The Moral Machine Analysis

Summary

Autonomous vehicles are going to be crusing on the road in the near future. However, what kind of moral standards should they hold when these machine encounter life-death situation remains a puzzle. In order to obtain public consensus, if any, on this type of question, MIT researchers have conducted a Moral Machine Project, where participatants conduct an online survey to decide whether an AV experience brake failure should spare the life of one group over the other. In order to investigte the relationship between committing omission error (AV staying on course) vs. commission error (make the AV swerve), we adopt a logistic-regression model, using omission character, respondants' age, country, education, income, and gender for predicting how much more likely it will be for some to swerve. The analysis finds that in fact, only omission characters are statistically significant predictors for preference of either omission or commission.

Introduction

In this analysis, we are interested in several questions: do people from some countrie more likely to swerve over others, and does this differ by omission attributes? What are some of the omission attributes being tested that are going to influence whether or not the respondant chooses to swerve? What are some of the demographic features of the respondant that are predictive of their decision? By investigating these questions, we hope to gain some insights about how this moral dilemma is framed by the public, as well as the possible frameworks that AV engineers wil need to take into account when trying to design a *moral machine*.

Data

This analysis uses data obtained from the data repository of the Moral Machine project. The orignal data contained 40 million decisions in 10 languages in 233 countries. In the research design, each participatant is presented with 13 different scenarios. Each scenario explores one of six dimensions of target attributes: species, social value, gender, age, fitness, and utilitarianism. Users are presented with two randomly sampled scenarios of each of the six dimensions, in addition to one completely random scenario that can have any number of characters on each side and in any combination of characters, making up of a total of 13 questions for each user. All 6 dimensions are thus mutually exclusive. For example, a randomly sampled scenario in fitness dimension is generated by randomly drawing from three sets of characters corresponding to three levels: 1) low fitness: Large Man, Large Woman 2) neutral fitness: Man, Woman 3) high fitness: Male Athlete, Female Athlete. After making sure gender-preserving bijections and equal number of characters, pairs of lower fitness are arranged in one side of the scenario and higher fitness the other side.

In the orignal research project, the outcome variable that the researchers were primarily interested in was **`spare`** which represents the character that was ultimately spared in each life-death situation. Given that this analysis has a different research interest which is the odds of omission vs. commission error, outcome variable –**`swerve`**–is engineered by selecting observations where: the default choice (which is an arbitrary benchmark choice in each question) is on the omission lane (where the AV starts with) and this default choice is saved, or the default choice is not omission and is not saved. For these observations, the level of swerve is 1, and 0 otherwise.

In order to make the omission attributes a direct predictor, another variable `omission` –a factor variable with 12 levels– is created by being equal to `DefaultChoice` if `DefaultChoiceIsOmission` is 1, or take on the value of `NonDefaultChoice` otherwise. As a result, `omission` directly pinpoints the attribute of the group that is on the omission lane in each scenario, and combined with **`swerve`**, it is clear whether the respondant committed omission or commission when faced with different attributes.

Considering limitation in computing power, this analysis chooses a subset of 10 million observations from the original data and focus on respondants who are from one of three countries in this subset: China, Russia, and the U.S. The reasoning behind choosing this three are two fold: first, they all have a large population and citizens are likely to have access to the Internet, which will decrease the likelihood of insufficient data in each; second, these three countries have very different political ideology orientation and culture, where Chinese culture emphasizes collectivity and the U.S. culture stresses individuality. Ideally, such differences may be reflected in whether or not their citizens are more likely to swerve when the inaction will lead to fatal consequence to a group that have some culturally preferred characteristics.

When doing data inspection, we find that there are 309202 observations with empty values, and 13378 observations with age larger than 100 or less than 0. After dropping these observations, the data has 2587162 observation and 30 variables. We also mean-centered age varialbe to facilitate interpretation. Considering the scope and interest of this analysis, 20 varialbes are dropped from analysis and only 9 variables remain: user country, reviewer's age review's education, reviewer's gender, reviewer's income, reviewer's political score, reviewer's religious score, omission, and swerve. Among these, scenario type, user country, reviewer's education, reviewer's gender,reviewer's income, omission, and swerve are converted into factor variables.

The EDA process is as follows: First, we check the balance of the response variable `swerve`, and find that there are 1293734 observations with swerve == 0 and 1293428 with swerve == 1, showing that the number of omission and commission errors are balanced in data. Next, we examine the balance of `omission` and find that although there are some differences in the quantity of observation in each omission level (e.g., omission == Fat vs. omission == Male), there are no level that have significantly less observations that need to be concerned. Then a chi-squared test is run on `swerve` and `omission`, and the result shows that p-value is 0.00050, meaning there may be statistically significant difference in omission vs. commission between different omission levels.

Then the balance of observation for country is examined. It is found that, although each country has fairly large number of observations, the number differs significantly among the three, where Russia has 10 times more observations than China (39743), and USA has about 50 times more observations than China. This might be something that can lead to bias in our analysis. Using a chi-squared test between `swerve` and `UserCountry3`, it is found that the result is not statistically significant at p-value = 0.9255, which demands further modeling to verify.

After that, the relationship between respondant demographics and `swerve` is explored. First it is found that the distribution of age between people who served vs. those who did not swerve is very similar. with median age at around 25. The visual inspections for religious score and swerve and political score swerve also show no significant difference in distrubtion. For other categorical demographics variables such as education, income, and gender, chi-squared test also shows no statistically significant results between them and `swerve`, with p-value all large than 0.05.

After examining all main effects, the interaction of interest– omission and swerve by country– is also explored. It turns out the relationship between omission and swerve are statistically significant for all three countries at p-value < 0.05, which might signal non-existance of interaction effect.

Model

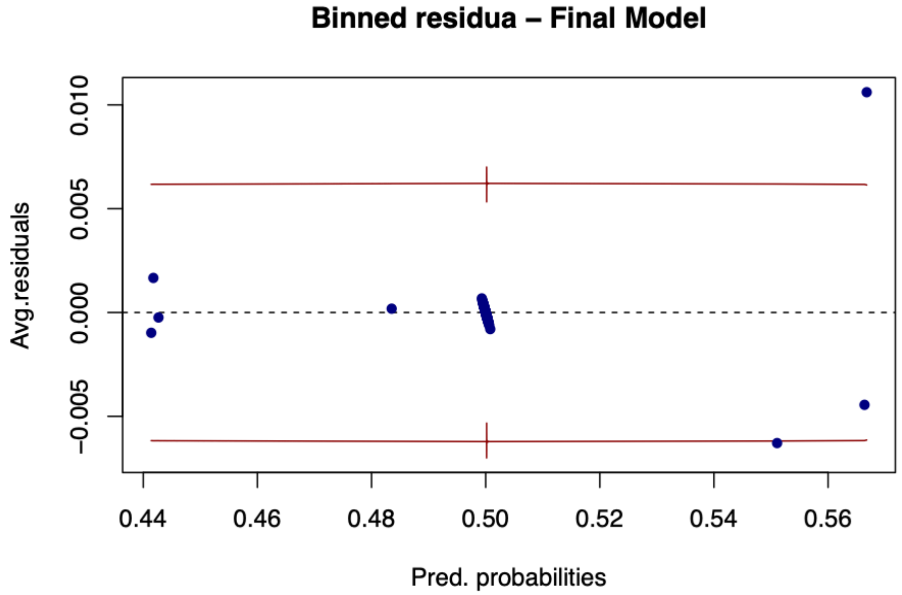
Considering our research interest, we first fit a baseline logistic regression model with all main effects as predictors and `swerve` as the response variable.The model output shows that among all main effects, only the two levels within `Omission` have coefficients that are statistically significant with p-value < 0.05: `omissionLow` and `omissionHigh`, which represent omission level of lower social status or higher social status. Binned residuals plots between predicted probability and each numerical predicotors show no clear pattern, meaning that residuals are randomly distributed, and transformation of variables is not needed. In terms of model assessment, this baseline base has an accuracy of 0.504, sensitivity of 0.552, and specificity of 0.456. With an AUC of 0.507, this model does not perform well in terms of prediction.

With a purpose to find a better model, we conduct model selection using AIC-stepwise, AIC-backward, and AIC-forward methods. It turns out that AIC-stepwise and AIC-backward methods suggest only omission as the predictor for `swerve`, while AIC-forward method returns the null model. Given the limitation of AIC metrics, we use ANOVA test to examine the necessity of each predictor variable not recommended by AIC model selection method. The results shows that none of the models that include country, income, education, gender, regligiou score, political score are statistically different from the model that uses `omission` as the only predictor, with p-values > 0.05. We are confident that our final model should include the main effect of `omission`. Other predictors are important to answer our inference questions of interest but may not be statistically significant. Since we are also interested in the interaction between omission and country, we also use ANOVA to test for a model that includes this interaction with one that does not. Although p-value between these two models are larger than 0.05, we will still need to include this interaction in the final model given that this interaction is one of our research focuses.

Final model:

$$log(\pi_i/1-\pi_i) = \beta_0 + \gamma_{0ji}Country + \beta_{Gender}Gender_i + \beta_{Omission}Omission_i + \beta_{Education}Education_i + \\ \beta_{Age}Age_i + \beta_{Income}Income_i + \beta_{Political}Political_i + \beta_{Religious}Religious_i + \beta_{Omission:Country}Omission_i : Country_i$$

Binned residual plots do not have any salient patterns, and most of dots are within the 95% boundary.



However, as we can observe from the model assessment metrics, the final model does not have a significant improvement compared to the baseline model. Although sensitivity has incerased to 0.570 from 0.552, specificity has slightly decreased from 0.456 to 0.437. AUC has stayed the same at 0.5073.

Finally, we check multicolinearity of our predictors, and there is no evidence of multicolinearity given that all have vif scores lower than 10.

term	estimate	std.error	statistic	p.value
(Intercept)	0.0001046	0.0340077	0.0030747	0.9975468
Review_religious	-0.0006535	0.0041561	-0.1572408	0.8750551
Review_political	-0.0005180	0.0048504	-0.1068061	0.9149428

term	estimate	std.error	statistic	p.value
Review_income15000	-0.0008551	0.0078540	-0.1088740	0.9133024
Review_income25000	-0.0013429	0.0081032	-0.1657283	0.8683707
Review_income35000	-0.0007862	0.0078042	-0.1007368	0.9197594
Review_income5000	-0.0003338	0.0073968	-0.0451251	0.9640076
Review_income50000	-0.0010415	0.0074257	-0.1402582	0.8884560
Review_income80000	-0.0020404	0.0087202	-0.2339890	0.8149935
Review_incomeabove100000	-0.0007250	0.0073506	-0.0986366	0.9214268
Review_incomedefault	-0.0010412	0.0068746	-0.1514616	0.8796116
Review_incomeover10000	-0.0033866	0.0255003	-0.1328064	0.8943465
Review_incomeunder5000	-0.0006082	0.0063913	-0.0951557	0.9241911
Review_genderfemale	0.0014447	0.0077391	0.1866745	0.8519159
Review_gendermale	0.0005525	0.0076407	0.0723077	0.9423570
Review_genderothers	-0.0002939	0.0105505	-0.0278523	0.9777800
Review_age	-0.0000360	0.0001408	-0.2555135	0.7983266
Review_educationcollege	0.0005200	0.0043911	0.1184136	0.9057400
Review_educationdefault	0.0018152	0.0080630	0.2251246	0.8218824
Review_educationgraduate	0.0006545	0.0043374	0.1509070	0.8800491
Review_educationhigh	0.0006948	0.0048665	0.1427666	0.8864745
Review_educationothers	-0.0000343	0.0072125	-0.0047548	0.9962063
Review_educationunderHigh	0.0017160	0.0048728	0.3521608	0.7247177
Review_educationvocational	0.0036868	0.0092032	0.4005966	0.6887171
omissionFemale	-0.0000019	0.0459513	-0.0000423	0.9999663
omissionFit	0.0000009	0.0488599	0.0000191	0.9999848
omissionHigh	0.2503584	0.0686634	3.6461676	0.0002662
omissionHoomans	0.0000162	0.0460256	0.0003517	0.9997193
omissionLess	0.0000169	0.0446716	0.0003779	0.9996985
omissionLow	-0.2143251	0.0659182	-3.2513793	0.0011485
omissionMale	0.0000118	0.0460630	0.0002569	0.9997950
omissionMore	-0.0000105	0.0447935	-0.0002343	0.9998130
omissionOld	0.0000078	0.0459450	0.0001696	0.9998647
omissionPets	-0.0000089	0.0459758	-0.0001947	0.9998447
omissionYoung	0.0000165	0.0468804	0.0003523	0.9997189
UserCountry3RUS	-0.0002034	0.0340282	-0.0059786	0.9952298
UserCountry3USA	-0.0003956	0.0327576	-0.0120765	0.9903646
omissionFemale:UserCountry3RUS	-0.0000107	0.0481781	-0.0002217	0.9998231
omissionFit:UserCountry3RUS	-0.0000120	0.0512111	-0.0002343	0.9998130
omissionHigh:UserCountry3RUS	-0.0463466	0.0718818	-0.6447608	0.5190822
omissionHoomans:UserCountry3RUS	-0.0000213	0.0482516	-0.0004424	0.9996470
omissionLess:UserCountry3RUS	-0.0000290	0.0468988	-0.0006184	0.9995066
omissionLow:UserCountry3RUS	-0.0084725	0.0689801	-0.1228254	0.9022454
omissionMale:UserCountry3RUS	-0.0000204	0.0482897	-0.0004228	0.9996627
omissionMore:UserCountry3RUS	0.0000021	0.0470051	0.0000439	0.9999649
omissionOld:UserCountry3RUS	-0.0000141	0.0481733	-0.0002928	0.9997664
omissionPets:UserCountry3RUS	-0.0000030	0.0481952	-0.0000614	0.9999510
omissionYoung:UserCountry3RUS	-0.0000289	0.0491441	-0.0005883	0.9995306
omissionFemale:UserCountry3USA	0.0000061	0.0463756	0.0001312	0.9998953
omissionFit:UserCountry3USA	0.0000097	0.0493071	0.0001965	0.9998432
omissionHigh:UserCountry3USA	0.0171923	0.0692835	0.2481445	0.8040226
omissionHoomans:UserCountry3USA	-0.0000152	0.0464480	-0.0003271	0.9997390
omissionLess:UserCountry3USA	-0.0000090	0.0450946	-0.0002005	0.9998400
omissionLow:UserCountry3USA	-0.0198894	0.0664989	-0.2990939	0.7648684
omissionMale:UserCountry3USA	-0.0000045	0.0464856	-0.0000979	0.9999219

term	estimate	std.error	statistic	p.value
omissionMore:UserCountry3USA	0.0000147	0.0452172	0.0003255	0.9997403
omissionOld:UserCountry3USA	-0.0000042	0.0463689	-0.0000909	0.9999275
omissionPets:UserCountry3USA	0.0000205	0.0463986	0.0004420	0.9996473
omissionYoung:UserCountry3USA	-0.0000106	0.0473104	-0.0002243	0.9998211

This model output can be interpreted as follows: For a person who is from China at an average age, default gender, with religious and political score at 0, and received bachelor degree and has income 10000, the odds to swerve when the omission level is Fat is about 0.1%.
Based on the model summary, only two coefficients are statistically significant: omissionLow and omission High, which aligns with our EDA finding. It means that compared to the intercept, when the omission level is lower social status, the odds to swerve decreases by 19.29%, holding all other variables constant. Similarly, compared to the intercept, when the omission level is higher social status, the odds to swerve increases by 28.45%, holding all other variables constant. However, it is necessary to keep in mind that omission levels are mutually exclusive, and omissionFat is not comparable to omissionHigh or omissionLow, given that different omission types are never presented and compared in the same scenario. As a major limitation of the analysis, more dicussion on this will be mentioned in the following section. According to this model, none of the demographics factors, including the interaction between country and omission, are statistically significant predictors.

Conclusions

This analysis aims to offer insights on the moral machine dilemma by extending the original research to explore a new outcome variable: swerve, which is a representation of omission/commission errors. According to the final model output, demographics factors such as education, age, religious score, political score, income are not statistically significant for predicting swerve, meaning that there is no proved relationship between these variables and whether or not a person will choose omission or commission. On the other hand, people from China, Russia, or USA do not have a significant preference for choosing omission or commission compared to the two other countries, and the fact that the interaction effect between country and omission is not significant further corroborates the findng that choosing omission or commission does not differ by country in different omission levels. However, one interesting phenomenon is that generally people are less likely to swerve when the omission level is a group with lower social status, and more likely to swerve when the omission level is a group with higher social status. This shows that these respondants are more willing to spare the lives of those with higher social status (e.g.,company executives) than those of ordinary people.

However, this analysis has some profound limitations: In the original study, researchers used multiple regressions to test specifically each scenario type, given that those types are mutually exclusive (i.e., type Fitness and type Social Status are never presented together in a scenario). In this model, we choose to generalize and compare all levels of omission levels at once, resulting in a fact that our coefficients are built on the premises of comparisons that may not hold practical validity. Thus, our significant finding on social status is based on comparing omissionLow and omissionHigh with omissionFat, and the exact significance in difference of odds between omissionLow and omissionHigh will require t-test or the impelementation of another model that compare only lower social status with higher social status.

The second limitation is that, our model AUC is roughly 0.5, showing that this model eseentially does not own predictive power. Although this analysis focuses mainly on inferences rather than on predictions, it is still an alarming signal that this model might not have been effective in capturing the relationship. This might be that swerve is better predicted by variables that were not recorded as part of the study–things that we have not observed. More importantly, however, it may also be the case that choosing which character to spare is statistically meaningful (just as what has been done in the orignal study), but chooing to swerve or not is not – it may just be an extraraneous representation of the underlying character preference, which is assigned either at omission or commission side at random.