BANA 7042: Final Assignment/Exam

Note: Please be sure to properly label all figures and include a caption for each!

Question 1. The data below gives the number of new AIDS cases each year in a country, from 1981 onwards (Source: Venables and Ripley (2003), slightly modified).

```
cases <- c(12, 14, 33, 50, 67, 74, 123, 141, 165, 204, 253, 246, 240, 246, 232)
year <- 1:15 + 1980
```

a. Plot the number of new AIDS cases against year. Comment on your plot.

b. Fit a simple Poisson regression model of the form

$$\log(\mu) = \beta_0 + \beta_1 \texttt{year}$$

for prediting the mean number of new AIDS cases. Plot the fitted mean response $\hat{\mu}$ on top of the previous scatter plot. Does the model seem adequate?

c. Modify and run the code below to perform a basic residual analysis to inspect the adequacy of the model. You'll need to change the name of the model object. Based on these results, how might you consider improving the model?

```
par(mfrow = c(2, 2))  # set up 2x2 plotting grid
plot(<my-model>, which = 1:4)
```

d. Based on your assessment of the residuals, does it seem like a quadratic model would provide a better fit? Fit a second degree polynomial model (i.e., a model using the formula `cases ~ year + I(year^2)`). How does this model comapre to the simpler model. Be sure to provide a plot of the fitted regression line on top of a scatter plot fo the raw data.

e. Use a likilhood ratio test to comapre the two models and interpret the results. Be sure to write out the associated null and alternative hypotheses. (Hint: use R's built-in `anova()` function with the previous two models and specify `test = "Chisq"` like we've done in class.)

Question2

Source: Bilder and Loughin (2015)

In order to maximize sales, items within grocery stores are strategically placed to draw customer attention. This exercise examines one type of item—breakfast cereal. Typically, in large grocery stores, boxes of cereal are placed on sets of shelves located on one side of the aisle. By placing particular boxes of cereals on specific shelves, grocery stores may better attract customers to them. To investigate this further, a random sample of size 10 was taken from each of four shelves at a Dillons grocery store in Manhattan, KS. These data are given in the `cereal.csv` file. The response variable is the shelf number, which is numbered from bottom (1) to top (4), and the explanatory variables are the sugar, fat, and sodium content of the cereals. Using these data, complete the following:

a. The explanatory variables need to be re-formatted before proceeding further. First, divide each explanatory variable by its serving size to account for the different serving sizes among the cereals. Second, re-scale each variable to be 190 Analysis of Categorical Data with R within 0 and 1.[1] The code below can be used to re-format the data after the data file is read into an object named `cereal`:

```
# Function to rescale a variable to be within (0, 1)
stand01 <- function (x) {
    (x - min(x)) / (max(x) - min(x))
}

# Standardize and rescale data for analysis
cereal2 <- data.frame(
  Shelf = cereal$Shelf,
  sugar = stand01(cereal$sugar_g / cereal$size_g),
  fat = stand01(cereal$fat_g / cereal$size_g),
  sodium = stand01(cereal$sodium_mg / cereal$size_g)
)
```

b. Construct side-by-side box plots with dot plots overlaid for each of the explanatory variables. Below is code that can be used for plots involving sugar:

```
boxplot(sugar ~ Shelf, data = cereal2, ylab = "Sugar", xlab = "Shelf",
        pars = list(outpch = NA))
stripchart(cereal2$sugar ~ cereal2$Shelf, lwd = 2, col = "red",
           method = "jitter", vertical = TRUE, pch = 1, add = TRUE)
```

[1]The author of `nnet::multinom()` suggests this can help with the convergence of parameter estimates when using the function. We recommend the 0-1 rescaling because slow convergence occurs here when estimating the model in part (d) without re-scaling.

Based on your visual analysis, discuss if possible content differences exist among the shelves.

c. The response (`Shelf`) has values of 1, 2, 3, and 4. Under what setting would it be desirable to take into account ordinality. Do you think this occurs here? If so, explain.

d. Estimate a multinomial regression model with linear forms of the sugar, fat, and sodium variables. Which variable seems most important in predicting shelf placement (e.g., using the LOVO method)?

e. Interpret the coefficients for `sugar` (there should be three of them since the response has four categories). Further, construct an effect plot (like we've done several times in class) that shows how `sugar` effects the predicted probability of shelf of placement for each of the four shelves (i.e., you should have four curves, one for each shelf). You can use the code below if needed (though, you'll need to at least modify the name of the fitted model). Descibe the plot.

```
#install.packages("pdp")
library(pdp)
pfun <- function(object, newdata) {
  probs <- predict(object, newdata = newdata, type = "probs")
  colMeans(probs)
}
# Assuming the name of your model is `fit`
pd <- partial(fit, pred.var = "sugar", pred.fun = pfun, plot = FALSE)
lattice::xyplot(yhat ~ sugar|yhat.id, data = pd, type = "l",
                ylab = "probability")
```

f. Kellogg's Apple Jacks (http://www.applejacks.com) is a cereal marketed toward children. For a serving size of 28 grams, its sugar content is 12 grams, fat content is 0.5 grams, and sodium content is 130 milligrams. Estimate the shelf probabilities for Apple Jacks. (Careful here, remember that you rescaled the original data. Any of those transformations would also have to be applied to new data before making predictions!)

g. Based on the background provided, along with your analysis, which shelf (i.e., 1, 2, 3, or 4) seems to be the most beneficial for targeting children and why?

Question 3 The failure of an O-ring on the space shuttle Challenger's booster rockets led to its destruction in 1986. Using data on previous space shuttle launches, Dalal et al. (1989) examine the probability of an O-ring failure as a function of temperature at launch and combustion pressure; in class, we looked at only temperature. Data from their paper is included in the `challenger.csv` file. Below are the variables:

• `Flight`: Flight number • `Temp`: Temperature (F) at launch • `Pressure`: Combustion pressure (psi) • `O.ring`: Number of primary field O-ring failures • `Number`: Total number of primary field O-rings (six total, three each for the two booster rockets)

The response variable is `O.ring`, and the explanatory variables are `Temp` and `Pressure`. Complete the following:

a. The authors use logistic regression to estimate the probability an O-ring will fail. In order to use this model, the authors needed to assume that each O-ring is independent for each launch. **Discuss why this assumption is necessary and the potential problems with it.** Note that a subsequent analysis helped to alleviate the authors' concerns about independence.

b. Estimate the logistic regression model using the explanatory variables in a linear form (i.e., `O.ring ~ Temp + Pressure`).

c. test whether or not `Pressure` can be dropped from the model. Be sure to specify which null and alternative hypotheses are implied by this test, and describe which test you used (e.g., marginal test or likelihood ratio test (LRT)) and the conclusion you reached using an $\alpha = 0.1$ level of significance.

d. The authors chose to remove Pressure from the model based on the LRTs. Based on your results, discuss why you think this was done. Are there any potential problems with removing this variable?

e. Refit the model using only `Temp` as a predictor. What is the estimated probability of an O-ring failure at 31 °F?

f. How does the probability change when you switch to using a probit link function?

Question 4. Consider the `Bikeshare` data set from the ISLR2 package. The data can be loaded into R using the following code:

```
# install.packages("ISLR2")  # run this first if you don't have ISLR2 installed
bike <- ISLR2::Bikeshare
```

Read the help page for the data set using `?ISLR2::Bikeshare` in the R console or by visiting the package documentation at the above link. Build an appropriate model using `bikers` as the response against the following predictors: `workingday`, `temp`, `weathersit`, `mnth`, and `hr`. Create a 1-2 page report describing the data, variables, and model you've built (e.g., what kind of model did you choose and why). Be sure to include a well-formatted regression table with coefficients and standard errors. Any evidence of overdispersion or zero-inflation? If so, how did you deal with it. Provide an interpretation for each coefficient (for categorical variables, you only need to provide an interpretation for one of the categories). Discuss which variables seem the most important in predicting the response. Under what conditions does bike rentals seem to be highest? Do these results make sense? Include effect plots for what you consider to be the two most important predictors and describe any trends you see (be sure to explain how you selected these predictors in the first place). What recommendations can you make to the bike rental agency to improve the rental sales? **DO NOT INCLUDE ANY R CODE**

OR DIRECT OUTPUT. Use well-formatted tables and graphics with captions. This is a report for a stakeholder, so simplicity, good grammar, and complete sentences are key. (Feel free to put R code and output in an appendix if you feel it's necessary. **Hint**: It's not.)