36-350: Data Mining

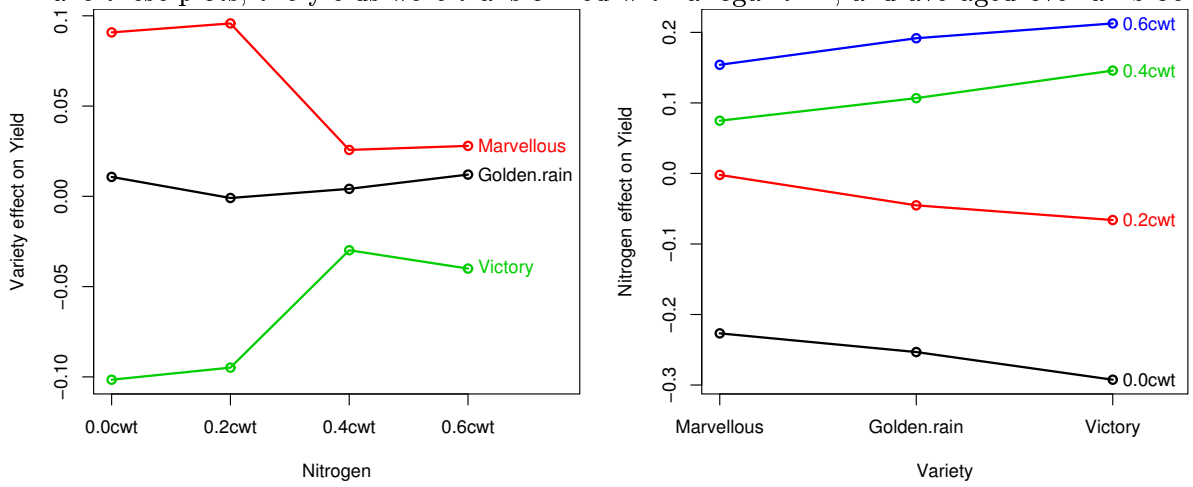
Homework 13
Date: November 20, 2002 **Due: start of class November 25, 2002**

---

1. An experiment was conducted to study how nitrogen treatment affects the growth of oats. A field was divided into six blocks, each block divided among three varieties of oats, and each variety split into four sub-plots receiving different levels of nitrogen treatment. The yield of each sub-plot was recorded, giving a data frame with entries like

```
Block     Variety Nitrogen Yield
   II Golden.rain   0.2cwt   108
   IV Golden.rain   0.0cwt    64
   IV  Marvellous   0.4cwt   104
   VI  Marvellous   0.6cwt   121
   IV     Victory   0.6cwt   122
...
```

   Below are row plots of Variety effect versus Nitrogen and Nitrogen effect versus Variety. To make these plots, the yields were transformed with a logarithm, and averaged over all blocks.



   (a) If the predictors were additive, what would that tell us about the dependence of Yield on Variety and Nitrogen (in non-technical terms)?

   (b) Describe, in a simple way, the interaction between Variety and Nitrogen in predicting Yield.

   (c) Variety is a categorical variable in the above data frame. Give a new data frame which uses a numeric indicator code for Variety.

   (d) Give a new data frame which uses a numeric effect code for Variety.

2. To improve advertising efficiency, a large survey was conducted to find out where people in different demographic groups are likely to get their news. The media categories were:

N_ NEWS	national newspaper
R_ NEWS	regional newspaper
MAGAZ	magazines
TVMAG	TV magazines
TV	TV news
RADIO	radio news

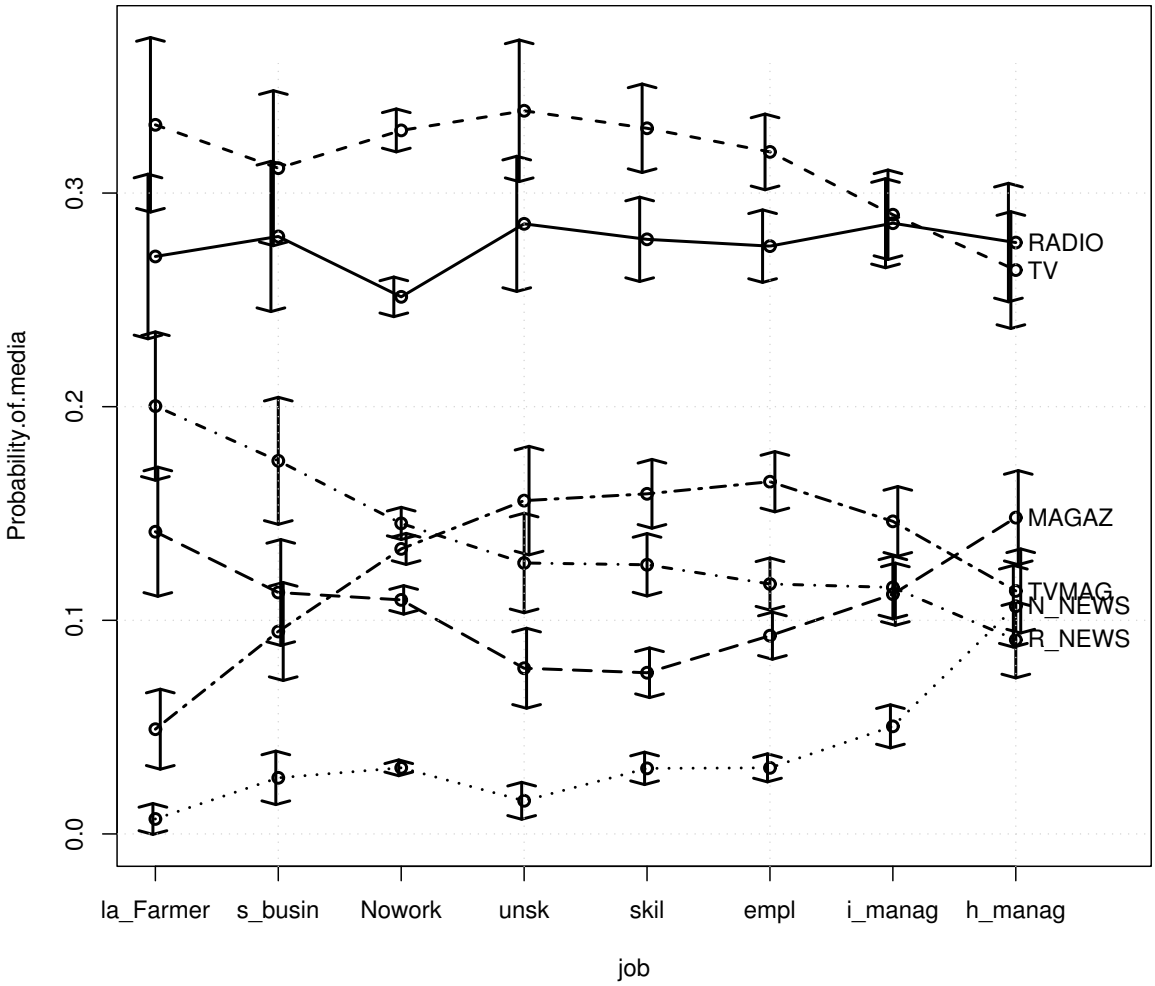
The job categories were:

h_ manag	high-level manager
i_ manag	intermediate-level manager
empl	employer
s_ busin	small business employee
skil	skilled labor
unsk	unskilled labor
la_ Farmer	farmer
Nowork	unemployed

The results are summarized by the following contingency table of job versus media:

	job							
media	la_Farmer	s_busin	h_manag	i_manag	empl	skil	unsk	Nowork
RADIO	96	122	193	360	511	385	156	1474
TV	118	136	184	365	593	457	185	1931
N_NEWS	2	11	74	63	57	42	8	181
R_NEWS	71	76	63	145	217	174	69	852
MAGAZ	50	49	103	141	172	104	42	642
TVMAG	17	41	79	184	306	220	85	782

Below is a row probability plot of this table, with "1.64" error bars:



Note that columns are not ordered the same way in this plot as they are in the table. They have been ordered to make the curves smooth.

(a) If media were independent of job, what would this graph look like?

(b) Where are the three most significant deviations from independence?

(c) Describe the trend for national newspapers vs. regional newspapers.

(d) Describe the trend for magazines vs. TV magazines.

(e) To simplify the description for ad executives, we may want to merge some of the job categories. Find three job categories which could be merged into one, without losing much information about media.

3. In the computer lab, you used slice plots to propose interaction terms for a linear regression.

(a) Which interactions ended up being used in the model?

(b) Interpret each coefficient in the linear model. What does it say about car safety?