Midterm exam, STAT 4/6C03

released 17 October 2018, due 24 October 2018

Rules:

- You may use any notes or books, but **not** web resources other than the course web site. Please do *not* speak (text, e-mail, etc.) with any one other than me about the exam. Please feel free to contact me for clarification.
- The exam is due in Dropbox by midnight on (i.e., at the end of) **Wednesday 24 October**. Please submit your solutions as a single `.R`, `.Rmd`, or `.Rnw` file, with descriptions and explanations as comments.
- Data are available from the course web page.
- When in doubt, show how you did something and explain your approach. (e.g., if I say "do X . . . ", I want you to include the code that you used.) Your solution should include *working* R code (points off for anything that doesn't work when I try it!) and an explanation of what you did.

1 Titanic

The classic data set on survivors of the Titanic.

Read in the `titanic_long_binary.csv` data set: the variables are

- `Class`: passenger class (1st/2d/3d) or crew
- `Sex`: male or female
- `Age`: age in years
- `survived`: whether a particular individual survived or not.

1. using `aggregate` from base R or `group_by` + `summarise` from the `dplyr` package, create a new data set that has `Class`, `Sex`, `Age` and two additional columns: `prop` (proportion survived) and `total` (total number in the category). You can compare your results to the `titanic_long.csv` data set (if you have trouble with this step, you can use `titanic_long` in the next part of the question).

2. using `ggplot2` and some combination of colour, point shape, facets, and y-axis position, plot the proportion survived as a function of all three explanatory variables (class, sex, age), in a single plot (multiple sub-plots/facets are OK).

3. *Going back to the original, disaggregated data set for this and following questions:* Define a set of custom contrasts for the `Class` variable that will define the parameters as β_0=overall average across classes; β_1=crew vs passengers (1st, 2nd, 3rd); β_2=1st vs (2nd and 3rd); β_3 = 2nd vs 3rd.

4. Fit a logistic regression including all of the two-way interactions, using sum-to-zero contrasts for all parameters.

5. What does the `Age1:Class3` coefficient mean? Why is it `NA`?

6. What do the `Age1:Class1` and `Age1:Class2` coefficients mean? Interpret the magnitude and sign of the coefficients.

7. Run `car::Anova` on the model with `test="LR"` and `test="Wald"`. Explain the meaning of these two kinds of tests. Which p-values differ (e.g. a difference between $p \ll 0.01$ and $p > 0.05$), and why? Which of these two sets of results should you trust more, and why?

8. Fit a logistic regression with the main effects of the three predictor variables only.

9. Compute the estimated odds ratio for female survival vs. average survival, and its 95% confidence intervals.

10. Compute the estimated probability of survival for a 1st-class passenger, and its 95% confidence intervals (use Wald intervals on the logit scale, then back-transform to the probability scale).

11. Based on the reduced (main effects only) model, interpret the meaning of each of the parameters in `summary()` (sign and statistical significance only; interpretation of the magnitude of the parameters is optional).

2 Income

The following question analyzes a data set on adult incomes from the UCI machine learning repository. Run the following R code to retrieve data on income categories in adults and simplify it:

```
## download.file("https://archive.ics.uci.edu/ml/machine-learning-databases/adul
library(dplyr)
adult <- read.csv("adult.csv",header=FALSE,strip.white=TRUE,
                  stringsAsFactors=FALSE)
```

```
nms <- c("age","workclass","fnlwgt","education","education.num",
  "marital.status","occupation","relationship","race","sex",
  "capital.gain","capital.loss","hours.per.week","native.country",
  "income")
names(adult) <- nms
adult2 <- (adult
    ## keep a subset of explanatory variables
    %>% select(age,education.num,marital.status,race,sex,
               native.country,income)
    ## US only
    %>% filter(native.country=="United-States")
    ## we don't need the native.country variable any more, drop it
    %>% select(-native.country)
    ## select only races with >500 observations
    %>% group_by(race)
    %>% filter(n()>500)
    ## select only marital status categories with >500 observations
    %>% group_by(marital.status)
    %>% filter(n()>1000)
    %>% ungroup()
    ## convert all character variables to factors
    %>% mutate_if(is.character,factor)
)
```

The data contain

- `age`: age in years
- `education.num`: number of years of education
- `marital status`: description of marital status
- `race`: Black or White
- `sex`: Female or Male
- `income`: less than or greater than US$50,000 (this is the response variable)

1. create a variable `income.num` within the data frame that is 0 for income $< \$50,000$ and 1 for income $\geq \$50,000$
2. for the three categorical predictor variables, use `aggregate` or `dplyr` functions to compute the *univariate* summaries of the probabilities in each category of having income $\geq \$50,000$.

3. for the two continuous predictor variables `age` and `education.num`, use ggplot to plot `income.num` with points along with a smooth function of the predictor

4. Fit a logistic regression including quadratic effects of `age` (use `poly(age,2)` so that the linear and quadratic terms are treated together in the following steps), linear effects of `education`, all three categorical predictors, and all of the two-way interactions among `poly(age,2)`, `education.num`, and the three categorical predictors (the resulting model will have 27 total parameters).

5. Use `drop1` to run a likelihood ratio test on all of the interaction terms in the model. Pick one of the statistically significant interactions; for one of the parameters associated with this interaction (there may be only one), explain what the sign and magnitude of the parameter mean in terms of the differences in log-odds of having an income $\geq \$50,000$ between particular groups (e.g. "the difference in log-odds between males and females decreases by (amount) when age increases by 1 year", or "the log-odds difference between Blacks and Whites is (amount) greater in males than for the population as a whole").

6. Use your model to compute the probability that a 50-year-old, Divorced, White Male with 12 years of education will have an income $> \$50,000$.

7. Compute 95% quantile bootstrap confidence intervals for the predicted value from the previous question. (Reminder: for each bootstrap replicate you'll need to (1) create a new data set with observations resampled with replacement from the original data set; (2) re-fit (update) the original fitted model to use the bootstrapped data; (3) compute and save the predicted value.) Since this may be a little slow, you can limit your computation to 100 bootstrap replicates.