Topics to be covered this week

Statistics 207

Winter Quarter, 2016

Monday, February 29 Ridge regression, Lasso.

Wednesday, March 2 Lasso, Partial Least Squares.

Homework 6 (due on Monday, March 7).

1. Consider a linear model of the form $Y = X\beta + \varepsilon$. Assume that all the variables have been centered so that there is no need to fit an intercept. The least squares estimate of β is $\hat{\beta} = (X^T X)^{-1} X^T Y$. We also know that $Cov(\hat{\beta}) = \sigma^2 D^{-1}$, where $D = X^T X$ and σ^2 is the common variance of ε_i's (ε_i's form the vector ε). Now one may wish to compare the variance of $\hat{\beta}_j$ to the the variance of $\hat{\beta}_j$ case when the independent variables are uncorrelated. This done via a quantity called the variance inflation factor (VIF). The VIF for the j^{th} variable is given by $VIF_j = 1/(1 - R_j^2)$, where R_j^2 is the coefficient of multiple determination when variable X_j is regressed on other independent variables. Note that $VIF_j \geq 1$ and it can be shown that $Var(\hat{\beta}_j) = (\sigma^2/D_{jj})VIF_j$. It is easy to see that σ^2/D_{jj} would be the variance of $\hat{\beta}_j$ if the independent variables were uncorrelated. If VIF_j is larger than then 10 (only a rule of thumb) then it is suggested in the literature that there is a strong case case of multicollinearity).

Consider the "Apartment Building" data:

Y: sale price, X_1:age of structure, X_2:number of apartment units, X_3:lot size (in sq. ft), X_4:number of on-site parking spaces, X_5:gross building area.

Standardize the variables before you start analyzing this data.

(a) Obtain a matrix plot of the data. Also obtain the correlation matrix. What features do you see about the relation between Y and the independent variables. Does it seem that multicollinearity is present? Explain.

(b) Obtain the eigenvalues of the matrix $X^T X$. What do these values suggest about multicollinearity? Explain.

(c) Run a multiple regression model, obtain the parameters, their standard errors and the ANOVA table. Summarize your findings.

(d) Obtain the variance-inflation factors for each of the independent variables. Summarize your findings.

(e) Use the generalized cross-validation criterion to select the penalty for the ridge regression. Use this penalty to run a ridge regression, and obtain the parameter estimates and their standard errors.

(f) Plot the observed response against the fitted Y values obtained from the ridge regression in part (e). Also plot the residuals against the fitted, and the normal probability plot of the residuals. Summarize your findings.

(g) One may obtain the values of the VIF for the estimated ridge regression (using the penalty $\hat{k}$) obtained in part (e). These values are now given by the diagonal elements of the matrix $(X^TX + \hat{k}I)^{-1}X^TX(X^TX + \hat{k}I)^{-1}$. Obtain these VIF values for the ridge regression and compare them to the ones obtained in part (d).

2. Consider the data" ratdrink" which consists of five weekly measurements of body weight for 27 rats. The first 10 rats were on a control treatment, seven rats had thyroxin and 10 rats had thiouacil added to their water. Thus the factor "treatment" has three levels. The goal is to model rat weights that shows the effect of the treatment. In this data set, the subject (rat) effect is random. In the questions below, "week" is considered to be a factor.

(a) For each treatment, plot the weights over week for the rats. Thus there are three plots. For comparison purposes use the same scale for the horizontal and vertical exes for these plots. Explain your findings.

(b) Obtain boxplots for weights over time for the three treatment (R command: "boxplot(y~week*treatment") What do these plots indicate about the mean weight over time (week)? How does the variability of weight change over time and treatment? Explain.

(c) Fit an ANOVA model that is additive in the factors subject (random), week, treatment and week*treatment. Plot the observed weights against the fitted values and residuals against the fitted values. Obtain the boxplots for weights over time for the three treatments. What do these plots indicate about the suitability of the model? Explain.

(d) For the model in part (c), obtain the ANOVA table, the parameter estimates and their standard errors. Explain all the effects that seem to be present in the model. If you feel some terms can be dropped from the model, drop them one at a time in a stepwise manner (using AIC) and arrive at a final model.

3. The plots in parts (a) and (b) of question 2 give an idea of how some of the modeling can be done. Create a time (week) variable T which takes values $0, \ldots, 5$.

(a) Fit a longitudinal model which has random subject effect, a linear growth in T with random slopes, the treatment effect. Plot the residuals against the fitted values, and the histogram of the residuals. Also, obtain boxplots of residuals over time for the three treatments. Explain your findings.

(b) Do the same as in part (a) except that now you fit a quadratic model with random slopes for both T and T^2.

(c) Of the models in part (a) and (b) which one seems to be superior? Use the AIC criterion to select the model. For this selected model, obtain ANOVA table, the parameter estimates and their standard errors. What do you conclude? Explain.

(d) Does it seem that the slopes may depend on the treatment? In that case, one needs to make the slopes depend on treatment. Thus there will be six slope parameters (three for T and three for T^2). Write down a model which has random subject effect, treatment effect and separate slopes for the three treatments (no random slopes).

(e) Fit the model described in part (d). Obtain the ANOVA table, parameter estimates and their standard errors. If you find some parameters not significant, drop them in a stepwsie manner (one at a time) till you arrive at a final model. For this final model, write down the ANOVA table, the parameter estimates and their standard errors.
(f) Compare the final model you have obtained in part (e) to the one in part (d) of question 2 by comparing the boxplots of residuals over weeks and treatments.
(g) For this part, no actual computations are necessary. Write down a model as in part (d) except that the slopes are random.

4. A random sample of s individuals were randomly selected and their incomes were recorded every three years over a total period of 15 years. Also known is the educational level of each individual. Let t_j be the j^{th} time period and $Y_{ij} = \log(income + 1)$ for the i^{th} individual at time $t_j, j = 1, \ldots, r = 5$. In these models below, let ρ_i be the random subject (person) effect, x_i is the educational level of person i. It is assumed that ρ_i 's are iid $N(0, \sigma^2_\rho)$, ε_{ij}'s are iid $N(0, \sigma^2)$, and that $\{\rho_i\}$ and $\{\varepsilon_{ij}\}$ are independent. For each of the following models obtain, the $E(Y_{ij})$, $Var(Y_{ij})$, $Cov(Y_{ij}, Y_{ij'})$ and $Corr((Y_{ij}, Y_{ij'}), j \neq j'$.
(a) $Y_{ij} = \mu + \rho_i + \beta_1 x_i + \gamma_1 t_j + \varepsilon_{ij}$,
(b) $Y_{ij} = \mu + \rho_i + \beta_1 x_i + \gamma_1 t_j + \gamma_{i1} t_j + \varepsilon_{ij}$, where γ_{i1} are iid $N(0, \sigma^2_{\gamma_1})$, and $\{\rho_i\}$, $\{\varepsilon_{ij}\}$, and $\{\gamma_{i1}\}$ are mutually independent.