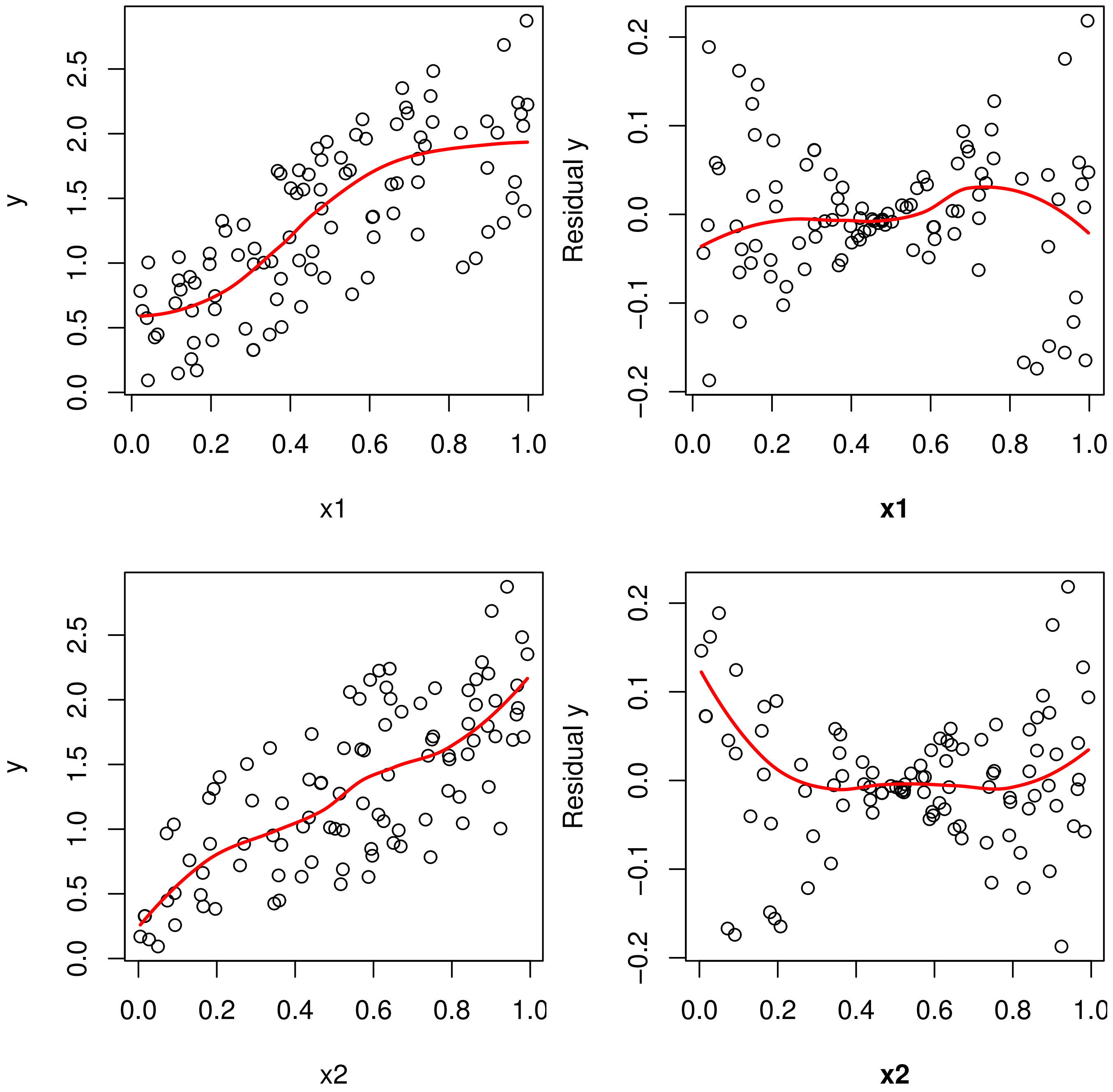
36-350: Data Mining

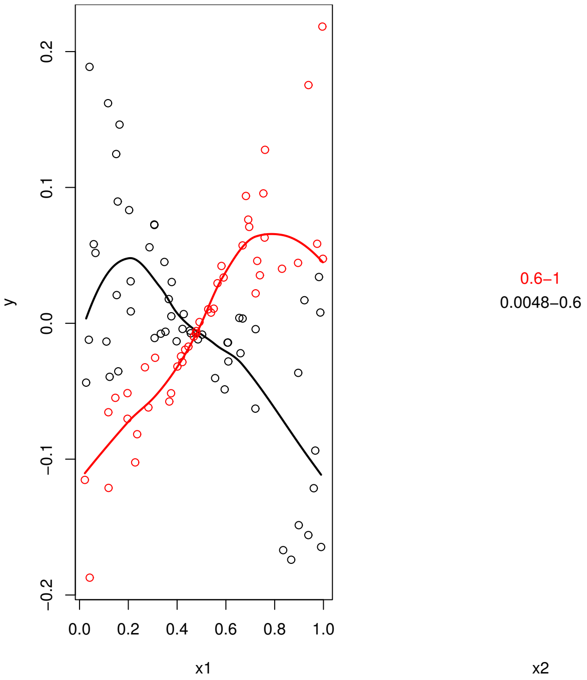
Handout 18
October 27, 2003

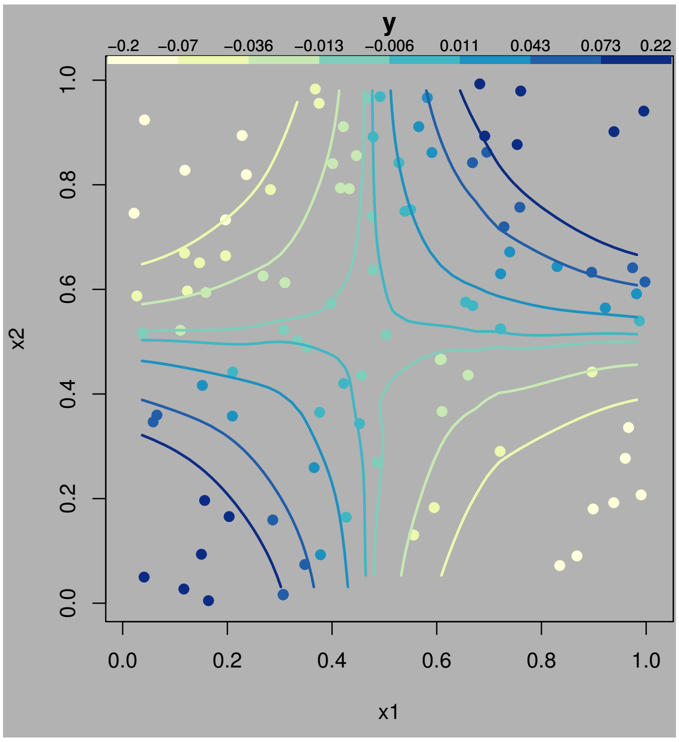
Adding interaction terms to a linear model

Slice and contour plots for interaction terms were introduced in handout 13. Now we apply them to find interactions in the grocery data.

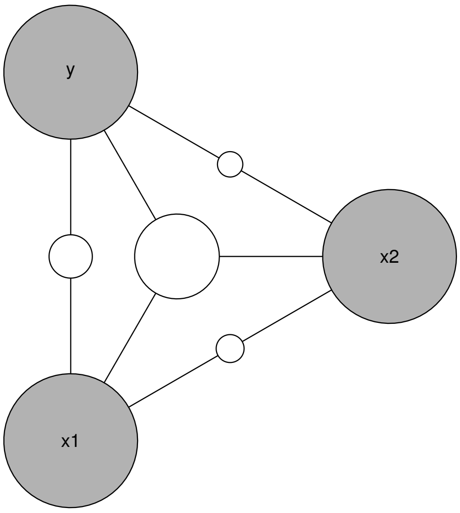
A simple dataset. Is an interaction term needed to predict y?





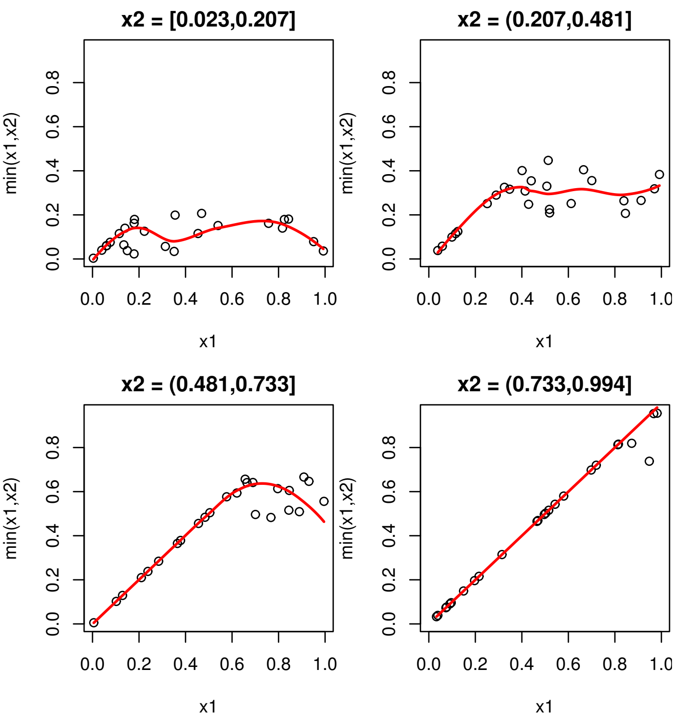


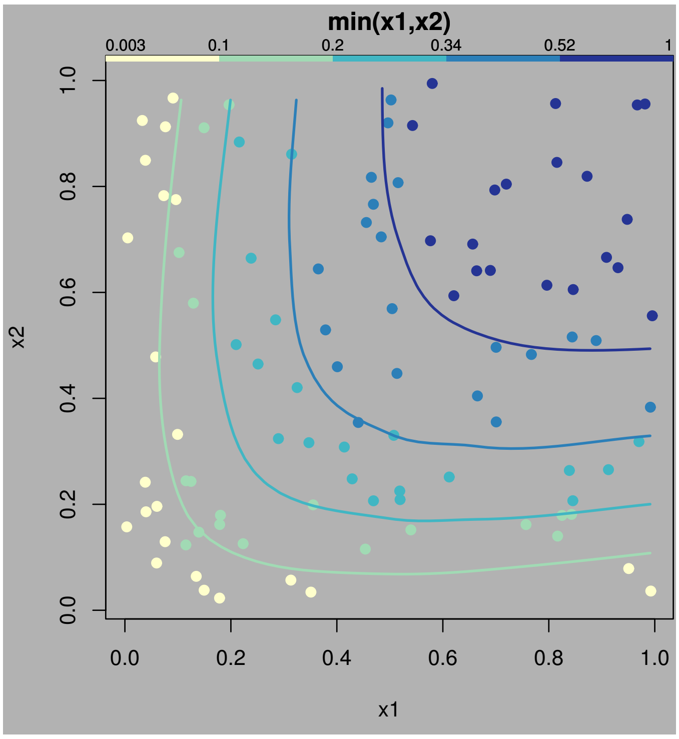
This data has formula $y = x_1 + x_2 + x_1 x_2$. This is known as a **bilinear** interaction, because it is linear in each variable (notice how the slices have linear trends). It is easy to recognize this type of interaction from a contour plot. Using the definition in handout 5, it is a positive interaction, because knowing one variable makes the other more important.



The danger of positive interactions is that they might involve predictors you previously discarded as unimportant. When searching for interactions, you must reconsider all predictors.

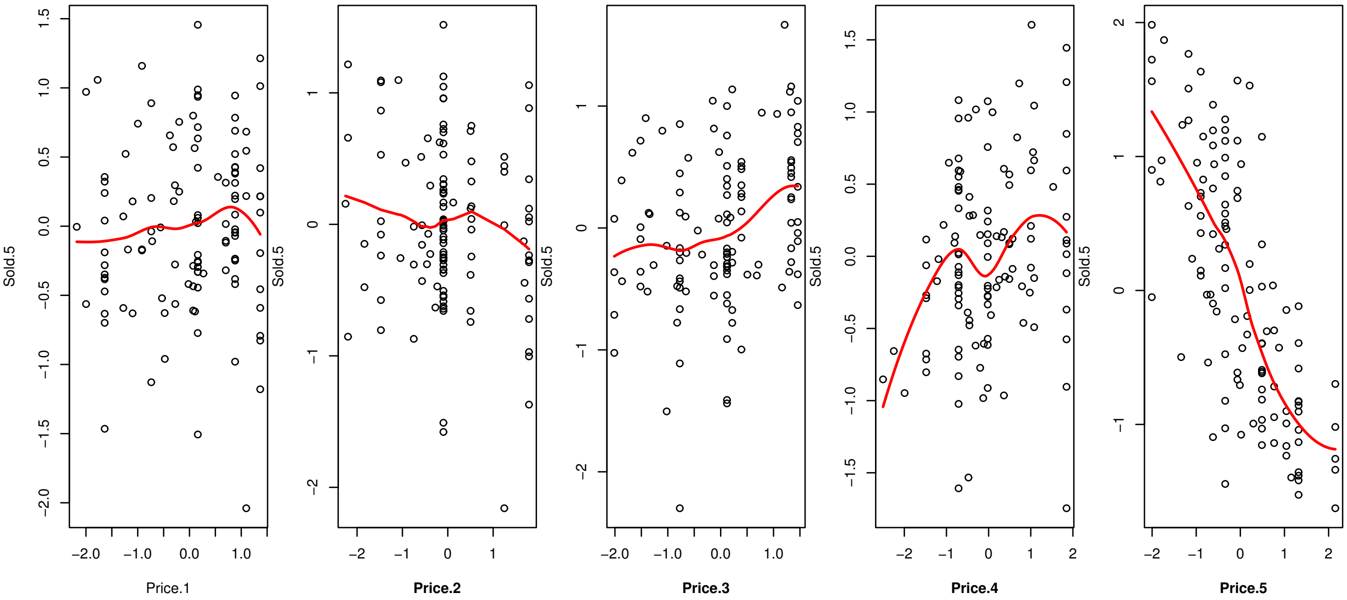
Another type of interaction is $\min(x_1, x_2)$. This function is also easy to recognize from a contour plot. Only the smaller variable is important. Slices clarify the difference from the bilinear interaction (the trends are not linear).



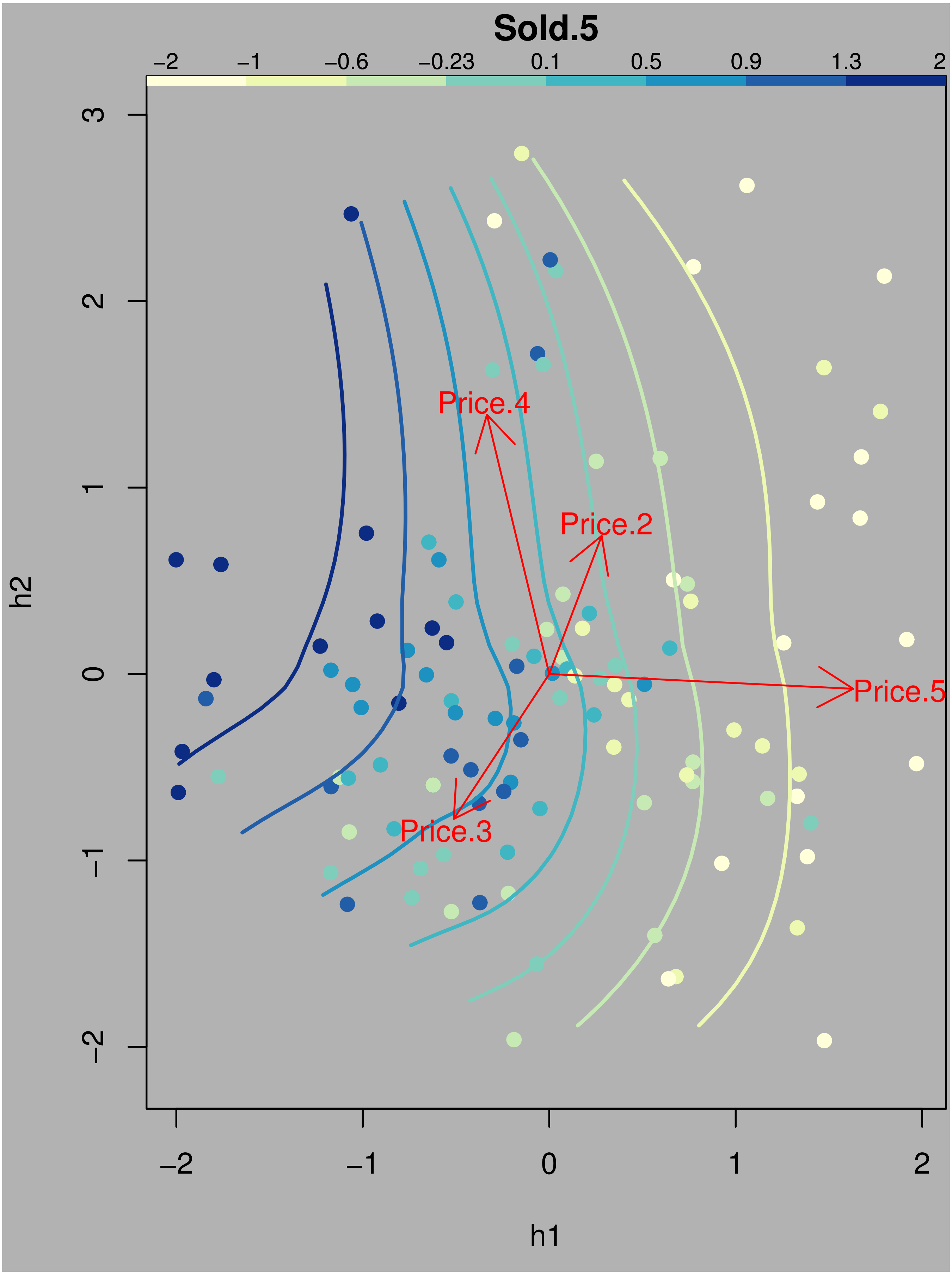


Grocery data, 5 products:

Price.1	Tropicana Premium 64 oz
Price.2	Tropicana Premium 96 oz
Price.3	Tropicana Regular 64 oz
Price.4	Minute Maid 64 oz
Price.5	Dominicks 64 oz
Sold.5	Number of units sold for Dominicks 64 oz

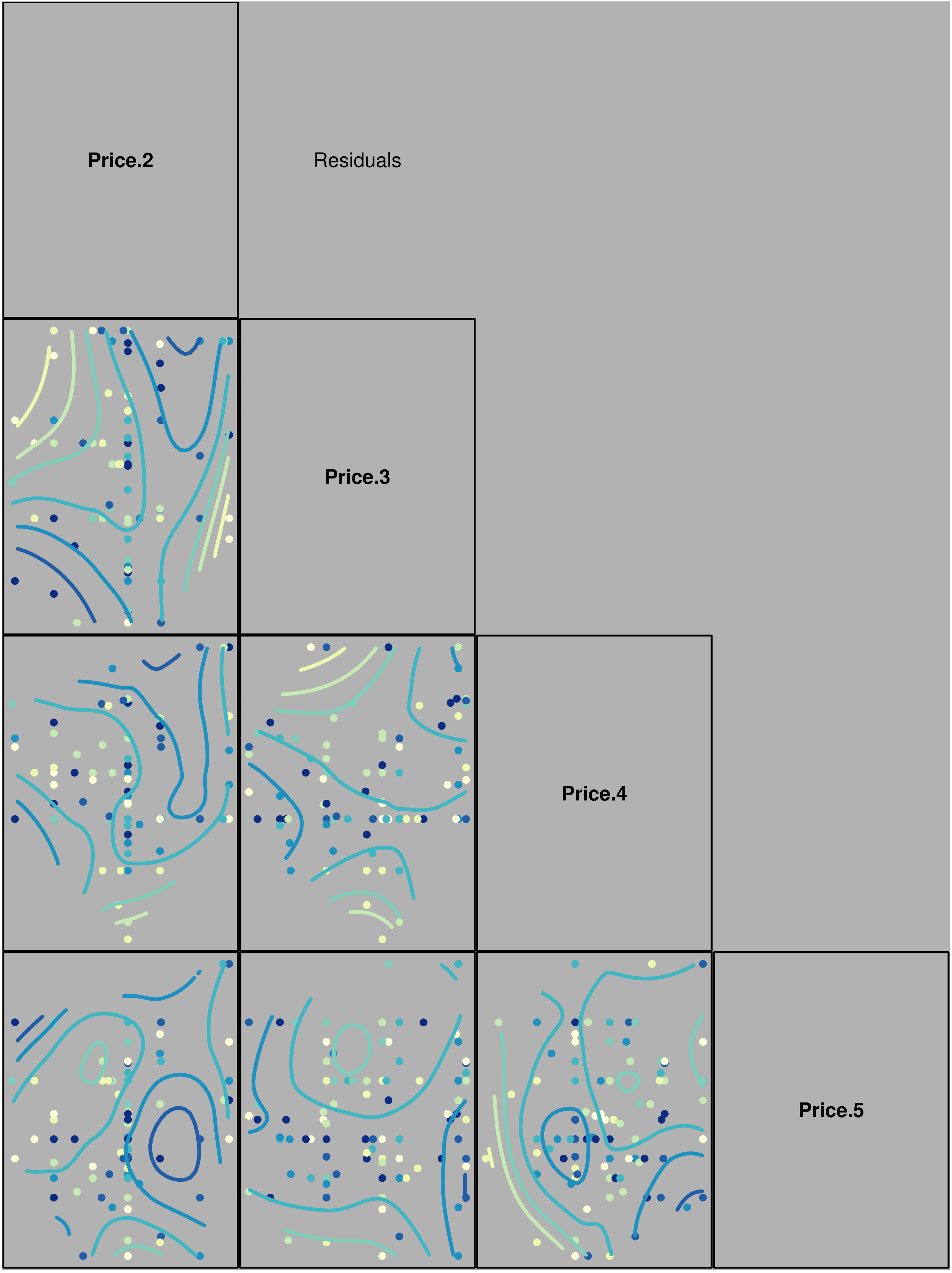


Regression projection:



Shows that an interaction exists, but hard to see what type.

To add interaction terms visually, look for structure in the residuals. Structure = a consistent bend in the contours, across multiple color groups. One contour line wiggling is not structure.



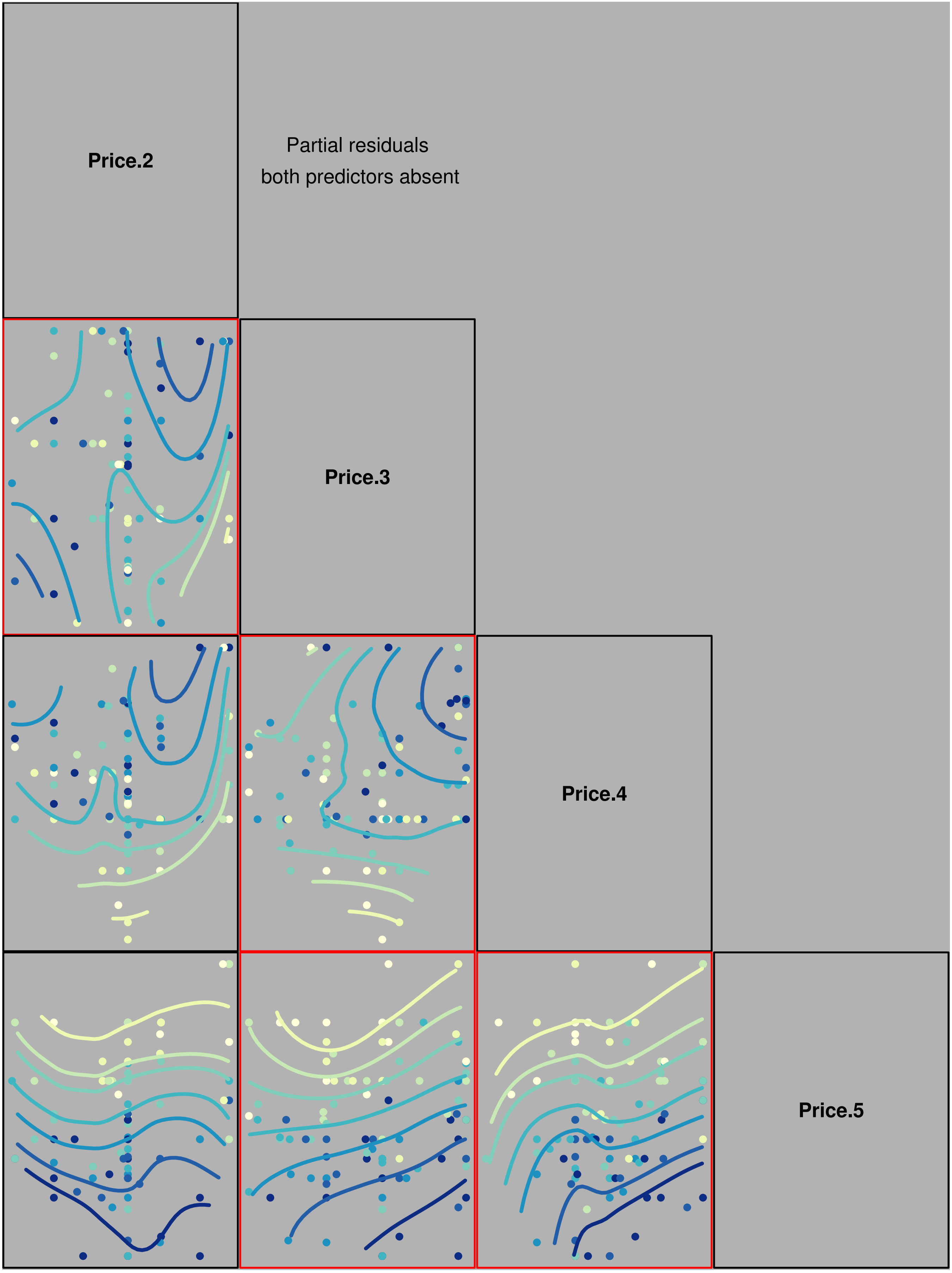
To find interactions automatically, add bilinear terms and use AIC to score them.

```
Start:  AIC= -104.03
 Sold.5 ~ Price.1 + Price.2 + Price.3 + Price.4 + Price.5

                  Df Sum of Sq      RSS      AIC
+ Price.2:Price.3  1     3.544   39.118 -112.094
+ Price.3:Price.4  1     3.070   39.591 -110.697
+ Price.1:Price.2  1     1.766   40.895 -106.939
+ Price.1:Price.3  1     1.361   41.301 -105.795
- Price.1          1     0.194   42.855 -105.508
<none>                           42.661 -104.035
+ Price.2:Price.4  1     0.342   42.319 -102.968
+ Price.1:Price.4  1     0.274   42.387 -102.782
+ Price.4:Price.5  1     0.266   42.395 -102.760
+ Price.2:Price.5  1     0.135   42.526 -102.403
+ Price.1:Price.5  1     0.074   42.588 -102.235
+ Price.3:Price.5  1     0.015   42.646 -102.075
- Price.2          1     1.652   44.313 -101.628
- Price.4          1     2.639   45.300  -99.072
- Price.3          1     4.318   46.979  -94.851
- Price.5          1    58.755  101.416   -5.585

...

Step:  AIC= -119.31
 Sold.5 ~ Price.1 + Price.2 + Price.3 + Price.4 + Price.5 + Price.2:Price.3 +
    Price.1:Price.2 + Price.3:Price.4 + Price.4:Price.5
```

In an R formula, `Price.2:Price.3` actually means `Price.2` × `Price.3`, while `Price.2*Price.3` is shorthand for `Price.2 + Price.3 + Price.2:Price.3` (three separate predictors).

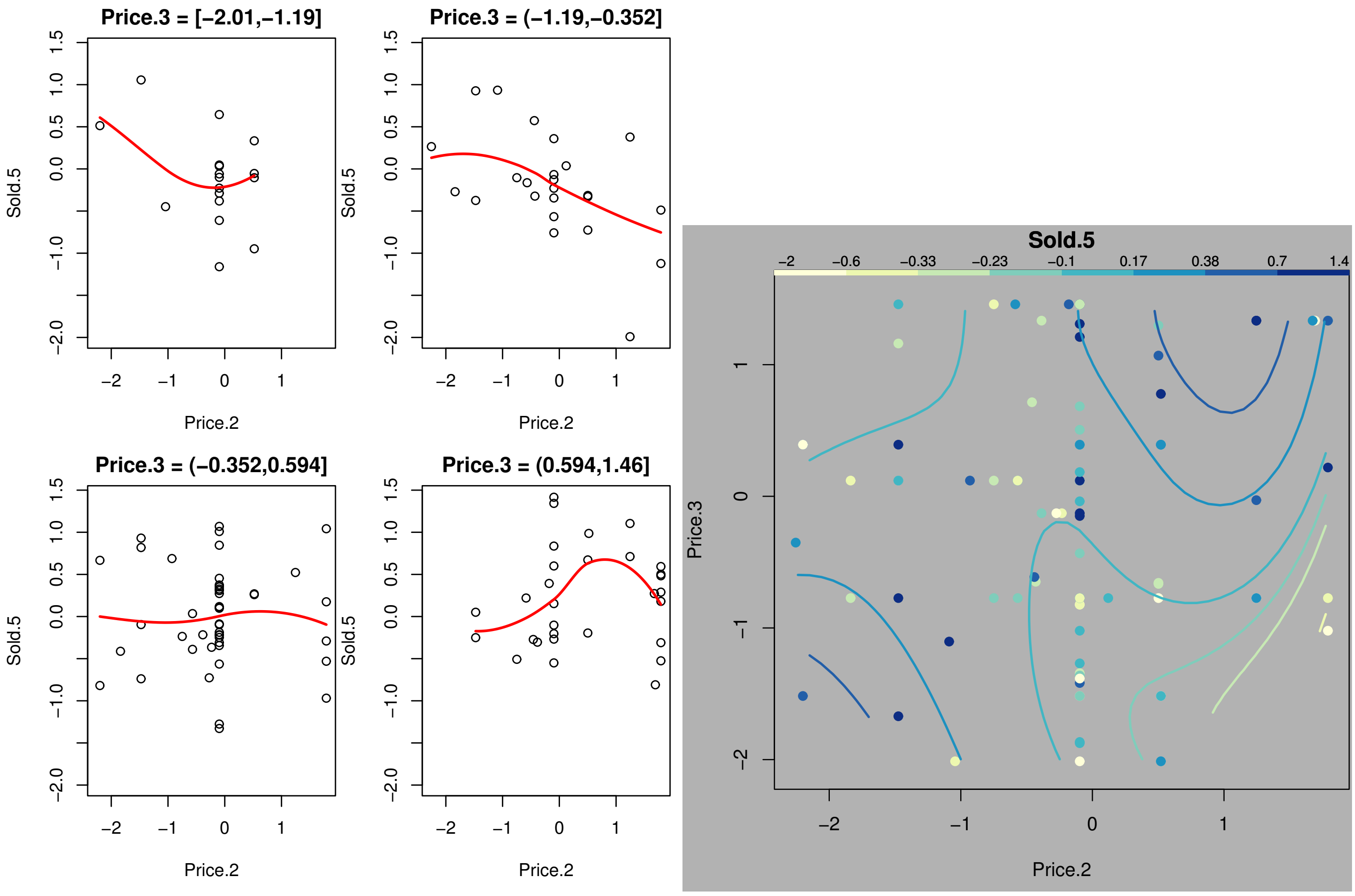
It is pointless to use `summary` to determine if these interactions are significant, since they probably aren't bilinear anyway.

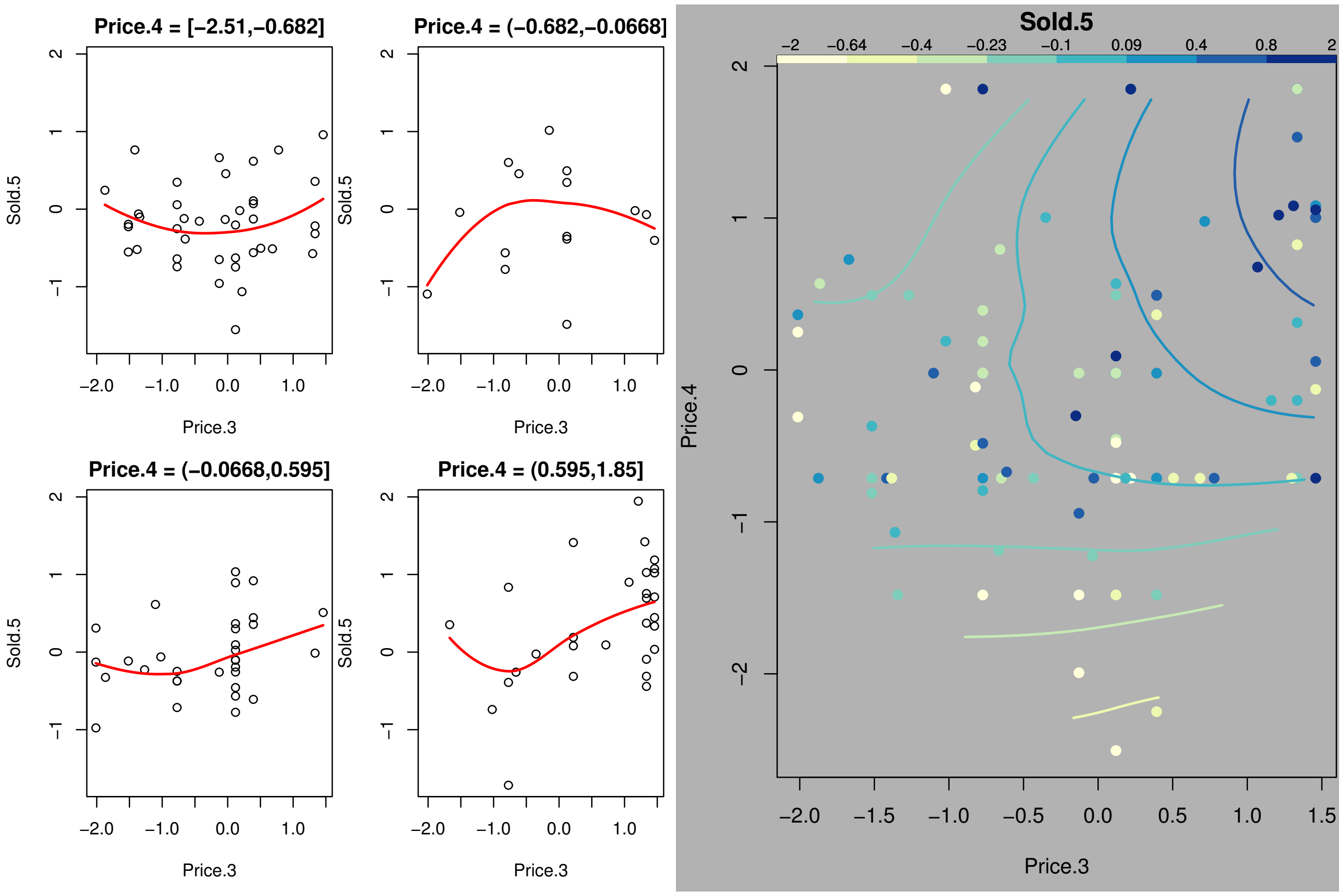
Use partials to determine the form of each interaction:



Predictors in the model are indicated in bold, and cross terms in the model are indicated by red boxes.

What kind of interaction is it?


Price.3 = [−2.01,−1.19]
Price.3 = (−1.19,−0.352]
Price.3 = (−0.352,0.594]
Price.3 = (0.594,1.46]
Sold.5
Price.2
Sold.5
Price.3
Price.2


Price.4 = [−2.51,−0.682]
Price.4 = (−0.682,−0.0668]
Price.4 = (−0.0668,0.595]
Price.4 = (0.595,1.85]
Sold.5
Price.3
Sold.5
Price.4
Price.3