Tutorial 9

1. Suppose Y has 5 covariates X_1, X_2, X_3, X_4, X_5 denote any model $Y = \beta_0 + \beta_i X_i + \beta_j X_j + \beta_k X_k + \varepsilon$ by (ijk). All the models can be listed as (0), (1), (2), (3), (4), (5), (12), (13), (14), (15), (23), (24), (25), (34), (35), (45), (123), (124), (125), (134), (135), (145), (234), (235), (245), (345), (1234), (1235), (1245), (1345), (2345), (12345). With $n = 40$, their respectively SSE are 95.2261, 73.3037, 95.2139, 90.8737, 69.2889, 73.6588, 73.0233, 70.9692, 53.2715, 42.4473, 90.8726, 69.2660, 73.2872, 63.9012, 72.6848, 25.0160, 70.5948, 52.8136, 42.4440, 49.9665, 42.4467, 0.2618, 63.7998, 72.4197, 24.7802, 24.4749, 49.3597, 42.4432, 0.2512, 0.2496, 24.3022, 0.2406. And their respectively BIC are 0.9596, 0.7902, 1.0517, 1.0050, 0.7338, 0.7950, 0.8786, 0.8500, 0.5632, 0.3361, 1.0972, 0.8257, 0.8822, 0.7451, 0.8739, -0.1927, 0.9370, 0.6468, 0.4282, 0.5914, 0.4283, -4.6600, 0.8358, 0.9625, -0.1099, -0.1223, 0.6714, 0.5204, -4.6093, -4.6157, -0.0372, -4.5600.

   (a) Based on BIC, which model should be preferred

   (b) Based on BIC and using forward selection, which model is selected. Please indicate which models are calculated

   (c) Based on BIC and using backward elimination, which model is selected. Please indicate which models are calculated

   (d) using backward elimination and F-test, start from the full model, test whether X_2 and X_3 can be removed simultaneously.

2. Consider the Gaoline consumption and automotive variables. Y: Miles/Gallon; X_1: Displacements; X_2: Horsepower; X_3: Torque; X_4: compression ratio; X_5: real axle ratio; X_6: Carburetor; X_7: Number of transmission speeds; X_8: overall length; X_9: width; X_{10}: weight; X_{11}: type of transmission (1=automatic; 0=manual). the data is available at **(data)** http://www.stat.nus.edu.sg/%7Estaxyc/CarGasoline.txt

   (a) Would you include all the variables to predict the gasoline consumption of the cars?

   (b) Six alternative models have been suggested:

      i. regression Y on X_1

      ii. regression Y on X_{10}

      iii. regression Y on X_1 and X_{10}

      iv. regression Y on X_2 and X_{10}

      v. regression Y on X_8 and X_{10}

      vi. regression Y on X_8, X_5 and X_{10}

      which model would you choose?

(c) Use stepwise selection method to choose the best model for the prediction of Y based on AIC

(d) Plot Y against X_1, X_2, X_8, X_{10}. Do the plot suggest that the relationship between Y and the predictors may not be linear?

(e) Consider $W = 100/Y$ (gallons per hundred miles). Plot W against X_1, X_2, X_8, X_{10}, is teh relationship between W and the predictors more linear than that of Y with the predictors?

(f) Use stepwise selection method to choose the best model for the prediction of W based on AIC

(g) regression Y on $X_{13} = X_8/X_{10}$

3. **Advertising agency** (data) The managing partner of an advertising agency is interested in possiblity of making accurate predictions of monthly billings. Monthly data on amount of billings (Y) and on number of hours of staff time (X) for the 20 most recent months. A simple linear regression model is believed to be appropriate, but positive autocorrelation in the error may be presented.

   (a) Fit a simple linear regression model by the ordinary least squares and obtain the residuals, also obtain $s(b_0)$ and $s(b_1)$.

   (b) conduct a formal test for positive autocorrelation using $\alpha = 0.01$.

   (c) Based on the estimated autocorrelation coefficient, estimated the transformed model and also obtain $s(b'_0)$ and $s(b'_1)$

   (d) Test whether any positive autocorrelation for the residuals of the transformed model with $\alpha = 0.01$

   (e) based on part (c), reestimate the coefficients in the original model

   (f) obtain the 99% confidence interval for β_1, what is your interpretation

   (g) Staff time in month 21 is expected to be 4.625, predict its EY in month 21.

4. Suppose for model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, ..., n$, the error terms follows

$$\varepsilon_i = \rho_1 \varepsilon_{i-1} + \rho_2 \varepsilon_{i-2} + u_i$$

   where u_i are IID $N(0, \sigma^2)$. How do you transform the model to use LSE?