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.

```
binom_cholesterol<-glm(cholCat~sex+fat, data = cholesterol,family = binomial(link = "logit"))
summary(binom_cholesterol)
Deviance Residuals:
  Min        1Q     Median       3Q       Max
-2.89662  -0.73093   0.07127   0.64186   2.23806

Coefficients:
  Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.759162   0.563834  -8.441   <2e-16 ***
  sex          1.356750   0.552130   2.457    0.014 *
  fat          0.065729   0.007826   8.399   <2e-16 ***
  ---
  Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1       1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 435.54  on 314  degrees of freedom
Residual deviance: 279.58  on 312  degrees of freedom
AIC: 285.58

Number of Fisher Scoring iterations: 5


#Ho: Neither sex nor fat are associated with wheter or not individual has high cholesterol.
#Ha: At least one of these variables (sex and fat) are associated with wheter or not individual has high cholesterol.
# The p value is 0.014 for sex and <2e-16 for fat, therefore we reject the null hypothesis that neither sex nor fat are useful predictors for whether or not an individual has high cholesterol. There is evidence to suggest that at least one of the variables (sex and fat) is a useful predictor of 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)

```
#For women, increasing fat by 1 gram increases the log odds of having high cholesterol by 0.065729.
```

(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 men, increasing fat by 1 gram increases the log odds of having high cholesterol by 1.422479.
```

(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(-(-4.759162+0.065729*100)))
#SOLUTION: 0.859813
```

(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.

```
#Yes, because increasing fat intake may affect men and women differently, so adding an interaction term could show how the relationship between fat intake and hich cholesterol differs by sex.
binom_cholesterol_multiplicative<-glm(cholCat~sex*fat, data = cholesterol,family = binomial(link = "logit"))
summary(binom_cholesterol_multiplicative)
 Call:
  glm(formula = cholCat ~ sex * fat, family = binomial(link = "logit"),
      data = cholesterol)

 Deviance Residuals:
  Min        1Q    Median       3Q       Max
 -2.86893  -0.72131  0.06984  0.65091  2.22120

 Coefficients:
  Estimate Std. Error z value Pr(>|z|)
 (Intercept) -4.674853  0.587978  -7.951 1.85e-15 ***
  sex         0.541829  1.924729  0.282   0.778
 fat         0.064513  0.008187  7.880 3.28e-15 ***
  sex:fat     0.012351  0.028011  0.441   0.659
---
  Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1    1

 (Dispersion parameter for binomial family taken to be 1)

 Null deviance: 435.54  on 314  degrees of freedom
 Residual deviance: 279.37  on 311  degrees of freedom
 AIC: 287.37

Number of Fisher Scoring iterations: 6

#Since the p value for the interaction between sex and fat is >0.05, fail to reject the null hypothesis that there is no association
```

```
    between sex:fat and high cholesterol. Adding this interaction 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 1984-86 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.

```
#Make ”GDWPdiff” a factor variable with the levels ”no change”, ”positive”, and ”negative.” Set ”no change” as the reference category.
gdpChange<-read.csv(”gdpChange.csv”)
gdpChange1<- gdpChange
gdpChange1$GDPWdiff<- factor(gdpChange1$GDPWdiff, levels = c(”no change”,
    ”negative”, ”positive”))
gdpChange1$GDPWdiff<- relevel(gdpChange1$GDPWdiff, ref = ”no change”)

library(nnet)
multinom_GDPWdiff<- multinom(GDPWdiff~REG+OIL, data = gdpChange1)
summary(multinom_GDPWdiff)

multinom(formula = GDPWdiff ~ REG + OIL, data = gdpChange1)

Coefficients:
  (Intercept)      REG      OIL
negative    3.805370 1.379282 4.783968
positive    4.533759 1.769007 4.576321

Std. Errors:
  (Intercept)       REG       OIL
```

```
negative   0.2706832 0.7686958 6.885366
positive   0.2692006 0.7670366 6.885097

Residual Deviance: 4678.77
AIC: 4690.77

#Being a democracy and having >50% oil exports increases the baseline
    odds that the GDPWdiff is positive relative to "no change".Being a
    democracy and having >50% oil exports also increases the baseline odds
     that the GDPWdiff is negative relative to "no change".The odds are
    highest for the positive relationship between democracatic regime and
    the negative relationship between >50% oil exports.
```

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

```
install.packages("MASS")
library(MASS)

ordered_GDPWdiff<- polr(GDPWdiff~REG+OIL, data = gdpChange1, Hess = T)
summary(ordered_GDPWdiff)
polr(formula = GDPWdiff ~ REG + OIL, data = gdpChange1, Hess = T)

Coefficients:
  Value Std. Error t value
REG  0.4102    0.07518   5.456
OIL -0.1788    0.11546  -1.549

Intercepts:
  Value    Std. Error t value
no change|negative  -5.3199   0.2523   -21.0865
negative|positive   -0.7036   0.0476   -14.7932

Residual Deviance: 4686.606

#Being a democracy increases odds of positive GDP growth, while having
    >50% oil exports decreases odds of positive GDP growth.
```