<u>**MA329 Statistical linear models**</u>

Assignment 5 (Due date: Dec 24 , 11pm. For late submission, each day costs 10 percent)

1. (20 marks) Consider a linear regression model,

   $$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_{p-1} x_i^{p-1} + \epsilon_i,$$

   $i = 1, ..., n$. Also, $\epsilon_1, \epsilon_2, ..., \epsilon_n$ are i.i.d. $N(0, \sigma^2)$. Let $P_k(x) = 1 + x + \cdots + x^k$ be a polynomial of order k. Then the above model could be rewritten as

   $$y_i = a_0 P_0(x_i) + a_1 P_1(x_i) + \cdots + a_{p-1} P_{p-1}(x_i) + \epsilon_i.$$

   Assume that

   $$\sum_{i=1}^{n} P_l(x_i) P_m(x_i) = 0, \quad l \neq m, \quad \text{for all } l \text{ and } m,$$

   (a) Derive the least squares estimator of $\hat{a}_j$, $j = 0, 1, ..., p-1$ and prove that $\hat{a}_j'$s are uncorrelated for $j = 0, 1, ..., p-1$.

   (b) Derive the test for testing the null hypothesis H_0: $a_j = 0$.

   (c) Construct a confidence interval for the mean of y^* at $x = x^*$.

2. (Dataset: 6data.cvs, you may use R to do this question, 40 marks)

   A commercial real estate company evaluates vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data are taken from 40 suburban commercial properties that are the newest, best located, most attractive, and expensive for five specific geographic areas. The response variable is rental rates (Y, in 10000 dollars) and the predictor variables are

   - X_1 – age, in years
   - X_2 – operating expenses and taxes, in 1000 dollars
   - X_3 – vacancy rates
   - X_4 – total square footage, in 1000 square feet

   (a) Use a multiple linear regression model to fit the data and conduct a hypothesis test for the overall utility of the model

   (b) Compute the following statistics (and tabulate your results)

      i. Studentized residuals

      ii. Studentized deleted residuals

iii. Leverage values

iv. Dffits

v. Cook's distance

(c) Identify outlying Y observations, if any. (Use Studentized residuals, considered a point to be an outlier if absolute value of studentized residual is greater than 2)

(d) Identify outlying X observations, if any. [state clearly your identification criterion]

(e) Identify influential observations, if any. [Use Dffits and Cook's distance, state clearly your identification criterion]