Stat 200b **Lab 2** **due: 2 Apr. 99**

In Lab 1, we studied the results of an experiment conducted in the early 1980's to develop a predictive equation for the amount of hydrocarbons released into the atmosphere when gasoline is pumped into the tank of an automobile. To refresh your memory, the data set is stored in /class/data/s200b/labs/data/vapor.dat as a 125 by 5 matrix. Each of the $n = 125$ rows of this matrix represents an observation on 5 variables. A description of each of these variables is given below.

Column	**Variable**	**Description**
1	x_1	Temperature of the dispensed gasoline (degrees Fahrenheit)
2	x_2	Vapor pressure of the dispensed gasoline (psi)
3	x_3	Initial temperature in the tank (degrees Fahrenheit)
4	x_4	Initial vapor pressure in the tank (psi)
5	Y	Emitted hydrocarbons (grams)

Again, the variable we are interested in estimating is the amount of emitted hydrocarbons, column 5. The first three observations or rows of **vapor.dat** are given below:

x_1	x_2	x_3	x_4	Y
33	3.49	28	3.00	22
48	3.22	24	2.78	27
53	3.42	33	3.32	29

In the first half of this lab, we will use computer simulation to explore the distributional results surrounding least squares estimation and goodness of fit tests presented in Chapter 10. It is hoped that this exercise will clarify somewhat the motivation behind hypothesis testing. In the second half of the lab, we will be concerned primarily with techniques for model selection known as backward deletion and stepwise addition. Although we are using this method to analyze experimental data, it is is more commonly used for model selection with observational data. Our goal will be to develop a predictive equation for Y based on some part of the full quadratic model in x_1, x_2, x_3, and x_4. We will be particularly interested in developing a *hierarchical* model. Basically, this implies that if higher order terms are included in the model (i.e., quadratics and interactions), then we must also include all associated linear terms. Finally, having developed a reasonable model, we will investigate the fit using hypothesis tests developed in Chapter 10.

I. Simulation

Assume for the moment that the regression function of Y based on x_1 and x_2 is given by

$$\mu(x_1, x_2) \quad = \quad \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Under the assumption of the normal linear model, for each fixed level of x_1 and x_2, the amount of hydrocarbon emissions has a normal distribution with mean $\mu(x_1, x_2)$ and variance σ^2. In this portion of the lab, we will empirically investigate the behavior of least squares estimates of the coefficients in the model given above as well as certain goodness of fit statistics discussed in Chapter 10 of your text. In particular, you will be making a large number of plots and you will be asked to comment on them. Comments should address not only what you have observed, but more importantly, what you expected to observe in the context of the normal linear model.

Note: For this first part of the lab, we will use x_1 and x_2 as they appear in **vapor.dat** and not standardize them as in Lab 1.

1. For the purpose of simulation, we will assume that our regression function is of the form given above, and we will set $\beta_0 = \beta_1 = 1$, and $\beta_2 = 0$. Let $\mathbf{X}$ denote the 125 by 3 design matrix corresponding to the functions 1, x_1, and x_2. Using $\mathbf{X}$, we can simulate an observation of Y, a 125 element column vector, by generating a sample of 125 independent normal random variables ϵ and setting $Y = \mathbf{X}(1,1,0)^t + \epsilon$. For each such Y we generate, we can calculate $\hat\beta_0$, $\hat\beta_1$, and $\hat\beta_2$, their standard errors, and the associated t-statistics.

Generate 100 observations, labeled $Y_1, \ldots, Y_{100}$, using the formula above and taking $\sigma^2 = 2$. Just to be completely clear, each Y_i is a 125 element column vector. For each of these 100 vectors Y_i, calculate an estimate of β_0, β_1 and β_2 using ordinary least squares. We will refer to these estimates as $\hat\beta_{0,i}$, $\hat\beta_{1,i}$, and $\hat\beta_{2,i}$, respectively. Next, calculate the associated standard errors for each of these estimates. We will refer to these quantities as $\mathrm{SE}(\hat\beta_{0,i})$, $\mathrm{SE}(\hat\beta_{1,i})$, and $\mathrm{SE}(\hat\beta_{2,i})$, respectively.

2. Make a histogram of your 100 observations for $\hat\beta_{2,i}$. What do you observe? Next, make a histogram of your 100 t-statistics $\hat\beta_{2,i}/\mathrm{SE}(\hat\beta_{2,i})$ for the test that $\beta_2 = 0$. What do you observe? How many would you expect to fail this test at the 5% level? Make a quantile-quantile plot of these quantities based either on an appropriate t-distribution or on the standard normal distribution. What do you observe? Should the choice of distribution make a difference in the quantile-quantile plots in this case? Why or why not? Finally, for each of your 100 t-statistics, calculate the P-value for the test that $\beta_2 = 0$ and make a histogram of these values. What do you observe?

3. Repeat the steps of question 2, this time replacing β_2, $\hat{\beta}_{2,i}$ and SE($\hat{\beta}_{2,i}$) with β_1, $\hat{\beta}_{1,i}$ and SE($\hat{\beta}_{1,i}$). What do you observe?

4. For each of your 100 datasets, compute the F-statistic associated with the test that both β_1 and β_2 are zero. What proportion of your datasets pass this test at the 5% level? Make a quantile-quantile plot of your F-statistics against an appropriate F-distribution. Finally, make a histogram of the P-values from these 100 F-statistics. What do you observe?

II. Model Selection

We now turn to the actual business of developing a predictive equation for Y based on x_1, x_2, x_3, and x_4.

1. Standardize each of the variables x_1, x_2, x_3, x_4, and Y by subtracting off their means and dividing by their standard deviations. For the rest of the lab, when we refer to a variable x_1, x_2, x_3, x_4 or Y, we will mean these standardized versions.

2. As mentioned previously, we will model the hydrocarbon emissions using the full quadratic model based on x_1, x_2, x_3, and x_4. This means that we will be using all linear predictors, quadratics, and pairwise interactions based on these variables. In all, there are 15 terms including the intercept. Form the design matrix for this model and get the least squares estimates for the coefficients in this model using **lm**. Look at the P-values corresponding to each term in the model. Are all of the terms significant? Note that the quadratics and interactions terms do not have mean 0, so it make sense to include the intercept.

3. We will restrict our attention to hierarchical models. Thus, we can only delete a linear term from the model if all higher order terms involving it have already been removed. For example, if x_3 has a higher P-value than $x_3 * x_4$, we cannot delete x_3 before removing $x_3 * x_4$ from the model. Also, if x_3 has a higher P-value than x_3^2, we cannot delete x_3 before removing x_3^2. We need to look for another term to delete. Think about the intercept as being a varaibale of order 0. At each step in the deletion process, those terms that could be removed without violating the hierarchical structure are termed removable.

In general, the variable deletion rules for model selection will be

(a) Determine the terms in your model and fit it using least squares.
(b) Look at the (two-sided) P-values for each term. If all P-values are 0.05 or less, no variable deletion is necessary. If not, go to (c).
(c) If the term with the largest P-value is a quadratic or an interaction, delete it and fit the reduced model and go back to step (a). If not, go to (d).
(d) If the term with the largest P-value is a linear term, delete it if it is not involved in any quadratic or interaction term. Otherwise, look for the term with the next largest P-value and repeat step (c) until either a term is deleted, and then go back to step (a), or all the P-values associated with the removable terms are 0.05 or less.

Do stepwise deletion starting from the model with all linear terms, quadratics and interactions until all the P-values associated with the removable terms are 0.05 or less. Which terms remain in the model?

4. Construct an F-test of the hypothesis that your final model (from quesion 4 above) is in fact linear. For such a test, you will need the fitted values from your final model from question 4 as well as those from the linear submodel. Calculate the latter in two ways. First, use ordinary least squares on Y and the relevant predictor variables. Next, replace Y with the fitted values based on your larger model. Each technique should yield the same result. Why? Compare your F-statistic with the appropriate F-distribution and calculate a P-value for this test. Present your results in ANOVA table and interpret them.

Part 5 is **optional**. I recomend you to do them, and turn them in with the rest of the lab, but it will not be part of the grade.

5. Now we will see if any of the deleted terms can be added back into the model to significantly improve the fit. Try adding terms to the current model one at a time and see if the fit is improved significantly. (Look at the P-values. The term is significant if the corresponding P-value is 0.05 or less.) Remove any terms that are not significant. Remember that under a hierarchical model, we cannot add a higher order term if the associated linear term is not already in the model. Which ones were added? What is your new model? What do you expect would happen if you left out x_4 from the beginning?

S Hints

In part I, generate Y as a 125 by 100 matrix. Each column represents a different data set. X is a matrix of 125 by 3. Using `out <- lm(y~ x)` will regress each column of Y on X. Now, `out$coef` and `out$res` will give you the residuals and coefficients of the 100 data sets.

Another S function that you might find helpful is **apply**. If A is a matrix of size n by m, the command `apply(A,1,sum)` will create a vector of length n. The i*th* entry in this vector is the sum of the i*th* row of A. The command `apply(A,2,sum)` will create a vector of length m, with the sum of the columns of A. Sum is not the only function you can use. Any other function that expect as argument a vector (or a matrix, for more complicated structures of data) can be used. For more information type `help apply`.