Adam Sperber
STAT 410 project paper

A geographic analysis of Chicago car crashes: 2013-2021

I ran a study of which factors increased or decreased the probable cost of damages for Chicago car crashes between the years of 2013 and 2021 using a .csv file of 495,416 police reports from the Chicago Data Portal. I fit the data to a proportional odds model in R with a response variable of crash damage split into three ordered tiers: under $500 of damage, between $500 and $1500, and over $1500. I chose major Chicago Police Department beat districts (north, central, and south), weather condition, temporary road surface condition, road defects, and posted speed limit as explanatory variables, with a baseline case being an accident in the central area with undisturbed roads and clear weather. The results showed a significant increase in odds of higher expense for crashes in the south district with unclear, non-defective roads at higher speed limits. Weather condition proved insignificant; I excluded it from the final model with proper justification.

This dataset, though interesting, came with its set of flaws, especially with respect to selecting a response variable. There were several potential most notably fatalities, injuries, police response, and damages, though only cash cost was included in all rows as well as subdivided into three neat, ordered categories. Explanatory variables were more complex, ranging from time and date to weather and functionality of traffic equipment. However, given that the dataset was from Chicago itself, I wanted to find a centerpiece variable which was uniquely Chicagoan—thus, I selected CPD beats to form a geographic component. I grouped the beats into their three-digit districts, from which I formed the aforementioned three official CPD district areas. I selected weather, road conditions, and road defects (reduced to binary form along a dichotomy of ideal conditions vs anything unsafe) from my own experience as a Chicago driver to complement the geographic theme, including speed limit so to have at least one continuous variable in my model. While I wished to include lane count as another continuous explanatory variable, the category was missing a significant (~ 60%) of its data, and what content it had proved to be factually unreliable, ranging from instances of zero-lane major roads classified as things other than parking lots to hundreds of instances of roads with over fifty lanes (one even numbering in six digits), so I concluded I had more to lose than to gain by imputation. Ultimately, I ignored the category. I grouped any unknown cells in binary variables in with ideal conditions under the assumption that were it relevant to the crash, law enforcement would have noted it.

I used the tidyverse and lubridate packages (as well as base R) for variable extraction and cleanup, while creating the proportional odds model itself with the vglm and polr functions from VGAM and MASS, respectively, verifying it with the brant package for proportionality. The study's scope was originally for just the calendar year of 2018[1], though after the initial model processed and I streamlined all 'for' loops out of my code, I expanded the model to the full dataset with one surprising result: not only did the model improve, but given that a crash occurred, weather conditions at the time of the crash proved irrelevant to the final cost. Ironically, removing weather from the model strengthened the significance of road surface condition while diminishing its standard error. To make sure this exclusion was not an error due to the obviously correlated nature,

[1] All excluded models included at the end of the appendix

I ran a vglm with an interaction term of weather and surface condition, yet this interactive model also proved insignificant. Surprisingly, weather was not a factor in crash cost.

The final results were more than I initially hoped for: the geographic areas formed a clear gradient of decreasing odds of cheap accidents from north to south. The same crash occurring in all three districts would have decreased odds of greater financial cost by 4% in the north compared to the center while increasing by nearly 10% in the south. By far the biggest indicator of an expensive-tier crash were negative temporary road conditions, increasing odds by nearly 13%. However, the single biggest factor in either direction proved to be road defects, increasing the odds of a cheap accident by over 45%—perhaps a better interpretation would be to note how people react to potholes and road fissures. Finally, each 1 mph increase in speed limit decreased the odds of a cheap accident by about 1.4%.

The biggest flaw in my work is obviously my impromptu method of mode imputation with unknown cells in my binary variables, given that the mode in each category was a neutral road (no surface obstruction, defects, etc.). Further study should extend to the individual beat districts—I happened to notice the single highest case fell in the area which contained Midway Airport—as well as focusing on specific causes of surface-related conditions which induce crashes. Additionally, the data file did not account for the make and model of car; for example, the same crash happening to an old Buick vs a new Mercedes would almost certainly fall in opposite categories of the response variable. Conversely, the data indicate citywide road repairs may not be a high priority as things stand given the ferociously high odds ratio for defects.

Given the nature of expensive crashes tilting heavily towards Chicago's South Side, this somewhat informal study serves to further illuminate the issues of city wealth inequality. A 2019 WBEZ investigation[2] highlighted the concentrated nature of income in the city, more or less correlating negatively with the amount of money spent on car accidents. It's easy to conclude that the shockingly high proportion of high-expense crashes clustered away from the city's top-earning households despite the lowest number of crashes cannot be a natural phenomenon—it would be worth looking into road maintenance given the significance of temporary road conditions increasing the odds of an expensive crash. Given that crashes cost more per instance in a part of the city with the lowest income per household, it seems fairly likely that this effect would translate beyond just car crashes, and it does: ProPublica's investigation with the Chicago Tribune into discrepancies in property tax rates[3] falls into this theme as well. Any future studies would do well to focus on the systemic cost of cost in low-income areas.

[2] https://www.wbez.org/stories/the-middle-class-is-shrinking-everywhere-in-chicago-its-almost-gone/e63cb407-5d1e-41b1-9124-a717d4fb1b0b

[3] http://apps.chicagotribune.com/news/watchdog/cook-county-property-tax-divide/assessments.html

APPENDIX

Total crashes table between the three districts

```
(area <- table(crash$AREA, crash$DAMAGE))
```

```
## 
##            $500 OR LESS $501 - $1,500 OVER $1,500
##   CENTRAL         28198         62856      126527
##   NORTH           24117         58373      107998
##   SOUTH           11122         23262       52963
```

Proportion tables

```
prop.table(area, 1)
```

```
## 
##            $500 OR LESS $501 - $1,500 OVER $1,500
##   CENTRAL     0.1295977     0.2888855   0.5815168
##   NORTH       0.1266064     0.3064393   0.5669543
##   SOUTH       0.1273312     0.2663171   0.6063517
```

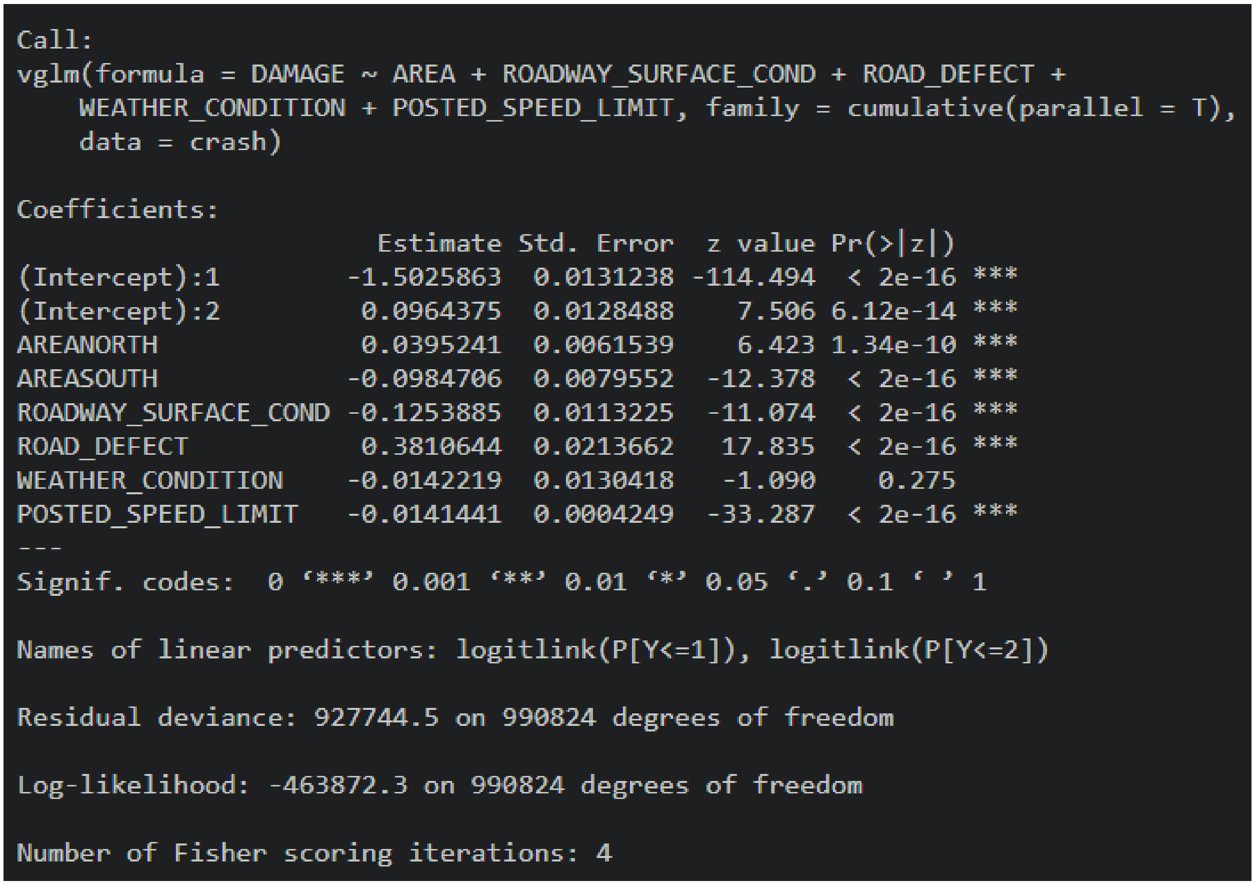
```
prop.table(area, 2)
```

```
## 
##            $500 OR LESS $501 - $1,500 OVER $1,500
##   CENTRAL     0.4445040     0.4350167   0.4401123
##   NORTH       0.3801725     0.4039906   0.3756609
##   SOUTH       0.1753235     0.1609927   0.1842268
```

Note how the greatest proportion of expensive accidents per area is in the South. (table 2)

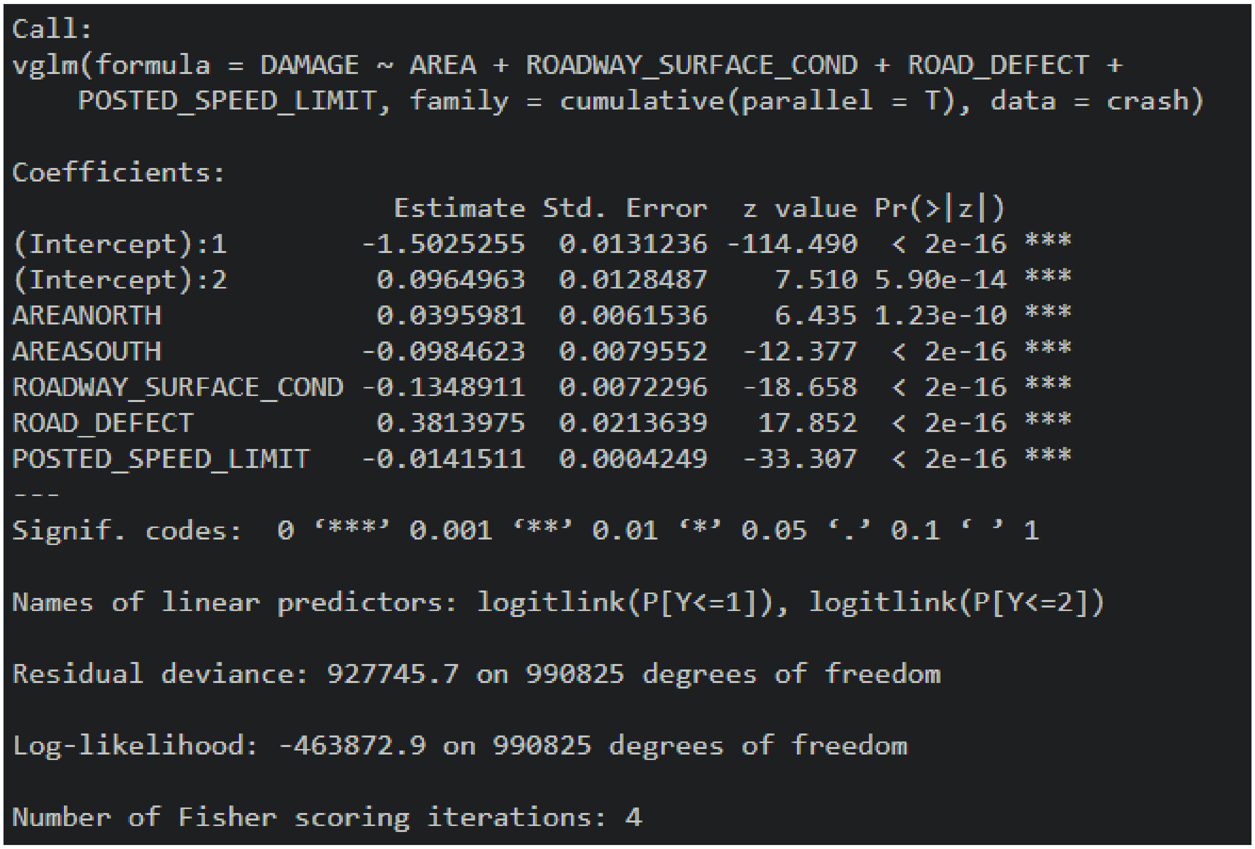
The proportion of overall crashes in the Central Area cements it as the baseline for the PO model. (table 3)

Initial model with weather included



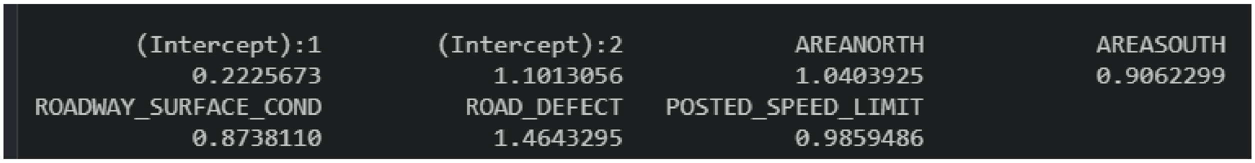
```
Call:
vglm(formula = DAMAGE ~ AREA + ROADWAY_SURFACE_COND + ROAD_DEFECT +
    WEATHER_CONDITION + POSTED_SPEED_LIMIT, family = cumulative(parallel = T),
    data = crash)

Coefficients:
                       Estimate Std. Error  z value Pr(>|z|)
(Intercept):1        -1.5025863  0.0131238 -114.494  < 2e-16 ***
(Intercept):2         0.0964375  0.0128488    7.506 6.12e-14 ***
AREANORTH             0.0395241  0.0061539    6.423 1.34e-10 ***
AREASOUTH            -0.0984706  0.0079552  -12.378  < 2e-16 ***
ROADWAY_SURFACE_COND -0.1253885  0.0113225  -11.074  < 2e-16 ***
ROAD_DEFECT           0.3810644  0.0213662   17.835  < 2e-16 ***
WEATHER_CONDITION    -0.0142219  0.0130418   -1.090    0.275
POSTED_SPEED_LIMIT   -0.0141441  0.0004249  -33.287  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])

Residual deviance: 927744.5 on 990824 degrees of freedom

Log-likelihood: -463872.3 on 990824 degrees of freedom

Number of Fisher scoring iterations: 4
```

Final model with all explanatory variables



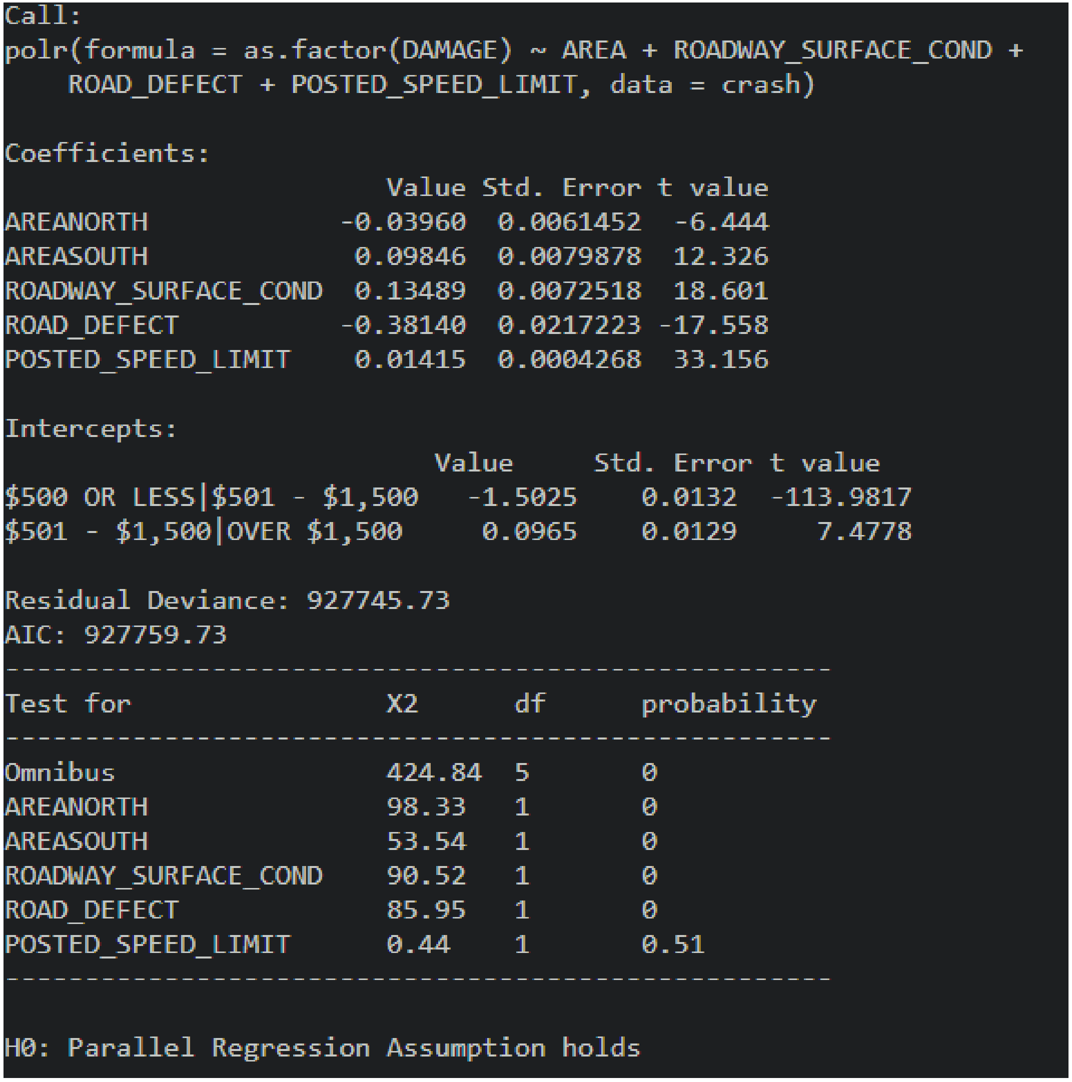
```
Call:
vglm(formula = DAMAGE ~ AREA + ROADWAY_SURFACE_COND + ROAD_DEFECT +
    POSTED_SPEED_LIMIT, family = cumulative(parallel = T), data = crash)

Coefficients:
                       Estimate Std. Error  z value Pr(>|z|)
(Intercept):1        -1.5025255  0.0131236 -114.490  < 2e-16 ***
(Intercept):2         0.0964963  0.0128487    7.510 5.90e-14 ***
AREANORTH             0.0395981  0.0061536    6.435 1.23e-10 ***
AREASOUTH            -0.0984623  0.0079552  -12.377  < 2e-16 ***
ROADWAY_SURFACE_COND -0.1348911  0.0072296  -18.658  < 2e-16 ***
ROAD_DEFECT           0.3813975  0.0213639   17.852  < 2e-16 ***
POSTED_SPEED_LIMIT   -0.0141511  0.0004249  -33.307  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])

Residual deviance: 927745.7 on 990825 degrees of freedom

Log-likelihood: -463872.9 on 990825 degrees of freedom

Number of Fisher scoring iterations: 4
```

Odds ratios for cheap crashes by geographic areas (against a clear road in the central area) and other explanatory variables



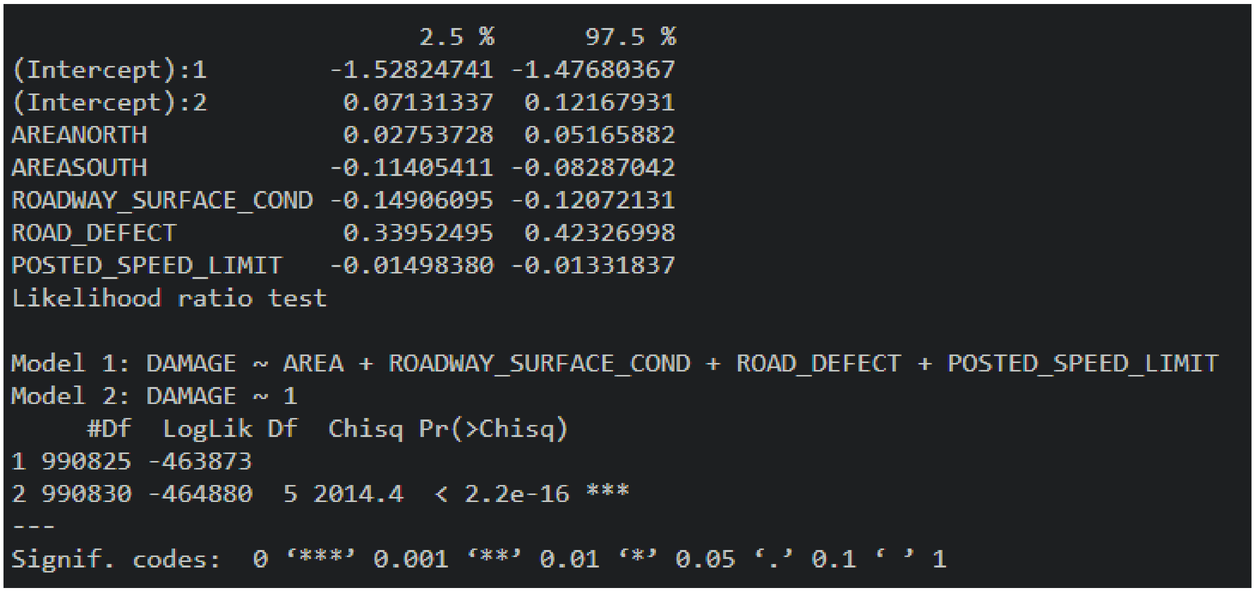
```
       (Intercept):1        (Intercept):2            AREANORTH            AREASOUTH
           0.2225673            1.1013056            1.0403925            0.9062299
ROADWAY_SURFACE_COND          ROAD_DEFECT   POSTED_SPEED_LIMIT
           0.8738110            1.4643295            0.9859486
```

Verification of PO model with polr and brant functions



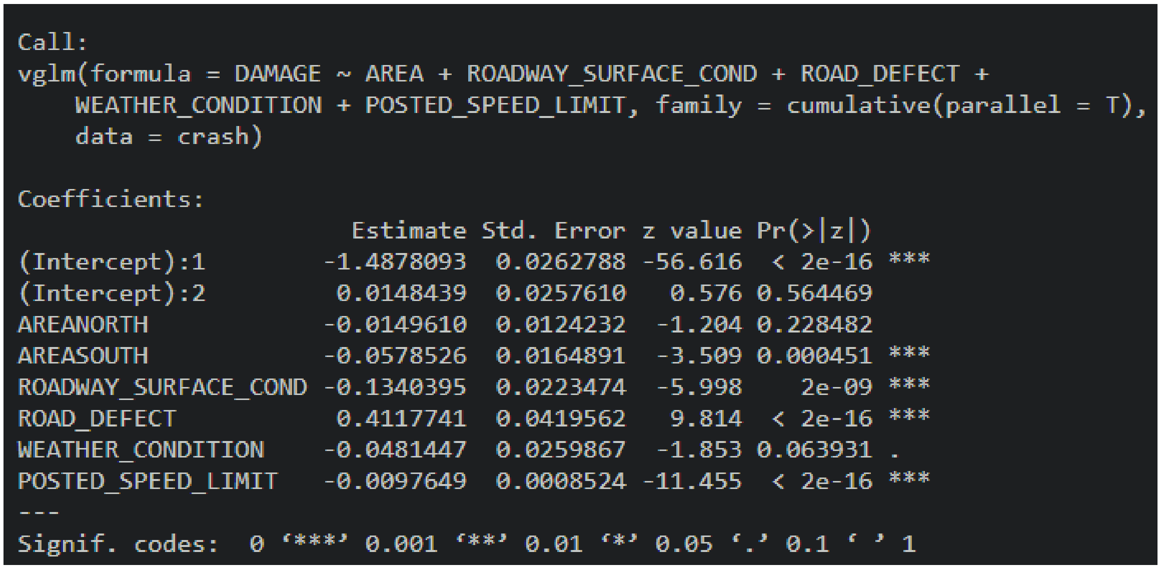
```
Call:
polr(formula = as.factor(DAMAGE) ~ AREA + ROADWAY_SURFACE_COND +
    ROAD_DEFECT + POSTED_SPEED_LIMIT, data = crash)

Coefficients:
                        Value Std. Error t value
AREANORTH            -0.03960  0.0061452  -6.444
AREASOUTH             0.09846  0.0079878  12.326
ROADWAY_SURFACE_COND  0.13489  0.0072518  18.601
ROAD_DEFECT          -0.38140  0.0217223 -17.558
POSTED_SPEED_LIMIT    0.01415  0.0004268  33.156

Intercepts:
                           Value     Std. Error t value
$500 OR LESS|$501 - $1,500   -1.5025    0.0132  -113.9817
$501 - $1,500|OVER $1,500     0.0965    0.0129     7.4778

Residual Deviance: 927745.73
AIC: 927759.73
----------------------------------------------------
Test for                X2      df      probability
----------------------------------------------------
Omnibus                 424.84  5       0
AREANORTH               98.33   1       0
AREASOUTH               53.54   1       0
ROADWAY_SURFACE_COND    90.52   1       0
ROAD_DEFECT             85.95   1       0
POSTED_SPEED_LIMIT      0.44    1       0.51
----------------------------------------------------

H0: Parallel Regression Assumption holds
```

95% confidence intervals for intercepts and variables and likelihood-ratio test against a null model



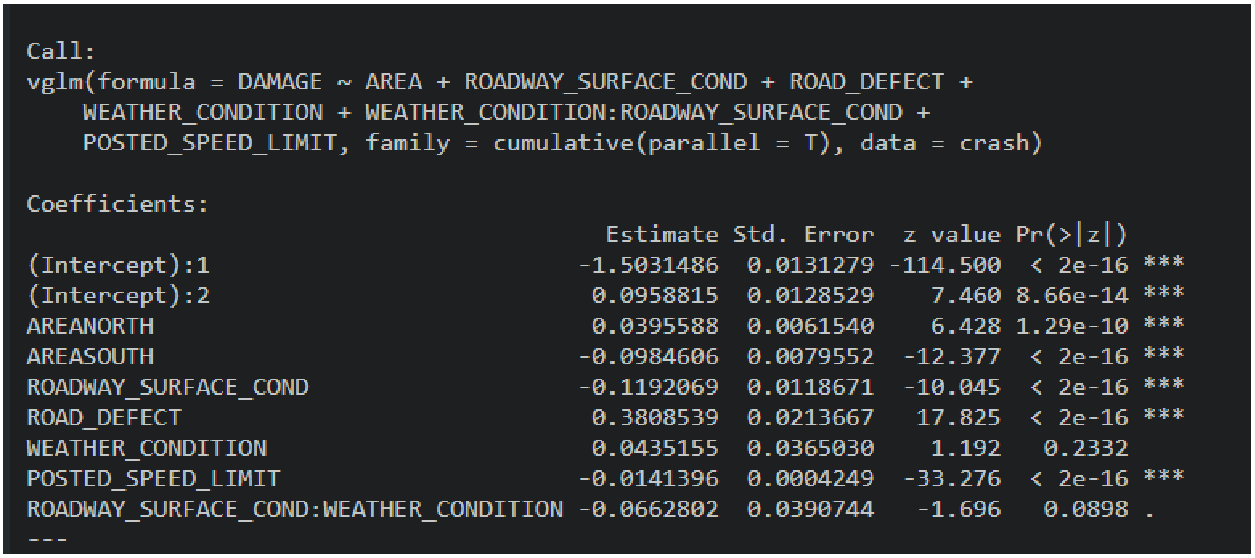
```
                           2.5 %      97.5 %
(Intercept):1        -1.52824741 -1.47680367
(Intercept):2         0.07131337  0.12167931
AREANORTH             0.02753728  0.05165882
AREASOUTH            -0.11405411 -0.08287042
ROADWAY_SURFACE_COND -0.14906095 -0.12072131
ROAD_DEFECT           0.33952495  0.42326998
POSTED_SPEED_LIMIT   -0.01498380 -0.01331837
Likelihood ratio test

Model 1: DAMAGE ~ AREA + ROADWAY_SURFACE_COND + ROAD_DEFECT + POSTED_SPEED_LIMIT
Model 2: DAMAGE ~ 1
     #Df  LogLik Df  Chisq Pr(>Chisq)
1 990825 -463873
2 990830 -464880  5 2014.4  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The initial model of data from only 2018—as you can see, one intercept and an entire geographic area were insignificant, as well as WEATHER_CONDITION’s middling p-value



```
Call:
vglm(formula = DAMAGE ~ AREA + ROADWAY_SURFACE_COND + ROAD_DEFECT + 
    WEATHER_CONDITION + POSTED_SPEED_LIMIT, family = cumulative(parallel = T), 
    data = crash)

Coefficients: 
                       Estimate Std. Error z value Pr(>|z|)    
(Intercept):1        -1.4878093  0.0262788 -56.616  < 2e-16 ***
(Intercept):2         0.0148439  0.0257610   0.576 0.564469    
AREANORTH            -0.0149610  0.0124232  -1.204 0.228482    
AREASOUTH            -0.0578526  0.0164891  -3.509 0.000451 ***
ROADWAY_SURFACE_COND -0.1340395  0.0223474  -5.998    2e-09 ***
ROAD_DEFECT           0.4117741  0.0419562   9.814  < 2e-16 ***
WEATHER_CONDITION    -0.0481447  0.0259867  -1.853 0.063931 .  
POSTED_SPEED_LIMIT   -0.0097649  0.0008524 -11.455  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Attempt to incorporate weather via in interactive term.



```
Call:
vglm(formula = DAMAGE ~ AREA + ROADWAY_SURFACE_COND + ROAD_DEFECT +
    WEATHER_CONDITION + WEATHER_CONDITION:ROADWAY_SURFACE_COND +
    POSTED_SPEED_LIMIT, family = cumulative(parallel = T), data = crash)

Coefficients:
                                         Estimate Std. Error  z value Pr(>|z|)
(Intercept):1                          -1.5031486  0.0131279 -114.500  < 2e-16 ***
(Intercept):2                           0.0958815  0.0128529    7.460 8.66e-14 ***
AREANORTH                               0.0395588  0.0061540    6.428 1.29e-10 ***
AREASOUTH                              -0.0984606  0.0079552  -12.377  < 2e-16 ***
ROADWAY_SURFACE_COND                   -0.1192069  0.0118671  -10.045  < 2e-16 ***
ROAD_DEFECT                             0.3808539  0.0213667   17.825  < 2e-16 ***
WEATHER_CONDITION                       0.0435155  0.0365030    1.192   0.2332
POSTED_SPEED_LIMIT                     -0.0141396  0.0004249  -33.276  < 2e-16 ***
ROADWAY_SURFACE_COND:WEATHER_CONDITION -0.0662802  0.0390744   -1.696   0.0898 .
---
```