4/26/09

Stat 430, HW Set 7, Due Friday 5/8/08

As usual, turn in the SAS code which generated your results, and an edited form of the tables and plots (only those that are asked for *or that play a role in your conclusions*). You may decide to remove variables from regressions in the p-value for the hypothesis test that the coefficient is 0 is greater than 0.05, but please make residual plots after each model-fitting step (for your own benefit — you need not hand all the plots in).*All of your problem solutions should include a narrative with reference to the plots and outputs you include, explaining how you obtain and how you justify the conclusions you reach.* **This problem set counts 15 points.**

(I). The dataset `state` contains listings of 1960 per capita state and local expenditures (EX), as well as estimates of the fraction of the state's inhabitants who live in large metropolitan areas (MET). You are to investigate how, on average, per capita expenditure depends on the fraction of a state's citizens who live in large urban area.

(a) Fit a linear regression and comment of the quality of fit based on R^2, residual plots, and coefficient estimates.

(b) Fit a cubic polynomial regression (with variables x, x^2, x^3). Are there any predictors that should be eliminated from the model ? Comment on the effects of any changes you make or on why no changes are needed. Again refer *specifically* in your answer to residual plots, R^2 and other regression output.

(c) Suppose that you wanted to predict 1960 per capita expenditures for a state omitted from this dataset which had 55% of its population living in big cities. Would it be reasonable to expect that your prediction would be within \$20 of the true value with probability of at least 0.95 ?

(d) Using automatic model selection features of SAS PROC REG, find the best linear regression model you can for EX as response variable versus MET, MET**2, MET**3, GROW, YOUNG, OLD, WEST, and MET*WEST. Explain what criterion you are using for choosing the "best" model, and justify your choice using residual plots and an ANOVA table in additional to your preferred criterion.

(II). In the dataset “ESOPH.data” is contained information on 1175 patientsfrom French hospitals, 200 of whom were selected as esophageal-cancer cases, and the remaining 975 of whom were selected from the comparable non-cancer patient populations as ‘controls’. Think of the controls as a typical random sample from a population of non-esophageal-cancer patients.

The objective is to examine which aspects of age and alcohol and tobacco use are post predictive of case versus control status. Imagine a binary variable “Status which is 1 for all of the cases and 0 for all of the controls: our objective will be to fit the best logistic regression models we can to the combined dataset.

(a) Using PROC FREQ, print tables of the proportions respectively of the cases and controls in the various age by tobacco groups for the patients in the dataset.

(b) Write a data-step to make a SAS file with 1175 records and columns AGE, ALC, TOBACC, STATUS containing exactly the same information as the `ESOPH.dat` dataset, but in a form which can be used by PROC LOGISTIC. Code AGE into groups identifiedbymidpoint of age-interval 30, 40, 50, 60, 70, 80; similarly code ALC into four group-identifying values 20, 60, 100, 140, and TOBACC into values 0, 10, 20, 30.

(c) Fit thebest Logistic Regression modelyou can from these data. Note that AGE, ALC, and TOBACC can be viewed either as group-labels (by specifying these as CLASS variables) or asnumeric values. Try it both ways, and also consider the possibility that (pairwise) interactions of the predictor variables should enter the model. Find the best model overall according to the AIC criterion.

(d)For your best model, tabulate the observed and expected numbers of cases within each of the AGE/ALC/TOBACC groups.

(III). The dataset `home` contains a random sample of records of resales of homes from Feb. 15 to Apr. 30, 1993, from the files maintained by the Albuquerque Boeard of Realtors. This type of data is collected by multiple listing agencies and is used by realtors as an information base. Ignore the predictor AGE, but use the other predictors (two measurements TAXES and SQFT, a count FEATS, and three categorical variables COR, NE, CUST).

(a) First use Proc GLM to check formally whether the mean resale price varies with any of the categorical variables and whether they seem to have an interaction effect.

(b) Next fit a separate regression model (with `log(price)` as response variable) to each set of measurements defined by different combinations of the levels of the two categorical variables COR, NE. This involves data-steps to set up dummy columns, investigation to see whether interaction terms (including the categorical variables) and/or quadratic terms are necessary, and variable selection done separately in the four groups. Do not assume initially that the same predictor variables will be important in all groups. Explain your findings: summarize and contrast the models you have fitted.

(c) Examine the dataset as a whole for outliers, both before model-fitting and after fitting the best combined model you can. How can you characterize the outlier(s) objectively ? Should the outlier(s) be removed ?

(d) Whether or not you decide to remove any outlier(s), contrast the model-fits you obtained in (b) and (c): should models be exhibited separately for the 4 groups, or together ? Note that a comprehensive model can always do the job of the separate models, **if** you include enough higher-order interaction terms, but these may be difficult to justify and interpret.