Problem Set 6

QTM 200: Applied Regression Analysis

Due: May 1, 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 Friday, May 1, 2020. No late assignments will be accepted.
- Total available points for this homework is 100.

Question 1 (50 points): Biology

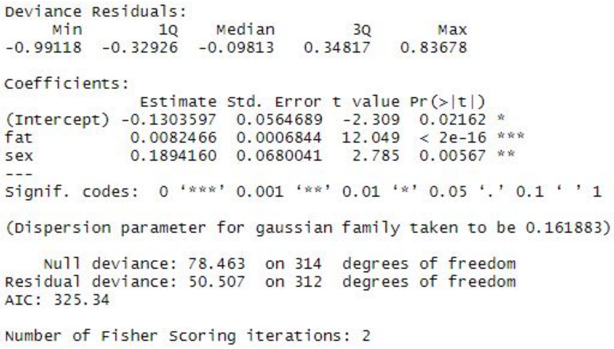
Load in the data labelled `cholesterol.csv` on GitHub, which contains an observational study of 315 observations.

- 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.



```
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
```

```r
#H0: the amount of fat consumed and one's sex have no effect on whether someone has cholesterol, Ha: the amount of fat consumed and one's sex do have an effect on whether or not someone has cholesterol

model1<-glm(cholCat~fat + sex, data=cholesterol)
summary(model1)

#Since both variables are significant with p-values less than .05 (and less than .01), we reject H0 in favor of Ha. It appears that both the amount of fat consumed and one's gender have an effect on whether someone has 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 their fat intake by 1 gram per day increases their odds of being in the high cholesterol group by about .8%
```

   (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)

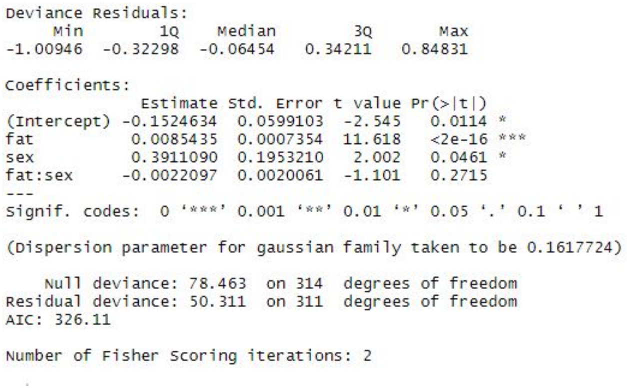
```
#For men, increasing their fat intake by 1 gram per day changes their odds on being in the high cholesterol group by about 19.7%
```

   (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 is about 80%
```

(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.



```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.00946  -0.32298  -0.06454   0.34211   0.84831

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.1524634  0.0599103  -2.545   0.0114 *
fat          0.0085435  0.0007354  11.618   <2e-16 ***
sex          0.3911090  0.1953210   2.002   0.0461 *
fat:sex     -0.0022097  0.0020061  -1.101   0.2715
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.1617724)

    Null deviance: 78.463  on 314  degrees of freedom
Residual deviance: 50.311  on 311  degrees of freedom
AIC: 326.11

Number of Fisher Scoring iterations: 2
```

```
model2<-glm(cholCat~fat + sex + fat:sex, data=cholesterol)
summary(model2)
#No, because the interaction term is not significant
```

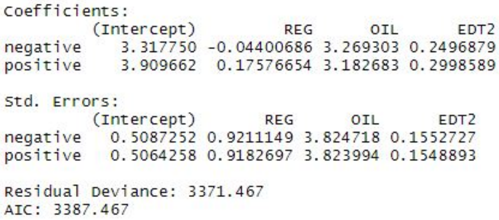
Question 2 (50 points): Political Economy

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 1984-86exceeded 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.



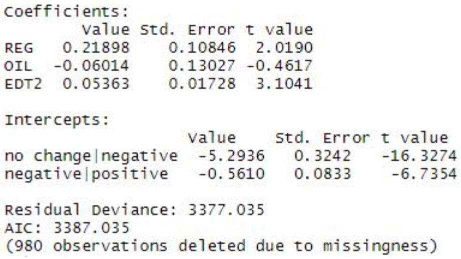
```
Coefficients:
         (Intercept)         REG      OIL      EDT2
negative    3.317750 -0.04400686 3.269303 0.2496879
positive    3.909662  0.17576654 3.182683 0.2998589

Std. Errors:
         (Intercept)       REG      OIL      EDT2
negative   0.5087252 0.9211149 3.824718 0.1552727
positive   0.5064258 0.9182697 3.823994 0.1548893

Residual Deviance: 3371.467
AIC: 3387.467
```

```
#create reference category and fix variable types
gdpChange$EDT2<-as.numeric(gdpChange$EDT)
gdpChange$GDPWdiff2<-as.factor(gdpChange$GDPWdiff)
gdpChange$GDPWdiff3 <- relevel(gdpChange$GDPWdiff2, ref = ”no change”)
#run regression
model3<-multinom(GDPWdiff3~REG + OIL + EDT2, data=gdpChange)
summary(model3)
#Interpretation: Not being a democracy increased the log odds of having
    negative growth vs. no change, as did exporting a lot of oil and
    having a more educated labor force. Meanwhile being a democracy
    increased the log odds of a country experiencing positive growth vs.
    no change, as did exporting a lot of oil (although not as much as with
    negative growth) and having a more educated workforce (more so this
   time). Since being a democracy was the only variable that is a
   different sign between positive and negative gdp growth, it is the
   only variable from this model that can be said to conclusively affect
   GDP growth.
```

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



```
Coefficients:
        Value Std. Error t value
REG   0.21898    0.10846  2.0190
OIL  -0.06014    0.13027 -0.4617
EDT2  0.05363    0.01728  3.1041

Intercepts:
                     Value    Std. Error t value
no change|negative  -5.2936   0.3242    -16.3274
negative|positive   -0.5610   0.0833     -6.7354

Residual Deviance: 3377.035
AIC: 3387.035
(980 observations deleted due to missingness)
```

```
model4<-polr(GDPWdiff3~REG + OIL + EDT2,data=gdpChange, Hess=T)
summary(model4)
#Interpretation: here we see that being a democracy, not being a major
   oil exporter, and having a more educated workforce all increase the
   odds of having positive GDP growth
```