Homework 4
Due: Reading Due Thursday, October 4, 2001
Technical Exercises Due Tuesday, October 9, 2001

- **Reading.** I have been skipping around a bit relative to the organization in RPD. For regression diagnostics (plotting residuals and the leverage values h_{ii}, see section 11.1 (pp. 341–359). Some general background is provided in Ch 10. I have also been discussing

  - The formal calculation of F statistics from differences of sums of squares (Ch 4, pp. 126–135), and
  - Dummy coding of categorical variables (Ch 9, pp. 269–277).

  However, it might be better to wait to read these until we have discussed quadratic forms and construction of the F-tests for typical linear regression hypotheses (since the RPD discussion flows from the theory much more than from the application, and right now I am emphasizing basic tools for applications in class).

- **On Writing.** Please read the following, papers, to be handed out in class Tuesday:

  - Alley, Chapter 2: Organizing your documents.
  - Huth, Chapter 4: The Critical argument and the structure of the scientific paper.
  - Spurrier, Chapter 12: Strategies for effective written reports.

  Next week we will examine one or more papers with the ideas of these chapters in mind; but for this week please use the chapters themselves as examples to see if the writers satisfy their own rhetorical advice. The papers will be available in class on Tuesday, and after that they will be available in the box by my office door. *To be discussed on Thursday.*

- **Problems to be turned in:**

1. (a) Show the following facts, in multiple regression (use facts discussed in lecture to save yourself work, where possible!):
   - $RSS = Y'(I - H)Y$, where I is an $n \times n$ identity matrix and H is the hat matrix.
   - $SYY = Y'(I - \frac{1}{n}J)Y$, where J is an $n \times n$ matrix filled with all 1's.
   - $SS_{Reg} = Y'(H - \frac{1}{n}J)Y$.
   - For any $n \times n$ matrices A and B, $Y'(A + B)Y = Y'AY + Y'BY$.

   and therefore $SYY = SS_{Reg} + RSS$ in multiple regression.

   (b) The trace of a square matrix $S = [s_{ij}]$ is the sum of diagonal elements, $tr(S) = \sum_i s_{ii}$.

   i. Show algebraically that for any $m \times n$ matrix A and any $n \times m$ matrix B, that $tr(AB) = tr(BA)$.

   ii. Use part (i) to show that when $X'X$ is invertible, the trace of the hat matrix $H = X(X'X)^{-1}X'$ is the number of regression coefficients, i.e. $tr(H) = k + 1$.

Homework 4

2. [continuing from problem #6, hw03] Last week we fit the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon,$$

where Y is the auction price, X_1 is the age of the clock in years, and X_2 is the number of bidders, to explore the relative importance of these two variables on aution price, using the data in `clocks.dat` in the 707 web area.

Now, suppose the collector, having observed many auctions, believes that the *rate of increase* of the auction price with age will be driven upward by a large number of bidders. In other words, he believes that there is an *interaction* between age and the number of bidders and hence that the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

is suitable.

   (a) State two ways in which you might test the hypothesis that the price–age slope increases as the number of bidders increases. Perform one of these tests. Would your results differ if you had used the other possible test? Why or why not?

   (b) Write a brief (one or two paragraph) report to the collector, summarizing your findings after performing the series of regression analyses described above. Remember, this collector is not a statistician, so keep your summary non–technical.

Notes/Hints: There are (at least) three ways to include the term X_1X_2 (called an "interaction term") in a linear model in Splus:

   - Construct a new variable `X3 <- X1*X2` and run the regression `lm(Y ~ X1 + X2 + X3)`;
   - Let Splus construct X3, via its special ":" notation for constructing interaction terms: `lm(Y ~ X1 + X2 + X1:X2)`;
   - Let Splus construct the whole model, via its special "*" notation for constructing an interaction and including all lower-order terms: `lm(Y ~ X1*X2)`.

Try all three and compare the results *(you do not have to turn anything in on this comparison, but you should try it).*

3. C. Davis obtained the data on measured and reported heights and weights of 200 Canadians engaged in regular exercise in `Davis.dat` and `Davis.cbk` in the 707 data area. In this exercise we will concentrate on the variables:

```
[2] Sex of subject:  M = male  F = female
[3] Measured weight in Kg
[5] Reported weight in Kg
```

Homework 4

Create a subset of the data as follows:

```
davis _ read.table("Davis.dat",na.strings="?")  # why na.strings??
davis20 _ davis[1:20,c(2,3,5)]
names(davis20) _ c("Sex","Meas.Wt","Rept.Wt")
```

This subset will be used for the illustrations below.

(a) *Numerical values of dummy variables.* In this problem we consider the simpler regression model `Meas.Wt ~ Sex`. We may

- Code a dummy variable as
  ```
  > DSex _ ifelse(davis20$Sex=="F",1,0)
  > DSex
   [1] 0 1 1 0 1 0 0 0 0 0 0 1 1 1 1 1 0 1 0 1
  ```
  and then do the regression `lm(Meas.Wt ~ DSex, data=davis20)`; or
- Code the dummy variable as
  ```
  > CSex _ ifelse(davis20$Sex=="F",1,-1)
  > CSex
   [1] -1  1  1 -1  1 -1 -1 -1 -1 -1 -1  1  1  1  1  1 -1  1 -1  1
  ```
  and then do the regression `lm(Meas.Wt ~ CSex, data=davis20)`; or
- `Splus` treats all non-numeric variables as "factors", for which it automatically codes dummy variables. To see what the result of Splus's dummy coding is, try the regression `lm(Meas.Wt ~ Sex, data=davis20)`.

i. Try all three regressions above and compare the results. *Turn in answers only to the following questions:* What parts of the regression analysis do the numerical values of the dummy variable affect? What parts do the numerical values not affect?

ii. Perform a test for whether Sex should be in this model. Do you "believe" the results of this test? Why or why not?

(b) *Testing for outliers.*

i. Regress reported weight (kg) on measured weight (kg); provide appropriate numerical and graphical summaries and diagnostics, and write a sentence describing any interesting features of the regression.

ii. You should have noticed an outlier in part (i). This particular outlier is pretty clear without further analysis but we will illustrate a way of testing for outliers using dummy variable coding. Create a dummy variable as follows:
```
> dim(davis20)
[1] 20  3
> D12 _ rep(0,length=20)
> D12[12] _ 1
> D12
 [1] 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
```

Compare the models below using one or more appropriate statistical tests.

```
Rept.Wt ~ Meas.Wt
Rept.Wt ~ Meas.Wt + D4
```

Explain why comparing these models is a test of whether observation 12 is an outlier.

4. Anscombe (1981) presents and analyzes the data on the US states and Washington DC given in `Anscombe.dat` (see also `Anscombe.cbk` in the data area of the 707 web pages. The variables are:

```
[1] STATE
[2] PER-CAPITA EDUCATION EXPENDITURES, dollars
[3] PER-CAPITA INCOME, dollars
[4] PROPORTION UNDER 18, per 1000
[5] PROPORTION URBAN, per 1000
```

   (a) Separately fit the simple regressions for per-capita school expenditures on each of the variables income, proportion under 18, and proportion urban. For each regression provide (i) estimates of the regression parameters and their standard errors, (ii) r, R^2 or F, (iii) a small set of appropriate diagnostic plots, and (iv) one sentence summarizing the evidence you have provided for goodness of fit of the simple regression model.

   (b) Now fit the mulitple regresssion of school expenditures on all three other variables; providing the same 4 items as in part (a). In addition, provide one or two sentences comparing the results of this mulitple regression with the three simple regressions you fit in part (a).

5. [Continuation.] Consider the general regression model

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

An alternative procedure for calculating the least-squares coefficient $\hat{\beta}_1$ is as follows:

- Regress y on $X_2, \ldots, X_k$
- Regress X_1 on $X_2, \ldots, X_k$
- Regress the residuals $e_{y \mid x_2 \cdots x_k}$ on the residuals $e_{x_1 \mid x_2 \cdots x_k}$

   (a) Apply this procedure to the data set of problem #4 to obtain the coefficient for income. Is it the same as in the multiple regression in problem #4?

   (b) Illustrate the regression in part (a) with an appropriate scatter plot, and comment on the usefulnesss of X_1 in the model, based on this scatter plot (sometimes called an "added variable plot" or "partial regression plot").

   (c) Notice that the intercept in your regression in part (a) is 0. Why is that?

Homework 4

Note: In light of this procedure it is reasonable to describe $\hat{\beta}_1$ as the "effect of X_1 on y after removing the influence of $X_2, \ldots, X_k$", and one often speaks of tests of $H_0: \beta_1 = 0$ as testing "whether X_1 has any effect on y after removing influence of $X_2, \ldots, X_k$".

6. [Continuation.]

   (a) Use the output from `anova()` to verify the value of the F statistic quoted in `summary()`, for the multiple regression of problem #4.

   (b) Use the Extra Sum of Squares Principle to test the hypothesis $H_0: \beta_2 = \beta_3 = 0$ from the output in `anova(myfit.514)`. Note that in Splus, `pf(2.46,3,55)` gives $P[F_{3,55} \leq 2.46]$.

   (c) An alternative approach in Splus is

   ```
   small <- lm(expend ~ income, data=anscombe)
   big <- lm(expend ~ income + under18 + urban, data=anscombe)
   anova(big,small)
   ```

   Do you get the same answer as in part (b)?

7. Return to Angell's data on the moral integration of US cities (see `Angell.dat` and `Angell.cbk` in the 707 data area). Construct dummy variables to represent the four regions East, Midwest, West and South. Regress moral integration on ethnic heterogeneity, geographic mobility and region. Use an appropriate F test to test the null hypothesis of no region effects. Attach appropriate summary and diagnostic information (and/or plots) from Splus.

   Notes/Hints: Construct indicators (dummy variables) by hand for each region using commands like

   ```
   > D1 _ ifelse(Reg=="E",1,0)
   > D2 _ ifelse(Reg=="MW",1,0)
   > D3 _ ifelse(Reg=="W",1,0)
   ```

   (Why not construct `> D4 _ ifelse(Region=="S",1,0)` also?)

8. [Continuation.] Continuing with the `angell.dat` data set, compare the models

   ```
   Int ~ Mob
   Int ~ Mob + D1 + D2 + D3
   Int ~ Mob + D1 + D2 + D3 + Mob:D1 + Mob:D2 + Mob:D3
   ```

   following the outline of the handout on Cyril Burt's "Twins" data. To get a visual sense of what you are doing, begin by making the following plot:

   ```
   > plot(angell$Mob,angell$Int,type="n")
   > text(angell$Mob,angell$Int,labels=as.vector(angell$Reg))
   ```