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.

```

lm.chol <- lm(cholCat ~ sex + fat, data=cholesterol)
summary(lm.chol)

# Yi = -0.13 + 0.189(sexXi) + 0.0082(fatXi)

# The global null hypothesis would be that the slope of the regression line
# is equal to zero (meaning that there is no significant linear relationship
# between sex and fat intake and high cholesterol).

# The p-value is <2.2e-16, which is much less than 0.05. This indicates that
# we can reject the null hypothesis and conclude that the slope is not equal
# to zero. There is support for a significant linear relationship between high
# cholesterol and sex and fat intake.
```

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)

   (b)

```
#     A 1-gram increase in fat intake would result in a 0.0082 percent increase
#     in the likelihood of a woman having high cholesterol.
```

   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)

   (c)

```
#     A 1-gram increase in fat intake would result in a 0.1972 percent increase
#     in the likelihood of a man having high cholesterol.
```

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

   (d)

```
-.13 + 0.189*0 + 0.0082*100
#     A woman with a fat intake of 100 grams per day is estimated to have a 69%
#     likelihood of high cholesterol.
```

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.

```r
# Make a new model with an interaction term.
lm.chol.2 <- lm(cholCat ~ sex*fat, data=cholesterol)
summary(lm.chol.2)

# The p-value for the interaction term is 0.2715. This is less than 0.05, indicating
# that it is significant and that there is an interaction relationship between sex
# and fat intake. This suggests that the answers to 2a and 2b could change with a new
# model that includes the interaction 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 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.

```
gdpChange <- read.csv(”gdpChange.csv”)
```

```
class(gdpChange$GDPWdiff)

gdpChange$GDPWdiff2 <- relevel(gdpChange$GDPWdiff, ref = ”no change”)

multi.gdp <- multinom(GDPWdiff2 ˜ REG + OIL, data=gdpChange)
summary(multi.gdp)

# The coefficient for REG is higher for ”positive”, indicating that if a country is
# a democracy, they are more slightly more likely to have a positive change in GDP.
# Meanwhile the coefficients for OIL indicate that if  oil export ratios exceeded 50%, countries
# are slightly more likely to have a negative difference in GDP.
```

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

```r
ord.gdp <- polr(GDPWdiff ~ REG + OIL, data=gdpChange, Hess=TRUE)
summary(ord.gdp)

# For democratic countries, the odds of having a negative difference in GDP is 0.41 points lower
# than countries that are non-democratic.
# For countries with an oil export ratio above 50%, the odds of having a negative difference in
# GDP is 0.179 points higher than countries with a ratio below 50%.
```