Chloe Sutter: 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.

```
# load data
chol <- (data=cholesterol)
table(chol)
```

- 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.

```
# fit an additive model
model1 <- lm(cholCat ~ sex + fat, chol)
model1
# Coefficients:
 # (Intercept)          sex          fat
 #   -0.130360     0.189416     0.008247
# global null hypothesis: B=0
# find p value, install car

anova(model1)
# Df Sum Sq Mean Sq F value    Pr(>F)
# sex          1  4.454  4.4543  27.516 2.882e-07 ***
 # fat          1 23.502 23.5017 145.177 < 2.2e-16 ***
 # Residuals 312 50.507  0.1619
```

   Since the adjusted p value for the model is substantially smaller than 0.05, we would reject the null hypothesis and conclude that sex and fat do have an effect on 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) Answer: For women, increasing fat intake by 1g/day increases the log odds of cholesterol by .189

   (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( 0.189416)
```

   Answer: For men, increasing fat intake by 1g/day increases the log odds of cholesterol by 1.208

   (c) What is the estimated probability of a woman with a fat intake of 100 grams per day being in the high cholesterol group? Answer: I would estimate that a woman with a fat intake of 100 grams per day is highly likely to be in the high cholesterol group.

   (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.

```
model2 <- lm(cholCat ~ sex + fat +sex:fat, chol)
model2
#(Intercept)          sex          fat      sex:fat
#-0.152463     0.391109     0.008544    -0.002210

anova(model2)
#Response: cholCat
#Df Sum Sq Mean Sq  F value    Pr(>F)
#sex          1  4.454  4.4543  27.5346 2.861e-07 ***
 # fat          1 23.502 23.5017 145.2762 < 2.2e-16 ***
 # sex:fat      1  0.196  0.1963   1.2133    0.2715
# $Residuals 311 50.311  0.1618
```

Answer: Yes, the answers would potentially change, because it is looking at the direct interaction between fat intake and sex on cholesterol. Changing sex from female to male also increases the log odds of high cholesterol.

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 1984-86 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.

```
#import gdpChange
gdpChange1
summary(gdpChange1)

#download nnet package
multinom_model1 <- multinom(GDPWdiff~REG+OIL, data=gdpChange1)
# weights:  12 (6 variable)
# initial  value 4087.936326
# iter  10 value 2339.387349
# final  value 2339.363928
# converged
multinom(formula=GDPWdiff~REG+OIL, data=gdpChange1)
#Coefficients:
#(Intercept)       REG        OIL
#no change  -3.8011902 -1.351703 -7.9240683
#positive    0.7284081  0.389905 -0.2076511

#Residual Deviance: 4678.728
#AIC: 4690.728
# for every one unit increase in the log-odds of
exp(coef(multinom_model1)[,c(1:3)])
```

```
#(Intercept)        REG           OIL
#no change  0.02234416 0.2587991 0.0003619269
#positive   2.07177984 1.4768404 0.8124904479
```

When there is a one unit increase in the baseline odds, where there is no change in GDPWdiff for a democracy, the oil ratio of fuel exports will increase by .03%

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

```
#import mass package
as.factor(gdpChange1$GDPWdiff)
as.ordered(gdpChange1$GDPWdiff)
ordered_logit <- polr(as.factor(gdpChange1$GDPWdiff)~REG+OIL, data=
    gdpChange1, Hess=T)
ordered_logit
#Coefficients:
#REG         OIL
#0.3984834 -0.1987177

#Intercepts:
 # negative|no change no change|positive
#-0.7311784          -0.7104851

#Residual Deviance: 4687.689
#AIC: 4695.689

#get odds ratios and CI's
exp(cbind(OR=coef(ordered_logit), confint(ordered_logit)))
# OR      2.5 %    97.5 %
#REG 1.4895639 1.2861520 1.727083
#OIL 0.8197813 0.6545844 1.030656
```

For democracy, the odds of a difference in GDP between years is 1.4 x higher than in a non-democracy, and if the average ratio of fuel exports to total exports in 1984-1986 exceeded 50%, the odds of GDP change are .81x higher.