When gasoline is pumped into the tank of an automobile, hydrocarbon vapors in the tank are forced out and into the 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 recovered, and, by means of a “sniffer,” whether any vapor escaped into the atmosphere. Therefore, to get an idea of the efficiency of a given device, it is necessary to estimate the amount of hydrocarbons that would have been emitted if the device was not in place. To be useful, this estimate must be based only on variables that can be measured in practice.

In the next two labs, you will be using the data collected during an experiment conducted in the early 1980’s to ascertain such a predictive relationship. The data set can be downloaded from the class web site. It is 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.

<table>
<thead>
<tr>
<th>Column</th>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( x_1 )</td>
<td>Temperature of the dispensed gasoline (degrees Fahrenheit)</td>
</tr>
<tr>
<td>2</td>
<td>( x_2 )</td>
<td>Vapor pressure of the dispensed gasoline (psi)</td>
</tr>
<tr>
<td>3</td>
<td>( x_3 )</td>
<td>Initial temperature in the tank (degrees Fahrenheit)</td>
</tr>
<tr>
<td>4</td>
<td>( x_4 )</td>
<td>Initial vapor pressure in the tank (psi)</td>
</tr>
<tr>
<td>5</td>
<td>( Y )</td>
<td>Emitted hydrocarbons (grams)</td>
</tr>
</tbody>
</table>

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:

\[
\begin{array}{cccc}
 x_1 & x_2 & x_3 & x_4 & Y \\
33 & 3.49 & 28 & 3.00 & 22 \\
48 & 3.22 & 24 & 2.78 & 27 \\
53 & 3.42 & 33 & 3.32 & 29 \\
\end{array}
\]

For questions 1 through 6, we will consider only those variables that record the state of the dispensed gasoline, \( x_1 \) and \( x_2 \). In the second part of
this lab, questions 7 through 9, we will make use of $x_3$ and $x_4$, the measured
temperature and pressure of the tank. A much more thorough investigation
of how all four variables can be used to estimate $Y$ will be taken up in lab
2.

1. Standardize each of the variables $x_1$, $x_2$, 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 $x_1$, $x_2$ or $Y$, we will mean these
standardized versions.

2. Let $X$ denote the design matrix formed from the basis functions 1, $x_1$
and $x_2$, where 1 represents the function that is the constant 1 over
the design set. Form the three-by-three Gram matrix associated with
these variables. We obtain the sample correlation matrix of $x_1$ and
$x_2$ by dividing the lower two-by-two submatrix of the Gram matrix
by $n−1 = 125 − 1$. What do you observe? Confirm the result by
checking with function cor(). Next, perform the same calculations
with the $4 \times 125$ matrix formed by adding the column $Y$ to the end
of $X$. In this case, we obtain the sample correlation matrix of $x_1$, $x_2$
and $Y$ by dividing the lower three-by-three submatrix of the result
by $n−1 = 125 − 1$. What do you observe? Use a series of pairwise
plots to confirm your observations. Please do not include any of these
plots with your write-up, but simply report your observations. You
might also want to examine plots of $x_1$, $x_2$ and $Y$ against observation
number.

In the following questions, we will explore how the high correlation be-
tween $x_1$ and $x_2$ manifests itself in standard error estimates and derived
confidence intervals. Unless otherwise indicated, you are encouraged to cal-
culate the required least squares estimates both directly and via the lm
command (see the attached S appendix).

3. Let $\mu(\cdot)$ represent the regression function of $Y$ on our two predictor
variables. For the moment, assume that

$$\mu(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$
Using $X$, $Y$ and formula (13) on page 501 of your text, form the ordinary least squares estimates of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$. Calculate RSS, the residual sum of squares, and form

$$S^2 = \frac{\text{RSS}}{n - p} = \frac{\text{RSS}}{125 - 3}.$$ 

Recall that under the assumptions of the homoskedastic linear model, $S^2$ is an unbiased estimate of of $\sigma^2$ (see Sections 10.2 of your text). Use $S$ and the Gram matrix you calculated in question 1 to form the variance-covariance matrix of $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$.

Next, try fitting the same model, this time by leaving out the $x_2$ term. Compare the new estimate of $\beta_1$ and its standard error to the values you obtained earlier. What do you find?

4. Observe that $x_1$ and $x_2$ each fall roughly in the interval $[-1.5, 2.5]$. Returning to the model with $x_2$, make a plot of the SE $(\hat{\mu}(x_1, -1.5))$ versus $x_1$ as $x_1$ ranges from -1.5 to 2.5. Contrast this with a similar plot for SE $(\hat{\mu}(x_1, 2.5))$. What do you observe? Explain this result intuitively based on a plot of $x_1$ versus $x_2$.

5. You have observed in question 2 that the Gram matrix for this fit is of the form

$$(n - 1) \begin{pmatrix}
\frac{n}{n-1} & 0 & 0 \\
0 & 1 & \rho \\
0 & \rho & 1
\end{pmatrix},$$

where $\rho$ is the correlation between $x_1$ and $x_2$. Therefore, the variance-covariance matrix of $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ is of the form

$$\frac{\hat{\sigma}^2}{(n-1)(1-\rho^2)} \begin{pmatrix}
\frac{n-1}{n} & 0 & 0 \\
0 & 1 & -\rho \\
0 & -\rho & 1
\end{pmatrix}.$$
Use this result to explain the effect observed in question 4. What would you expect to happen if $\rho = 0$? What does it mean for $\rho$ to be zero?

6. One way to remove the dependence between $x_1$ and $x_2$ would be to model $x_2$ as a function of $x_1$ and use the residuals from that fit in our model. That is, let

$$\hat{x}_2 = \hat{\gamma}_0 + \hat{\gamma}_1 x_1$$

be the least squares “predictor” of $x_2$, where $\hat{\gamma}_0$ and $\hat{\gamma}_1$ are obtained in the usual manner by the method of least squares. Construct a new variable $x_2^\ast = x_2 - \hat{x}_2$, and repeat the steps in question 2, substituting $x_2^\ast$ for $x_2$ for in the model. The Gram matrix formed from $1$, $x_1$ and $x_2^\ast$ should be of what form? Why? The new variable $x_2^\ast$ ranges over the interval $[-0.7, 2.5]$. Using -0.7 as the low value and 2.5 as the high value, construct plots similar to those from question 4. Does this agree with your answers to question 5?

Next, we will consider just the variables relating to the state of the tank at the time of the experiment, $x_3$ and $x_4$, the tank’s initial temperature and pressure, respectively. Here, we will consider these two variables on their original scale (that is, without standardizing).

7. Make simple plots of $x_3$ and $x_4$ versus observation number, as well as a plot of $x_4$ versus $x_3$. Rather than use these variables as they are, we are going to recode them as indicator functions. That is, we are interested in defining $I_l(x_3)$, $I_m(x_3)$, and $I_h(x_3)$ from $x_3$ so that each takes on only the values of 0 or 1 (here, $l$, $m$, and $h$ refer to values of $x_3$ that are low, medium and high, respectively). For example, we will set

$$I_l(x_3) = \begin{cases} 
1 & \text{if } x_3 < 45 \\
0 & \text{otherwise}
\end{cases}$$

Similarly, we define $I_m(x_3)$ to be 1 if $x_3$ is between 45 and 75, and $I_h(x_3)$ to be 1 only if $x_3$ is larger than 75.
Suppose we make similar definitions for indicator functions based on $x_4$, taking a low value to be 3.7 or below, a medium value to be between 3.7 and 6, and a high value to be larger than 6. Do you anticipate any problems creating a model which includes both of the three indicator functions for each of $x_3$ and $x_4$?

8. Suppose we are interested in the effect that different levels of tank temperature has on hydrocarbon emissions. Using the indicator functions given above, we could propose the following model for the mean of $Y$

$$
\mu(x_3) = \beta_l I_l(x_3) + \beta_m I_m(x_3) + \beta_h I_h(x_3).
$$

Why is the constant function not included in this model? Find estimates of $\beta_l$, $\beta_m$, and $\beta_h$ using ordinary least squares. Because this sort of model gives rise to particularly simple normal equations, it is important in this case that you consider how you would arrive at these estimates directly. Find a 95% confidence interval for $\beta_h$ and $\beta_l - \beta_h$, and interpret these quantities.

9. Consider instead the following model

$$
\mu(x_3) = \beta_0 + \beta_m I_m(x_3) + \beta_h I_h(x_3).
$$

Why does $I_l$ not make an appearance in this model? Using ordinary least squares, find estimates of $\beta_0$, $\beta_m$, and $\beta_h$. Again, it is important that you consider the form of the normal equations for this model, while the actual fitting can be done via the `lm` command. Find a 95% confidence interval for $\beta_h$ and interpret this quantity.

When would an experimenter be more interested in the model as specified in question 8? When is the form presented in this question more useful? Is there any difference in the estimates given by these models?

**Hints.** As suggested above, you should carry out the OLS fits in this lab both directly by matrix calculations and by using the `lm` (linear model) command, until you are sure you understand how (in principle)
all the lm output is obtained. After that, use whatever methods suit you.

Matrices: Your most useful resource will be the Matrices section of the Index to S Functions in the back of the S manual, plus the S function documentation also available at the end of the manual or on-line. Some particularly useful commands:

- \%\%: Matrix multiplication (usage: a\%\%b)
- t: Matrix transpose
- solve: Invert a matrix or solve a matrix equation

Of course, matrices and vectors may be added, subtracted, and multiplied by scalars just like single numbers.

lm: This command has lots of bells and whistles, which you are welcome to explore, but it can be used very simply for the purposes of this lab. Suppose you have column vectors \(x_1\), \(x_2\), and \(y\) of equal length. To “fit the model”

\[
\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2
\]

you could type: `lm1 <- lm(y ~ x1+x2)` . This produces an \textit{lm} object, \texttt{lm1}, that contains everything you might ever want to know relevant to the fit. The argument to \texttt{lm} is called a \textit{formula}; you should think of it as saying, “regress \(y\) on the factors \(x_1\) and \(x_2\).” (Note: you do not need to set up the arguments \(x_1\), \(x_2\), and \(y\) as separate column vectors - it will often be more convenient to have all your regressors [here \(x_1\) and \(x_2\)] as the columns of a matrix, say \(X\), and use the formula \(y \sim X\) .) Notice that an intercept term is included by default: to suppress it, you could use the formula \(y \sim x1+x2-1\). Several functions extract information from \texttt{lm} objects, the most immediately useful being \texttt{summary} (type \texttt{summary(lm1)}). To see a list of other such functions, type \texttt{help(lm.object)}.

Miscellaneous: The functions \texttt{var}, \texttt{cor}, and \texttt{mean} will have obvious uses. \texttt{pairs} provides all pairwise scatter-plots of the columns of a matrix. (To do this for, say, a 3-column matrix, you will probably want to reformat your graphics page with \texttt{par(mfrow=c(3,3))}.)

To read the data into S-PLUS you can use the function \texttt{scan}.
To construct the indicator functions for questions (7-9), you will probably find it easiest to use subscripting. E.g., to form $I_l(x_3)$, you can do something like

```r
I1 <- rep(0,125)
I1[x3 < 45] <- 1
```

Some general advice: you are now entering the fun world of statistical computing, where everything you do once, you will probably do a thousand times. The one key word here is *functions*. Take a few moments to think what the repetitive task will consist of, and write a function to do it. You will save yourself much time and trouble.