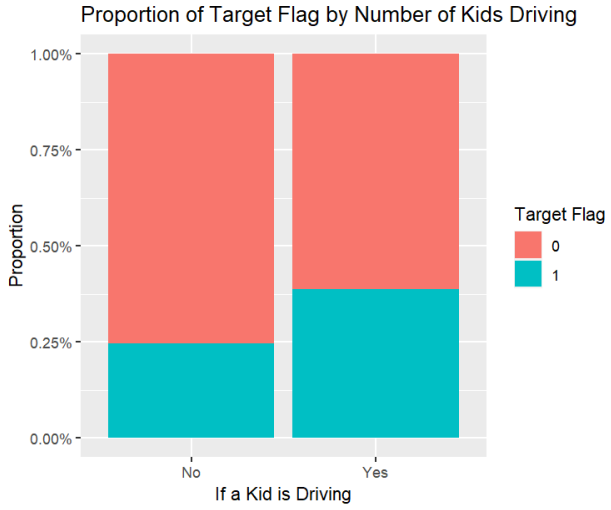


Figure 2: Proportion of Target_Flag depending on if Kids are Driving

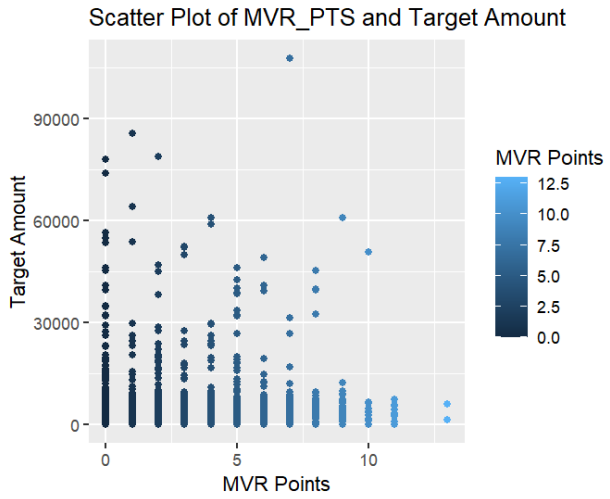


Figure 3: Proportion of Target_Flag depending on if Kids are Driving

| Predictor | Correlation matrix or Chi-Square | Interpretation |
|---|---|---|
| AGE | -0.1366 | Weak negative correlation |
| INCOME | 0.1142 | Weak positive correlation |
| HOMEKIDS | 0.2188 | Moderate positive correlation |
| MVR_PTS | < 0.001 | Significant association |
| MSTATUS | < 0.001 | Significant association |
| CAR_USE | 0.0608 | Marginally significant |
| SEX | < 0.001 | Significant association |
| EDUCATION | < 0.001 | Significant association |
| KIDSDRIV | < 0.001 | Significant association |
| REVOKED | 0.4624 | Not significant |
| RED_CAR | | |

**Background:** These are the possible associations between predictors of interest and Target_AMT using an ANOVA test.

| Predictor | ANOVA p-value | Significance |
|---|---|---|
| EDUCATION | < 0.001 | Significant |
| CAR_USE | < 0.001 | Significant |
| SEX | 0.215 | Not significant |
| REVOKED | < 0.001 | Significant |
| KIDSDRIV | < 0.001 | Significant |
| MSTATUS | < 0.001 | Significant |
| HOMEKIDS | 0.006946 | Significant |
| INCOME | 0.007439 | Significant |

**Background:** These are the possible associations between predictors of interest and Target_Flag using either correlation matrix (numerical variables) or Chi-Square test (categorical)

Table 2: Associations with TARGET_FLAG

Table 3: Predictors Affecting TARGET_AMT

| **Predictor** | **Pr($>|z|$)** |
|---|---|
| EDUCATIONHigh School | 0.435 |
| EDUCATIONBachelors | $2.05 \times 10^{-13}$ |
| EDUCATIONMasters | $1.15 \times 10^{-13}$ |
| EDUCATIONPhD | $8.03 \times 10^{-15}$ |
| MSTATUSYes | $< 2 \times 10^{-16}$ |
| CAR_USECommercial | $< 2 \times 10^{-16}$ |
| KIDSDRIVYes | $< 2 \times 10^{-16}$ |

**Background:** This was the summary output for the logistic regression for model 1.

Table 4: Significant Predictors for Model 1 of Logistic Regression

| | |
|---|---|
| (Intercept) | -0.747 |
| EDUCATIONHigh School | -0.073 |
| EDUCATIONBachelors | -0.615 |
| EDUCATIONMasters | -0.689 |
| EDUCATIONPhD | -0.925 |
| MSTATUSYes | -0.671 |
| CAR_USECommercial | 0.649 |
| KIDSDRIVYes | 0.676 |
| REVOKEDYes | 0.900 |

Table 5: Coefficients for the Logistic Regression Model 2

| | |
|---|---|
| **(Intercept)** | -0.6773180 |
| **EDUCATIONHigh School** | -0.0343026 |
| **EDUCATIONBachelors** | -0.3627590 |
| **EDUCATIONMasters** | -0.3971267 |
| **EDUCATIONPhD** | -0.5238834 |
| **MSTATUSYes** | -0.3832966 |
| **CAR_USECommercial** | 0.3563177 |
| **KIDSDRIVYes** | 0.3864763 |
| **REVOKEDYes** | 0.5203932 |
| **MVR_PTS** | 0.1195776 |

Table 6: Coefficients of the Logistic Probit Link Model