

WRITING ASSIGNMENT (HAND IN SEPARATELY)**1. NENANA ICE CLASSIC (10 POINTS)**

In 1917, railroad workers were building a bridge across the Tanana River in Alaska. As a diversion, they placed bets on when the ice in the river would start to break up, and a betting pool was started with a pot of \$800. Over the years, the contest has grown, and the Nenana Ice Classic (as it has come to be known) now sports a pot of over \$300,000 and is regulated by the state of Alaska as a legalized game of chance.

Because of the large amount of money at stake, the exact moment of ice breakup has been recorded carefully each year, providing a consistent and high-quality source of data on local climatic change. The data are available in the usual place on my web site. The dataset gives the breakup date for each year from 1917 to 2007. The breakup date is given as the month/date/time, and also as a corresponding Julian date, which is a way of expressing the date in a numerical form that adjusts for leap years.

Analyze and explore this dataset using whatever methods you think are appropriate. Some questions you may want to explore: Is there any evidence for a warming trend? If so, over what time period has the trend occurred? Is the trend linear or nonlinear? Is there evidence for any short-term fluctuations or cycles? **Your answer should be about 1–2 typed pages of text; also hand in any plots you found informative.**

(To be clear, this is just one dataset on one location; obviously by itself this cannot prove or disprove the existence of global warming. However, it may provide one piece of evidence that can be combined with other evidence to increase our understanding of the complex phenomenon of global warming.)

For more information on the Nenana Ice Classic, see the following:

official web site: <http://www.nenanaakiceclassic.com>
news article: <http://news-service.stanford.edu/news/2001/october31/alaskabet-1031.html>
research article: R Sagarin and F Micheli (2001): *Science* vol. 294, p. 811.

Some helpful Data Desk commands: Make a scatterplot of the breakup date vs year. To fit a lowess smoother, go to the HyperView menu (top left corner of the scatterplot) and choose **Smoothing** ♦ **Add Lowess Smooth**. To change the *bandwidth* or *span* of the smoother (i.e., the size of the smoothing window, which controls the degree of smoothing), choose **Smoothing** ♦ **Smoothing Options** and change the **Lowess Span** % value. Higher values result in more smoothing; lower values result in less smoothing. To fit the *4255H, twice* smoother, go to the HyperView menu and choose **Smoothing** ♦ **Add Median Smooth**.

BUT WAIT, THERE'S MORE...

2. SMOKING AND LONGEVITY (10 POINTS)

As part of a study on women's health conducted in the UK in the mid-1970's, information was obtained on whether each of 1314 British women smoked (among many other questions). A follow-up study was conducted 20 years later, in which it was determined whether each woman was still alive or had died. The data are available in the usual place on my web site.

Use MANET to explore this dataset. The program, which runs only on Mac OS X, can be downloaded from the author's home page at <http://www.public.iastate.edu/~hofmann/research.htm>. (use the second link on the page). You can find more information about the program at <http://stats.math.uni-augsburg.de/Manet> (but don't download the software from this page, as it is not up-to-date). Note that this is a developmental version of the software and may be buggy; copy and paste your plots into a word processor and save your work often.

(Personally, I find the default color settings unattractive; if you'd like to change the colors, go to **MANET OS X** ♦ **Preferences** and click on the color bars.)

Drag the `smoking.txt` icon onto the MANET icon to open the dataset. The program should open with a small window in the upper right corner of your screen, which lists the variables age (given as one of five age groups), status (smoker/non-smoker), and survival (alive/dead after 20 years). (There will also be a small window in the bottom left corner, which you can ignore.)

(a) Make bar charts of each variable. To do this, select all three variables in the list (use command-click to select multiple variables) and choose **Plots** ♦ **Bar Chart**. You should get three bar charts in separate windows. Now click on the bar for "alive" and observe that the living women are selected in all other plots as well. Is a higher proportion of smokers alive after 20 years, or a higher proportion of non-smokers? Do you find this surprising?

(b) Convert each bar chart to a spine plot by clicking near the bottom of each window (the cursor should change into a two-sided arrow; you can click to toggle between bar charts and spine plots). Again observe whether a higher proportion of smokers or non-smokers is still alive after 20 years. Briefly describe whether it is easier to determine this using a bar chart or a spine plot, and why. Hand in a copy of whichever plot you found more effectively answered the question.

(c) Make a mosaic plot (essentially a multi-dimensional spine plot) of age and status. To do this, select age and status in the variable list and choose **Plots** ♦ **Mosaic Plot**. Hand in a copy of this plot and **label it** with each age and status category (you can just write in the labels by hand on your paper). Click on the bar for "alive" and observe that the living women are selected in the mosaic plot. After controlling for age, is a higher proportion of smokers alive after 20 years, or a higher proportion of non-smokers? Does this seem to contradict your finding in parts (a) and (b)? If so, what explains this apparent paradox? What do you conclude about the association between smoking and longevity in this population?

Your answer should be about 1–2 typed pages of text; also hand in the plots requested above.

AND THAT'S NOT ALL...

PROBLEMS (HAND IN SEPARATELY)

1. SPAM (14 POINTS)

Ever wonder how your email program is able to distinguish spam from real messages? One way to do this is by using logistic regression.

The dataset `spam` (in the usual place on my web site) has data from 1000 email messages sent to George Forman, a Hewlett-Packard computer scientist. For each message, Forman recorded how often (as a percentage of the total words in the message) the words `meeting` and `credit` appeared in the email.

(This is actually excerpted from a much larger dataset, with 4601 messages and 57 predictors. If you're curious, you can download the entire dataset from <http://archive.ics.uci.edu/ml/datasets/Spambase>.)

- 1) First, make a picture. Select `spam` as X and `meeting` and `credit` as Y variables. Then select **Plot ▸ Dotplot y by x**. (Hand in these two dotplots.) How does the distribution of each word differ in spam vs real messages?
- 2) Now we'll fit a logistic regression model. Select `spam` as Y and `credit` and `meeting` as X 's. Then choose **Calc ▸ Linear models**. In the window that opens, select `Logistic` under **Type of analysis** in the **Dependent variables** panel. Under **Factors**, in the column labeled **Kind**, set both variables to be `Continuous`. Underneath that, set interactions to include up to 2-way interactions. Now open the **Results** panel by clicking on the \triangleright symbol next to the word **RESULTS**. An ANOVA table should open. Is the interaction term significant? Why or why not? If not, delete the interaction term by setting the interactions back to 1-way. (Hand in the entire **Linear Model** window, including the ANOVA table results, for your final model.)
- 3) What is the coefficient giving the effect of `credit`? To find this, open the panel labeled **Results for factor crt**, and then open the **Coefficients** panel. Is `credit` a statistically significant predictor? Calculate the odds ratio for `credit` and explain what this quantity means.
- 4) What is the coefficient giving the effect of `meeting`? To find this, set the pop-up menu in **Results for factor crt** to `mtg`, and then open the panel. Is `meeting` a statistically significant predictor? Calculate the odds ratio for `meeting` and explain what this quantity means.
- 5) Using your model, predict the probability that a message is spam if `credit` makes up 0.2% of the total words in the message and `meeting` does not appear at all. (It's easiest to calculate this by hand rather than by using Data Desk. To find the intercept, set the pop-up menu in **Results for factor mtg** to `Const`.)
- 6) How successful is your model at distinguishing spam messages? To determine this, first select **Compute ▸ Predicted** from the **HyperView** menu of the **Linear Models** window. This should create a new variable titled `Predicted(LM)`, which holds the predicted logits. We must now convert these to predicted probabilities. To do this, create a new derived variable (**Data ▸ New ▸ Derived Variable**) and name it `pred spam`. In the blank formula window that opens, type the formula $1/(1+\exp(-\text{predicted(LM)})) > .5$ and close the window. This derived variable says that we predict a message to be spam if it has greater than 50% chance of being spam. Select the icon for `pred spam` as Y and the icon for `spam` as X , and then make a table by choosing **Calc ▸ Contingency Tables**. (Hand in this table.) Of the messages your model predicted to be spam, what percent actually are spam? Of the messages your model predicted to be non-spam, what percent actually are not spam?
- 7) Is your spam filter conservative (incorrectly misses many spam messages) or aggressive (incorrectly classifies real messages as spam)? How can you easily change the conservativeness/aggressiveness of your spam filter?