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.

```r
cholestrol <- read.csv("cholesterol.csv") # Importing cholestrol dataset

# 1a)
lm1 <- lm(cholCat ~ sex + fat, data = cholestrol) # Additive model for predicting cholestrol category based on sex and fat intake
lm1 # Equation: Yi = -0.130 + 0.189 * sex + 0.008 * fat
summary(lm1) # Summary output
```

output.png

```
Call:
lm(formula = cholCat ~ sex + fat, data = cholestrol)

Residuals:
     Min       1Q   Median       3Q      Max
-0.99118 -0.32926 -0.09813  0.34817  0.83678

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.1303597  0.0564689  -2.309  0.02162 *
sex          0.1894160  0.0680041   2.785  0.00567 **
fat          0.0082466  0.0006844  12.049  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4023 on 312 degrees of freedom
Multiple R-squared:  0.3563,    Adjusted R-squared:  0.3522
F-statistic: 86.35 on 2 and 312 DF,  p-value: < 2.2e-16
```

Figure 1: Summary Output

The summary of the linear model shows that the global null hypothesis is rejected, as the individual null hypotheses for B0 = 0, B1 = 0, B2 = 0 are all rejected since the p-value is less than the alpha value of 0.05. According to the linear model, males are more likely to have higher cholestrol since the coefficient attached to the sex variable is 0.189. Moreover, higher fat intake contributes to higher cholestrol since the coefficient attached to the fat variable is 0.008. Therefore, an individual's sex and fat intake are related to the cholestrol category of the individual.

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)

Based on the coefficient, increasing fat intake by 1 gram per day for women increases their odds of being in the high cholestrol group by 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)

Based on the coefficient, increasing fat intake by 1 gram per day for men increases their odds of being in the high cholestrol group by 0.189 + 0.008 = 0.197.

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

The estimated probability of a woman with a fat intake of 100 grams per day being in the high cholestrol group is -0.130 + 0.189 * 0 + 0.008 * 100 = 0.67.

(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.

```
lm2 <- lm(cholCat ~ sex + fat + sex:fat, data = cholestrol) #  model
    including the interaction term
lm2 # Equation: Yi = -0.152 + 0.391 * sex + 0.008544 * fat + 0.002 *
    sex * fat
summary(lm2) # Summary output
```

Based on the output, the p-value of the coefficient for the interaction term is 0.272. Since this is greater than 0.05, we fail to reject the null hypothesis that the coefficient is 0, meaning that there is no significant interaction between the two explanatory variables that the answers to 2a and 2b change drastically

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.

```
GDP <- read.csv("gdpChange.csv", stringsAsFactors = F) # Importing
    dataset
GDP$GDPWdiff <- as.factor(GDP$GDPWdiff) # Changing to factor variable
GDP$GDPWdiff <- relevel(GDP$GDPWdiff, ref = "no change")
multi_model1 <- multinom(GDPWdiff ~ REG + OIL, data = GDP) # Constructing
    unordered multinomial logit
summary(multi_model1) # Summary
exp(coef(multi_model1)[,c(1:3)]) # Exponentiate coefficients
```

   The coefficient REG for countries with a GDP decrease is 3.972, while the coefficient is 5.865 for countries with GDP growth. Therefore, democratic countries are more likely to have GDP growth than undemocratic ones. On the other hand, coefficient OIL for countries with a GDP decrease is 119.578, while the coefficient is 97.156 for countries with positive GDP growth. Therefore, countries that export on average more than half of their exports in fuel are more likely to have a decrease in GDP

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

```
multi_model2 <- polr(GDPWdiff ~ REG + OIL, data = GDP, Hess = T) #
    Constructing ordered multinomial logit
summary(multi_model2) # Summary
exp(cbind(OR = coef(multi_model2), confint(multi_model2))) # Odds ratios
    and CI
```

For countries that are democratic, the odds of having positive GDP growth is 1.507 times greater than for countries that are not democratic. For countries that export more than 50% of total export in oil, the odds of having positive GDP growth is 0.836 times greater than for countries that export less than 50%.