401-2\_Homework2

Cat Dang Ton

2024-04-14

## Question 1

I am still using data from the 2022 Health and Retirement Survey. The sample consists of respondents to the survey’s subsection on Widowhood and Divorce, who experienced the death of a spouse between 2020 and 2022.

Dependent variable: whether or not the widow had to sell assets, withdraw money that normally would not be touched, get help from a relative, or from a church or other institution, or do anything else special to find the money to cover the deceased spouse’s death expenses (funeral expenses, legal fees, expenses for any illness that led to death). (Yes = 1, No = 0).

Independent variables:

1. Whether the living spouse is US-born or foreign born (US-born = 1, foreign-born = 0)
2. Total death expenses that are NOT covered by insurance or the deceased spouse’s estate (Exact dollar amounts, or if not available, the midpoint of approximated estimates)
3. The widow’s level of education (note: widow’s income calculations are forthcoming. Education is a stand-in for now.)

### A.

# Convert dependent variable and continuous variable to numeric so the model takes it  
df$deathexpense\_special <- as.numeric(df$deathexpense\_special)  
df$deathexpense\_usd <- as.numeric(df$deathexpense\_usd)  
# convert unit of analysis to thousands of dollars  
df$deathexpense\_1k <- df$deathexpense\_usd/1000   
# run model  
model <- lm\_robust(deathexpense\_special ~ usborn + deathexpense\_1k + degree, data = df)  
  
modelsummary(model)

|  |  |
| --- | --- |
| (Intercept) | 0.363 |
|  | (0.074) |
| usbornus-born | -0.127 |
|  | (0.054) |
| deathexpense\_1k | 0.006 |
|  | (0.004) |
| degreeDegree unknown/Some College | 0.037 |
|  | (0.098) |
| degreeGED | -0.046 |
|  | (0.078) |
| degreeHigh school diploma | -0.071 |
|  | (0.046) |
| degreeNo degree | -0.028 |
|  | (0.057) |
| degreePostgraduate degree | -0.066 |
|  | (0.073) |
| Num.Obs. | 741 |
| R2 | 0.031 |
| R2 Adj. | 0.022 |
| AIC | 853.5 |
| BIC | 894.9 |
| RMSE | 0.43 |

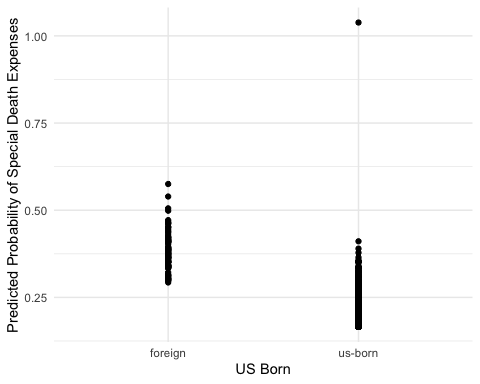
Reference category: College-educated foreign-born person, who has no death expenses beyond what is covered by insurance or by their deceased spouse’s estate.

The number of cases in the model is 741.

The slope for US-born respondents suggests that, after controlling for the amount final expenses not covered by insurance and for their level of education, a widow’s US-born status renders them 23% less likely to have to resort to special means to finance final expenses, compared to foreign-born widows.

### B.

# Generate predicted values  
df$predicted\_prob <- predict(model, newdata = df)  
  
# Create a scatter plot of the predicted probabilities  
ggplot(df, aes(x = usborn, y = predicted\_prob)) +  
 geom\_point() +  
 labs(x = "US Born", y = "Predicted Probability of Special Death Expenses") +  
 theme\_minimal()



### C.

# Estimate a logistic regression model  
model\_logit <- glm(deathexpense\_special ~ usborn + degree + deathexpense\_1k,   
 data = df,   
 family = binomial(link = "logit")) #specify that we want logistic regression(link = "logit"))   
modelsummary(model\_logit) # print logit regression estimates

|  |  |
| --- | --- |
| (Intercept) | -0.590 |
|  | (0.313) |
| usbornus-born | -0.601 |
|  | (0.240) |
| degreeDegree unknown/Some College | 0.176 |
|  | (0.457) |
| degreeGED | -0.252 |
|  | (0.436) |
| degreeHigh school diploma | -0.384 |
|  | (0.235) |
| degreeNo degree | -0.147 |
|  | (0.291) |
| degreePostgraduate degree | -0.369 |
|  | (0.402) |
| deathexpense\_1k | 0.032 |
|  | (0.012) |
| Num.Obs. | 741 |
| AIC | 824.7 |
| BIC | 861.5 |
| Log.Lik. | -404.329 |
| F | 2.910 |
| RMSE | 0.42 |

# Calculate and print odds ratios  
rbind(coef(model\_logit), # get coefficients  
 exp(coef(model\_logit))) # get odds ratios

(Intercept) usbornus-born degreeDegree unknown/Some College degreeGED  
[1,] -0.5898708 -0.6005425 0.1762282 -0.2516803  
[2,] 0.5543989 0.5485140 1.1927102 0.7774932  
 degreeHigh school diploma degreeNo degree degreePostgraduate degree  
[1,] -0.3842166 -0.1468196 -0.3689075  
[2,] 0.6809839 0.8634497 0.6914893  
 deathexpense\_1k  
[1,] 0.03173747  
[2,] 1.03224647

Reference category: College-educated foreign-born person, who has no death expenses beyond what is covered by insurance or by their deceased spouse’s estate.

US-born: the logged odds of resorting to special means to finance death expenses are lower by 0.60 for US-born widows than for foreign-born widows, holding all other variables constant. The odds of resorting to special means to finance death expenses are 46% lower for US-born widows than for foreign-born widows.

Degree: Compared to those with a college degree, the logged odds of resorting to special means to finance death expenses are lower for those with no degree, those with a GED, a high school diploma, or a postgraduate degree. The odds of resorting to special means to finance death expenses are 14% lower for those with no degree, 23% lower for those with a GED, 32% lower for those with a high school diploma, 31% lower for those with a postgraduate degree, and 19% higher for those who enrolled in college without obtaining a degree.

Death expenses: the logged odds of resorting to special means to finance death expenses increase by 0.03 for each thousand-dollar increase in such expenses. Each additional thousand dollars in death expenses not covered by insurance increases the odds of resorting to special means of financing them by 3%.

### D.

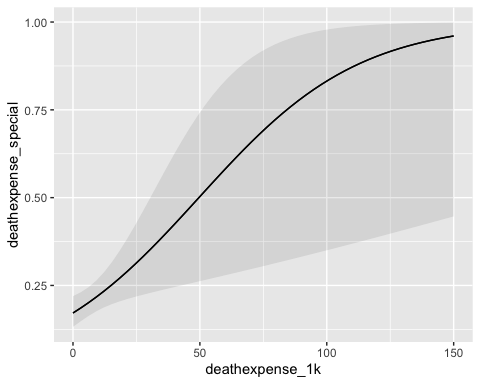
# Average Marginal Effects is:  
# The average of all observation-specific marginal effects  
avg\_slopes(model\_logit)

Term Contrast Estimate  
 deathexpense\_1k dY/dX 0.00574  
 degree Degree unknown/Some College - College degree 0.03689  
 degree GED - College degree -0.04832  
 degree High school diploma - College degree -0.07153  
 degree No degree - College degree -0.02884  
 degree Postgraduate degree - College degree -0.06893  
 usborn us-born - foreign -0.12048  
 Std. Error z Pr(>|z|) S 2.5 % 97.5 %  
 0.00216 2.654 0.00795 7.0 0.0015 0.00997  
 0.09755 0.378 0.70530 0.5 -0.1543 0.22809  
 0.08102 -0.596 0.55095 0.9 -0.2071 0.11048  
 0.04555 -1.570 0.11634 3.1 -0.1608 0.01775  
 0.05697 -0.506 0.61261 0.7 -0.1405 0.08281  
 0.07196 -0.958 0.33811 1.6 -0.2100 0.07211  
 0.05217 -2.309 0.02092 5.6 -0.2227 -0.01823  
  
Columns: term, contrast, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high   
Type: response

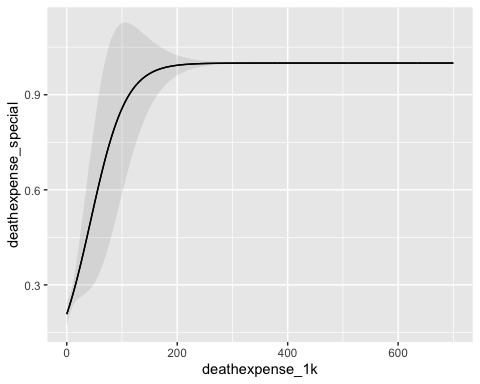
Among the variables in the model, it seems the most important predictor of whether or not a widow turns to special means to finance their deceased spouse’s final expenses is the widow’s US-born status because the estimate is statistically significant at 0.05, the coefficient estimate for usborn has the largest absolute value among the coefficients, the confidence interval does not include 0, and the standard error is relatively small.

### E.

# 1. Graph the probability of the outcome for values of one of the independent variables,   
# holding the other variables at their mean or mode.  
# Also known as the default setting of plot\_predictions.  
# condition = independent variable of interest.  
# if condition var = continuous, plot\_predictions plots the mean.  
plot\_predictions(model\_logit, condition = "deathexpense\_1k") # continuous indep. var.



# 2. Plot average predictions   
# graph average predictions based on letting controls retain original values,  
# but changing X rather than holding other variables at mean/mode.  
# this is known as a "counterfactual" grid type in plot\_predictions()  
plot\_predictions(model\_logit, by = "deathexpense\_1k", newdata = datagrid(deathexpense\_1k = 0:700, grid\_type = "counterfactual"), type = "response")



## Question 2

### A.

Outcome variable: whether or not the speech contains populist claims, based on an iteratively defined dictionary of populist terms. It is a binomial variable with 0 and 1 values.

Main independent variables:

* Party (Republican or Democratic),
* party incumbency,
* prior office,
* political career length,
* phase of campaign in relation to election day,
* interaction term between field position and campaign duration (?) (from Hypothesis 6; but regression table 2 did not have an interaction term?)
* Geographic region

### B.

The odds ratio of an incumbent’s speech being populist is 0.279, which means that for every political speech from an incumbent party member that contains a populist claim, there would be nearly 4 speeches from challengers containing a populist claim. Incumbents are thus less likely to make populist claims than challengers.

Controls were included for length of speech, because the longer the speech, the more likely it is to contain a populist claim.

### C.

For a baseline probability of 0.3, the effect of being part of the incumbent party on the proability of using populist rhetoric is

### D.

Given the hypotheses listed in Table 1, I’m inclined to think that the only control variable is speech length. Each of the models in table 2 corresponds to one of the hypotheses in table 1 (e.g. models 1 & 2 tests hypothesis 1, models 5-7 tests hypothesis 4), which means that these are separate models for separate hypotheses. There are no hypotheses here that would require running a model with all variables included as controls.

### E.

Significance levels: 0.05, 0.01, 0.001

One constraint comes from a confounding variable. It is possible that carrying out a populist speech increases (or decreases) the likelihood that a candidate stays in the running and makes more speeches. This latent “success rate” relates to the independent variable (duration of campaign) and the dependent variable (presence of populist claims).

### F.

The status of having never been in political office (“field position: none”) is the strongest predictor of populism in a speech, because its odds ratios are greatest across 11 models, and their estimates are statistically significant at the 0.001 threshold. Compared to members of the previous administration, those who held no previous office had 3.6 to 5 times higher odds of making populist claims.

### G.

The measurement of populist claims-making using only the presence or absence of populist claims does not capture the salience or prevalence of such claims within a given speech. Measuring salience is important not only in consideration of the role of rhetorical emphasis in reinforcing political claims, but is also crucial for a contemporary analysis. Given the rise of populism, it may well be that populist claims exist in a majority of political speeches in the 21st century, but some speeches mention a single populist sentence, while others make populism their entire platform. Classifying speeches on presence-absence only, thus, erases this potential variance in the sample, which could result in less precise regression estimates.

It is difficult to measure the prevalence of some type of claim using only a dictionary model (which relies on exact keyword matches). Structural topic models could offer a solution to this issue.

Validation also involves testing the dictionary algorithm on an untrained dataset to avoid overfitting results. If I understand correctly, it requires splitting the original data set into at least two parts, a training set and a test set. We develop the dictionary on one of them, then apply the dictionary on the other, then manually review/validate and report only the results of the test set. The authors of the article didn’t specify if they’d done this, so we can’t rule out the possibility that the resulting estimates were overfitted–that the estimates were contingent on what happens to come up in those speeches used to develop the dictionary.