Problem Set 6

QTM 200: Applied Regression Analysis

Due: May 1, 2020

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 the course GitHub page in `.pdf` form.
- This problem set is due before midnight on Friday, May 1, 2020. No late assignments will be accepted.
- Total available points for this homework is 100.

Question 1 (50 points): Biology

Load in the data labelled `cholesterol.csv` on GitHub, which contains an observational study of 315 observations.

- Response variable:
  - `cholCat`: 1 if the individual has high cholesterol; 0 if the individual does not have high cholesterol
- Explanatory variables:
  - `sex`: 1 Male; 0 Female
  - `fat`: grams of fat consumed per day

Please answer the following questions:

1. We are interested in predicting the cholesterol category based on sex and fat intake.

   (a) Fit an additive model. Provide the summary output, the global null hypothesis, and p-value. Please describe the results and provide a conclusion.

```
###Part 1 Question 1###
#Part A#

fit <- glm(cholCat ~ sex+fat, data = cholesterol, family=gaussian())
summary(fit) #has a p-value of .02
plot(fit)

table(cholesterol$sex, cholesterol$cholCat)
```

   The obtained p-value is 2e-16, meaning it is statistically significant. The estimates for the intercept, 'sex' variable, and 'fat' variable are -.13, .19, and 0.008. We see that at a baseline of 0, meaning the individual is female and also 0 grams of fat consumed per day, they would not have high cholesterol (arguably a negative value of cholesterol measurement). However, as we increase sex (or go towards male which is sex=1), the estimate for cholesterol category goes up .18. Furthermore, if we increase 1 gram of fat intake per day, there is an increase of .008 towards the cholesterol category.

2. If explanatory variables are significant in this model, then

   (a) For women, how does increasing their fat intake by 1 gram per day change their odds on being in the high cholesterol group? (Interpretation of a coefficient)
   Only increasing their odds by .008 for 1 gram per day intake, in begin in the high cholesterol group.

   (b) For men, how does increasing their fat intake by 1 gram per day change their odds on being in the high cholesterol group? (Interpretation of a coefficient)
   For men however, you must add 0.19 in addition to the .008 for 1 gram per day intake. This is due to the observed variable coefficient for 'sex'.

   (c) What is the estimated probability of a woman with a fat intake of 100 grams per day being in the high cholesterol group?

```
###PArt 1 Question 2###
#Part c#
-.1303597 + (.1894160*0) + (0.0082466*100)
```

$$EstimatedProbability = -0.013 + 0.189(sex) + 0.008(fat) \tag{1}$$

Estimated Probability = 0.6943003

(d) Would the answers to 2a and 2b potentially change if we included the interaction term in this model? Why?

There would potentially be a change due to the interaction between female and male odds of the high cholesterol group. We see in the initial data in 1 that depending on if they have high cholesterol or does not with the variable cholCat, there is a significant change.

Question 2 (50 points): Political Economy

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.

- Response variable:
  - `GDPWdiff`: Difference in GDP between year t and $t-1$. Possible categories include: "positive", "negative", or "no change"
- Explanatory variables:
  - `REG`: 1=Democracy; 0=Non-Democracy
  - `OIL`: 1=if the average ratio of fuel exports to total exports in 1984-86exceeded 50%; 0= otherwise
  - `EDT`: Cumulative years of education of the average member of the labor force

Please answer the following questions:

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.

```
###Part 2 Question 1###
library(nnet)
mydata = gdpChange
mydata$GDPWdiff2 = relevel(mydata$GDPWdiff, ref = "no change")

multil = multinom(GDPWdiff2 ~ REG + OIL, data=mydata)
summary(multil)
coef(summary(multil))
```

   Keeping all variables constant, when looking at REG, that a one unit increase in REG (which at REG=1, it is categorized as a Democracy) can lead to a GDPWdiff of 1.379 in the negative category. With that in mind, when compared to the middle range of "no change" ([1.379+1.769]/2) that is 1.574, REG negative is lower by .195. As for an increase of one unit in REG can lead to a GDPWdiff of 1.769 in the positive category. That is the GDPWdiff can go as change as high as 1.769 when increasing REG by one unit (or changing to a Democracy according to the variable outputs).

Interestingly, OIL has flipped increments, as an increase in 1 unit is a lower bound for the positive GDPWdiff compared to the negative GDPWdiff. As OIL observes the average ratio of fuel exports to total exports exceeding 50%, this may show that a countries ability to export other materials more so than oil, can lead to more GDP-Wdiff.

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

```
###Part 2 Question 2###
library(MASS)
ordinal = polr(GDPWdiff ~ REG + OIL, data=mydata)
summary(ordinal)
coef(summary(ordinal))
```

The output illustrates that based on REG that a one unit increase leads to .398 increase in GDPWdiff. Based on OIL however, a one unit increase leads to a (-)0.199 decrease in GDPWdiff. Those coefficients do not change in relation to the changes in GDPWdiff range of negative to no change, where it is -0.731, or no change to positive where it is -0.710.