Problem Set 6: QTM 200: Applied Regression Analysis

Farris Sabir

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.

```
binom_model1 <- glm(cholCat ~ sex + fat, data=chol, family=binomial(link="logit"))
summary(binom_model1)

#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
```

   The global null hypothesis is that: $beta_{sex} = beta_{fat} = 0$ meaning that neither explanatory variable is associated with the response variable.

```
binom_null1 <- glm(cholCat ~ 1, data=chol, family=binomial(link="logit"))

summary(binom_null1)

# Model 1: cholCat ~ 1
# Model 2: cholCat ~ sex + fat
# Resid Df Resid.  Dev Df    Deviance    Pr(>Chi)
# 1      314    435.536
# 2      312    279.5785   2   155.9575    <2e-16
```

   Because p-value ¡ 0.01, we can reject the global null hypothesis and conclude at least one predictor is reliable in the model.

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)

```
exp(0.065729)
```

For women, increasing their fat intake by 1 gram per day changes their odds on being in the high cholesterol group by a multiplicative factor of 1.067937 or increases the odds by 6.7937

(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)

As with women, for men, increasing their fat intake by 1 gram per day changes their odds on being in the high cholesterol group by a multiplicative factor of 1.067937 or increases the odds by 6.7937

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

```r
prediction1 = 1/(1 + exp(-(-4.759162 + 1.356750*0 + 0.065729*100)))
prediction1
```

The estimated probability of a woman with a fat intake of 100 grams per day being in the high cholesterol group is 85.9813

(d) Would the answers to 2a and 2b potentially change if we included the interaction term in this model? Why?

It could potentially change if for men and women, increase in fat intake predicate significantly different changes in odds of being in the high cholesterol group.

- Perform a test to see if including an interaction is appropriate.

```r
#We can determine this with a multiplicative model.
binom_model2 <- glm(cholCat ~ sex * fat, data=chol, family=binomial(link="logit"))
summary(binom_model2)
#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
```

Because the slope of interaction is not significantly different than zero, the change in odds by change in fat intake is not significantly different between men and women. In other words, an interaction is not necessary or appropriate to consider.

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.

```r
gdpchange <- read.csv("gdpChange.csv")
gdpchange_modified <- gdpchange
gdpchange_modified$GDPWdiff <- gsub("no change", "constant", gdpchange_modified$GDPWdiff)
```

- 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.

   ```r
   multinom_model1 <- multinom(GDPWdiff ~ REG + OIL, data = gdpchange_modified)
   summary(multinom_model1)
   #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
   ```

   After having constructed the unordered multinomial logit, the above summary provides enough information to interpret it, including estimated cutoffs and coefficients.

```
exp(coef(multinom_model1))
#Exponentiated Coefficients:
#          (Intercept)      REG        OIL
#negative     44.94186 3.972047 119.57794
#positive     93.10789 5.865024  97.15632
5.865024/3.972047 #for REG
97.15632/119.57794 #for OIL
```

The estimated cutoff points are 44.94186 for a negative difference in GDP and 93.10789 for a positive difference in GDP. Transforming from a non-democracy to democracy, the odds of GDPWdiff being more positive increases by multiplicative factor of 1.476575. Transforming from less than 50% average ratio of fuel expoerts to total exports to above 50%, the odds of GDWPdiff being more positive decreases by multiplicative factor of 0.8124937.

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

```
ordered_model1 <- polr(GDPWdiff ~ REG + OIL, data = gdpchange, Hess=T)
summary(ordered_model1)
#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
```

After having constructed the ordered multinomial logit, the above summary provides enough information to interpret it, including estimated cutoffs and coefficients.

```
exp(coef(ordered_model1))
#   REG        OIL
#1.4895639 0.8197813
```

Switching from a non-democracy to democracy, the odds of having a more positive or less negative change in GDP increases by factor 1.4895639 or increases by 48.96%. Switching from less than 50% average fuel to total exports ratio to above 50%, the odds of having a more positive or less negative change in GDP decreases by a factor of 0.8198.

```
exp(-0.7312) #negative to no change
exp(-0.7105) #no change to positive
```

If odds are below 0.481331, then predict switch to negative change in GDP. If odds are between 0.481331 and 0.4913984, then predict no change in GDP. If odds are above 0.4913984, then predict switch to positive change in GDP.