Answer Key: Problem Set 3

Applied Stats II

Jeffrey Ziegler

Instructions

- *Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in* `R`*, please include the code you used to get your answers. Please also include the* `.R` *file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.*
- *Your homework should be submitted electronically on GitHub in* `.pdf` *form.*
- *This problem set is due before 23:59 on Sunday March 26, 2023. No late assignments will be accepted.*

Question 1

We are interested in how governments' management of public resources impacts economic prosperity. Our data come from Alvarez, Cheibub, Limongi, and Przeworski (1996) and is labelled `gdpChange.csv` *on GitHub. The dataset covers 135 countries observed between 1950 or the year of independence or the first year forwhich data on economic growth are available ("entry year"), and 1990 or the last year for which data on economic growth are available ("exit year"). The unit of analysis is a particular country during a particular year, for a total > 3,500 observations.*

1. *Construct and interpret an unordered multinomial logit with* `GDPWdiff` *as the output and "no change" as the reference category, including the estimated cutoff points and coefficients.*

   First, let's load in our data and change our outcome to an unordered categorical variable in which the baseline is "no change". Then, we can estimate our unordered multinomial logit.

```r
# load data
gdp_data <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2023/main/datasets/gdpChange.csv", stringsAsFactors = F)
# create factor variable for outcome
gdp_data[gdp_data$GDPWdiff==0, "GDPcat"] <- "no change"
gdp_data[gdp_data$GDPWdiff>0, "GDPcat"] <- "positive"
gdp_data[gdp_data$GDPWdiff<0, "GDPcat"] <- "negative"
gdp_data$GDPcat <- relevel(as.factor(gdp_data$GDPcat), ref="no change")

# (1)
# run unordered multinom logit
unordered_logit <- multinom(GDPcat ~ REG + OIL, data=gdp_data)
summary(unordered_logit)
```

```
Coefficients:
(Intercept)      REG      OIL
negative    3.805370 1.379282 4.783968
positive    4.533759 1.769007 4.576321
Residual Deviance: 4678.77
AIC: 4690.77
```

Looking at the above output, we can see that we have essentially estimated two different regression lines in which we model the likelihood of shifting from "no change" to "negative", and "no change" to "positive". So, we can interpret the estimated coefficient for the intercept in `negative` as: when REG and OIL both $= 0$, the estimated log odds of going from no change to negative is ≈ 3.8. When we shift REG from 0 to 1 (a one unit change in x) there is an estimated change in log odds of going from no change to negative by ≈ 1.38, holding OIL constant at the observed mean. We can then make the same statement about OIL: going from 0 to 1 coincides with an estimated change in log odds of going from no change to negative by ≈ 4.78, holding REG constant at the observed mean.

2. *Construct and interpret an ordered multinomial logit with* ***`GDPWdiff`*** *as the outcome variable, including the estimated cutoff points and coefficients.*

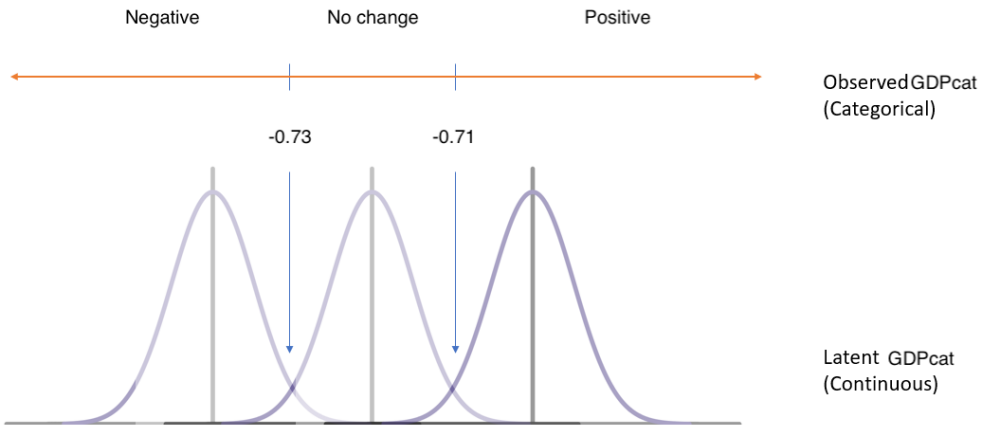
Before we run our ordered model, we need to re-order our outcome so that the categories are in increasing order (negative → no change → positive).

```r
# (2)
# re-level factor to impose ordering
gdp_data$GDPcat <- relevel(gdp_data$GDPcat, ref="negative")
# run ordered multinom logit
ordered_logit <- polr(GDPcat ~ REG + OIL, data=gdp_data)
summary(ordered_logit)
```

```
Coefficients:
Value Std. Error t value
REG  0.3985    0.07518   5.300
OIL -0.1987    0.11572  -1.717
Intercepts:
Value    Std. Error t value
negative|no change  -0.7312   0.0476   -15.3597
no change|positive  -0.7105   0.0475   -14.9554
Residual Deviance: 4687.689
AIC: 4695.689
```

Now, when we look at the output from the ordered multinomial logit we need to interpret the estimated coefficients for our intercepts (or cutpoints) as well as the other covariates differently than the unordered model. So, a shift in REG from 0 to 1 (a one unit change in x) corresponds to an estimated change in the log odds of going from negative to no change AND from no change to positive by ≈ 0.39, holding OIL constant at the observed mean. We can then make the same statement about OIL: going from 0 to 1 coincides with an estimated change in the log odds of going from negative to no change and no change to positive by $\approx$ -0.19, holding REG constant at the observed mean. Notice in Figure 1 that the cutpoints are the values in a latent continuous space in which we transition from negative to no change, or no change to positive. The coefficients for our covariates are now interpretable as linear associations across categories (i.e., a 1 unit increase in x leads to a $\hat{\beta}$ shift in the log odds of going to the next category, rather than expressing a shift in the log odds of going from the baseline to another outcome category). Based on the model in part #1, we should be skeptical of forcing an ordered structure on the outcome (since the covariates seem to be associated with any change, not any "ordered" change).

Figure 1: Categorical cutpoints along latent dimension.



Question 2

Consider the data set ***MexicoMuniData.csv****, which includes municipal-level information from Mexico. The outcome of interest is the number of times the winning PAN presidential candidate in 2006 (****PAN.visits.06****) visited a district leading up to the 2009 federal elections, which is a count. Our main predictor of interest is whether the district was highly contested, or whether it was not (the PAN or their opponents have electoral security) in the previous federal elections during 2000 (****competitive.district****), which is binary (1=close/swing district, 0="safe seat"). We also include* ***marginality.06*** *(a measure of poverty) and* ***PAN.governor.06*** *(a dummy for whether the state has a PAN-affiliated governor) as additional control variables.*

(a) *Run a Poisson regression because the outcome is a count variable. Is there evidence that PAN presidential candidates visit swing districts more? Provide a test statistic and p-value.*

```
# load data
mexico_elections <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsII_Spring2023/main/datasets/MexicoMuniData.csv")
mexico_elections <- mexico_elections[!is.na(mexico_elections$pan.vote.09) & mexico_elections$pan.vote.09 < 1, ]
mexico_elections$competitive.district <- ifelse(.45 < mexico_elections$pan.vote.09 & mexico_elections$pan.vote.09 <.55, 1, 0)
# estimate poisson model
mex_poisson <- glm(PAN.visits.06 ~ competitive.district + marginality.06 + PAN.governor.06, data = mexico_elections, family=poisson)
summary(mex_poisson)
```

```
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)          -3.9304     0.1747 -22.503   <2e-16 ***
competitive.district -0.4594     0.3276  -1.402    0.161
marginality.06       -2.0981     0.1210 -17.343   <2e-16 ***
PAN.governor.06      -0.2073     0.1660  -1.249    0.212
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 1433.83  on 2392  degrees of freedom
Residual deviance:  963.57  on 2389  degrees of freedom
AIC: 1255.9
Number of Fisher Scoring iterations: 7
```

We can see above that we do not have a large enough test statistic (-1.4) to reject the null hypothesis that the estimated association between competitive.district and the number of visits from the winning PAN presidential candidate in 2006 is zero (p-value $= 0.16 > 0.05$).

(b) *Interpret the* ***marginality.06*** *and* ***PAN.governor.06*** *coefficients.*

We can interpret the estimated coefficient of marginality.06 as: holding PAN governor and competitive.district constant at their observed means, a one unit increase in poverty leads to a change in the expected number of visits by a multiplicative factor of ≈ 0.122 (exp(-2.0981)). The same can be said of PAN governor: having a PAN governor versus having a non-PAN governor changes the expected number of visits by a multiplicative factor of ≈ 0.813 (exp(-0.2073)).

(c) *Provide the estimated mean number of visits from the winning PAN presidential candidate for a hypothetical district that was competitive (*$\mathbf{competitive.district}=1$*), had an average poverty level (*$\mathbf{marginality.06} = 0$*), and a PAN governor (*$\mathbf{PAN.governor.06}=1$*).*

```
# option #1: manually
exp(coef(mex_poisson)[1] + coef(mex_poisson)[2]*1 + coef(mex_poisson)[3]*
    0 + coef(mex_poisson)[4]*1)

# option #2: predict() function
predict(mex_poisson, newdata=data.frame(competitive.district=1,
    marginality.06 = 0, PAN.governor.06=1), type="response")
```

```
0.01008022
```

The expected number of visits from the winning PAN candidate is ≈ 0.01 given that a district is competitive, had an average poverty level, and a PAN governor.