5.1 This problem uses the data set P134, given on page 132 of the text. The model (5.6) is on page 138 of the text, and the variables there are: $Y$ (or $y_{ij}$) for the job performance, $X$ (or $x_{ij}$) for the test score, and $Z$ (or $z_{ij}$) as an indicator variable for race (1 for a minority applicant and 0 for a white applicant, defined on page 132).

a. The usual least squares assumptions can be checked as described in the Chapter 4 Homework Supplement.

b. The $F$-test can be conducted by defining a reduced model with $\gamma = 0$ and comparing it to the full model using the \texttt{anova()} function.

c. This test is conducted as part of the \texttt{summary()} output for the full model.

d. You do not need to do this part of the problem, although if the $p$-values in parts (b) and (c) do not agree, you did something wrong.

5.2 This problem uses the data set P134, given on page 132 of the text. The model (5.8) is on page 139 of the text, and the variables are as in Problem 5.1. Parts (a)-(c) are as for Problem 5.1 also, and you do not need to do Part (d) of this problem (although see the note for Part (d) of Problem 5.1).

5.3 This problem uses the data set P143, given on page 142 of the text.

The instructions are essentially telling you to fit and interpret the model:

$$
\text{Sales} \sim \text{PDI} + q,
$$

where $q$ is a factor variable indicating which of the four quarters the data point is for. Since R does things alphabetically by default, it will treat the first quarter as the “control” and define indicator variables for each of the other three quarters (instead of the $Z_1$, $Z_2$, and $Z_3$ that they recommend, indicating the first three quarters and treating the “control” as the fourth quarter).

Before fitting this model though, some preparation is necessary for the data (as is often the case in statistics). Inconveniently enough, the quarters and years are indicated in the same column, which makes it impossible to use that column as a factor indicating the quarter directly. While this can be modified by hand since the data set is small, it would be a hassle to do so. However, R can automate the process quite readily.

In R, we can define a new variable $q$ by:

$$
q = \text{as.factor(substring(P143$Date,first=1,last=2))}
$$

(You might also type $q$ into R to verify that it was defined correctly.) This \texttt{substring} command extracts the characters from position 1 to position 2 in each entry of \texttt{P143$Date}, and the \texttt{as.factor} ensures that it will be treated as a factor in the linear model (which it probably would be anyway, since it contains text, but it is not a bad idea to be safe on this).

Now fit the model above, check the usual regression diagnostics, and interpret the regression output. In your interpretation, you should note that the shift in sales is very similar for some of the groups (point out which ones), suggesting that perhaps a model with two seasons


instead of four might be more appropriate. Create a scatterplot of sales versus PDI with different symbols `pch` (and/or colors) for each quarter, and plot the four different regression lines you obtained in fitting the model (one for each quarter) on this scatterplot with different line types `lty` (and/or colors). This should show very clearly what the two seasons ought to be, in case you didn’t see them solely from the regression output. Be sure to state explicitly which quarters make up the two seasons you arrive at.

5.4 This problem uses three data sets: P145 for 1960, P146 for 1970, and P147 for 1975. This is quite inconvenient for analysis, so you should combine them, making an extra column to indicate which year the data are from. For this, use:

```r
> new.P145 = cbind(P145,year=1960)
> new.P146 = cbind(P146,year=1970)
> new.P147 = cbind(P147,year=1975)
> ed.ex$year = as.factor(ed.ex$year)
```

The first three bind to the original data frames a column (`year`) consisting of a vector full of the appropriate year. The next binds the three new dataframes together rowwise one after the other, and the last converts the `year` column into a factor (which is important, since the default is to treat numbers as numerics).

Now the “thorough analysis” instructions are essentially telling you to fit the model consisting of all terms including interactions and compare it to the model without the `year` factor variable. In other words, the full model is

\[ Y \sim (X_1 + X_2 + X_3)*\text{year}, \]

and the reduced model is

\[ Y \sim X_1 + X_2 + X_3. \]

Remember that the * in a linear model stands for all individual terms to the left and right of the * along with all interactions between those terms. You should see this in looking at a summary of the full model.

For this problem then, you should set up the full model, check the usual regression diagnostics for it, compare the full model to the reduced model (with an \( F \)-test as usual), and interpret the result of this comparison. Keep in mind that the reduced model is one for which there is no change in intercept or slope for the three time periods; the full model is one for which there is a change (in intercept and/or slope). In your interpretation, you should discuss whether or not you detect a statistically significant change in slope and/or intercept among the three periods. (Unfortunately there are too many variables to illustrate what you see with a single scatterplot though.)

5.5 This problem uses the data set P148, given on page 147 of the text.

a. It would probably be easier to do this problem without creating indicator variables and just allowing R to do so by making the Fertilizer Type into a factor. However, it is good to know how to make indicator variables anyway, so you should follow their instructions. Doing so is straightforward in R. For example, try

\[ f1 = \text{as.numeric(P148$Fertilizer == 1)}, \]

and similarly for the others.
b. As usual, “fit the model” means “find the regression coefficients”.

c. Recall that such an $F$-test is part of the model summary. Be sure to state the null and alternative hypotheses being tested, the test used, and your conclusions at the 5% significance level, as they ask.

d. For this, you should create a suitable reduced model and conduct an $F$-test as usual.

e. This can, of course, be read from the regression coefficients of the summary (taking into account the variability in those coefficients, as indicated by their standard errors or by confidence intervals for them).