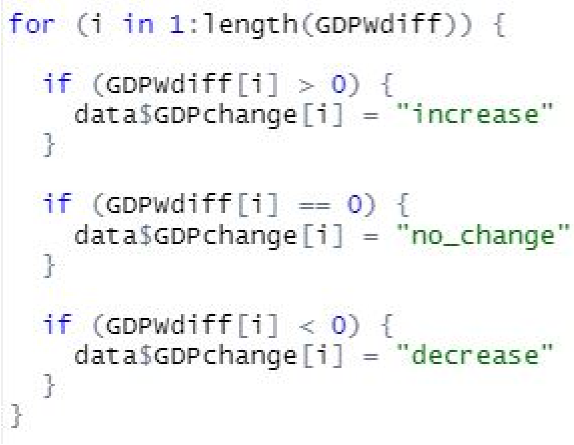
Applied Stats II, PS3

Luke Duggan

27 March 2022

1 Question One

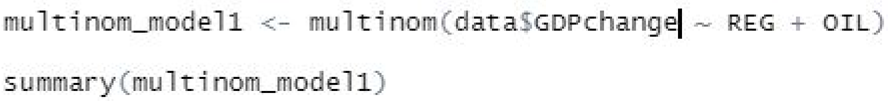
1. First, we run a multinomial logit model. The outcome variable is supposed to be `GDPWdiff`: however, in the dataset this variable is not a categorical variable. It records the actual numerical difference between the GDP in a year and GDP in the previous year, as opposed to being a factor variable with three levels for an increase, decrease, or no change in GDP. Accordingly, the first thing I did was create a categorical variable `GDPchange` with levels for ”increase”, ”decrease”, and ”no change”.

   ```
   for (i in 1:length(GDPWdiff)) {

     if (GDPWdiff[i] > 0) {
       data$GDPchange[i] = "increase"
     }

     if (GDPWdiff[i] == 0) {
       data$GDPchange[i] = "no_change"
     }

     if (GDPWdiff[i] < 0) {
       data$GDPchange[i] = "decrease"
     }
   }
   ```

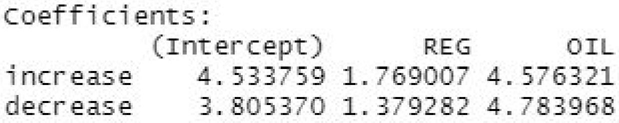
   A little bit of jiggery-pokery was required to make sure the levels of this factor were in the correct order: by default, the multinom() function treats the first listed level as the reference category, so I made sure this was ”no change”. See the attached R script file.

   We can then fit a multinomial logit model with this categorical variable as the outcome and two binary variables representing whether the country is a democracy and whether fuel predominates among exports as the explanatory variables:



```
multinom_model1 <- multinom(data$GDPchange ~ REG + OIL)

summary(multinom_model1)
```

The coefficient estimates from the model were as follows:

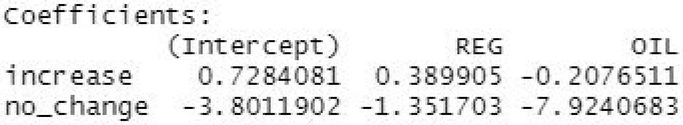


```
Coefficients:
         (Intercept)      REG      OIL
increase    4.533759 1.769007 4.576321
decrease    3.805370 1.379282 4.783968
```

The interpretation of the intercepts is that they give the additive effect on the log odds of GDP’s, e.g., increasing as opposed to staying the same when both explanatory variables are zero, i.e., a country is neither a democracy or primarily an oil exporter. The issue of ”log odds”, and the interpretation of the ”slope” coefficients, is somewhat involved. Briefly: if the variable `REG` increases by one (that is, we consider a democracy as opposed to a non-democracy), this is associated with an addition of about 1.77 to the *log odds* of an increase in GDP compared to GDP not changing. Equivalently, being a democracy as opposed to a non-democracy is associated with a multiplicative effect of $e^{(1.77)} \approx 5.87$ on the odds of GDP increasing as opposed to staying the same: that is, the odds of an increase in GDP as opposed to no change improve nearly sixfold.

This doesn’t necessarily tell us very much, however: going from being a non-democracy to being a democracy is also associated with a multiplicative effect of $e^{1.38} \approx 3.97$ on the odds of GDP *decreasing* as opposed to staying the same. Thus, we can conclude that going from being a non-democracy to being a democracy is strongly associated with an increase in the odds of GDP *changing*; it’s not as clear whether the change will be positive or negative, however.

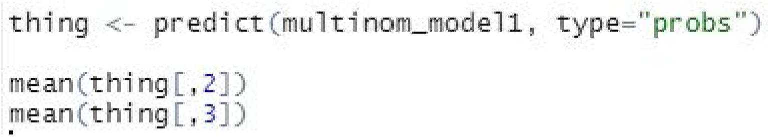
It seems to me it would be more profitable to investigate this question with one of ”increase” or ”decrease” as the reference category. If we choose ”decrease” as the reference category, the coefficient estimates become:



```
Coefficients:
          (Intercept)       REG        OIL
increase    0.7284081  0.389905 -0.2076511
no_change  -3.8011902 -1.351703 -7.9240683
```

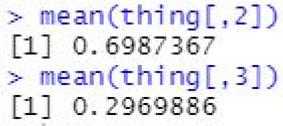
Similarly to before, being a democracy as opposed to a non-democracy is associated with large fall in the odds that GDP will stay the same as opposed to decrease (this is indicated by the negative sign on the estimated parameter). However, perhaps more interestingly, there estimated multiplicative effect of being a democracy on the odds of GDP increasing as opposed to decreasing is $e^{0.39} \approx 1.48$. Thus, the odds improve by about 50%. Another interesting finding is that being a predominantly oil exporting country is associated with a *negative* effect on the odds of GDP increasing as opposed to decreasing.

So much for the interpretation of the coefficients. Let us now turn to the estimated "cutoff points".



```
thing <- predict(multinom_model1, type="probs")

mean(thing[,2])
mean(thing[,3])
```

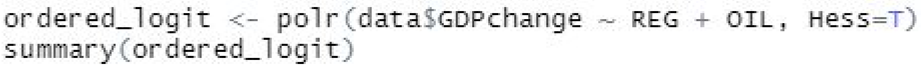
This gives estimated cutoff points:



```
> mean(thing[,2])
[1] 0.6987367
> mean(thing[,3])
[1] 0.2969886
```

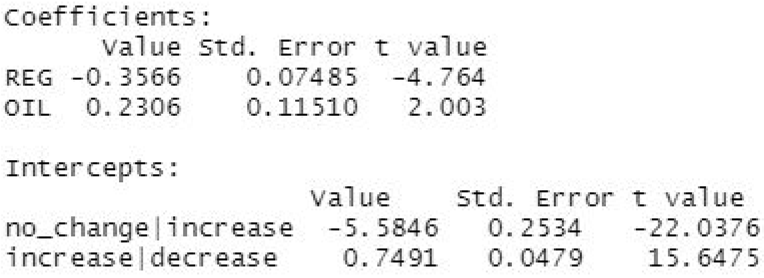
So, for example, if the model's predicted probability for an observational unit is above about 70%, we predict an increase in GDP. These predicted values are to be understood with reference to the logistic probability density function.

2. We now fit an *ordered* multinomial logit with the same outcome and explanatory variables.



```
ordered_logit <- polr(data$GDPchange ~ REG + OIL, Hess=T)
summary(ordered_logit)
```

The estimated coefficients from this model are:



```
Coefficients:
      Value Std. Error t value
REG -0.3566    0.07485  -4.764
OIL  0.2306    0.11510   2.003

Intercepts:
                      Value    Std. Error t value
no_change|increase  -5.5846    0.2534    -22.0376
increase|decrease    0.7491    0.0479     15.6475
```

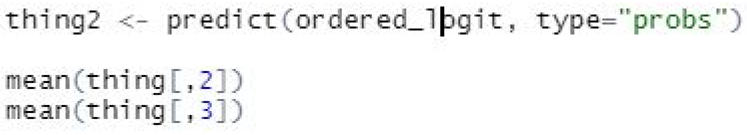
The interpretation of the coefficients is similar to the unordered multinomial model, with one important difference. Again, the coefficients give the additive effect on the log odds of a certain event's occurring; or, equivalently, exponentiating the estimates gives the multiplicative effect on the odds of the event happening. In an ordered model, however, the event is not that the outcome falls into one category as opposed to some reference category: instead, it is the event that the outcome falls into some category j

as opposed to the *next* category $j+1$; where the notion of the ”next category” makes sense under the assumption that the categories are ordered.

Note that there is only one estimated parameter for each explanatory variable. This encapsulates the ”proportionate odds” assumption: the effect of a one-unit increase in an explanatory variable on the probability an outcome falls into category j as opposed to category $j+1$ cannot depend on j; the effect must be uniform across categories.

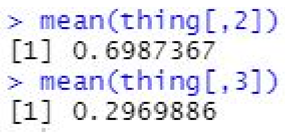
In this case, we have an estimate of about -0.36 on the democracy indicator. This suggests that being a democracy is associated with a multiplicative effect of $e^{-0.36} \approx .70$ on the odds of the outcome being in category j as opposed to $j+1$. Since I ordered the levels of the factor ”no change” / ”increase” / ”decrease”, this means that being a democracy is associated with a decrease in the probability of no change in GDP as opposed to an increase (as we saw in the unordered model) and in the probability of an increase in GDP as opposed to a decrease. This latter finding is contradictory to what we earlier saw, but may be an artifact of the unnatural ordering I gave the factor levels.

As for the estimated cutoff values:



```
thing2 <- predict(ordered_logit, type="probs")

mean(thing[,2])
mean(thing[,3])
```

From which the output is:

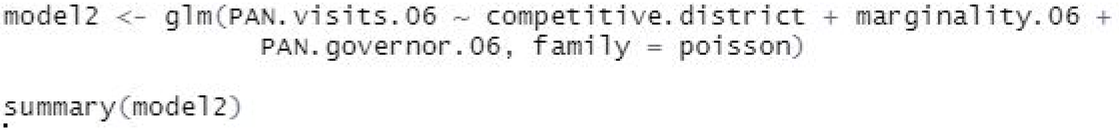


```
> mean(thing[,2])
[1] 0.6987367
> mean(thing[,3])
[1] 0.2969886
```

We see that the cutoff values are the same as in the unordered model.

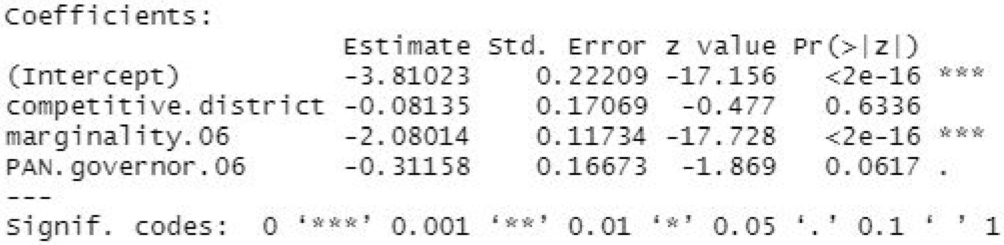
2 Question Two

1. First of all, we run a Poisson regression:

   ```
   model2 <- glm(PAN.visits.06 ~ competitive.district + marginality.06 +
                 PAN.governor.06, family = poisson)

   summary(model2)
   ```

   The output of this model is:

   ```
   Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
   (Intercept)          -3.81023    0.22209 -17.156   <2e-16 ***
   competitive.district -0.08135    0.17069  -0.477   0.6336
   marginality.06       -2.08014    0.11734 -17.728   <2e-16 ***
   PAN.governor.06      -0.31158    0.16673  -1.869   0.0617 .
   ---
   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
   ```

   We ask the question: is there evidence that PAN presidential candidates visit swing districts more? We can state a formal hypothesis as follows. The interpretation of the estimated coefficient for competitive.district is that its value under e^x gives the multiplicative effect of a one variable increase in the explanatory variable on the expected count of the outcome. In this case, if a district is a swing district, exponentiating the estimate gives the estimated multiplicative effect of being a swing district on the expected number of visits by a PAN presidential candidate.

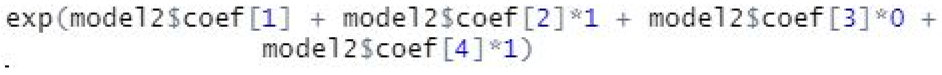
   Then, our formal null hypothesis might be: the coefficient on this variable is non positive, i.e., being a swing district has no effect or a negative effect on the expected number of visits by a presidential candidate. Of course, our theoretical intuition is probably the opposite: but we want to see if our data is powerful enough to reject the null. We see that the point estimate is indeed negative: however, using a z-test under the assumption that the null is true, we see that the result is not statistically significant, with a p-value of 0.6336. Thus, we cannot at any conventional level of significance reject the hypothesis that that being a swing district is irrelevant or negatively correlated with the number of visits by a presidential candidate.

2. Next, we interpret the coefficients on the marginality.06 and PAN.governor.06 variables. As mentioned before, the (exponentiated) coefficients in a Poisson regression signify the estimated multiplicative effect of a one unit increase in the explanatory variable on the expected count of the outcome. In this case, the PAN governor variable is a dummy; it seems that having a PAN governor is associated with a multiplicative effect of $e^{-0.31158} \approx 0.73$ on the expected number of visits by a presidential candidate: that

is, having a PAN governor cuts the number of visits you can expect to about 75% of if the district had no such governor. This probably makes theoretical sense.

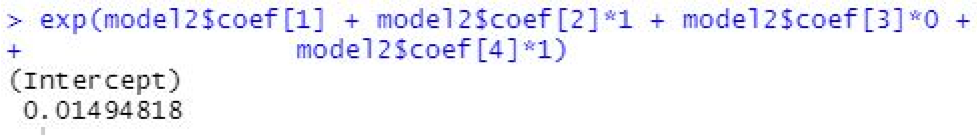
The marginality variable is continuous, and it suggests that a one unit increase in this variable strongly decreases the expected number of visits, reducing it to about 13% of its prior value. It's hard to interpret this without knowing more about the measure of poverty employed, e.g., does a high value of this measure mean greater or lesser poverty?

3. The fitted value for the combination of explanatory variable values described in the question is:



```
exp(model2$coef[1] + model2$coef[2]*1 + model2$coef[3]*0 +
                    model2$coef[4]*1)
```

Which is:



```
> exp(model2$coef[1] + model2$coef[2]*1 + model2$coef[3]*0 +
+                    model2$coef[4]*1)
(Intercept)
 0.01494818
```

Thus, the expected number of visits from the winning candidate for a district having these characteristics is about 0.01; not very high.