

Problem Set 6

QTM 200: Applied Regression Analysis

Due: May 6, 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 Wednesday, May 6, 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.

```
1 ##1)
2 AdditiveModel <- lm(cholCat ~ sex + fat, data=cholesterol)
3 summary(AdditiveModel)
4 confint(AdditiveModel, level=0.95)
```

H0: there is never a relationship between cholesterol ~ sex + fat

H1: The relationship does exist

P value is smaller than 2.2e-16

The probability that there is never going to be a relationship between response variable and explanatory variable is very low

There is a relationship between cholesterol and sex +fat

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

$$\text{cholesterol} = -0.1303597 + 0.1894 * \text{sex} + 0.0082466 * \text{fat}$$

- (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)

For females, when their fat increases by 1 gram, the probability that they will be in the high cholesterol group will increase 0.8 percent

- (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 males, when their fat increases by 1 gram, the probability that they will be in the high cholesterol group will increase 18.8 percent

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

$$\text{cholesterol} = -0.1303597 + 0.1894 * 0 + 0.0082466 * 100 = 0.6943003$$

The probability of a woman with a fat intake of 100 grams per day being in the high cholesterol group is 69.43 percent

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

- Perform a test to see if including an interaction is appropriate.

```
1 ##d)
2 InteractionModel <- lm(cholCat ~ sex + fat+ sex*fat, data=
  cholesterol)
3 summary(InteractionModel)
```

No, by incorporating a new interaction term, the coefficients for sex and fat are not changed and the explanatory power of my regression model is not increased compared with my model without the interaction term. Thus, there is no need to incorporate an interactive term.

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 for which 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 1984-86 exceeded 50%; 0= otherwise

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.

```
1 gdpChange <- read.csv("https://raw.githubusercontent.com/jeffreyziegler/
  QTM200Spring2020/master/problem_sets/PS6/gdpChange.csv")
2
3 str(gdpChange)
4
5
6 gdpChange$out<-relevel(gdpChange$GDPWdiff, ref="no change")
7
8 install.packages("nnet")
9 library(nnet)
10
11 ##1)
12 myModel<-multinom(out~REG+OIL, data=gdpChange, Hess = TRUE)
13 summary(myModel)
```

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

```

1 ##2)
2 install.packages("MASS")
3 library(MASS)
4 gdpChange$out<-ordered(gdpChange$out, levels=c("negative", "no change", "
    positive"))
5 polr(out ~ REG+OIL, gdpChange)

```