36-350: Data Mining

Lab 9
Date: October 25, 2002 **Due: end of lab**

---

1 Introduction

This lab teaches you how to construct a linear model relating the variables in a dataset.
There are 4 questions. For each one, submit your commands and a response from R demonstrating that they work. (Only hand in commands relevant to the question.) To submit a plot, click on the plot window and select

```
File -> Save as -> Postscript...
```

This saves the plot to a file which can be printed, incorporated into a Word document, or mailed to us as an attachment.

2 Starting R

Start R as in lab 1. On the class web page, go to "computer labs" and download the files for lab 9 into your work folder. Read the special functions into your running R application via the commands

```
source("lab5.r")
source("lab9.r")
```

If this fails, check that the files were downloaded correctly.

3 The data

The dataset used in this lab 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

Your job is to determine how the price variables affect `Sold.4`.
Load this data via

```
load("lab9.rda")
```

This defines a matrix called `x`.

4 Standardizing

At this point, it would be good to get out lab 7 since many of the commands will be similar. For example, you should start by making histograms:

```
hist.data.frame(x)
```

The response variable `Sold.4` needs to be transformed for symmetry—the others can be left alone.

> **Question 1:** Submit code to transform `Sold.4` and standardize the variables to have zero mean and unit variance. `sx` should end up with the standardized data.

5 Linear regression

The first step in building a linear model is to make pairwise scatterplots against the response variable. This is done with `predict.plot`, which adds loess curves for convenience:

```
predict.plot(sx)
```

You should find that `Price.4` is the most important single predictor.
The function to fit a linear model is `lm`, which is called the same way as `loess`. You provide a formula and a dataset, and it returns an object representing the fit:

```
fit <- lm(Sold.4 ~ Price.4,sx)
```

This command gives various information about the fit:

```
summary(fit)
```

However, you won't be using this information—you will use visualizations instead.
Linear models are visualized via residuals or partial residuals. `predict.plot` can do both:

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

The first line plots the residuals of the fit against the predictors in the model. This is useful for checking whether a linear fit is appropriate for the data. If it is appropriate, the residuals should have no structure with respect to the predictors in the model (the loess curve should be flat). The second line plots the partial residuals of the fit without each predictor versus the predictors in the model. You will use this later to decide which predictors to drop.

6 Adding predictors

To plot the residuals of the model versus all of the variables in `sx`, provide `sx` as a second argument:

```
predict.plot(fit,sx)
```

This tells you which variables to add to the model. The variables currently in the model are bold.

Question 2: (a) When the model only contains `Price.4`, which additional predictors appear relevant to predicting `Sold.4`? (There should be about four.) (The homework uses 'influential' instead of 'relevant'. They mean the same thing.)
(b) Pick one of the relevant predictors from question 2 and use `lm` to construct a new model with it included. Plot the residuals of this new model against the other variables. (You do not need to turn in the plot, only your code for making the plot and the new model.) Which predictors are relevant now? Have any ceased to be relevant? Have any become more relevant?
(c) Add another relevant predictor, make another plot, and answer the same questions from part (b).
(d) Keep adding predictors until no more predictors are relevant. For this part, just report your final model, with the predictors in the order that you added them. (We understand that it gets difficult to make relevance judgements near the end, and the grading will reflect this.)

There is also a command which tries to add predictors automatically. It is called `step`, and called like this:

```
fit <- step(fit,formula(sx))
```

It keeps adding and removing predictors until it finds a model with small AIC value. The result should not be perceived as the 'right' answer, only the computer's best guess at the answer. The quality of this guess can be influenced by many things, especially nonlinearity and outliers.

Question 3: Starting again from the model with only `Price.4`, use `step` to add predictors. Type `fit` or `summary(fit)` to get the coefficients of the fit, and save them for the homework. The set of predictors obtained this way should be different from the ones you chose above. How is it different? Do you think it is better, worse, or just as good?

7 Removing predictors

To decide which of the two models from the previous section is better, or if they can be improved, it helps to plot partial residuals and see if any predictors can be dropped. This is done in one of two ways:

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

The first line focuses on dropping predictors in the model. The second line shows partial residuals for predictors in the model, and ordinary residuals for variables not in the model, which allows you to either drop or add variables.

Question 4: Use `predict.plot` to show partial residuals for the model from `step`. Turn in your plot and keep a copy for the homework.

The homework asks about predictors which can removed from the model. You may find it instructive to construct a model using `lm` that has one less predictor, and see how it differs.