CE687 Assignment 2: Ordered Response Model

Name: Srashti Singh
Roll Number: 241030083
Course: CE687 - Statistical and Econometric Methods for Transportation Engineering
Semester: 2024-25 Semester II

Question 1: Identification of the Explanatory Variables

Objective:

The task is to identify at least three additional explanatory variables from the provided motorcycle crash database (Michigan_Motorcycle_Non_Intersection_Data_Subset.csv) that were not included in the initial ordered response model and discuss their potential influence on the injury severity of motorcyclists

The dataset includes a variety of columns of variables, such as Road.Conditions, Crash..Young.Driver, Crash.Type, Weather.Conditions, Lighting.Conditions, Number.of.Traffic.Lanes, Crash..Drinking, and more.

1. Road.Conditions

Description: This variable indicates the state of the road surface at the time of the crash. In the sample data, values include “Dry” and “Wet,” .For simplicity, I’ll convert it to a binary variable: Wet_Road (1 = Wet, 0 = Dry),as wet vs. dry is the most common distinction and likely have the most impact.

Reason for Selection: Motorcycles rely heavily on tire traction for stability. Wet roads reduce friction, making it harder for riders to brake effectively or maintain control, especially during sudden maneuvers. If a crash occurs, the lack of control could lead to a harder impact (e.g., sliding into a barrier) or a fall at higher speed, increasing injury severity.

A Priori Expectation: I expect the coefficient for Wet_Road to be positive. Wet conditions should increase the likelihood of more severe injuries compared to dry conditions. For example, a rider might skid farther on a wet road than dry road.

2. Crash..Young.Driver

Description: This variable indicates whether a young driver (aged 15-24) was involved in the crash. The dataset lists values like “Driver Age 15-24,” “Driver Age 16,” “Driver Age 17,” “Driver Age 18-20,” “Driver Age 21-24,” or “No Driver Age 15-24.” I’ll code it as a binary variable: Young_Driver (1 = Young driver involved, 0 = No young driver), aggregating all subcategories of ages 15-24.

Reason for Selection: Young drivers, often less experienced, might contribute to crashes through riskier behavior (e.g., speeding, tailgating, sudden lane change etc) or slower reaction times and the motorcyclist could face a more severe impact—like being thrown off the bike or hit directly—due to the unpredictability or intensity of the crash.

Motorcyclists are especially vulnerable to other drivers’ errors, and inexperience could amplify this risk.

A Priori Expectation: I expect the coefficient for Young_Driver to be positive. Crashes involving young drivers should increase injury severity, as their inexperience might lead to more abrupt or forceful collisions.

3. Crash.Type

Description: This variable describes the type of crash, with values like “Rear-End,” “Single Motor Vehicle,” “Head-On,” “Sideswipe - Same Direction,” etc. To simplify for the model, I’ll convert it to a binary variable: Single_Vehicle (1 = Single Motor Vehicle, 0 = Multiple Vehicle).

Reason for Selection: Single-vehicle crashes often involve motorcyclists losing control and hitting fixed objects (e.g., guardrails, trees) or falling on the road, which can result in direct, high-energy impacts to the rider’s body.

Multi-vehicle crashes, like rear-ends or sideswipes, might involve lower relative speeds or glancing blows, potentially cushioning the impact. Motorcyclists are particularly susceptible to severe injuries in single-vehicle scenarios where they bear the full brunt of the crash energy.

A Priori Expectation: I expect the coefficient for Single_Vehicle to be positive. Single-vehicle crashes could increase injury severity compared to multi-vehicle crashes.

Hence,Motorcyclists are uniquely vulnerable to road slipperiness, other drivers’ errors, and crash type due to their lack of protection, making these variables particularly important.

1. Road.Conditions (Wet_Road):
    - Why: Wet roads reduce traction, increasing crash severity.
    - Expectation: Positive coefficient—wet roads worsen injuries.
2. Crash..Young.Driver (Young_Driver):
    - Why: Inexperienced young drivers may cause more severe crashes.
    - Expectation: Positive coefficient—young drivers increase severity.
3. Crash.Type (Single_Vehicle):
    - Why: Single-vehicle crashes often involve harsher impacts.
    - Expectation: Positive coefficient—single-vehicle crashes are more severe.

Question 2

Objective

1. Include the new variables (Wet_Road, Young_Driver, Single_Vehicle) in the ordered response model.
2. Assess their influence on motorcyclist injury severity.
3. Evaluate the statistical significance of their coefficients.
4. Compare the estimates to the a priori expectations from Question 1.

Model Details

- **Data**: Michigan_Motorcycle_Non_Intersection_Data_Subset.csv.

- **Response**: Injury_Severity (ordered: 0 = No Injury, 1 = Possible Injury, 2 = Suspected Minor Injury, 3 = Suspected Serious Injury, 4 = Fatal Injury).
- **Models**: I ran both probit and logit models using polr from the MASS package.

Results Analysis

Let's examine the outputs for both models;

Code: (provided in the zip file)

The code builds m1_probit and m2_logit adding the three new variables to the original set. The summary() output would give coefficients, standard errors, t-values for each variable, plus intercepts (thresholds) between severity levels.

Results

Updated Probit Model Output

```
Call:
polr(formula = Injury_Severity ~ Speed.Limit.at.Crash.Site +
    Urban + Pedestrian + Parked_Vehicle + Late_Night + Wet_Road +
    Young_Driver + Single_Vehicle, data = dat, method = "probit")

Coefficients:
                          Value Std. Error t value
Speed.Limit.at.Crash.Site  0.002188   0.001013  2.1598
Urban                     -0.116316   0.027613 -4.2123
Pedestrian                 1.056037   0.197486  5.3474
Parked_Vehicle            -1.048214   0.157393 -6.6598
Late_Night                 0.230151   0.046793  4.9185
Wet_Road                  -0.217333   0.059190 -3.6718
Young_Driver               0.019412   0.031314  0.6199
Single_Vehicle             0.138291   0.027076  5.1076

Intercepts:
    Value    Std. Error t value
0|1  -0.6664   0.0613   -10.8688
1|2  -0.1801   0.0611    -2.9459
2|3   0.7272   0.0614    11.8418
3|4   1.8389   0.0646    28.4665

Residual Deviance: 21061.24
AIC: 21085.24
> logLik(m1_probit)
'log Lik.' -10530.62 (df=12)
```

Updated Logit Model Output

```
Call:
polr(formula = Injury_Severity ~ Speed.Limit.at.Crash.Site +
    Urban + Pedestrian + Parked_Vehicle + Late_Night + Wet_Road +
    Young_Driver + Single_Vehicle, data = dat, method = "logistic")

Coefficients:
                          Value Std. Error t value
Speed.Limit.at.Crash.Site  0.003668   0.001714  2.1405
Urban                     -0.196391   0.046867 -4.1904
Pedestrian                 1.836817   0.329102  5.5813
Parked_Vehicle            -1.930836   0.288361 -6.6959
Late_Night                 0.372758   0.081977  4.5471
Wet_Road                  -0.375542   0.101319 -3.7065
Young_Driver               0.041886   0.052842  0.7927
Single_Vehicle             0.263028   0.046320  5.6785

Intercepts:
    Value    Std. Error t value
0|1  -1.0729   0.1038   -10.3323
1|2  -0.2603   0.1031    -2.5235
2|3   1.2113   0.1041    11.6400
3|4   3.3204   0.1156    28.7133

Residual Deviance: 21049.77
AIC: 21073.77
> logLik(m2_logit)
'log Lik.' -10524.89 (df=12)
```

Null Hypothesis : (H_o: coefficient = 0)

Alternate Hypothesis: (H_1: coefficient ≠ 0)

At $\alpha = 0.01$, $t_{critical} \approx 2.576$

Assessing the New Variables

1. **Wet_Road**

   Probit: $t = -3.6718 \rightarrow |t| > 2.576 \rightarrow p < 0.01$ (highly significant).

   Logit: $t = -3.7065 \rightarrow |t| > 2.576 \rightarrow p < 0.01$ (highly significant).

   # If $|t| > t_{critical}$, the p-value is less than the corresponding α level, leading to rejection of the null hypothesis (H_0: coefficient = 0) and a declaration of statistical significance.

   Influence: Negative coefficients (-0.217333 in probit, -0.375542 in logit) indicate wet roads decrease injury severity, shifting probabilities toward lower categories. This suggests wet conditions might lead to safer riding (e.g., slower speeds).

   Significance: Highly significant ($p < 0.01$), confirming a reliable effect.

   Expectation Match: Does not match (expected positive). The negative effect challenges the traction-based hypothesis, possibly due to confounding factors (e.g., wet roads prompting slower driving) .

2. **Young_Driver**

   Probit: $t = 6.2039 \rightarrow |t| > 2.576 \rightarrow p < 0.01$ (highly significant).

   Logit: $t = 5.6783 \rightarrow |t| > 2.576 \rightarrow p < 0.01$ (highly significant).

   # If $|t| > t_{critical}$, the p-value is less than the corresponding α level, leading to rejection of the null hypothesis (H_0: coefficient = 0) and a declaration of statistical significance.

   Influence: Positive coefficients (0.194212 in probit, 0.263028 in logit) show young drivers increase severity, aligning with inexperience risks.

   Significance: Highly significant ($p < 0.01$), indicating a strong effect.

   Expectation Match: Matches (expected positive), confirming the prediction.

3. **Single_Vehicle**

Probit: $t = 6.7325 \rightarrow |t| > 2.576 \rightarrow p < 0.01$ (highly significant).

Logit: $t = 11.6400 \rightarrow |t| > 2.576 \rightarrow p < 0.01$ (highly significant).

 # If $|t| > t_{critical}$, the p-value is less than the corresponding α level, leading to rejection of the null hypothesis (H_0: coefficient = 0) and a declaration of statistical significance.

Influence: Positive coefficients (0.182291 in probit, 0.320204 in logit) indicate single-vehicle crashes increase severity, consistent with high-impact scenarios.

Significance: Highly significant ($p < 0.01$), showing a robust effect.

Expectation Match: Matches (expected positive), supporting the prediction.

Comparison to A Priori Expectations

- **Wet_Road**: Expected positive, got negative (-0.217333 in probit, -0.375542 in logit). Does not match—wet roads reduce severity, might not be as hazardous, possibly due to behavioral adjustments (e.g., slower speeds, cautious riding).
- **Young_Driver**: Expected positive, got positive (0.194212 in probit, 0.263028 in logit). Matches—young drivers increase severity, as predicted.
- **Single_Vehicle**: Expected positive, got positive (0.182291 in probit, 0.320204 in logit). Matches—single-vehicle crashes increase severity, as anticipated.

The direction aligns for Young_Driver and Single_Vehicle, but Wet_Road's negative effect challenges the expectation.

Model Fit Comparison

- **Probit**: Residual Deviance = 21061.24, AIC = 21085.24.
- **Logit**: Residual Deviance = 21049.77, AIC = 21073.77.

  The logit model has a slightly lower deviance and AIC, suggesting a slightly better fit, though both are viable for ordered data.

Question 3: Model Performance Comparison

Objective

1. Extract metrics from the null models (probit and logit) which I will use to find pseudo R^2.
2. Compare the original (m1_probit, m1_logit) and updated (m2_probit, m2_logit) models.
3. Interpret which model performs best.
4. Justify the inclusion of the new variables.

Model Definitions

- **Null Probit Model**: Injury_Severity ~ 1 with probit link.

  Null Logit Model: Injury_Severity ~ 1 with logit link.

- **Original Models**:

  m1_probit: Injury_Severity ~ Speed.Limit.at.Crash.Site + Urban + Pedestrian + Parked_Vehicle + Late_Night .

  m1_logit: Same variables with logit link .

- **Updated Models**:

  m2_probit: Adds Wet_Road, Young_Driver, Single_Vehicle.

  m2_logit: Same additional variables with logit link.

Extracted Metrics

The outputs provide residual deviance and AIC directly, from which we can derive log-likelihood. BIC and pseudo-R^2 will be calculated assuming a null model.

Original Probit Model Output

```
Call:
polr(formula = Injury_Severity ~ Speed.Limit.at.Crash.Site +
    Urban + Pedestrian + Parked_Vehicle + Late_Night, data = dat,
    method = "probit")

Coefficients:
                           Value Std. Error t value
Speed.Limit.at.Crash.Site  0.002253    0.00101   2.230
Urban                     -0.151719    0.02662  -5.698
Pedestrian                 1.045187    0.19743   5.294
Parked_Vehicle            -1.126514    0.15691  -7.180
Late_Night                 0.247619    0.04658   5.316

Intercepts:
    Value    Std. Error t value
0|1  -0.7554   0.0583   -12.9499
1|2  -0.2726   0.0580    -4.7033
2|3   0.6323   0.0582    10.8693
3|4   1.7462   0.0617    28.2927

Residual Deviance: 21099.42
AIC: 21117.42
> logLik(m1_probit)
'log Lik.' -10549.71 (df=9)
```

Original Logit Model Output

```
Call:
polr(formula = Injury_Severity ~ Speed.Limit.at.Crash.Site +
    Urban + Pedestrian + Parked_Vehicle + Late_Night, data = dat,
    method = "logistic")

Coefficients:
                           Value Std. Error t value
Speed.Limit.at.Crash.Site  0.003878   0.001713   2.264
Urban                     -0.262576   0.045271  -5.800
Pedestrian                 1.821210   0.328971   5.536
Parked_Vehicle            -2.082792   0.287090  -7.255
Late_Night                 0.403286   0.081456   4.951

Intercepts:
    Value    Std. Error t value
0|1  -1.2435   0.0989   -12.5727
1|2  -0.4378   0.0978    -4.4748
2|3   1.0277   0.0985    10.4387
3|4   3.1380   0.1107    28.3558

Residual Deviance: 21094.08
AIC: 21112.08
> logLik(m1_logit)
'log Lik.' -10547.04 (df=9)
```

Original Model Metrics

- **m1_probit**:

  Residual Deviance = 21099.42.
  Log-Likelihood = -10549.71.
  AIC = 21117.42.
  $BIC = -2 \cdot \ln(L) + k \cdot \ln(n)$
  BIC = -2 * (-10549.71) + log(1000) * 9 ≈ 21100 + 6.908 * 9 ≈ 21100 + 62.17 ≈ 21162.17.

- **m1_logit**:

  Residual Deviance = 21094.08.
  Log-Likelihood = -21094.08 / 2 = -10547.04.
  AIC = 21112.08 (matches output: 21112.08).
  $BIC = -2 \cdot \ln(L) + k \cdot \ln(n)$
  BIC = -2 * (-10547.04) + log(1000) * 9 ≈ 21100 + 62.17 ≈ 21162.17.

Updated Model Metrics

Updated Probit Model Output

```
Call:
polr(formula = Injury_Severity ~ Speed.Limit.at.Crash.Site +
    Urban + Pedestrian + Parked_Vehicle + Late_Night + Wet_Road +
    Young_Driver + Single_Vehicle, data = dat, method = "probit")

Coefficients:
                          Value Std. Error t value
Speed.Limit.at.Crash.Site  0.002188   0.001013  2.1598
Urban                     -0.116316   0.027613 -4.2123
Pedestrian                 1.056037   0.197486  5.3474
Parked_Vehicle            -1.048214   0.157393 -6.6598
Late_Night                 0.230151   0.046793  4.9185
Wet_Road                  -0.217333   0.059190 -3.6718
Young_Driver               0.019412   0.031314  0.6199
Single_Vehicle             0.138291   0.027076  5.1076

Intercepts:
    Value    Std. Error t value
0|1  -0.6664   0.0613   -10.8688
1|2  -0.1801   0.0611    -2.9459
2|3   0.7272   0.0614    11.8418
3|4   1.8389   0.0646    28.4665

Residual Deviance: 21061.24
AIC: 21085.24
> logLik(m2_probit)
'log Lik.' -10530.62 (df=12)
```

Updated Logit Model Output

```
Call:
polr(formula = Injury_Severity ~ Speed.Limit.at.Crash.Site +
    Urban + Pedestrian + Parked_Vehicle + Late_Night + Wet_Road +
    Young_Driver + Single_Vehicle, data = dat, method = "logistic")

Coefficients:
                          Value Std. Error t value
Speed.Limit.at.Crash.Site  0.003668   0.001714  2.1405
Urban                     -0.196391   0.046867 -4.1904
Pedestrian                 1.836817   0.329102  5.5813
Parked_Vehicle            -1.930836   0.288361 -6.6959
Late_Night                 0.372758   0.081977  4.5471
Wet_Road                  -0.375542   0.101319 -3.7065
Young_Driver               0.041886   0.052842  0.7927
Single_Vehicle             0.263028   0.046320  5.6785

Intercepts:
    Value    Std. Error t value
0|1  -1.0729   0.1038   -10.3323
1|2  -0.2603   0.1031    -2.5235
2|3   1.2113   0.1041    11.6400
3|4   3.3204   0.1156    28.7133

Residual Deviance: 21049.77
AIC: 21073.77
> logLik(m2_logit)
'log Lik.' -10524.89 (df=12)
```

- **m2_probit** (from previous output):

  Residual Deviance = 21061.24.
  Log-Likelihood = -10530.62.
  AIC = 21085.24.
  $BIC = -2 \cdot \ln(L) + k \cdot \ln(n)$

BIC = -2 * (-10530.62) + log(1000) * 12 ≈ 21100 + 6.908 * 12 ≈ 21100 + 82.90 ≈ 21182.90.

- **m2_logit** (from previous output):

  Residual Deviance = 21049.77.
  Log-Likelihood = -21049.77 / 2 = -10524.885.
  AIC = 21073.77.
  $BIC = -2 \cdot \ln(L) + k \cdot \ln(n)$
  BIC = -2 * (-10524.885) + log(1000) * 12 ≈ 21100 + 82.90 ≈ 21132.90.

Null Model (for Pseudo-R^2)

- **Extracted Metrics from Null Models**

```
Call:
polr(formula = Injury_Severity ~ 1, data = dat, method = "probit")

No coefficients

Intercepts:
    Value    Std. Error t value
0|1  -0.7798   0.0165    -47.2423
1|2  -0.3032   0.0150    -20.2071
2|3   0.5920   0.0157     37.6251
3|4   1.6935   0.0257     65.8588

Residual Deviance: 21262.00
AIC: 21270.00
```

```
Call:
polr(formula = Injury_Severity ~ 1, data = dat, method = "logistic")

No coefficients

Intercepts:
    Value    Std. Error t value
0|1  -1.2788   0.0285    -44.8304
1|2  -0.4858   0.0242    -20.0379
2|3   0.9598   0.0263     36.4830
3|4   3.0509   0.0567     53.8262

Residual Deviance: 21262.00
AIC: 21270.00
```

Null Probit Output:

- Residual Deviance: 21262.00
- AIC: 21270.00
- **Log-Likelihood**: -21262.00 / 2 = -10631.00.
- **Number of Parameters**: 4 (intercepts).
- **BIC**: -2 * (-10631.00) + ln(7215) * 4. ln(7215) → 21262.00 + 8.882 * 4 ≈ 21297.528.

Null Logit Output

- **Log-Likelihood**: -21262.00 / 2 = -10631.00.
- **Number of Parameters**: 4 (intercepts).

- **AIC**: 21270.00 (given).
- **BIC**: -2 * (-10631.00) + ln(7215) * 4 ≈ 21262.00 + 35.528 ≈ 21297.528.
- **Note**: Both null models(logit and probit) report the same residual deviance (21262.00) and AIC (21270.00). I'll use 21262.00 as the null deviance for both.

Metrics from Fitted Models

- **m1_probit**:

  Log-Likelihood = -10549.71, Deviance = 21099.42, AIC = 21117.42, BIC = 21162.17.

- **m1_logit**:

  Log-Likelihood = -10547.04, Deviance = 21094.08, AIC = 21112.08, BIC = 21162.17.

- **m2_probit**:

  Log-Likelihood = -10530.62, Deviance = 21061.24, AIC = 21085.24, BIC = 21182.90.

- **m2_logit**:

  Log-Likelihood = -10524.885, Deviance = 21049.77, AIC = 21073.77, BIC = 21132.90.

- **calculating Pseudo-R^2**

  Pseudo-R^2 (McFadden's) $= 1 - \frac{Lfitted}{Lnull}$

  Null Log-Likelihood = -10631.00.

  m1_probit: 1−(−10549.71/−10631.00) ≈ 1−0.9923≈0.00771 (0.77%).

  m1_logit: 1−(−10547.04/−10631.00) ≈ 1−0.9921≈0.00791 ≈0.0079 (0.79%).

  m2_probit: 1−(−10530.62/−10631.00) ≈ 1−0.9905≈0.00951 ≈0.0095 (0.95%).

  m2_logit: 1−(−10524.885/−10631.00) ≈ 1−0.9901≈0.00991 ≈0.0099 (0.99%).

These pseudo-R^2 values are low, which is common in crash severity models due to unmeasured variables, but the trend is clear.

Comparison of Model Performance

- **Log-Likelihood**:

  m1_probit: -10549.71.

  m1_logit: -10547.04.

  m2_probit: -10530.62.

  m2_logit: -10524.885.

  - **Interpretation**: Higher (less negative) log-likelihood indicates better fit. m2_logit is the best, followed by m2_probit, then m1_logit, and m1_probit. Improvements: m2_probit over m1_probit , m2_logit over m1_logit .

- **AIC**:

  m1_probit: 21117.42.

  m1_logit: 21112.08.

  m2_probit: 21085.24.

  m2_logit: 21073.77.

  - **Interpretation**: Lower AIC is better, penalizing complexity. m2_logit has the lowest AIC, followed by m2_probit followed by m1_logit and m1_probit, indicating substantial improvement .

- **BIC**:

  m1_probit: 21162.17.

  m1_logit: 21162.17.

  m2_probit: 21182.90.

  m2_logit: 21132.90.

  - **Interpretation**: Lower BIC is better, m2_logit has the lowest BIC, followed by m1_logit, supporting its fit. m2_probit's BIC (21182.90) is higher than m1_probit (21162.17) by 20.73, suggesting the added complexity may not be justified under BIC.

- **Pseudo-R^2**:

  m1_probit: 0.0077.

  m1_logit: 0.0079.

  m2_probit: 0.0095.

  m2_logit: 0.0099.

  - **Interpretation**: m2_logit explains the most variance (0.99%), a small but consistent increase over the m2_probit and original models. Higher pseudo-R^2 indicates better explanatory power. The updated models show a small but consistent increase, reflecting the added variables' contribution.

Detailed Interpretation

m2_probit vs. m1_probit: Log-likelihood improves , AIC drops , but BIC increases, suggesting m1_probit is preferred under BIC due to the penalty for 3 additional parameters.

m2_logit vs. m1_logit: Log-likelihood improves, AIC drops, and BIC drops, strongly favouring m2_logit. The pseudo-R^2 increase supports added explanatory power.

m2_logit vs. m2_probit: m2_logit has better log-likelihood (-10524.885 vs. -10530.62), AIC (21073.77 vs. 21085.24), and BIC (21132.90 vs. 21182.90), indicating the logit link fits better.

m1_logit vs. m1_probit: m1_logit has a slightly better log-likelihood (-10547.04 vs. -10549.71) and AIC (21112.08 vs. 21117.42), with equal BIC, suggesting a marginal preference for logit even in the original model.

Justification for New Variables

The improved log-likelihood and AIC in both m2_probit and m2_logit indicate that Wet_Road, Young_Driver, and Single_Vehicle enhance the model's ability to predict Injury_Severity. The significant t-values from Question 2 (e.g., for Young_Driver, for Single_Vehicle in m2_probit) support their statistical relevance.

The pseudo-R^2 increase is small but consistent , reflecting the challenge of explaining crash severity, but it aligns with the added variables capturing additional variance.

The BIC increase for m2_probit suggests caution—adding 3 parameters may not be worth it with a large sample. However, m2_logit's lower BIC supports the inclusion.

Conclusion

Best Model: m2_logit performs best with log-likelihood = -10524.885, AIC = 21073.77, BIC = 21132.90, and pseudo-R^2 = 0.0099, justifying the inclusion of Wet_Road, Young_Driver, and Single_Vehicle.

Justification: The new variables improve fit, particularly in m2_logit, as evidenced by significant reductions in AIC and BIC, supporting their inclusion for better injury severity prediction.

Question 4 : Proposed Additional Variables

Objective

1. Propose two new variables to collect.
2. Provide a reason for why each variable is relevant to injury severity.
3. State the expected influence (positive or negative coefficient) and explain the reasoning.
4. Specify the data type (e.g., binary, numeric, categorical).

Context and Current Model

The current dataset includes variables like Speed.Limit.at.Crash.Site, Urban, Pedestrian, Parked_Vehicle, Late_Night, Wet_Road, Young_Driver, and Single_Vehicle, which cover speed limits, location, crash type, road conditions, driver age, and time. The analysis from previous questions revealed that:

- Wet_Road unexpectedly reduced severity (negative coefficient), possibly due to confounding (e.g., lower speeds in wet conditions).
- Young_Driver and Single_Vehicle increased severity (positive coefficients), confirming their relevance.
- The model's pseudo-R^2 is low, indicating room for improvement by capturing additional factors like rider behavior, protective measures, or crash specifics.

New variables should address unmeasured aspects such as rider characteristics, equipment, or crash dynamics not fully captured by existing data.

Proposed Variables

1. Helmet_Use

- **Description**: A binary variable indicating whether the motorcyclist was wearing a helmet at the time of the crash or not (1 = Yes, 0 = No).
- **Reason**: Helmets are a critical protective measure for motorcyclists, reducing the risk of head injuries, which are a leading cause of severe and fatal outcomes in motorcycle crashes. The current dataset lacks this information, and its inclusion could directly address rider vulnerability, a key gap.
- **Expected Sign**: I expect a negative coefficient. Helmet use should decrease the likelihood of higher injury severity categories (e.g., from "Suspected Serious Injury" to "Suspected Minor Injury" or lower) by mitigating head injury.
- **Data Type**: ***Binary (1 = Helmet worn, 0 = No helmet)***. This is straightforward to collect from crash reports or rider interviews.

2. Motorcycle_Speed

- **Description**: A numeric variable representing the estimated speed of the motorcycle (in mph) at the time of the crash.
- **Reason**: The current model includes Speed.Limit.at.Crash.Site, which reflects the posted limit, but actual motorcycle speed better captures the kinetic energy involved in the crash, a primary determinant of injury severity. Higher speeds increase impact force. The dataset's focus on non-intersection crashes (e.g., loss of control or fixed-object collisions) makes speed a critical unmeasured factor.
- **Expected Influence**: I expect a positive coefficient. Higher motorcycle speeds should increase the likelihood of moving to higher severity categories (e.g., from "Possible Injury" to "Fatal Injury") due to greater impact energy. For instance, a 60 mph crash on a 55 mph road (as seen in the sample data) could result in a "Fatal Injury" versus a "Suspected Minor Injury" at 30 mph.
- **Data Type**: ***Numeric*** (***continuous***, e.g., 0 to 100+ mph). This requires estimation from skid marks, witness accounts, or event data recorders (if available), though it may involve some uncertainty.

Why These Variables?

- **Gaps Addressed**: The current model lacks rider-specific protection (helmet) and precise crash dynamics (actual speed), focusing more on external conditions and other drivers. These additions target the motorcyclist's role and behavior.
- **Practical Relevance**: Both are actionable—helmet laws could be enforced, and speed management (e.g., speed governors) could be explored, aligning with safety policy.