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.

```
cholesterol <- read.csv("cholesterol.csv", stringsAsFactors = F, header=T
    )
model1 <- lm(cholCat~sex+fat, data=cholesterol)
summary(model1)
```

   The null hypothesis is that neither sex nor fat significantly predict if an individual has high cholesterol. The p-value is 2.2e-16. Thus we reject the null hypothesis that neither sex nor fat are not significant predictors of if an individual has high cholesterol. At least one of these variables significantly predicts if an individual has high cholesterol.

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)

```
exp(0.0082466)
```

   For women, increasing their fat intake by 1 gram per day decreases their odds on being in the high cholesterol group by a multiplicative factor of 1.008281, or about 0.008

   (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)

```
exp(0.1894160 + 0.0082466)
```

   For men, increasing their fat intake by 1 gram per day increases their odds of being in the high cholesterol group by a multiplicative factor of 1.218551, or about 22

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

```
1/(1+exp(-(-0.1303597 + 0.1894160*0 + 0.0082466*100)))
```

The estimated probability of a woman with a fat intake of 100 grams per day being in the high cholesterol group is 0.6669229.

(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.

```
cholesterol_reg1 <-lm(cholCat~sex_d1 + sex_d2 + fat, data=cholesterol
    )
cholesterol_reg2 <- lm(cholCat~sex_d1 + sex_d2 + fat + fat:sex_d1 +
    fat:sex_d2, data=cholesterol)
anova(cholesterol_reg1, cholesterol_reg2)
```

The answers to 2a and 2b would potentially change if we included the interaction term because the model would be taking into account the effect of sex on fat consumed per day. However, the partial F-test for interaction was not significant, so including an interaction term is not appropriate.

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.

```r
#change reference group for DV, adjust variable type
gdpChange <- read.csv("gdpChange.csv", stringsAsFactors = F, header=T)
gdpChange$GDPWdiff.f <- factor(gdpChange$GDPWdiff,  labels = c("positive", "negative", "no change"))
is.factor(gdpChange$GDPWdiff.f)
gdpChange$GDPWdiff.fr <- relevel(gdpChange$GDPWdiff.f, ref = "no change")
#adjust variable types for IVs
is.numeric(gdpChange$EDT)
gdpChange$EDT.n <- as.numeric(gdpChange$EDT)
is.numeric(gdpChange$EDT.n)

#question 2 part 1
library(nnet)
unordered_logit <- multinom(GDPWdiff.fr~REG + OIL + EDT.n, data=gdpChange)
summary(unordered_logit)
exp(coef(unordered_logit)[1:4])
exp(cbind(OR = coef(unordered_logit), confint(unordered_logit)))
```

The difference in GDP increased for those whose fuel exports exceeded 50

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

```
#question 2 part 2
library(MASS)
ordered_logit <- polr(GDPWdiff.fr ~ REG + OIL + EDT.n, data=gdpChange,
    Hess=T)
summary(ordered_logit)
exp(coef(ordered_logit)[1:3])
exp(cbind(OR = coef(ordered_logit), confint(ordered_logit)))
```

There seems to be a decrease in GDP for those who are in a democracy by 0.8 times as well as those with fewer years of education by 0.95 times.