STAT-S 352 Problem Set 9

Upload your answers through the Assignments tab on Canvas by 11:59 pm,

Tuesday, November 4, 2025.

Important Note: Answer all questions and include R code when necessary. In general in this course, give explanations and/or working for all answers unless otherwise stated. Show your work for full credit.

Reminder: As a student at IU, you are expected to uphold and maintain professional and academic honesty and integrity. Academic integrity violations include: cheating, fabrication, plagiarism, interference, violation of course rules, and facilitating academic dishonesty. When you submit an assignment with your name on it, you are signifying that the work contained therein is yours, unless otherwise cited or referenced. Any ideas or materials taken from another source must be fully acknowledged.

1. In Problem Set 7 (#3), you fit four models to the variable `alldeaths` in the Atlantic hurricanes data (`hurricanes.csv`). Find the AIC in each case. Which model gives the lowest AIC? Is it the best or worst in these four models? (Note: Ch 4 slides and/or s352-R-week10-b R files.)

2. Using the data set `epl1819.csv`, create a variable `TotalGoals` that's the sum of the columns `FTHG` and `FTAG`. Perform a G-test of the hypothesis that `TotalGoals` follows a Poisson distribution, giving a P-value and conclusion. (Note: See Chapter 4 slides and/or s352-R-week10-a R files for the setup for this test.)

3. For the `alldeaths` variable in `hurricanes.csv`, a statistician wants to do a formal test of the hypothesis that the data follows a Negative Binomial distribution. They find MLEs of 0.4462 for the size parameter and 0.0211 for the probability parameter. They suggest using the categories given in the following table.

Number of deaths	Times observed	Times expected
0		
1–4		
5–9		
10–14		
15–19		
20–29		
30–49		
50–74		
75 or more		
Total	92	92

(a) Using the raw data, fill in the "Times observed" column.

(b) Using the MLEs and R functions such as `dnbinom()` and `pnbinom()`, fill in the "Time expected" column.

(c) Calculate the G-statistic for a likelihood ratio test.

(d) Find a P-value by comparing the G-statistic to a chi-squared distribution. As far as you can tell from the test, is the data compatible with the negative binomial? (Remember, the degrees of freedom is number of categories minus 1 minus the number of parameters estimated from the data.)

4. In the NFL, it is thought that in general, the windier it is, the fewer points are scored. The file `NFL.csv` contains several variables of interest recorded for NFL games since 1999:

   - `total`: the total number of points scored in the game (both teams combined.)
   - `total_line`: the total number of points that bookmakers predicted would be scored before the game began.
   - `wind`: average wind speed during the game (miles per hour.)

   One issue is that wind speed is missing for some games. (This is a particular problem when using a criterion like AIC, as then a comparison between a model that uses all rows and one that leaves some out wouldn't be fair.) Let's first clean the data by omitting missing values as follows:

```
NFL <- read.csv("NFL.csv")
NFLclean <- NFL[complete.cases(NFL$wind),]
```

   This should leave you with a 4569 row data frame with no missing values for the three variables we're interested in.

   (a) Fit a simple linear regression model on the cleaned data using `wind` to predict `total` (Model 1.) Write down an equation for the model, and make an argument that wind does help to predict total points scored. (You can make this argument using either a P-value or a criterion like AIC.)

   (b) Fit a simple linear regression model on the cleaned data using `total_line` to predict `total` (Model 2.) Write down an equation for the model, and make an argument that this model will give you much more accurate predictions than Model 1.

   (c) Fit a multiple linear regression model on the cleaned data using both `total_line` and `wind` to predict `total` (Model 3.) Write down an equation for the model. Find the AICs for Models 1, 2, and 3. Which model gives the lowest AIC?

   (d) Now split your data into a training set and a test set. We want close to a 70-30 split, so randomly select 3200 rows to be the training set and the rest to be the test set. Re-fit all three models on your training set, then use these re-fitted models to make predictions on the test set. Measure the accuracy of each of the three models on the test set by mean squared error. What are your mean squared errors for each of the three models? Which model gives the lowest mean squared error? (See Chapter 4 slides (pp. 50-52) and s352-R-week10-c R files for an example.)