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.

```
chol<-read.csv(”cholesterol.csv”)
```

- 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.

   ```
   model11a<-glm(cholCat~fat+sex, data=chol)
   summary(model11a)
   ```

   ```
   Call:
   glm(formula = cholCat ~ fat + sex, data = chol)

   Deviance 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 *
   fat          0.0082466  0.0006844  12.049  < 2e-16 ***
   sex          0.1894160  0.0680041   2.785  0.00567 **
   ---
   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

   (Dispersion parameter for gaussian family taken to be 0.161883)

   Null deviance: 78.463  on 314  degrees of freedom
   Residual deviance: 50.507  on 312  degrees of freedom
   AIC: 325.34

   Number of Fisher Scoring iterations: 2
   ```

   The global null hypothesis is $\beta sex = \beta fat = 0$. This means that is no significant associatation between the explanatory variables (fat and sex) and the response variable (cholesterol). Both p-values are: 1) less than 2e-16 2) 0.00567. Both are less than the standard 0.05 significance level and we reject the null hypothesis. This means that at least one of the variables, sex or fat, are associated with 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 their fat intake by 1 gram per day changes their odds of being in the high cholesterol group by 0.0082466 units.

(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) Answer: For men, increasing their fat intake by 1 gram per day changes their odds of being in the high cholesterol group by 0.1976626 units (0.0082466+0.1894160).

(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: The estimated probability of a woman with a fat intake of 100 grams per day being in the high cholesterol group is 0.6943003, which is calculated by -0.1303597+0.0082466(100)+0.1894160(0).

(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.

```
model12d<−glm(cholCat˜fat∗sex, data=chol, family=binomial(link=”logit”))
summary(model12d)
```

```
Call:
glm(formula = cholCat ~ fat * sex, family = binomial(link = "logit"),
data = chol)

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 ***
fat          0.064513   0.008187   7.880 3.28e-15 ***
sex          0.541829   1.924729   0.282    0.778
fat:sex      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
```

The answers to 2a and 2b would probably not change if we included the interaction term in this model. This is because the p-value (0.659) is not significant.

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.

```
gdpChange1<-gdpChange
gdpChange1$GDPWdiff<-gsub(”no change”,”constant”,gdpChange1$GDPWdiff)
model21<-multinom(GDPWdiff~REG+OIL, data=gdpChange1)
summary(model21)
```

Interpretation:

```
Call:
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
```

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

```
ordered_model22<-polr(GDPWdiff~REG+OIL, data=gdpChange, Hess=TRUE)
summary(ordered_model22)
```

```
Call:
polr(formula = GDPWdiff ~ REG + OIL, data = gdpChange, Hess = TRUE)

Coefficients:
Value Std. Error t value
REG  0.3985    0.07518   5.300
OIL -0.1987    0.11572  -1.717

Intercepts:
Value    Std. Error t value
negative|no change  -0.7312   0.0476   -15.3597
no change|positive  -0.7105   0.0475   -14.9554

Residual Deviance: 4687.689
AIC: 4695.689
```