Final Project, due December 8th, 9:30am

November 29, 2011

This project uses the `bodyfat` dataset with `fat` as the response and the remaining variables except `id`, `siri` and `density` as predictors.

1. Perform variable selection on the entire dataset using the Adjusted R^2 criterion. Report the selected variables.(1 point)

   selection = adjrsqr

2. Using the model from 1, draw the QQ plot of the residuals and test for normality. (2 points)

   ods graphics and proc univariate on resid

3. Using the model from 1, test for correlated errors. (1 point)

   dwprob

4. Divide the data into four subsets, the k-th subset containing the observations $4i + k, k = 1, 2, 3, 4$. Thus the first subset contains observations $1, 5, 9, ...$, second subset contains $2, 6, 10, ...4$, and so on.

   Build the models described below. For each model report the RMSE:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

   using 4-fold cross-validation on the four subsets you created. This means that you obtain the model from the union of any three of the subsets and compute the RMSE on the remaining one.

   Report the four RMSE's and their average in a table. Where variable selection is required, perform the variable selection each time.

   a) Linear regression with all predictors. (4 points)

   b) Linear regression with the variables selected by backward elimination, at the 95% confidence level. (4 points)

   c) Linear regression with the variables selected by forward selection at the 99% confidence level. (4 points)

   d) Linear regression with the variables selected by the Adjusted R^2 criterion. (4 points)

   e) Linear regression with the variables selected by the AIC criterion. (4 points)

   f) Principal Component Regression with 7 principal components. (6 points)

```
data bodyfat1;
do i=1 to N by 4;
set bodyfat point=i;
output;
end;
stop;
run;

data bodyfat2;
do i=2 to N by 4;
set bodyfat point=i;
output;
end;
stop;
run;

data bodyfat3;
do i=3 to N by 4;
set bodyfat point=i;
output;
end;
stop;
run;

data bodyfat4;
do i=4 to N by 4;
set bodyfat point=i;
output;
end;
stop;
run;
```

then merge three of them and compute models and compute the RMSE on the other dataset (4 times) e.g. Train set 2-4 Test on set 1