36-350: Data Mining

Lab 9
Date: October 24, 2003 **Due: end of lab**

Interspersed throughout this lab are questions that you will have to answer at check-off.

1. Download the files for this lab from the course web page to the desktop:

   `http://www.stat.cmu.edu/~minka/courses/36-350/lab/`

2. Open a Word or Notepad document to record your work.

Start R

3. `Start -> All Programs -> Class software -> R 1.7.0`

4. Load the special functions for this lab:

   `File -> Source R code...`

   Browse to the desktop and pick `lab9.r` (it may have been renamed to `lab9.r.txt` when you downloaded it). Another window will immediately pop up for you to pick the `mining.zip` file you downloaded.

The dataset

5. The dataset is 196 weeks of grocery sales, similar to that used in class, but for a different store. The variables are:

Price.1	DOLE PINEAPPLE ORANG 64 OZ
Price.2	FIVE ALIVE CTRUS BEV 64 OZ
Price.3	HH FRUIT PUNCH 64 OZ
Price.4	HH ORANGE JUICE 64 OZ
Price.5	MIN MAID O J CALCIUM 64 OZ
Price.6	MIN MAID O J PLASTIC 96 OZ
Price.7	MM PULP FREE OJ 64 OZ
Price.8	SUNNY DELIGHT FLA CI 64 OZ
Price.9	TREE FRESH O J REG 64 OZ
Price.10	TROP PURE PRM HOMEST 64 OZ
Price.11	TROP SB HOMESTYLE OJ 64 OZ
Sold.4	Number of units sold for HH ORANGE JUICE 64 OZ

   It is stored in a matrix `x`. Look at the first few rows via `x[1:3,]`.

Standardizing

6. Transform `Sold.4` appropriately (as in lab 5) and standardize all variables to have zero mean and unit variance.

Adding predictors

7. Construct a linear model to predict `Sold.4` as a function of `Price.4`.

8. Plot all variables versus the residuals of this model. *Which other variables appear important to* `Sold.4`*?* (There should be about four.)

9. Pick one of the important predictors from the last step and include it in a new model. *In light of the new model, which other variables are important? Have any ceased to be important, or become more important, due to the change?*

10. Add another important predictor and answer the same questions.

11. Keep adding predictors until no more predictors seem useful. Report your final model, with the predictors in the order that you added them.

Automatic selection

12. Starting again from a model with only `Price.4`, use `step` to add predictors. Save the coefficients of the fit for the homework. *How is the resulting model different from the one you constructed?*

13. Make a partial residual plot for the model from `step`. Keep a copy for the homework.

14. (Optional) The homework asks about predictors which can removed from the model. You may find it useful to also save the p-values, or construct a condensed model using `lm`, and see how it differs.

15. You can now get checked off.

Linear regression If `x` is a matrix with response `r` and predictors `p1,p2`:

```
fit = lm(r ~ p1 + p2, x)
summary(fit)
```

The model is in `fit`. `summary` shows p-values and other information.

Residual plots If `fit` is a model and `x` is a matrix:

```
predict.plot(fit,x)
predict.plot(fit,x,layout=c(4,3))
predict.plot(lm(r ~ p1 + p2,x),x)
```

The second version forces a 4×3 layout of panels. The third version fits a model and plots residuals in one step (useful for quickly comparing models).

Partial residual plots

```
predict.plot(fit,partial=T)
```

Automatic selection Given an existing model `fit` and a matrix `x` of potential predictors:

```
fit = step(fit,formula(x))
```

It keeps adding and removing predictors until it finds a model with small AIC. (`formula(x)` automatically creates a formula involving all predictors in `x`.)