Homework 11

Problem 1: Let $\{X_1, X_2, \ldots, X_n\}$ be an orthonormal basis of R^n (that is $X_i^T X_j = 0$ for $i \neq j$, and 1 for $i = j$). Then $Z_j = X_j^T Y$, $j = 1, 2, \ldots, n$, are the coefficients in the representation of a vector Y in this basis, that is $Y = \sum_{j=1}^n Z_j X_j$, and the norm square of a vector in R^n is equal to the sum of squares of its coefficients in an orthonormal basis, e.g., $\|Y\|^2 = \sum_{j=1}^n Z_j^2 = \|Z\|^2$. Now consider the linear regression model

$$Y_{n\times 1} = X_{n\times k}\beta_{k\times 1} + \epsilon_{n\times 1}$$

where $\epsilon_{n\times 1} \sim N\left(0, \sigma^2 I_{n\times n}\right)$ with σ unknown, and assume that columns of $X_{n\times k}$ are $X_1, X_2, \ldots, X_k$. Derive the following conclusions.

(i). The least squares estimator $\widehat{\beta}$ is given by

$$\widehat{\beta}_i = X_i^T Y = Z_i,\ i = 1, 2, \ldots, k.$$

(ii).

$$\begin{aligned} Var\left(\widehat{\beta}_i\right) &= \sigma^2 \\ cov(\widehat{\beta}_i, \widehat{\beta}_j) &= 0,\ \text{for } i \neq j. \end{aligned}$$

(iii). In the orthonormal basis the coefficients of the lest squares fit $\widehat{Y}$ has coefficients $(Z_1, Z_2, \ldots, Z_k, 0, \ldots, 0)^T$, and residual vector $Y - \widehat{Y}$ has coefficient $(0, \ldots, 0, Z_{k+1}, Z_{k+2}, \ldots, Z_n)^T$ and is independent of $\widehat{Y}$.

(iv).

$$RSS_k = \left\|Y - \widehat{Y}\right\|^2 = \sum_{i=k+1}^n Z_i^2,$$

and $RSS_k \sim \sigma^2 \chi^2_{n-k}$, and

(v). Let $\widehat{\sigma^2} = RSS_k/(n-k)$, then it is an unbiased estimator of σ^2 and $\left(\widehat{\beta}_i - \beta_i\right)/\widehat{\sigma} \sim t_{n-k}$.

Problem 2 (optional):

Let $\widehat{Y}$ be the lest square fit in the model $Y_{n\times 1} = X_{n\times k}\beta_{k\times 1} + \epsilon_{n\times 1}$ with $\epsilon_{n\times 1} \sim N\left(0, \sigma^2 I_{n\times n}\right)$. Imagine that we could have future observation $\widetilde{Y} = X_{n\times k}\beta_{k\times 1} + \widetilde{\epsilon}_{n\times 1}$ with $\widetilde{\epsilon}_{n\times 1} \sim N\left(0, \sigma^2 I_{n\times n}\right)$ independent of $\epsilon_{n\times 1}$. We use Akaike's final prediction squared error $FPE_k = \frac{n+k}{n(n-k)} RSS_k$ to estimate the average squared prediction error $\frac{1}{n} E\left\|\widetilde{Y} - \widehat{Y}\right\|^2$. Show that

$$\frac{1}{n} E\left\|\widetilde{Y} - PY\right\|^2 = \left(1 + \frac{k}{n}\right)\sigma^2 = E(FPE),$$

We call $se_{\widetilde{Y}-\widehat{Y}} = \sqrt{FPE}$ the standard error of prediction.

Problem 3: when gasoline is pumped into the tank of an automobile, hydrocarbon vapors in the tank are forced out and into atmosphere, producing a significant amount of air

pollution. For this reason vapor-recovery devices are often installed on gasoline pumps. It is difficult to test a recovery device in actual operation, since all that can be measured is the amount of vapor actually recovered and, by means of a "sniffer", whether any vapor escaped into atmosphere. To estimate the efficiency of the device, it is thus necessary to estimate the total amount of vapor in the tank by using its relation to the values of variables that can actually be measured. In this exercise you will try to develop such a predictive relationship using data that were obtained in a laboratory experiment. In the data file "vapor.txt", the columns of the data matrix are initial tank temperature, temperature of the dispensed gasoline, initial vapor pressure in the tank, vapor pressure of the dispensed gasoline, and emitted dispensed hydrocarbons, in that order

(i). Look at the relationships among the variables by pairs plot. Comment on which relationships look strong. Based on this information, what variables would you conjecture will be important in the model? What additional transformed variables or interactions do you suggest to consider?

R codes:

```
vapor=read.table("vapor.txt")
names(vapor)=c("tanktemp", "gastemp", "tankpress", "gaspress", "pollutant")
pairs(vapor)
attach(vapor)
```

(ii). Fit a linear model with four independent variables above. Comment on the coefficients with negative values and residuals plot.

(iii). Try including additional transformed variables or interactions and remove variables that are found to be insignificant, until you find a fitted model for which the residuals show no pattern and the coefficients are all significant.

(iv). We found a model with standard error of prediction $\sqrt{FPE} = \sqrt{\frac{n+k}{n(n-k)} RSS_k} = 2.46$ where k is the number of independent variables we use. Can you build a model with standard error of prediction less than 2.50 or even better than we got? [For this part, you may use additional transformed explanatory variables and interactions, but do not transform the dependent variable "pollutant".]