Chapter 11 of *The Sleuth* is about model checking and refinement. New concepts in the chapter include leverage and influence. Leverage is a function of the explanatory variables alone and measures the potential for a data point to affect the model parameter estimates. Influence is a measure of how much a data point actually does affect the estimated model. Leverage and influence both may be defined in terms of matrices. This handout will give the matrix description of leverage and influence not found in *The Sleuth* and will provide the R commands to compute them.

I will use the first case study of the chapter on alcohol metabolism in men and women as a running example. The response variable is *metabol*, the difference in alcohol metabolism when injected directly into the bloodstream as compared to when consumed orally and passed through the stomach first. The single quantitative explanatory variable is *gastric*, a measure of the activity of enzymes in the stomach that partially metabolize alcohol. There are two categorical variables, *SEX* and *ALCOHOL*, the second of which indicates whether or not the individual is an alcoholic. The following R commands (with the output supressed) replicate the scatter plot in Display 11.2 and the residual plot in Display 11.7. The residual plot includes two smoothed local regression lines, one using all the data and a second that excludes two outliers. Examination of a residual plot in this case can pick out two points that may be influential, the 31st and 32nd observations.

```
> case1101 = read.table("sleuth/case1101.csv", header = T, sep = ",")
> attach(case1101)
> female <- SEX == "FEMALE"
> male <- !female
> alc <- ALCOHOL == "ALCOHOLIC"
> nalc <- !alc
> plot(GASTRIC, METABOL, type = "n")
> points(GASTRIC[female & nalc], METABOL[female & nalc], pch = 17)
> points(GASTRIC[female & alc], METABOL[female & alc], pch = 16)
> points(GASTRIC[male & alc], METABOL[male & alc], pch = 1)
> points(GASTRIC[male & nalc], METABOL[male & nalc], pch = 2)
> legend(1, 12, c("Female nonalcoholic", "Female alcoholic", "Male nonalcoholic",
+     "Male alcoholic"), pch = c(17, 16, 2, 1))
> fit <- lm(METABOL ~ GASTRIC * SEX * ALCOHOL)
> plot(fitted(fit), residuals(fit))
> abline(h = 0, lty = 2)
> keep <- fitted(fit) < 7
> lines(lowess(fitted(fit), residuals(fit)), lty = 3)
> lines(lowess(fitted(fit)[keep], residuals(fit)[keep]), lty = 2)
```

The `subset` option `lm` includes only part of the observations from the analysis. Here we exclude the two outliers, both observations with *gastric* more than 4, and compare the two summaries, as in Display 11.9.

```
> fit2 <- lm(METABOL ~ GASTRIC * SEX * ALCOHOL, subset = (GASTRIC <
+     4))
> summary(fit)
> summary(fit2)
```

Section 11.4.1 gives a formula for the leverage of an observation, a measure of the distance of its explanatory variable values from the other observations, in the simple linear regression case. Here is the matrix representation. Recall from earlier in the course that we set up a regression problem with a response vector Y, a design matrix X that included a column for an intercept, one column for each quantitative explanatory variable, and $\ell - 1$ columns for each categorical variable with ℓ levels. The model is $Y = X\beta + \epsilon$ where β is the vector of regression coefficients and ϵ is the vector of unobserved errors. The least squares estimate of β is $\hat{\beta}$ and satisfies $\hat{\beta} = (X^T X)^{-1} X^T Y$ where the T superscript indicates matrix transposition. The fitted values are then

$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY$$

for the *hat* matrix $H = X(X^T X)^{-1} X^T$. H is an $n \times n$ matrix that orthogonally projects vectors into the space spanned by the columns of X. The leverage of the ith observation is the ith element of the diagonal of H, that we will call h_i. Individual leverage values are always at least $1/n$ and the average of them all is p/n if there are p regression coefficients.

In R, we can use the function `model.matrix` to return the model matrix from a fitted linear model. (This is useful if you want to see exactly how categorical variables are parameterized with dummy variables.) There is a function called `hat` that

returns the diagonal of the hat matrix from the model matrix. Here we compute the leverage for each observation and plot it versus observation number with a vertical line drawn from the x-axis to the leverage. (The `type="h"` command makes this type of plot.)

```
> h <- hat(model.matrix(fit))
> plot(h, type = "h")
```

The plot created with this command does not, however, indicate that observations 31 and 32 have the most leverage. Observations 1 and 23 have even more leverage. Let's see if we can understand this.

```
> case1101[c(1, 23), ]
```

```
   SUBJECT METABOL GASTRIC    SEX   ALCOHOL
1        1     0.6     1.0 FEMALE ALCOHOLIC
23      23     3.7     2.7   MALE ALCOHOLIC
```

Both of these observations are on alcoholics, of which there are only a few of each sex. Let's look at all of these points.

```
> case1101[alc, ]
```

```
   SUBJECT METABOL GASTRIC    SEX   ALCOHOL
1        1     0.6     1.0 FEMALE ALCOHOLIC
2        2     0.6     1.6 FEMALE ALCOHOLIC
3        3     1.5     1.5 FEMALE ALCOHOLIC
19      19     1.5     1.3   MALE ALCOHOLIC
20      20     1.9     1.2   MALE ALCOHOLIC
21      21     2.7     1.4   MALE ALCOHOLIC
22      22     3.0     1.3   MALE ALCOHOLIC
23      23     3.7     2.7   MALE ALCOHOLIC
```

We have fit a model with *gastric*, *sex*, and *alcohol* along with of the two- and three-way interactions. This, in effect, is the same as fitting four separate regression lines, one for each sex/alcohol combination (but there is a common estimate of σ so the SEs are different than a one group at a time analysis). Notice that there is one *gastric* measurement quite different from the rest for both the female and male alcoholics. These points are the high leverage observations, 1 and 23.

Now, the scatter plot and the summaries indicate that *alcohol* does not seem to have much of an effect. I fitted a model that took out all of the interactions involving *alcohol*, but left in *alcohol* as a main effect along with *gastric* and *sex* and their interaction. This model fits separate lines for each sex with a common offset in the intercept for alcoholics of each sex.

```
> fit3 <- lm(METABOL ~ GASTRIC * SEX + ALCOHOL)
> summary(fit3)
> plot(hat(model.matrix(fit3)), type = "h")
```

The summary indicates that we can ignore *alcohol* all together. This plot shows that observation 31 has a lot of leverage, but so does observation 17.

```
> case1101[17, ]
```

```
   SUBJECT METABOL GASTRIC    SEX       ALCOHOL
17      17     2.5       3 FEMALE NON-ALCOHOLIC
```

We see here that observation 17 is the woman with the highest *gastric* measurement of 3, whereas all the other women have gastric measurements between 0.8 and 2.2. In a model that fits women separately, the 17th observation has the potential to be influential. It uses

Here is some R code to make plot analogous to that in Display 11.12 for our original model. It uses the R functions `cooks.distance` to find Cook's Distances, `hat` to find leverages, and `rstudent` to find studentized residuals.

```
> par(mfrow = c(3, 1))
> plot(cooks.distance(fit), type = "h")
> title("Model with Gastric, Sex, Alcohol, and all interactions")
> plot(hat(model.matrix(fit)), type = "h")
> plot(rstudent(fit), type = "h")
```

Here is R code to replicate Display 11.12.

```
> fit4 <- lm(METABOL ~ GASTRIC * SEX)
> par(mfrow = c(3, 1))
> plot(cooks.distance(fit4), type = "h")
> title("Model with Gastric, Sex, and their interaction")
> plot(hat(model.matrix(fit4)), type = "h")
> plot(rstudent(fit4), type = "h")
```

Finally, here are the plots after removing the 31st and 32nd observations and not using *alcohol* as an explanatory variable.

```
> fit5 <- lm(METABOL ~ GASTRIC * SEX, subset = (GASTRIC < 4))
> par(mfrow = c(3, 1))
> plot(cooks.distance(fit5), type = "h")
> title("Model with Gastric, Sex, and their interaction, dropping outliers")
> plot(hat(model.matrix(fit5)), type = "h")
> plot(rstudent(fit5), type = "h")
```