Assignment 2

1. The data set gives data on per capita output (output) in Chinese yuan, number (SI) of workers in the factory, land area (SP) of the factory in square meters per worker, and investment (I) in yuans per worker for 17 factories in Shanghai.
   a. Using least squares, fit a model expressing output in terms of the other variables.
   b. In addition to the variables in part a, add SI^2 and $SP \times I$ and obtain another model.
   c. Using the model of part b, find the values of SP, SI, and I that maximize per capita output.

2. The dataset prostate comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. This data contains the following variables:
   lcavol: log(cancer volume)
   lweight: log(prostate weight)
   age: age
   lbph: log(benign prostatic hyperplasia amount)
   svi: seminal vesicle invasion
   lcp: log(capsular penetration)
   gleason: Gleason score
   pgg45: percentage Gleason scores 4 or 5
   lpsa: log(prostate specific antigen)

   a. Fit a model with lpsa as the response and lcavol as the predictor. Report the residual standard error and the R^2.
   b. Now add lweight, svi, lbph, age, lcp, pgg45, and gleason to the model *one at a time*. For each model record the residual standard error and the R^2. Plot the trends in these two statistics and comment on any features that you find interesting.
   c. Plot lpsa against lcavol. Fit the simple regressions of lpsa on lcavol and lcavol on lpsa. Display both simple regression lines on the plot and comment on any features that you find interesting. At what point do the two lines intersect?

3. The data set gives information on capital, labor and value added for each of three economic sectors: Food and kindred products (20), electrical and electronic machinery, equipment and supplies (36) and transportation equipment (37). For each sector:

   a. Consider the model $V_t = \alpha K_t^{\beta 1} L_t^{\beta 2} \varepsilon_t$, where the subscript t indicates year, V_t is value added, K_t is capital, L_t is labor, and ε_t is an error term with $E(log(\varepsilon_t))=0$ and $var(log(\varepsilon_t))$ a constant. Assuming that the errors are independent, and taking logs of both sides of the above model, estimate β_1 and β_2.
   b. The model given in part a above is said to be of the Cobb-Douglas form. It is easier to interpret if $\beta_1 + \beta_2 = 1$. Estimate β_1 and β_2 under this constraint.
   c. Sometimes the model $V_t = \alpha \gamma^t K_t^{\beta 1} L_t^{\beta 2} \varepsilon_t$ is considered, where γ^t is assumed to account for technological development. Estimate β_1 and β_2 for this model.
   d. Estimate β_1 and β_2 in the model in part c, under the constraint $\beta_1 + \beta_2 = 1$.