Applied Stats II, PS2

Luke Duggan

27 February 2022

1 Question One

First of all, we fit an additive model. Since the outcome variable is binary, a linear specification estimated by OLS would be inappropriate for a number of reasons, of which I mention two. Firstly, the point estimates for the coefficients would not necessarily lie in [0,1] and would therefore make no sense interpreted as contributions to the *probability* of the outcome variable taking the value 1. Secondly, the distribution of the error term would be non-normal and heteroskedastic, so OLS would be an inefficient estimator.

Instead, we use a generalized linear / additive specification: a logistic model. Trying to fit this model in R turned out to be a bit of a headache. My first attempt was as follows:

```
results <-glm(choice ~ countries + sanctions, data= climateSupport,
              family=binomial(link="logit"))
summary(results)
```

The estimated coefficients from the model are provided below.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.005665   0.021971  -0.258 0.796517
countries.L  0.458452   0.038101  12.033  < 2e-16 ***
countries.Q -0.009950   0.038056  -0.261 0.793741
sanctions.L -0.276332   0.043925  -6.291 3.15e-10 ***
sanctions.Q -0.181086   0.043963  -4.119 3.80e-05 ***
sanctions.C  0.150207   0.043992   3.414 0.000639 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The explanatory variables are coded in R as *ordered* factors since, e.g., a 5% sanction is more than None. For some reason, when you include ordered factors in a `lm` or `glm` call in R, it automatically includes certain polynomial terms: thus, "L" stands for linear in this output, "Q" for quadratic, and "C" for cubic. I can't honestly say I know why.

This doesn't resemble anything we saw in the lectures, so I tried estimating the model after removing the order from the factors.

```
countries_unordered <- factor(climateSupport$countries, ordered = FALSE )
sanctions_unordered <- factor(climateSupport$sanctions, ordered = FALSE )

results2 <-glm(choice ~ countries_unordered + countries_unordered,
               data= climateSupport, family=binomial(link="logit"))
summary(results2)
```

The coefficient estimates from this model were:

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -0.27266    0.05360  -5.087 3.64e-07 ***
countries_unordered80 of 192   0.33636    0.05380   6.252 4.05e-10 ***
countries_unordered160 of 192  0.64835    0.05388  12.033  < 2e-16 ***
sanctions_unordered5%           0.19186    0.06216   3.086  0.00203 **
sanctions_unordered15%         -0.13325    0.06208  -2.146  0.03183 *
sanctions_unordered20%         -0.30356    0.06209  -4.889 1.01e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This makes a bit more sense: the various values of the categorical variables have been re-coded as dummies, with a baseline or implied case (20 of 192 countries and 0% sanctions) for either. With some trepidation, I will use the output of this model to answer the questions below.

Let us interpret the coefficients of this model. Firstly, the baseline case is when only 20 countries participate and there are no sanctions. The actual quantity itself is the log odds ratio: to interpret it, we supply it as an argument to the exponential function. This is associated with a exp(-0.27) = 0.76 odds ratio: that is, the expected odds for a person's accepting the policy under these conditions is 0.76. This is less than one, so the odds are against the person's accepting the policy.

To take two of the other coefficients as illustrative: exponentiating the estimated log odds ratios for the 5% sanctions and the 20% sanctions, we get that an individual falling into these categories is associated with a multiplicative effect on the odds ratio of their accepting of either 1.21 and 0.74. In the former case, the odds move in favour of acceptance; in the other, against. Thus, the odds of a person's accepting the policy are greater for *smaller* sanctions, which probably accords with intuition.

The p-values for individual coefficients are calculated using a standard normal distribution, which is asymptotically valid. All our estimates are statistically significant, as the rightmost column shows.

What about the "global null hypothesis" that all coefficients are statistically insignificant? For this, we need the null and residual deviance, which we can use to perform a likelihood ratio test. These quantities are also provided in the summary output:

```
    Null deviance: 11783  on 8499  degrees of freedom
Residual deviance: 11568  on 8494  degrees of freedom
AIC: 11580
```

The difference between these quantities is Chi-squared distributed with degrees of freedom equal to the number of coefficients being tested. Alternatively, to get R to do the test for us, we can estimate a "reduced" model with all the coefficients dropped and perform a likelihood ratio test. The reduced model is:

```
reduced <- results <-glm(choice ~ 1, data= climateSupport,
                         family=binomial(link="logit"))
summary(reduced)
```

The output is:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.006588   0.021693  -0.304    0.761

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11783  on 8499  degrees of freedom
Residual deviance: 11783  on 8499  degrees of freedom
AIC: 11785
```

Then, to perform the likelihood ratio test, we use the following command:

```
anova(reduced, results2, test = "Chisq")
```

The output of which is:

```
Model 1: choice ~ 1
Model 2: choice ~ countries_unordered + sanctions_unordered
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      8499      11783
2      8494      11568  5   215.15 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is very small, we reject the null hypothesis that no explanatory variable is significant.

Summarizing these results, we would say that there is a statistically significant association between a larger number of countries participating and the odds of a person's supporting the policy, and a statistically significant association between the size of the sanction and a decrease in the odds of a person's supporting the policy.

2 Question Two

1. (a): Changing sanctions to 15% sanctions multiplies the odds of accepting by exp(-0.133) = 0.875, that is to say, it decreases the odds of a person's accepting, in line with what we said earlier. If the countries category variable is at 160, we should first multiply the baseline odds ratio by exp(0.336) = 1.4 to get the odds of accepting in that case: then, we multiply the baseline by 0.875 to compare the effects of this change.

2. (b): If there are no interaction effects in the model, an increase in the level of the `sanctions` categorical variable has an identical effect regardless on the number of countries participating: it multiplies the odds of a person's accepting by 0.875. We will consider whether there *ought* to be an interaction effect in part (d).

3. (c): The baseline case is -0.272 log odds or .761 odds of accepting. If we have 80 countries participating, this is associated with a change of about 1.4 in the odds of a person's accepting: multiplying our old odds, we get odds of around 1.065.

4. (d): The inclusion of an interaction term could alter the answers to (a) and (b): if there is a statistically significant interaction, it means that changes in, e.g., the levels of sanctions will have different effects on the odds of accepting the policy for different numbers of countries participating. Putting the same point negatively: if there's no statistically significant interaction, a change in the sanctions has the same effect on the odds regardless of how many countries are participating.

   To check if there is a statistically significant interaction, we include it in our original model:

```
results3 <-glm(choice ~ countries_unordered + sanctions_unordered
               + countries_unordered*sanctions_unordered,
               data= climateSupport, family=binomial(link="logit"))
summary(results3)
```

   The p-values from this model are:

```
                                                    Pr(>|z|)
(Intercept)                                         0.000267 ***
countries_unordered80 of 192                        0.000408 ***
countries_unordered160 of 192                       1.41e-08 ***
sanctions_unordered5%                               0.246909
sanctions_unordered15%                              0.370723
sanctions_unordered20%                              0.019412 *
countries_unordered80 of 192:sanctions_unordered5%  0.534071
countries_unordered160 of 192:sanctions_unordered5% 0.389063
countries_unordered80 of 192:sanctions_unordered15% 0.730262
countries_unordered160 of 192:sanctions_unordered15% 0.735136
countries_unordered80 of 192:sanctions_unordered20% 0.191675
countries_unordered160 of 192:sanctions_unordered20% 0.711279
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The inclusion of interaction terms has caused two of the levels for the `sanctions` categorical variable to become statistically insignificant: more importantly, however, none of the interaction variables are themselves significant. Thus, we cannot reject the hypothesis that there are no interaction effects between number of countries participating in the policy and the level of sanctions.