36-350: Data Mining

Handout 16
October 20, 2003

Linear regression for marketing research

Linear regression—Modeling the response as a linear combination of the predictors. Particularly useful when you are looking for small effects, and want a simple description of how each variable affects the response. (Regression trees are the opposite.)

The importance of a predictor is typically measured by the p-value of the hypothesis that the coefficient is zero. A small p-value means this hypothesis should be rejected (the predictor is important). Each p-value assumes that the other predictors are in the model. This creates the danger that if two predictors are important but the same, they will both have large p-value, making it look like they are not important.

Marketing research—Modeling sales based on prices, promotions, and product features.

The price of A usually has a negative impact on the sales of A. If the price of another product B also has a **negative** cofficient, then it **cooperates** with A. If the price of B has a **positive** cofficient, then it **competes** with A. In other words, **same sign = cooperates, different sign = competes**.

The signs of the coefficients can be distorted if the predictors are highly correlated. For example, if A is the same as B then $0.4A - 0.5B = -0.2A + 0.1B$. Removing the redundant predictors will give correct signs.

Cooperation is well known in marketing. It is why stores put milk in the back. It isn't necessarily symmetric. If B helps sales of A but A doesn't help sales of B, then A is somehow a more `preferable` product. A preference ordering on the products can thus be deduced from the coefficients of every sales/price combination.

References

[1] Sales data for "Dominick's Finer Foods."
`http://gsbwww.uchicago.edu/kilts/research/db/dominicks/`

[2] GSIA course on Marketing Research (45-822). `http://www.gsia.cmu.edu/45-822/`

Grocery data

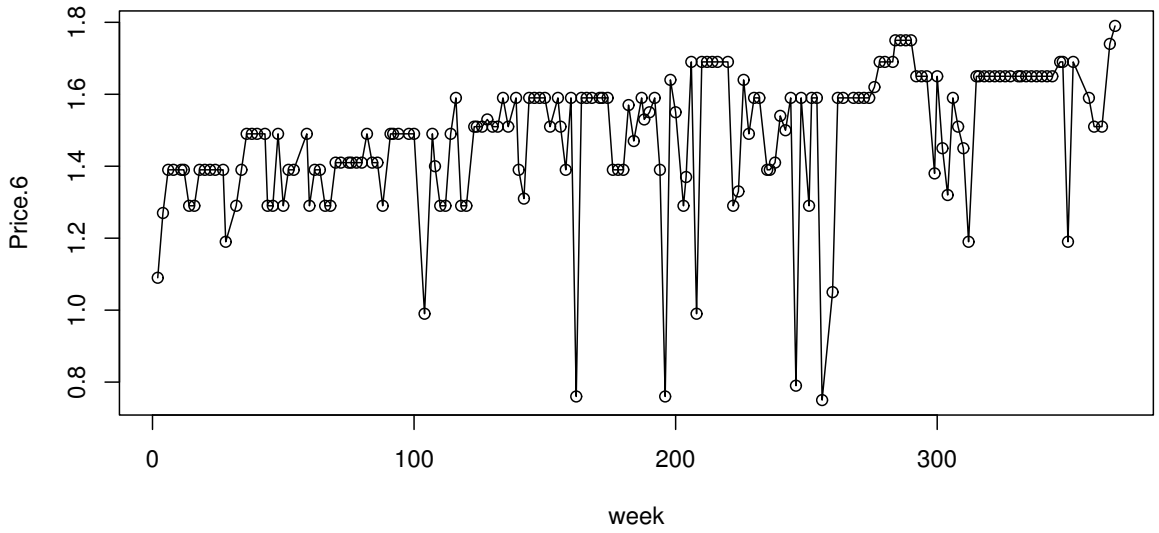
This dataset is a large, detailed record of the pricing and sales at a Chicago grocery store. The data spans five years from 1989 to 1994 and covers all 100 stores in the chain with over 3500 different products represented. The original data consists of individual product/week records of the form:

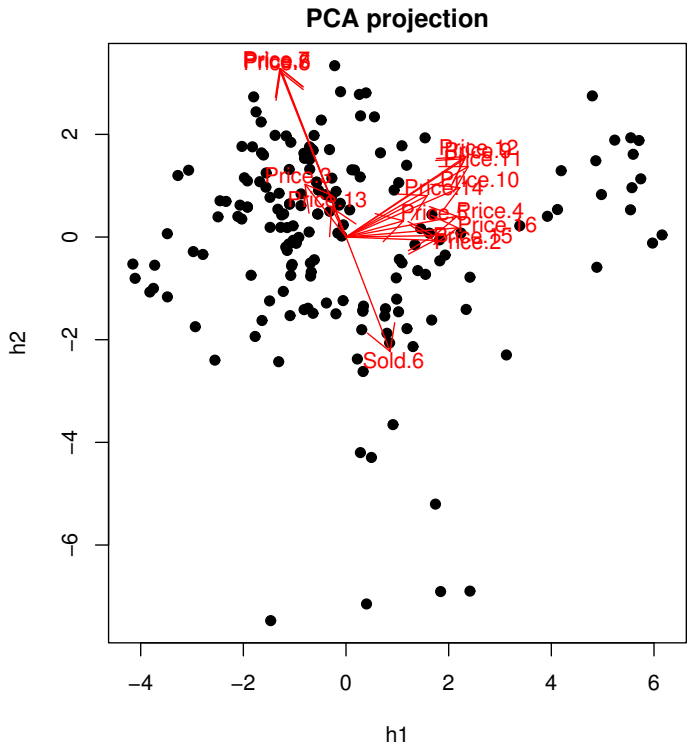
STORE	Store number
WEEK	Week number
UPC	UPC number of product
MOVE	Number of units sold
PRICE	Retail price

But this form is not easy to work with.

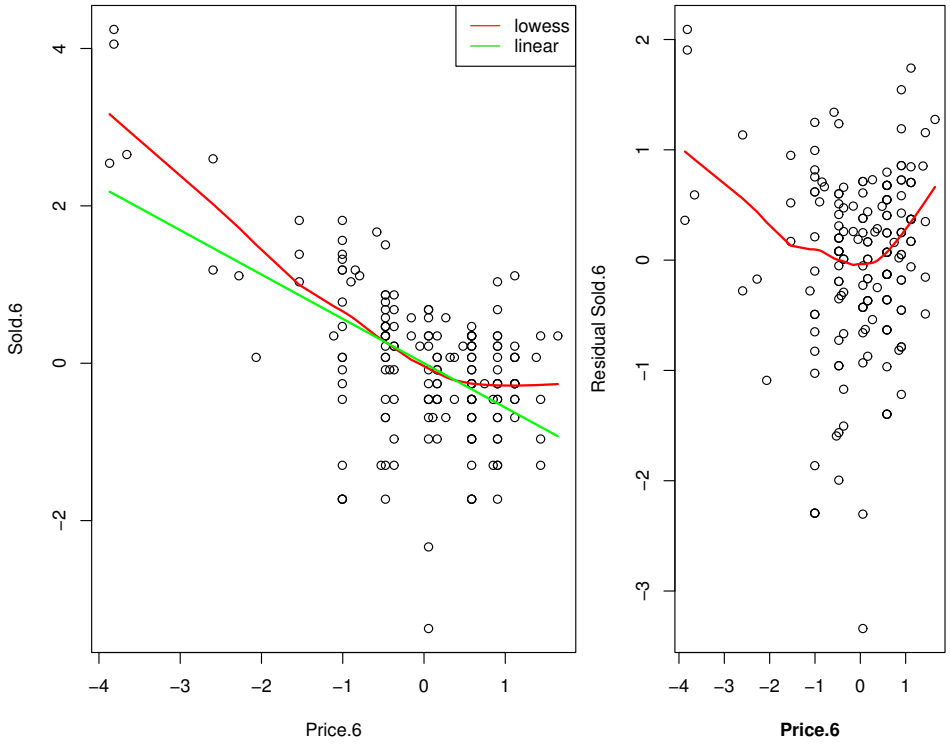
In the new format, all data for a given week appears together. The data was also restricted to 12 refrigerated juice products for one store.

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 128 OZ
Price.5	HH ORANGE JUICE 64 OZ
Price.6	MIN MAID CITRUS PUNC 64 OZ
Price.7	MIN MAID FRUIT PUNCH 64 OZ
Price.8	MIN MAID LEMONADE 64 OZ
Price.9	MIN MAID O J CALCIUM 64 OZ
Price.10	MIN MAID O J PLASTIC 96 OZ
Price.11	MIN MAID O J REGULAR 64 OZ
Price.12	MIN MAID OJ CNTRY ST 64 OZ
Price.13	SUNNY DELIGHT FLA CI 64 OZ
Price.14	TROP PURE PRM OJ 32 OZ
Price.15	TROP PURE PRM OJ 64 OZ
Price.16	TROP PURE PRM OJ 96 OZ
Sold.6	Number of units sold for MIN MAID CITRUS PUNC 64 OZ





This figure shows how the prices typically vary. Two price clusters are evident, as well as a high sales cluster. `Price.6,7,8` move together so we won't be able to distinguish their effects. `Price.6` is highly correlated with `Sold.6`—a lower price leads to higher sales.



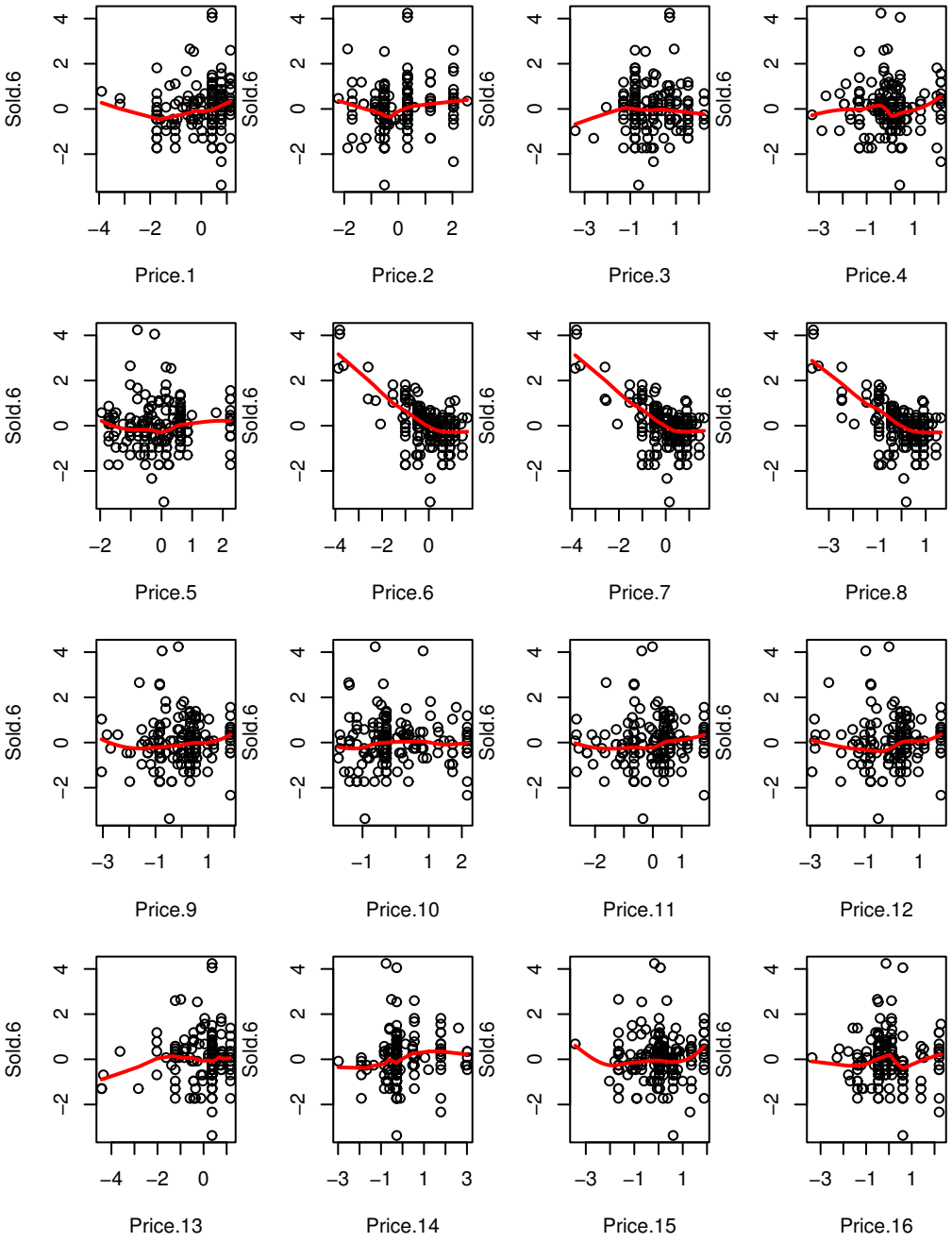

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.948e-17  6.280e-02 3.1e-16        1
Price.6     -5.636e-01  6.299e-02  -8.948 5.65e-16 ***
```

Result of fitting a full model:

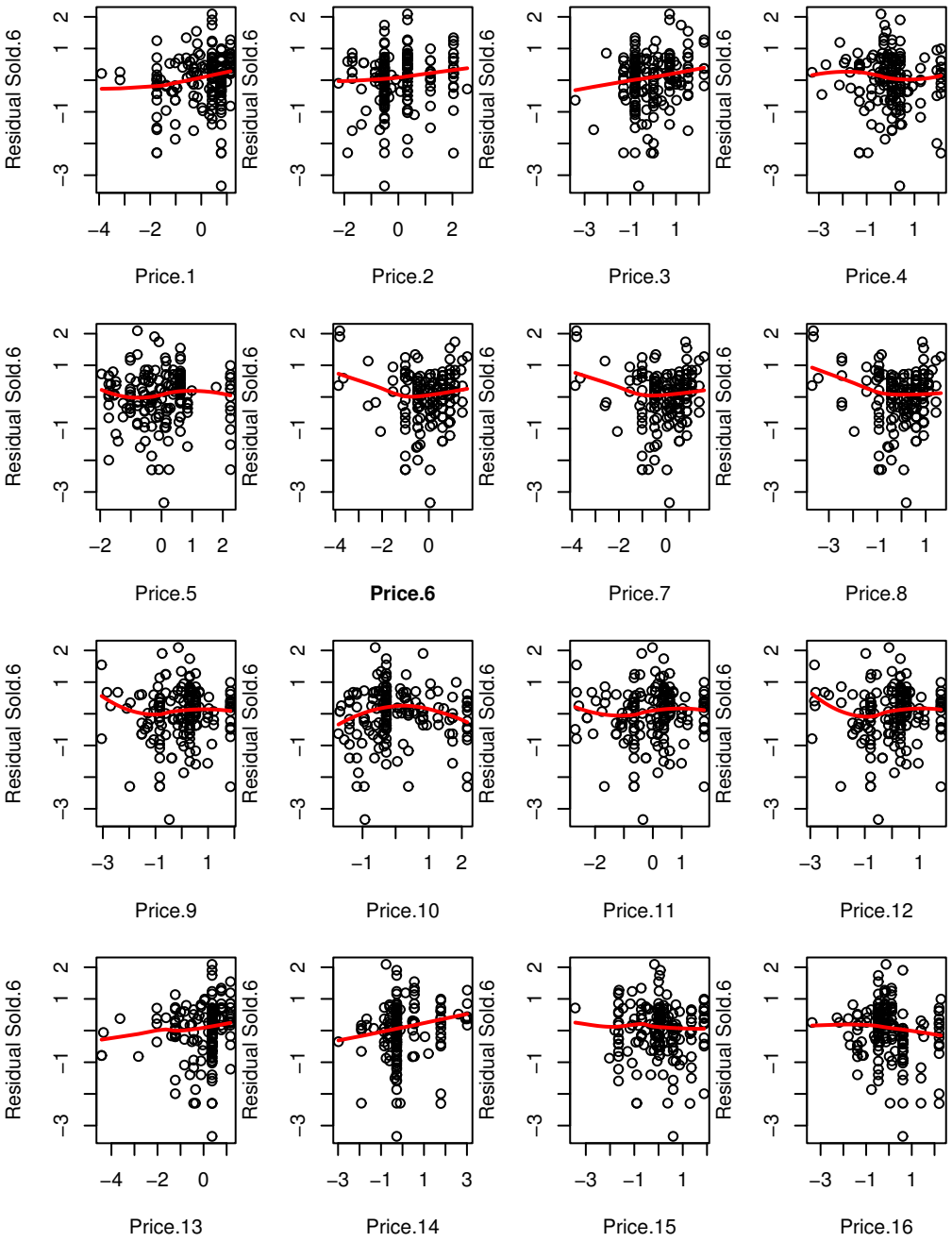
```
Coefficients:
              Estimate Std. Error   t value Pr(>|t|)
(Intercept) -8.862e-15  5.803e-02 -1.53e-13  1.00000
Price.1      9.496e-02  6.055e-02     1.568  0.11884
Price.2      1.809e-01  7.058e-02     2.563  0.01132 *
Price.3      1.675e-01  6.521e-02     2.569  0.01112 *
Price.4     -3.743e-02  8.235e-02    -0.455  0.65008
Price.5     -1.307e-03  6.309e-02    -0.021  0.98350
Price.6     -1.045e-01  3.152e-01    -0.332  0.74063
Price.7     -3.252e-01  3.404e-01    -0.955  0.34088
Price.8     -2.600e-01  1.731e-01    -1.502  0.13503
Price.9      1.636e-02  1.926e-01     0.085  0.93243
Price.10     8.852e-02  8.481e-02     1.044  0.29817
Price.11    -2.142e-02  1.498e-01    -0.143  0.88647
Price.12     1.350e-02  2.181e-01     0.062  0.95072
Price.13     9.044e-02  6.167e-02     1.466  0.14456
Price.14     1.361e-01  6.975e-02     1.951  0.05286 .
Price.15    -6.714e-02  7.224e-02    -0.929  0.35413
Price.16    -2.442e-01  8.205e-02    -2.977  0.00338 **
```

`Price.6` should be the most important predictor of `Sold.6`. Why does it have a large p-value?

Trend line for all predictors:



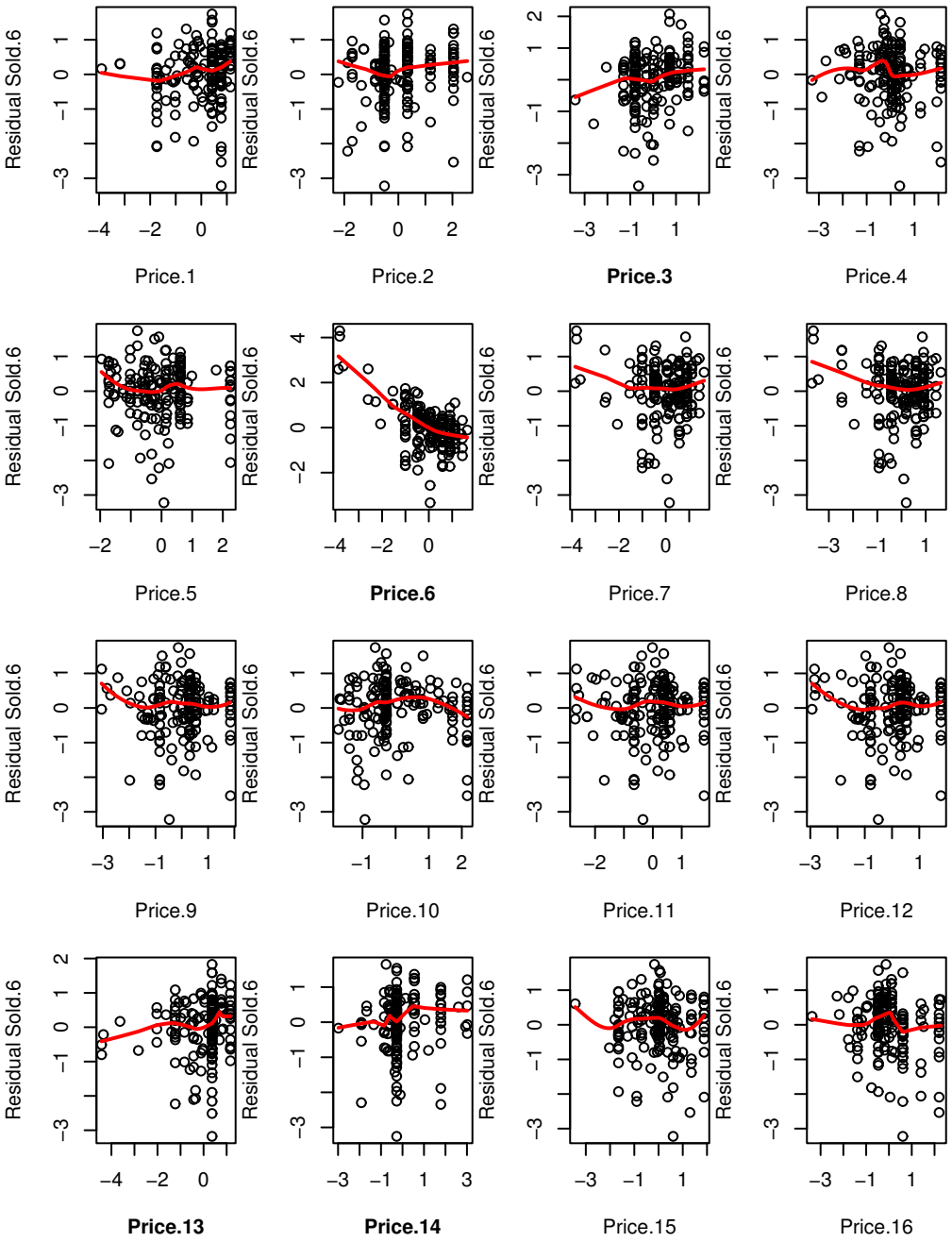
Residuals of the model `Sold.6 ~ Price.6`:



This tells you what predictor need to be added to the model. Which one?

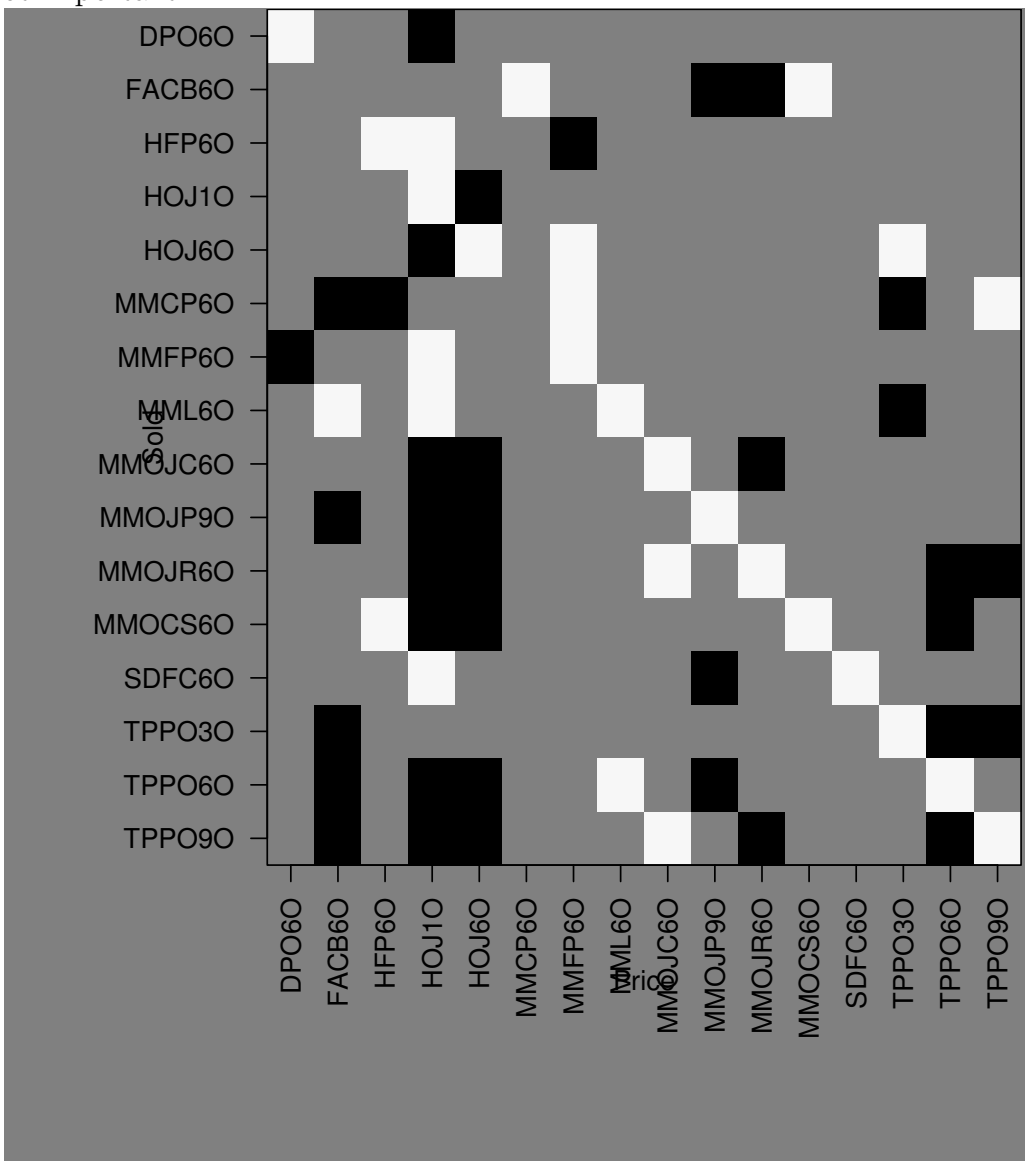
In a partial residual plot (next page), each predictor is plotted versus the residuals of the model with that predictor left out. This shows the importance of each predictor in the model (like a visual p-value), as well as what predictors need to be added.

Partial residuals of the model `Sold.6 ~ Price.6 + Price.3 + Price.13 + Price.14`:




```
Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  1.460e-17  6.036e-02 2.42e-16  1.00000
Price.6     -6.294e-01  6.328e-02   -9.945  < 2e-16 ***
Price.3      1.922e-01  6.311e-02    3.045  0.00270 **
Price.13     1.070e-01  6.143e-02    1.742  0.08341 .
Price.14     1.085e-01  6.077e-02    1.785  0.07603 .
```

Each row corresponds to a regression.
White = negative coefficient
Black = positive coefficient
Gray = not important



`HH OJ` is most preferred. The different sizes mutually compete. `FIVE ALIVE` and `MIN MAID OJ PLASTIC` also mutually compete, suggesting they are exchangeable. `HH OJ 128` is a controller: its price influences sales of almost every other product.