Homework 2 Statistics 8330 (DAIII; Wikle) September 6, 2016

DUE: September 13, 2016

Instructions: Write your answers clearly on a separate answer sheet. To minimize the grading effort, ONLY report the results that are asked for. Please type your answers using a word processor; if plots are asked for, imbed them in the document.

1. JWHT (2013), Chap 5, Prob 8
2. JWHT (2013), Chap 5, Prob 9
3. Newspapers that publish daily must know how many pages to expect for their classified advertising section in order to determine how much "news" can be included in the paper, and how much newsprint inventory must be on-hand. Classified advertisements consist of so-called "liner" ads (which are contained within the newspaper column) and "display" ads (which are like regular newspaper ads and can occupy more than one column). For the newspaper under consideration, the "liner" ads can vary in size between 0.15 inches and 22 inches in length and there is a 0.021 inch space between any two of them. Thus, if one knew the column inches of "display" ads, the column inches of "liner" ads, and accounted for the space between "liners", then it should be possible to determine the total number of pages needed for the classified section of the paper. However, for various reasons (related to how a liner ad can be broken at the end of a column, and how headings are created and sized) this relationship cannot be predicted deterministically. Thus, in order to build a statistical learning algorithm to generate a prediction, a major metropolitan newspaper collected a month of data on three variables:

   COUNT LINERS: the number of liner ads; INCHES LINERS: the number of column inches taken up by the liners; LINES DISPLAY: the number of column lines in the display ads (note: one column line equals 0.0764 column inches); and PAGES: total number of pages (1 page equals 220 column-inches).

   There is an obvious cycle to the number of pages since, for example, there are traditionally many more ads in the Sunday paper than any other day.

   The newspaper would like to know whether the total number of PAGES can be reasonably predicted by the information in the three variables: COUNT LINERS, INCHES LINERS, and LINES DISPLAY. These data can be found in the file `pages.dat` and are shown below.

   Considering a normal error regression model with an intercept, evaluate each possible 1 variable, 2 variable and 3 variable model through 5-fold cross-validation (with MSE as the predictive evaluation) using the `cv.glm` routine in R (as described in the JWHT (2013) Chap. 5 lab). Report the CV values for each model and select the "best" model. Do the residuals from this best model suggest any problems with the normal error regression assumptions?

DATE	PAGES	COUNT	INCHES	LINES
		LINERS	LINERS	DISPLAY
10/1	12	5874	1659	831
10/2	14	7205	2096	3644
10/3	37	8436	2509	56200
10/4	62	12719	7862	44760
10/5	12	6613	1891	2003
10/6	12	6080	1720	732
10/7	30	6579	2138	5941
10/8	11	6331	2075	2792
10/9	14	7006	2083	4240
10/10	43	8332	2498	71821
10/11	68	12503	7487	46820
10/12	12	6646	1929	1903
10/13	11	6208	1767	403
10/14	19	6610	2212	11891
10/15	11	6439	2612	6673
10/16	16	7000	2790	7946
10/17	39	8090	3469	68716
10/18	67	12157	8327	47180
10/19	12	6442	1900	1653
10/20	11	5976	1690	309
10/21	19	6421	2157	9043
10/22	11	6153	1783	1051
10/23	14	6177	2084	4933
10/24	40	8430	2556	76240
10/25	69	6070	4010	52267
10/26	12	6823	2049	1831
10/27	12	6416	1828	367
10/28	20	6783	2169	11870
10/29	11	6401	1830	651
10/30	15	6983	2178	4966
10/31	44	8049	2449	80142

4. The problem of interest concerns, ultimately, prediction of presence/absence for plant species in Missouri, based on readily available covariate information. In particular, we are interested in two species of plants, *Desmodium glutinosum* and *Desmodium nudiflorum.* The Desmodiums are a plant genus in the legume family (*Fabaceae*) that are important ecologically because they fix nitrogen and make it available for other nearby plants. The Missouri Ozarks have a fairly nutrient poor substrate and the ability for nitrogen fixing plants to succeed there in the abundance they do is an interesting ecological topic. [You may know Desmodium because it's the plant that is responsible for all those small triangular "velcro" type seeds which will attach to your pants and socks when you walk through the woods in late summer and early fall.] As much of a nuisance as it is to humans, it's a very important part of most forested and prairie ecosystems in Missouri. A very intensive large-scale field experiment (MOFEP) was conducted in Missouri and many species of plant were measured yearly at specified locations. Ideally, landscape ecologists would like to use the information

from these studies to predict the presence or absence of these plant species over very large spatial domains, at fairly high resolution. Clearly, it is too expensive to conduct field studies over such domains, so it is necessary to build statistical models to predict presence/absence given easily obtained covariates.

The file

```
des_site1and2sp_2.dat
```

contains measurements of plant presence/absence and associated covariates at plot locations in two separate experiment regions:

- *Column 1:* Presence (1), absence (0) for *Desmodium glutinosum*
- *Column 2:* Presence (1), absence (0) for *Desmodium nudiflorum*
- *Column 3:* UTM Northing coordinate (meters) [north/south location]
- *Column 4:* UTM Easting coordinate (meters) [east/west location]
- *Column 5:* “southwestness” variable; a transform of ”aspect” or ”which side of a hill you are on”. Aspect is important because many species either prefer hotter, sun-exposed slopes or cooler/wetter sun-protected slopes. The “southwestness” variable refers to how “southwest” you are on a hill, where 1 is the most SW and -1 is the most NE.
- *Column 6:* relative elevation; value between 0-1; a measure of where the observation is taken on a hill, vertically. The top of a hill is 1 and the bottom of a hill is 0.
- *Column 7:* slope; value between 0-200 (45 degree slope=100; 90degree slope =200)
- *Column 8:* geology (categorical); Geology refers to the geologic period in which the underlying surficial rock originates. This gives some measure of the quality of the growing substrate for plants, as well as how deep in a valley you are. Category 0g takes the value 1 and category 0Ce takes the value 0 for this variable.
- *Column 9:* land type association (LTA) (categorical, 2 categories); LTA is a bit more complicated. Essentially the entire continent was divided up into regions based on numerous ecological characteristics. Things related to geology and geographic location, plant species and community abundance as well as local factors such as which watershed it is in and fluvial properties. Basically, this is an overall descriptor of the area of the US you are in. There are several hundred in the state of Missouri but only a couple in this data set.
- *Column 10:* Ecological land type (ELT), categorical (10 types); You can think of ELT similarly to LTA in that it is made up of classified categories which are based on several site defining features. ELTs are determined by slope, aspect, geology, soil depth, and geolandform (the type of geologic feature; i.e. backslope, footslope, shoulder, shoulder ridge, and summit).
- *Column 11:* site (categorical); Corresponds to the two experimental sites from which measurements were taken. This is probably not useful as a covariate, but one might be interested if things are different between sites.
- *C olumn 12:* subplot number; this is just an index to delineate the measurement location within the site (not useful for us).

Consider a logistic regression trying to predict the presence of *Desmodium glutinosum*. You are to consider a model with an intercept and all possible two-variable models with covariates from (ONLY) Column 5,6,7,8,9,10. Use 10-fold cross-validation with the cost function for classification given in the R help associated with the `cv.glm` function (note: don't use the default cost function here!). Report results from these 15 models along with your choice for the best model. Interpret this cost function in terms of classification and logistic regression. Finally, find the best classification model possible using any of the Column 5,6,7,8,9,10 variables based on the cost function used above. Justify that your model is better than the best 2-variable model you identified above.