Predicting Belief in Climate Change with Logistic Regression

Evan Wagner

12/16/2021

Contents

1 Introduction

The facts of climate change are quite straightforward: the burning of fossil fuels is releasing enough heat-trapping gases into the atmosphere to significantly raise the average global temperature, unleashing consequences that will worsen if the trend is not reversed. However, not everyone is on the same page on these facts. For example, in a September 2017 survey of 1000 American adults,[1] only 64.6 percent believed that climate change is both real and caused by humans.

If we wish to reverse the climate crisis without sacrificing democracy, we need to get everyone on the same page. For starters, we should be able to explain belief in climate change with other factors. To do so, we will construct a logistic regression model using the aforementioned survey data to predict whether or not American adults believe in climate change.

Our goal is to answer the following questions of interest.

1. How do changes in our variables affect the odds that an American adult believes in anthropogenic climate change?
2. How accurately can our model predict the likelihood that various profiles of American adults believe in anthropogenic climate change?

1.1 Data Preparation

Our data was collected in a live-caller poll (including both landlines and cell phones) by Survey Sampling International, a research firm, and Cards Against Humanity, a card game company. Table 2 lists the questions asked in the survey, along with variable names created by the author.

We first eliminate the variables $PartyAffil$ and $TrumpJobApprove$. Climate change belief is highly polarized, so we would already expect these variables to be highly predictive and potentially crowd out other significant variables. On top of that, politics are more contextual and temporary than the more apolitical variables.

Next, we remove rows with missing values for one or more of our quantitative variables: $Income$, Age, $BooksInPastYear$, and $EstPctFedBudgetForScience$. Unfortunately, the data was missing over half the entries for $Income$ and about 300 for $EstPctFedBudgetForScience$, so including either of those variables would sacrifice too much predictive power.

We convert each of our categorical variables to binary by removing the level “DK/REF”, which indicates either indecision or refusal to answer, and sorting the remaining levels into two groups. Encountering too many “DK/REF” answers for $EarthFartherInWinter$ and $FedFundingScienceTooHighOrLow$, we drop them both. We also scrap the 6-level categorical variable $AgeRange$, as Age is a quantitative predictor that we can expect to have similar or better predictive power than just two age groups. Our final trimmed dataset includes 656 observations of 13 variables and is presented in Table 1.

Variable	**Levels**
Age	quantitative
$BooksInPastYear$	quantitative
$Male$	1 if $Gender$ = “Male”, 0 if $Gender$ = “Female”
$College$	1 if respondent holds $\geq$ BA, 0 otherwise
$Married$	1 if $Married$ = “Married”, 0 otherwise
$JobWillBeAutomated$	1 if $JobWillBeAutomated$ = “Likely”, 0 if $JobWillBeAutomated$ = “Unlikely”
$AnthClimateChangeReal$	1 if $ClimateChangeReal$ = “Real and Caused by People”, 0 otherwise
$HasSeenTransformers$	1 if $TransformersMoviesSeen \geq 1$, 0 if $TransformersMoviesSeen = 0$
$ScientistsGood$	1 if $ScientistsGood$ = “Somewhat Agree” or “Strongly Agree”, 0 otherwise
$VaccinesGood$	1 if $VaccinessGood$ = “Somewhat Agree” or “Strongly Agree”, 0 otherwise
$BelievesInGhosts$	1 if $BelievesInGhosts$ = “Yes”, 0 if $BelievesInGhosts$ = “No”
$SmartSadDumbHappy$	1 if $SmartSadDumbHappy$ = “Smart and Sad”, 0 if $SmartSadDumbHappy$ = “Dumb and happy”
$ShowerPeeingOk$	1 if $ShowerPeeingOk$ = “Acceptable”, 0 if $ShowerPeeingOk$ = “Unacceptable”

Table 1: Our final dataset

Variable	**Question**
$Gender$	Gender
Age	Age
$PartyAffil$	Political Affiliation
$TrumpJobApprove$	Do you approve or disapprove of how Donald Trump is handling his job as president?
$Educ$	What is your highest level of education?
$Race$	What is your race?
$Married$	What is your marital status?
$JobWillBeAutomated$	What would you say is the likelihood that your current job will be entirely performed by robots or computers within the next decade?
$ClimateChangeReal$	Do you believe that climate change is real and caused by people, real but not caused by people, or not real at all?
$TransformersMoviesSeen$	How many Transformers movies have you seen?
$ScientistsGood$	Do you agree or disagree with the following statement: scientists are generally honest and are serving the public good.
$VaccinesGood$	Do you agree or disagree with the following statement: vaccines are safe and protect children from disease.
$BooksInPastYear$	How many books, if any have you read in the past year?
$BelievesInGhosts$	Do you believe in ghosts?
$EstPctFedBudgetForScience$	What percentage of the federal budget would you estimate is spent on scientific research?
$FedFundingScienceTooHighOrLow$	Is federal funding of scientific research too high, too low, or about right?
$EarthFartherInWinter$	True or false: the earth is always farther away from the sun in the winter than in the summer.
$SmartSadDumbHappy$	If you had to choose: would you rather be smart and sad or dumb and happy?
$ShowerPeeingOk$	Do you think it is acceptable or unacceptable to urinate in the shower?

Table 2: Potential variables from our raw dataset

2 Methods

We begin with a graphical exploration of our data. Box plots of Age, $BooksInPastYear$, and $log(BooksInPastYear)$ against $AnthClimateChangeReal$ are shown in Figure 1. The log transformation on $BookInPastYear$ appears to improve its distribution, so we use this transformed variable for our analysis.[1] Both pairs of means look somewhat different, so we will pay close attention to Age and $log(BooksInPastYear)$ going forward.

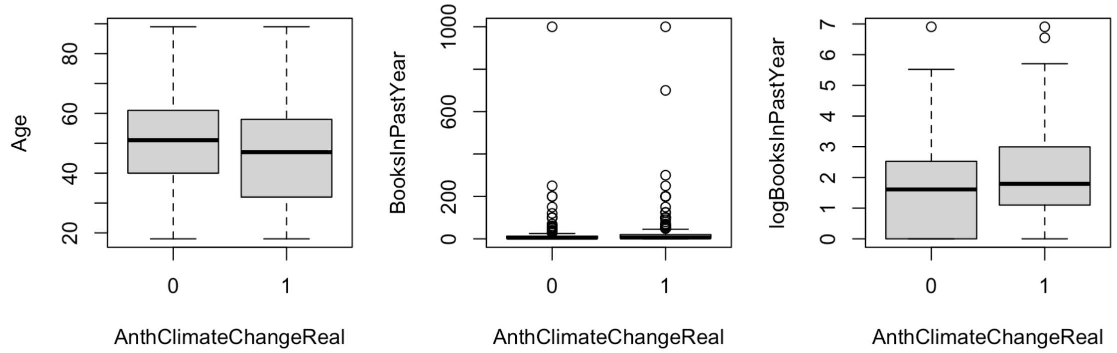


Figure 1: Box plots of Age, $BooksInPastYear$, and $log(BooksInPastYear)$ vs. $AnthClimateChangeReal$

We analyze potential effects of our binary variables by looking at tables of each against $AnthClimateChangeReal$. The percent of each level that reported belief in anthropogenic climate change for each variable is shown in Table 3. We see a significant gap in these percentages for $ScientistsGood$ and $SmartSadDumbHappy$, and moderately large gaps for $Male$, $VaccinesGood$, $College$, and $Married$. Even $HasSeenTransformers$ and $BelievesInGhosts$ look like they have potential to be significant. We will keep these rough rankings in mind as we proceed.

Since we have only two quantitative variables left in consideration, we can plot each against all binary variables to check for potential interactions. As Figure 2 shows, only four of these plots seem to have significantly different slopes for the two levels: Age against $Male$, $HasSeenTransformers$, and $BelievesInGhosts$; and $log(BooksInPastYear)$ against $Married$. We will include these four interactions in the variables we test as we build our model.

[1] The uppermost values of $BooksInPastYear$ are quite suspect, but we have no other reason to believe they are false, so we will leave them be.

Variable	**0**	**1**
$ScientistsGood$	0.421	0.715
$SmartSadDumbHappy$	0.572	0.717
$Male$	0.689	0.589
$VaccinesGood$	0.577	0.653
$College$	0.582	0.657
$Married$	0.684	0.601
$HasSeenTransformers$	0.610	0.667
$BelievesInGhosts$	0.620	0.676
$JobWillBeAutomated$	0.633	0.677
$ShowerPeeingOk$	0.655	0.625

Table 3: Ratio of anthropogenic climate change belief for each level of our binary variables

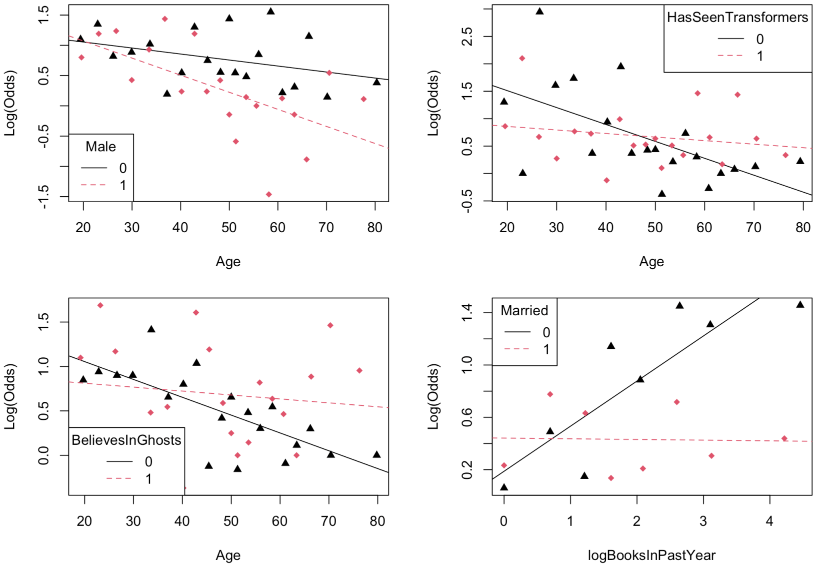


Figure 2: Significant-looking empirical logit plots of Age and $log(BooksInPastYear)$ split by binary variables

2.1 Model Development

To whittle down our variables into a suitable subset, we employ several methods. First, we run a backward step function on the full model, including all our terms and the four possible interactions we identified. This function identifies the least significant variable in the set, then checks if Akaike's Information Criterion (AIC)[2] a for the model without that variable is significantly higher. If it is, the step function drops that variable. It repeats this process until the increase in AIC is statistically significant.

Our backward stepwise function on the full set returns a model with many variables: Age, $log(BooksInPastYear)$, $Male$, $Married$, $HasSeenTransformers$, $ScientistsGood$, $SmartSadDumbHappy$, $Age : HasSeenTransformers$, and $log(BooksInPastYear) : Married$. Running this model confirms that almost every variable is indeed significant at level $\alpha = .05$. The exceptions are $Married$, which is significant at $\alpha = .1$, and the intercept, which is insignificant. The AIC for this model is 783.49, and the residual deviance is 763.49. The null deviance is 855.94, so this model provides a decent reduction in deviance. While the size of this model is remarkable, we may ultimately want to choose a more simple model if its predictive power is similar to this one.

We also run a forward stepwise function, which starts with the null model including only the intercept, adds the most significant term, checks that AIC was reduced significantly, and repeats until a fully significant set is found. The forward step returns a smaller model that includes $ScientistsGood$, $SmartSadDumbHappy$, Age, $Male$, $JobWillBeAutomated$, and $log(BooksInPastYear)$. We run a summary of this model and see that $log(BooksInPastYear)$ is not actually significant; removing it, we obtain a 5-term model with full significance besides the intercept. Its AIC is 793.42 and its residual deviance is 781.42.

The R function `bestglm()` offers a third way of testing our predictors all at once, based on BIC, a different criterion than AIC. At first, `bestglm()` returns a model with three predictors: $ScientistsGood$, $SmartSadDumbHappy$, and $Age : Male$. We check the summary of this model and see that each variable is highly significant. However, when we use interaction terms, we should include both original variables in the model. Checking the model using these variables plus Age and $Male$, we see that those terms are not significant, and neither is $Age : Male$ with them included. While this interaction is intriguing, we cannot use it.

Removing $Age : Male$ from the input, we run `bestglm()` again, which returns a model including $ScientistsGood$, $SmartSadDumbHappy$, $log(BooksInPastYear)$, and $log(BooksInPastYear) : Married$. Throwing $Married$ into the mix, we run this model and find that all but $Married$, including the intercept, are significant at level $\alpha = .01$. This model has an AIC of 793.78 and a residual deviance of 781.78. Despite

[2] AIC is a balanced measurement of a model's goodness of fit. It increases with more variables and decreases with smaller residual deviance, so smaller AIC values generally indicate better models.

Married being insignificant, this model is promising due to its simplicity and the fact that the intercept is significant.

2.2 Model Selection

We now have three models under consideration, each with their own merits.

$$\text{Model 1: } log(odds) = \beta_0 + \beta_1 ScientistsGood + \beta_2 SmartSadDumbHappy + \beta_3 Age + \beta_4 log(BooksInPastYear) + \beta_5 Male + \beta_6 Married + \beta_7 HasSeenTransformers + \beta_{37} Age : HasSeenTransformers + \beta_{46} log(BooksInPastYear) : Married + \varepsilon$$

$$\text{Model 2: } log(odds) = \beta_0 + \beta_1 ScientistsGood + \beta_2 SmartSadDumbHappy + \beta_3 Age + \beta_4 Male + \beta_5 JobWillBeAutomated + \varepsilon$$

$$\text{Model 3: } log(odds) = \beta_0 + \beta_1 ScientistsGood + \beta_2 SmartSadDumbHappy + \beta_3 Age + \beta_4 log(BooksInPastYear) + \beta_5 Married + \beta_{45} log(BooksInPastYear) : Married + \varepsilon$$

To help make our final decision, we will run them as training models on a subset of our data, then see how accurately they can predict whether the remaining subjects believe in anthropogenic climate change.

We first randomly sample the integers from 1 to 656, the number of observations in the full dataset, 500 times without repeats. We designate the subset of our data with those row indices our training set. After finding the coefficients for our three models with the training set, we use the models to obtain $log(odds)$ for the 156 remaining observations, then convert the $log(odds)$ to probabilities of belief in anthropogenic climate change.

Now that we have our predictions, how do we assess their accuracy? Consider the observations for which a model returns a probability greater than 0.5; if the value of *AnthClimateChange* for that row is 1, we could consider it a "correct" prediction. We can consider a prediction "correct" also if our probability guess is less than 0.5 and $AnthClimateChange = 0$. Unfortunately, this measurement does not give us the entire picture about our model accuracy. For instance, it does not measure how strong (i.e. far from 0.5) the predicted probabilities are. That said, it is still a useful measure of our model accuracy.

If we count up the number of "correct" predictions for each model and divide them by 156, we get the proportion of "correct" guesses for each model. After resampling the training set and repeating the process 1000 times, we obtain three 1000-length vectors of these accuracy proportions. Figure 3 displays histograms of these three vectors. Seeing that their distributions are unimodal and fairly symmetrical, we can safely assume that their means are a good representation of the true accuracy proportions for the three models.

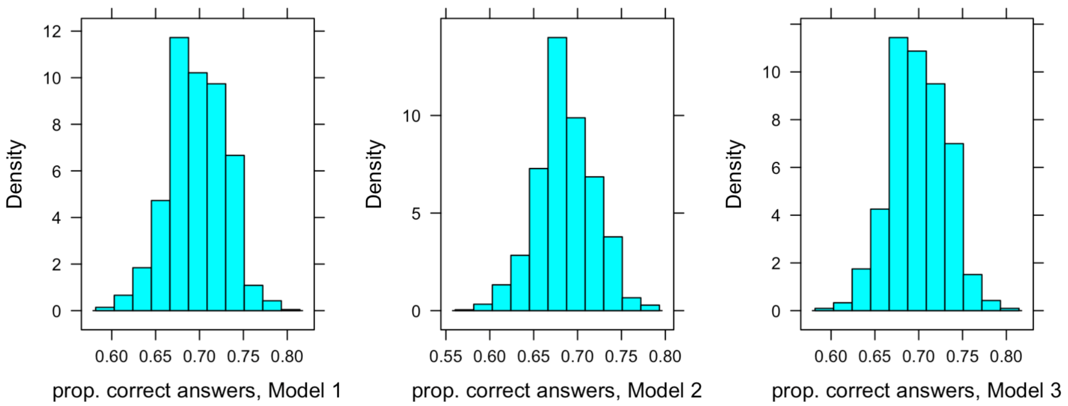


Figure 3: Histogram plots for the proportion of correct guesses made by the three models

Running the whole process a few times, we obtain means that average to 0.697 for Model 1, 0.686 for Model 2, and 0.699 for Model 3. These proportions are clearly very similar and our decision is not much easier. We might be inclined to eliminate Model 1 from consideration because it is much more complex than the others, but its AIC, which is designed to punish complexity, is the lowest of the three. Also, complexity might be good for certain purposes because each variable is a new opportunity to differentiate individuals from each other.

Perhaps controversially, the author decides to publish results for all three models. Each one is simply too interesting to ignore.

2.3 Checking Conditions

Logistic regression models are valid when their variables (a) are random, (b) are independent, and (c) have a linear relationship with the $log(odds)$ returned. We conclude that condition (a) is satisfied because the survey was conducted by a reputable firm using sound randomization procedures.

To check condition (b), we find the correlation coefficients between each of the variables in our three models. We see relatively high values for $cor(log(BooksInPastYear), SmartSadDumbHappy)$, $cor(log(BooksInPastYear), Male)$, $cor(Age, HasSeenTransformers)$, and $cor(Age, Married)$, though none are greater than 0.33. Pearson's correlation test tells us that all four correlations are highly significant, calling Models 1 and 3 into question but not Model 2, which does not include any of those variable combinations.

We can check the collinearity of our models individually with the variance inflation factor (VIF) of each variable other than the interaction terms, which we expect to be correlated with the original terms. Doing so, we find VIFs very close to 1 for all three models and decide that independence is a reasonable-enough assumption for each.

Finally, we check condition (c). Binary variables necessarily have a linear relationship with $log(odds)$ because there are only two possible values of $log(odds)$. Shown in Figure 4, plots of $log(odds)$ against our two quantitative variables display fairly linear relationships, though some points stray from the estimated linear function. We conclude that all our conditions are sufficient for each model.

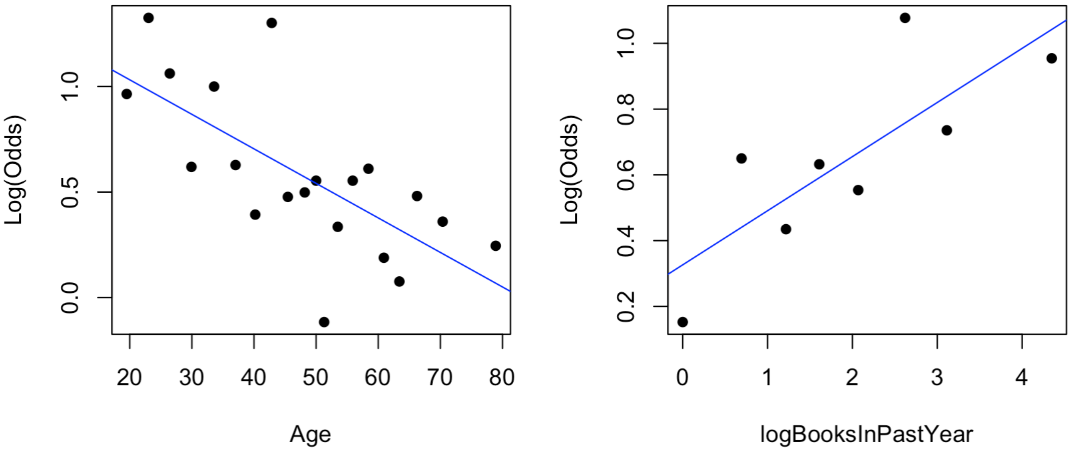


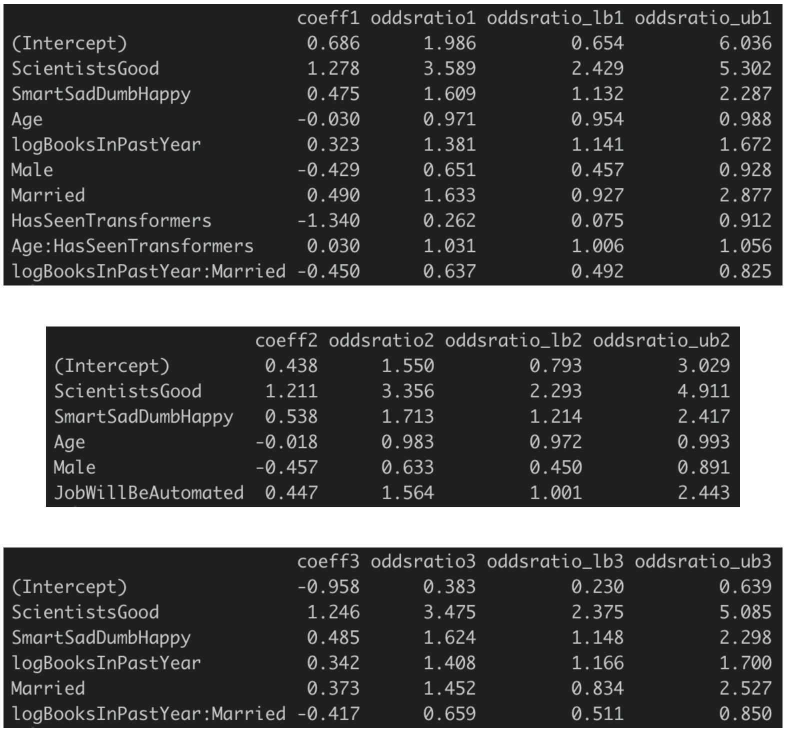
Figure 4: Plots of $log(odds)$ against Age and $log(BooksInPastYear$

3 Results

With our three models in hands, we turn to our questions of interest.

Question 1: How do changes in our variables affect the odds that an American adult believes in anthropogenic climate change?

We can answer this question by finding the odds ratio for each coefficient β_k, given by e^{β_k}, as well as the upper and lower bounds of its 95 percent confidence interval, given by $e^{\beta_k \pm z_{.025} SE_{\beta_k}}$ where $z_{.025}$ is the quantile q of the standard normal distribution for which $P(x < q) = .025$. Results are listed in Figure 5. For examples of how to interpret these numbers, consider the odds ratios for *Age* and *ScientistsGood* in Model 1. We estimate that an American adult is 0.971 times less likely to believe in climate change for every year they have been alive, and are 95 percent confident that the true shrinkage factor is between 0.954 and 0.988. We also estimate that an American adult who believes scientists are generally honest and serve the public good is 3.589 times more likely to believe in anthropogenic climate change than an American adult who does not, and we are 95 percent confidence that the true growth factor is between 2.429 and 5.302.



```
                            coeff1 oddsratio1 oddsratio_lb1 oddsratio_ub1
(Intercept)                  0.686      1.986         0.654         6.036
ScientistsGood               1.278      3.589         2.429         5.302
SmartSadDumbHappy            0.475      1.609         1.132         2.287
Age                         -0.030      0.971         0.954         0.988
logBooksInPastYear           0.323      1.381         1.141         1.672
Male                        -0.429      0.651         0.457         0.928
Married                      0.490      1.633         0.927         2.877
HasSeenTransformers         -1.340      0.262         0.075         0.912
Age:HasSeenTransformers      0.030      1.031         1.006         1.056
logBooksInPastYear:Married -0.450      0.637         0.492         0.825
```

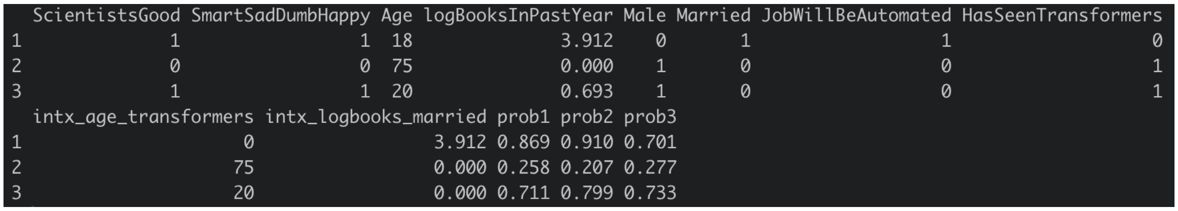
```
                  coeff2 oddsratio2 oddsratio_lb2 oddsratio_ub2
(Intercept)        0.438      1.550         0.793         3.029
ScientistsGood     1.211      3.356         2.293         4.911
SmartSadDumbHappy  0.538      1.713         1.214         2.417
Age               -0.018      0.983         0.972         0.993
Male              -0.457      0.633         0.450         0.891
JobWillBeAutomated  0.447      1.564         1.001         2.443
```

```
                            coeff3 oddsratio3 oddsratio_lb3 oddsratio_ub3
(Intercept)                 -0.958      0.383         0.230         0.639
ScientistsGood               1.246      3.475         2.375         5.085
SmartSadDumbHappy            0.485      1.624         1.148         2.298
logBooksInPastYear           0.342      1.408         1.166         1.700
Married                      0.373      1.452         0.834         2.527
logBooksInPastYear:Married -0.417      0.659         0.511         0.850
```

Figure 5: Tables of the estimated odds ratios for our model coefficients, with associated 95 percent confidence intervals

Question 2: How accurately can our model predict the likelihood that various profiles of American adults believe in anthropogenic climate change?

We answered this question in its grandest sense with our training models in the previous section. With accuracy proportions of 0.697 for Model 1, 0.686 for Model 2, and 0.699 for Model 3, all three models turn out to have decent accuracy with pure up-or-down predictions. For the sake of illustration, we use the full models to generate probabilities for a handful of respondent profiles, the results of which are shown in Figure 6. Based on the signs on each coefficient, the first row is designed to represent someone who is very likely to believe in climate change, and the second row someone who is very unlikely to believe in climate change. The third row contains answers provided by a friend of the author.



```
  ScientistsGood SmartSadDumbHappy Age logBooksInPastYear Male Married JobWillBeAutomated HasSeenTransformers
1              1                 1  18              3.912    0       1                  1                   0
2              0                 0  75              0.000    1       0                  0                   1
3              1                 1  20              0.693    1       0                  0                   1
  intx_age_transformers intx_logbooks_married prob1 prob2 prob3
1                     0                 3.912 0.869 0.910 0.701
2                    75                 0.000 0.258 0.207 0.277
3                    20                 0.000 0.711 0.799 0.733
```

Figure 6: Model predictions for three profiles of American adults

We see that, of the three models, Model 1 returns the highest probability for our likely believer and the lowest for our unlikely believer. In contrast, Model 3 gives the most conservative predictions for those cases. This is probably because Model 1 has the most predictors and we stacked the deck in favor of climate change belief as much as possible, while Model 3 has the fewest. If the reader wants more aggressive predictions, they should use Model 1; for more cautious predictions, Model 3 is best.

Each of our models correctly predicted with moderate certainty that the author's friend believes in climate change. Interestingly, his first three answers were consistent with the believer, while the other five were consistent with the nonbeliever, and yet he was predicted to believe in climate change. This emphasizes the outsized contributions of *ScientistsGood* and *Age* to our model.

3.1 Further Discussion

Clearly, our inferences are limited to the population of American adults. These models should not be used at times far from the year 2017, either. While we can probably expect our variables to remain somewhat constant over a few years, the dynamics of belief in climate change among the public were very likely different 5-10 years before this survey. Even four years later, we would want to use a survey conducted more recently to make conclusions about climate change belief in 2021.

The reader should also keep in mind that our model conditions are not as robust as we would like them to be, and there are likely some relationships between the variables that artificially reinforce predictions. Recall that Model 2 appears to be the least collinear of the three, and Model 3 is the simplest. The cautious statistician should prefer these two models to Model 1, even though its variables offer more chances to differentiate between individuals.

3.2 Conclusion

In this paper, we sought to develop a logistic regression model for the probability that an American adult believes in human-caused climate change. Using three versions of stepwise regression, we found three suitable models, each with benefits and downsides. We then used our models to derive meaningful statistical statements about how our variables affect the odds of climate change belief.

Note that the coefficient on *ScientistsGood* is by far the largest in each of our models. Our study is observational and we cannot conclude whether trust in scientists causes greater climate change belief or vice versa, nor whether a third factor causes both. That said, it is worth exploring further how much the fostering of trust in scientists among the public can increase the number of adults who believe in climate change.

References

[1] Cards Against Humanity. Pulse of the Nation [dataset]. Chicago, IL: Cards Against Humanity, 2017. Retrieved 12/05/2021 from https://www.kaggle.com/cardsagainsthumanity/pulse-of-the-nation.