Problem Set 6

Vanessa Wong

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.

   - Fit an additive model. Provide the summary output, the global null hypothesis, and p-value. Please describe the results and provide a conclusion.

   - prediction equation: y = -4.759 + 1.357sex + 0.066fat

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

   - summary output

```
# summary output:
# 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
```

   - global null: $\text{beta}_{\text{fat}} = \text{beta}_{\text{sex}} = 0$

   - p <2.2e-16. p <0.01 therefore at least one beta is not equal to 0

```
rlnull <- glm(cholCat ~ 1, data=chol, family=binomial(link="logit"))
summary(rlnull)
anova(rlnull, reg_logit, test="Chisq")
```

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 intake by one gram is associated with an increase in the odds of being in the high cholesterol group by a multiplicative factor of 1.068 ($e^{0.0657}$)

(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 intake by one gram is associated with an increase in the odds of being in the high cholesterol group by a multiplicative factor of 1.068 ($e^{0.0657}$)

(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 cholesterol group is 86.3%

```
# B0 + B1X
(-4.759 + (0.066 * 100))
# = 1.841
1/(1+exp(-1.841))
# = 0.863
```

(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.

- p = 0.6454 >0.05. There is no evidence that including an interactive term of sex and fat intake would be a significant predictor of the odds of being in the high cholesterol group. Therefore, the answers to 2a and 2b would likely not change if the interactive term were included.

```
reg_logit <- glm(cholCat ~ sex + fat, data=chol, family=binomial(link
    ="logit"))
reg_logit2 <- glm(cholCat ~ sex * fat, data=chol, family=binomial(
    link="logit"))
anova(reg_logit, reg_logit2, test="Chisq")
```

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.

- In a given country, there is an increase in baseline odds of a positive GDP difference by 5.865 times when that country has a democratic government

- In a given country, there is an increase in baseline odds of a negative GDP difference by 3.972 times when that country has a democratic government

- In a given country, there is an increase in baseline odds of a positive GDP difference by 97.156 times when that country has a ratio of fuel to total exports greater than 50% in 1984-86

- In a given country, there is an increase in baseline odds of a negative GDP difference by 119.577 times when that country has a ratio of fuel to total exports greater than 50% in 1984-86

- cutoff points, coefficients

```
#########################################################################
# cutoff points, coefficients:
# multinom(formula = GDPWdiff_new ~ REG + OIL, data = gdp)

# Coefficients:
#       (Intercept)   REG      OIL
# 1 (+)    4.533759 1.769007 4.576321
# 2 (-)    3.805370 1.379282 4.783968

# Std. Errors:
#         (Intercept)  REG       OIL
# 1 (+)    0.2692006 0.7670366 6.885097
# 2 (-)   0.2706832 0.7686958 6.885366

# Residual Deviance: 4678.77
# AIC: 4690.77
#########################################################################
#> exp(coef(gdp_reg)[,c(1:3)])
#     (Intercept)     REG        OIL
# 1 (+)    93.10789 5.865024  97.15632
# 2 (-)    44.94186 3.972047 119.57794
#########################################################################
```

```
require(nnet)
# 1
gdp$GDPWdiff_new[gdp$GDPWdiff=="no change"] <- 0
gdp$GDPWdiff_new[gdp$GDPWdiff=="positive"] <- 1
gdp$GDPWdiff_new[gdp$GDPWdiff=="negative"] <- 2
gdp_reg <- multinom(GDPWdiff_new ~ REG + OIL, data=gdp)
summary(gdp_reg)
exp(coef(gdp_reg)[,c(1:3)])
```

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

- The odds of having a positive GDP change when switching from a non-democratic to a democratic government multiply by a factor of 1.489.

- The odds of having a positive GDP change, if that country has an average ratio of fuel:total exports that is greater than 50%, multiplies by a factor of 0.8197.

- coefficients:

```
exp(0.3984834)
exp(-0.1987177)
# coefficients (exp)
# REG : 1.489564
# OIL : 0.8197
```

- intercepts:

```
exp(-0.7311784)
exp(-0.7104851)
# intercepts (exp)
# negative|no change     no change|positive
#      0.4813414              0.4914058
```

```
ordered_logit <- polr(GDPWdiff ~ REG + OIL, data=gdp, Hess=T)
summary(ordered_logit)
ordered_logit
```