Math 150 - Methods in Biostatistics - Homework 5

your name here

Due: Wednesday, February 22, 2023

Assignment Summary (Goals)

- fluent use of the multiple logistic model for prediction and for coefficient interpretation
- working with variables that interact and variables which are multicollinear
- practice using `ggplot()` so that visualizations can inform the larger analysis

Note that if you don't know the R code either check my notes or ask me!!! Happy to scaffold, debug, send resources, etc. Don't go down a rabbit hole trying to figure out an R function or syntax.

Also, note that you'll need to get the data from Sakai and use it for this analysis. Look back to your own HW1 file to see the line of code **you** used to import the `games1.csv` dataset. Ask me if it isn't obvious to you after you look at your own HW1. And just like in HW4, you'll need to deal with the missing variables coded as "*".

Q1. Collaborative Learning Describe one thing you learned from someone (a fellow student or mentor) in our class this week (it could be: content, logistical help, background material, R information, etc.) 1-3 sentences.

Q2. Chp 7, E1 Bird Nest study The file `Birdnest` contains data for 99 species of North American passerine birds. Passerine are "perching birds" and include many families of familiar small birds (e.g., sparrows and warblers), as well as some larger species like crows and ravens, but do not include hawks, owls, water fowl, wading birds, and woodpeckers. One hypothesis of interest was about the relationship of body size to type of nest. Body size was measured as average length of the species. Although nests come in a variety of types (see the `Nesttype` variable), in this data set nest type was categorized into either closed or open. "Closed" refers to nests with only a small opening to the outside, such as the tree cavity nest of many woodpeckers or the pendant-style nest of an oriole. "Open" nests include the cup-shaped nest of the American robin. (Note: `Closed?` = 1 for closed nests; `Closed?` = 0 for open nests.)

```
birdnest <- read_csv("~/Dropbox/teaching/MA150/PracStatCD/Data Sets/Chapter 07/CSV Files/C7 Birdnest.cs
                     na="*")
glm(`Closed?` ~ Length, data=birdnest, family="binomial") %>% tidy()
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   0.457     0.753      0.607   0.544
## 2 Length       -0.0677    0.0425    -1.59    0.112
```

(a) Create a logistic regression model using bird length (`Length`) to estimate the probability that a bird species has a closed net type. Interpret the model in terms of the odds ratio.

(b) Use the Wald statistic to create a 95% confidence interval for the odds ratio. (Wald just means normal distribution, use Z. You can do it by hand using the standard output, or you can find the CI for the slope, using `tidy(conf.int = TRUE)`, and exponentiate. You could do it both ways and check to make sure you get the same answer!)

(c) Test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ using both Wald's test (from `tidy()` output) and the likelihood ratio test (from `glance()` output). State your conclusions based on these tests.

For the LRT in R, the first thing you'll need to do is pipe the glm into `glance()`. That is: `glm(...) %>% glance()`. The "deviance" value with the MLEs is called `deviance`, the "deviance" value with the null value of $\beta_1 = 0$ is called `null.deviance`. You have to use R as a calculator to subtract the values. Then use `pchisq()` to find the p-value. See page 223 in your book.

skip (d), (e), (f)

Q3. Chp 7, E9 (no (d)) Donner Party In 1846, a group of 87 people (called the Donner Party) were heading west from Springfield, Illinois, for California. The leaders attempted a new route through the Sierra Nevada and were stranded there throughout the winter. The harsh weather conditions and lack of food resulted in the death of many people within the group. Social scientists have used the data to study the theory that females are better able than men to survive harsh conditions.

(a) Create a logistic regression model using `Gender` and `Age` to estimate the probability of survival. Create a plot of the data plus the estimated probability of survival using `Age` as the explanatory variable and grouping the data by `Gender`. Use the plot and the model to interpret the coefficients in terms of the odds ratios.

```
donner <- read_csv("~/Dropbox/teaching/MA150/PracStatCD/Data Sets/Chapter 07/CSV Files/C7 Donner.csv",
                   na="*")

names(donner) <- c("name", "gender", "age", "survived", "familysize", "X6", "X7",
                   "X8", "X9", "adultname", "adultgender", "adultage",
                   "adultsurvived", "adultfamilysize")
```

The code will look something like this. Fill in the blanks. And run the code one line at a time so that you know exactly what each line is doing. [Recall: what is the difference between the output of `tidy()`, `glance()`, and `augment()`? That is, all are data frames. And they give output that is of dimension 1, p (p is the number of variables), and n (n is the number of observations). Which is which?]

```
glm(___ ~ ___ + ___, data = ___, family="___") %>%
  ___()


glm(___ ~ ___ + ___, data = ___, family="___") %>%
  ___(type.predict = "response") %>%
  arrange(age) %>%
  ggplot() +
  geom_point(aes(x = ___, y = ___)) +
  geom_line(aes(x = ___, y = .fitted, group = ___, color = as.factor(___)))
```

(b) Create and interpret a logistic regression model using `Gender`, `Age`, and `Gender*Age` to estimate the probability of survival. Plot the observations, add lines representing the estimated probability of survival using `Age` as the explanatory variable and grouping the data by `Gender`. [Code from above almost identical.]

(c) Explain any key differences between the plots created in parts (a) and (b). Discuss how adding the interaction term `Gender*Age` impacts the model.

Q4. Chp 7, E10 Variable Selection Techniques and Multicollinearity Wolberg and Mangasarian developed a technique to accurate diagnose breast masses using only visual characteristics of the cells within the tumor (PNAS 1990). A sample is placed on a slide, and characteristics of the cellular nuclei within the tumor, such as size, shape, and texture are examined under a microscope to determine with the cancer cells are benign or malignant. Benign tumors are scar tissue or abnormal growths that do not spread and

are typically harmless. Malignant (or invasive) cancer cells are cells that can travel, typically through the bloodstream or lymph nodes, and begin to replace normal cells in other parts of the body. If a tumor is malignant, it is essential to remove or destroy all cancerous cells in order to keep them from spreading. If a tumor is benign, surgery is not needed and the harmless tumor can remain.

(a) Crate a logistic regression model using `Radius`, `Concave`, and `Radius*Radius`, and `Radius*Concave` as explanatory variables to estimate the probability that a mass is malignant. Submit the logistic regression model and the likelihood ratio test results (testing whether any variables at all are significant), including the log-likelihood (or deviance) values and a statement of the null hypothesis. [Note that you need to create the `Radius*Radius` variable before running the `glm`.]

```
cancer <- read_csv("~/Dropbox/teaching/MA150/PracStatCD/Data Sets/Chapter 07/CSV Files/C7 Cancer2.csv",
                   na="*")
cancer <- cancer %>%
  mutate(Radius2 = Radius*Radius)

glm(`Malignant?` ~ Radius*Concave + Radius2, data=cancer, family="binomial") %>%
  tidy()
```

```
## # A tibble: 5 x 5
##   term           estimate std.error statistic p.value
##   <chr>             <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)      -9.36      7.78    -1.20    0.229
## 2 Radius            0.346     3.78     0.0915  0.927
## 3 Concave           6.54      3.03     2.16    0.0310
## 4 Radius2           0.351     0.465    0.757   0.449
## 5 Radius:Concave   -0.806     0.749   -1.08    0.282
```

(b) Even though in part (a) Wald's test shows the highest p-value for `Radius`, it is typically best to attempt to keep the simplest terms in the model. Generally, keeping simpler terms in the model makes the model easier to interpret. Thus, we suggest as a first attempt keeping `Radius` in the model and eliminating the variable with the next highest p-value. Create a logistic regression model using `Radius`, `Concave`, and `Radius*Concave` as explanatory variables to estimate the probability that a mass is malignant. Submit the logistic regression model and the likelihood ratio test results, including the log-likelihood (or deviance) values. Conduct the drop-in-deviance test (aka LRT) to determine if `Radius*Radius` should be included in the model, state the null hypothesis for your test. [Use `glm(...) %>% glance()` on models with and without the squared term.]

(c) Use a scatterplot to compare `Radius` to `Radius*Radius` and calculate the correlation between these two terms. Are the two variables highly correlated?

(d) Chapter 3 discusses **multicollinearity** (highly correlated explanatory variables). Explain whether you believe `Radius` is important in the logistic regression model. Why is the p-value for `Radius` so large in part (a) but very small

(e) Create a logistic regression model using `Radius` and `Concave` as explanatory variables to estimate the probability that a mass is malignant. Submit the logistic regression model and the likelihood ratio test results, including the log-likelihood (or deviance) values. Conduct the drop-in-deviance test to determine if `Radius*Concave` should be included in the model, state the null hypothesis for your test.

(f) Create a logistic regression model using only `Concave` as an explanatory variable to estimate the probability that a mass is malignant. Submit the logistic regression model and the likelihood ratio test results, including the log-likelihood (or deviance) values. Conduct the drop-in-deviance to test to determine if `Radius` should be included in the model.

(g) Submit a final model and provide a justification for choosing that model.

```
praise()
```

```
## [1] "You are groovy!"
```