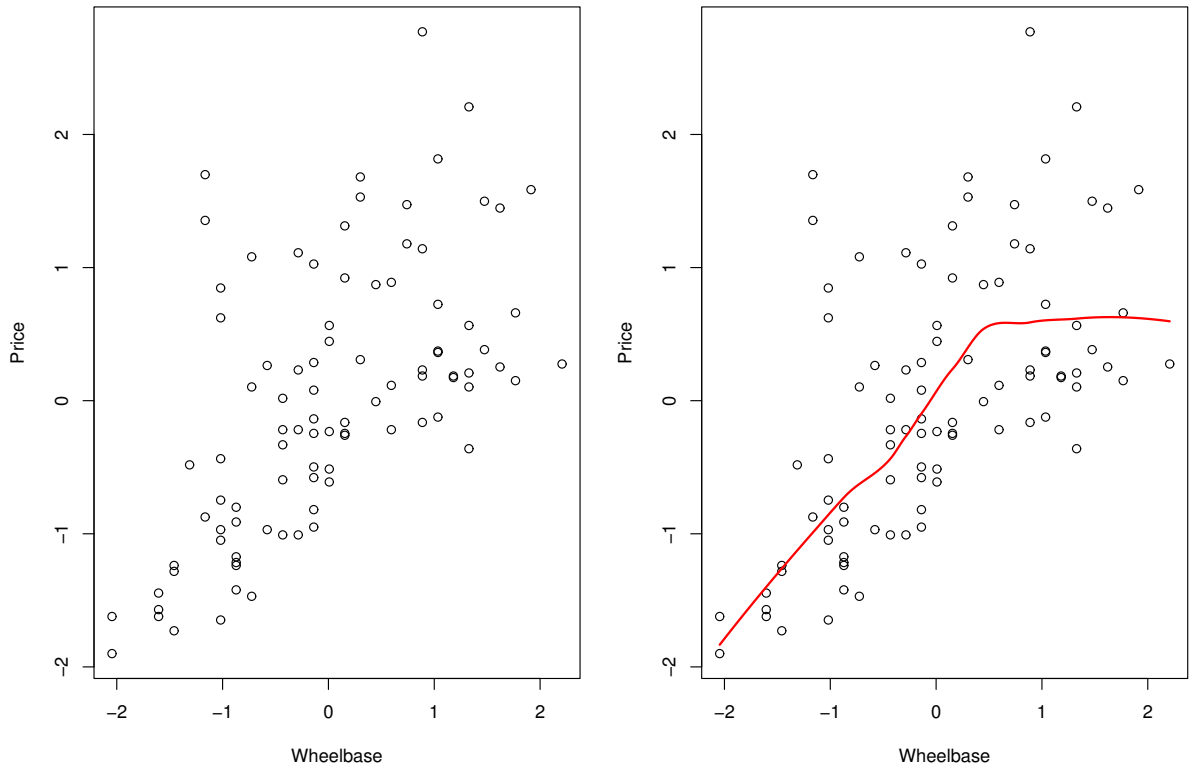
36-350: Data Mining

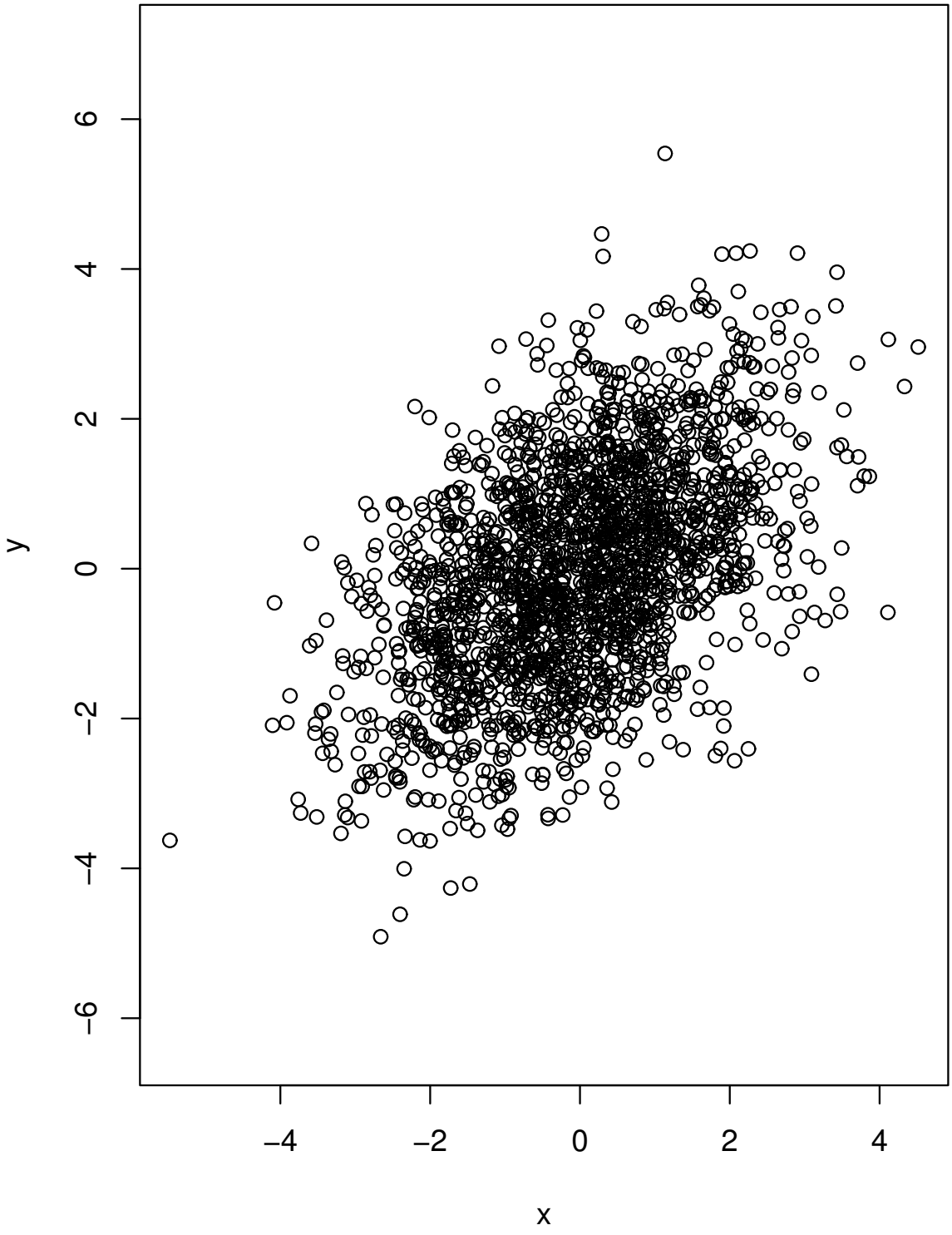
Handout 12
October 6, 2003

Visualization for predictive modeling

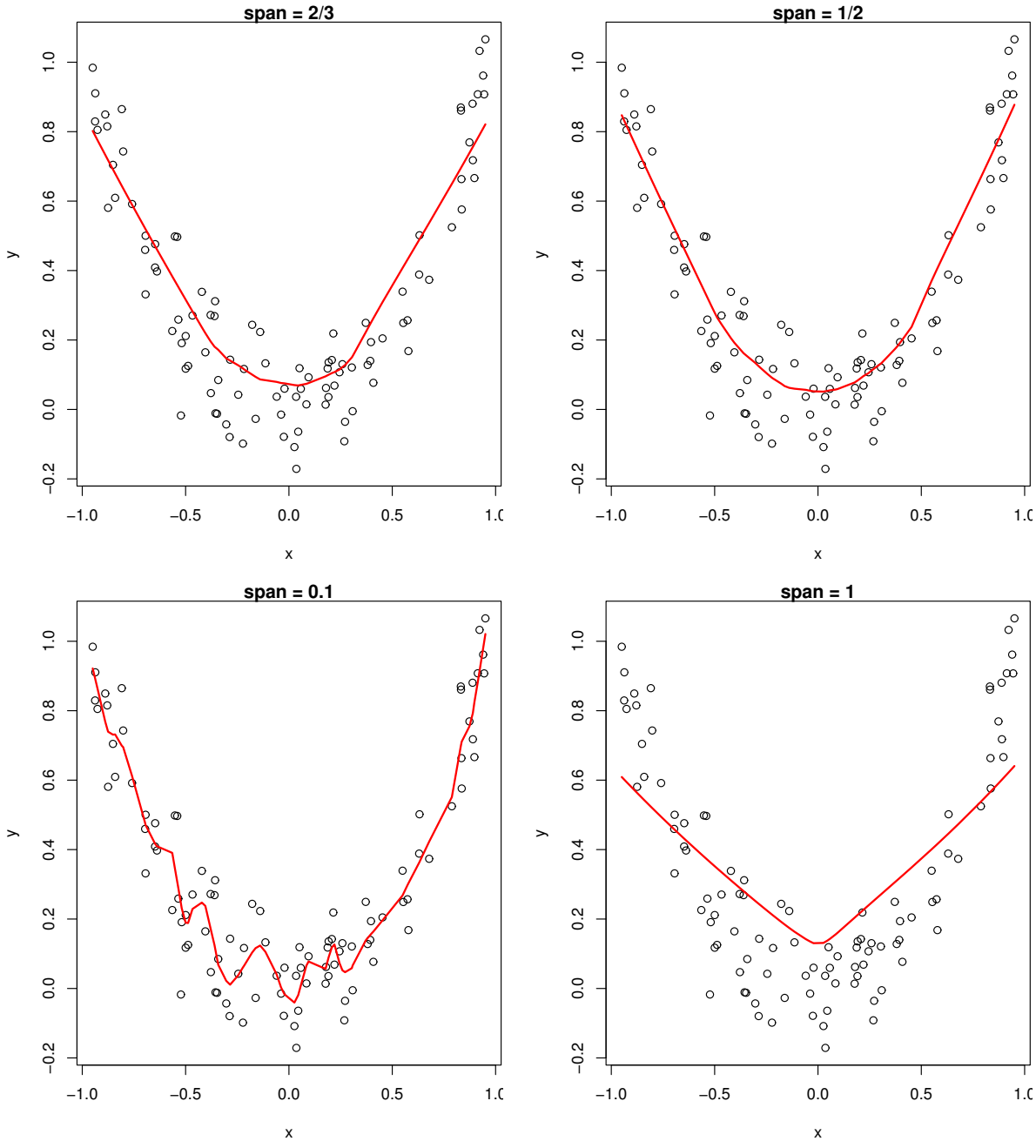
Trend line, a.k.a. **prediction line** or **regression line** - Depicts the mean of y as a function of x. If this line is flat, y is not predictable from x.



Can you draw the correct prediction line?

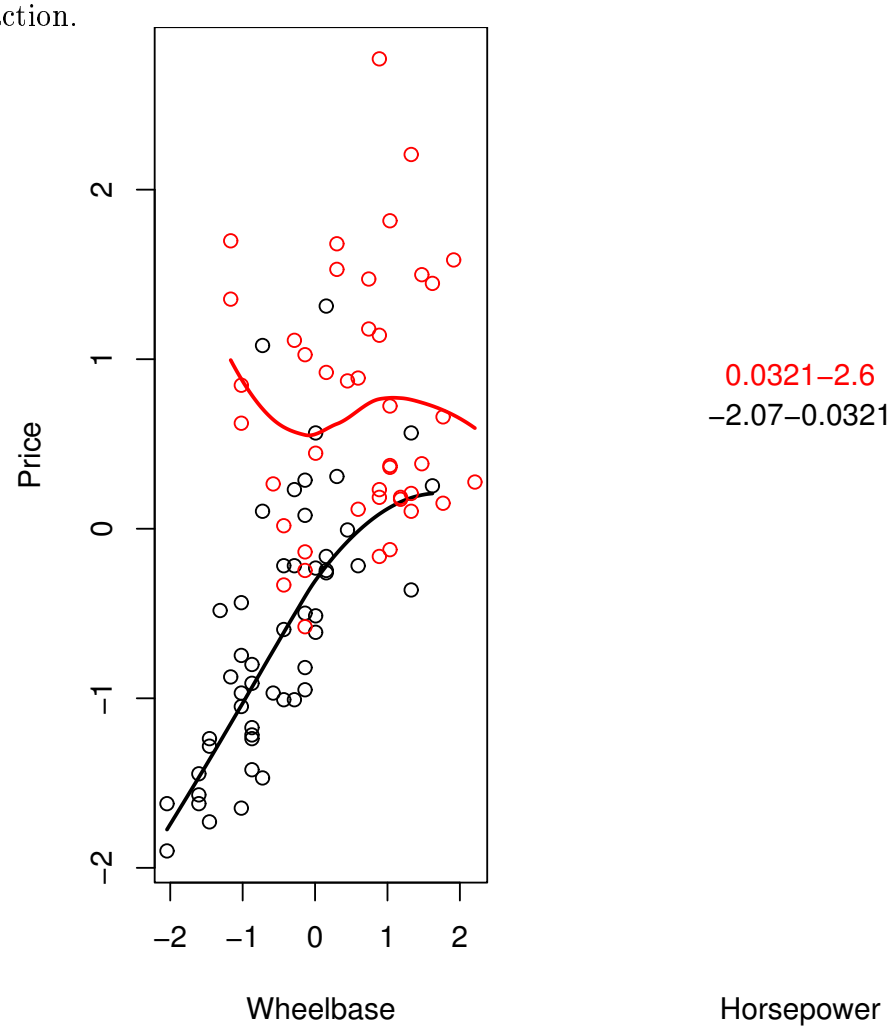


Smoothing span:



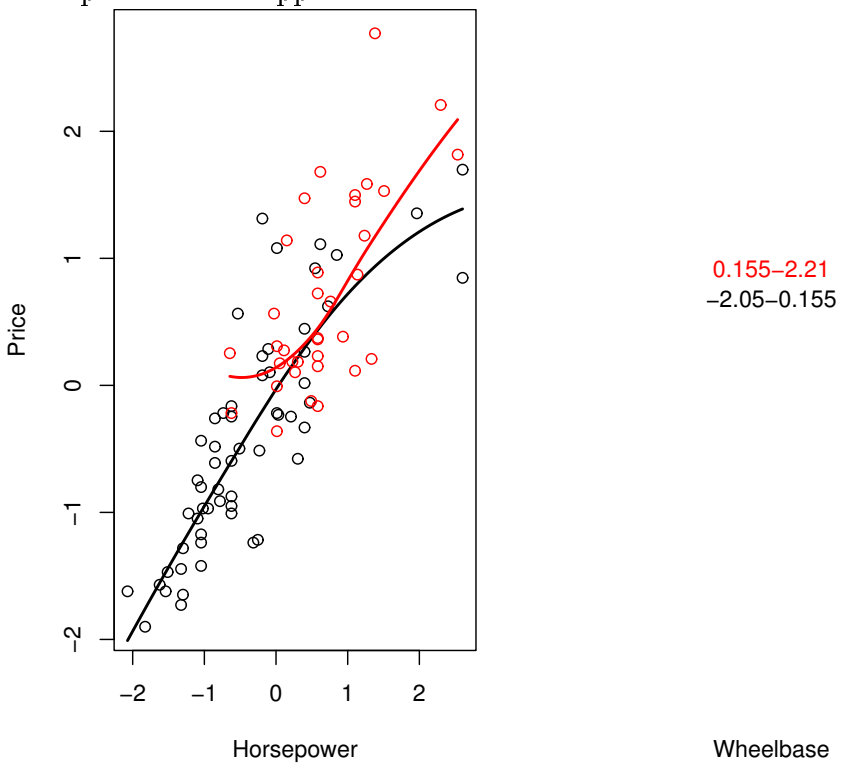
Visualizing interactions

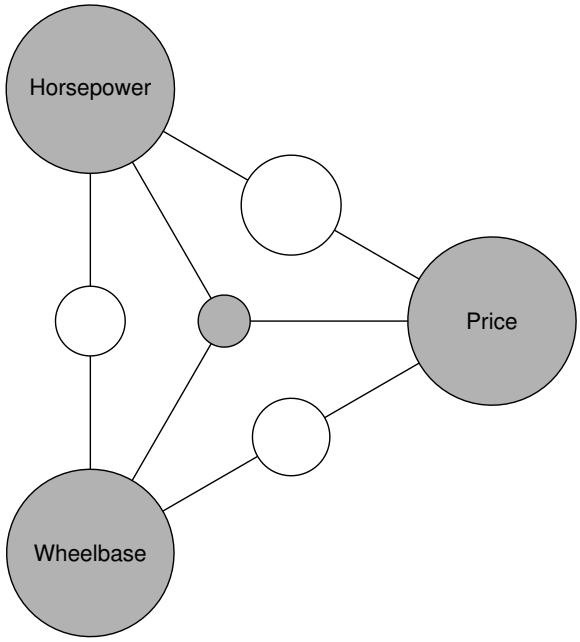
Slice plot, a.k.a. **conditioning plot**—Response versus predictor 1, with points colored according to predictor 2. If the regression curve changes, besides just shifting up or down, there is an interaction.



Colors are chosen by the **equal-count rule**: each color covers the same number of points, in rank order. This makes the plot identical under any transformation that preserves ranks.

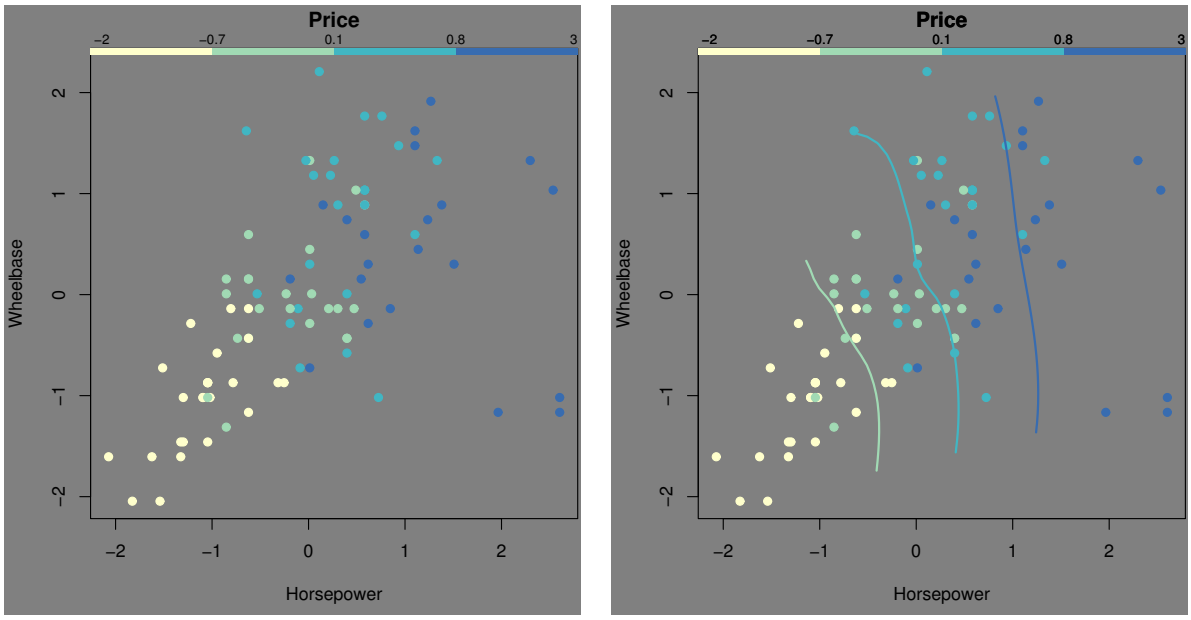
Same thing, with predictors swapped:





Interaction is symmetric, but `Wheelbase` takes away a small fraction of the information in `Horsepower`, while `Horsepower` takes away most of the information in `Wheelbase`.

Color plot—Predictor 1 versus predictor 2, with points colored according to the response. You are looking at the slice plots from "above" rather than from the "sides." Each row or column of points is a slice, allowing you to see the interaction from either direction.



As you change a predictor, look at how the *average* response varies, in other words, the average color in the region. Contour lines (on the right) delineate regions of different average color, allowing you to see the trends more clearly. When you cross a contour line, the average response is changing. As you move along a contour line, the average response is not changing.

The interaction in the previous slice plots is visible as a change in the slope of the contour lines. (More on this next time.)

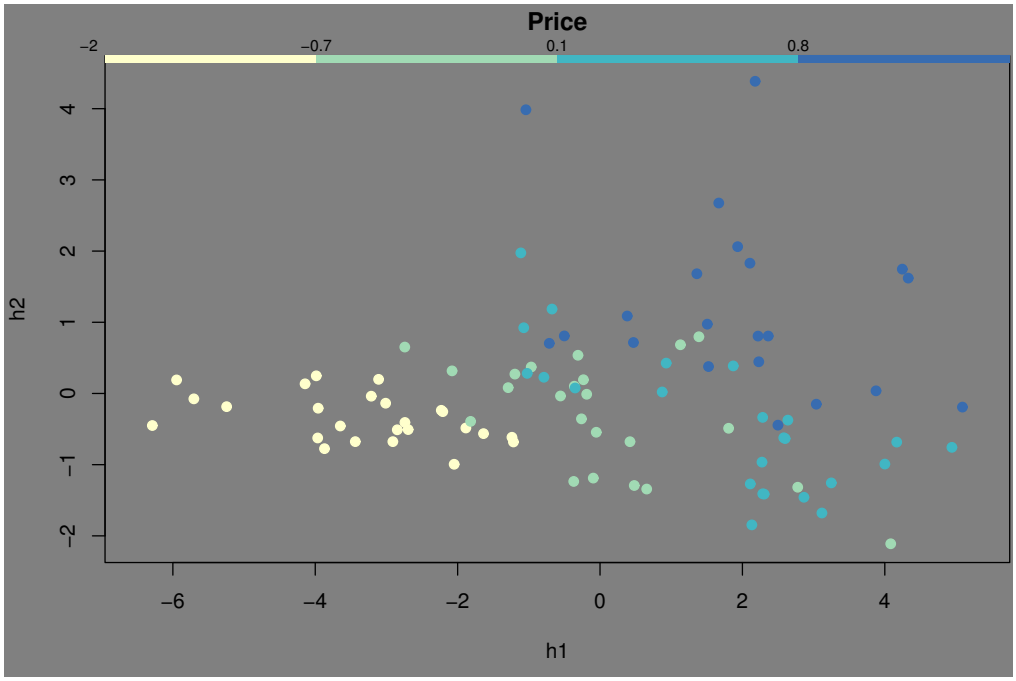
Finding important attributes—Turn the response into a subgroup variable, by dividing the data into groups of similar responses (equal-count rule). Now apply previous methods to compute the information between the response and each predictor. Does *not* require assumptions about the response function (e.g. linear).

```
       Var1     Score
 Passengers 0.1277944
Turn.circle 0.3860456
      Width 0.4247228
     Length 0.4399956
  Wheelbase 0.4411857
MPG.highway 0.5075420
 EngineSize 0.6004170
     Weight 0.7264830
 Horsepower 0.8889632
```

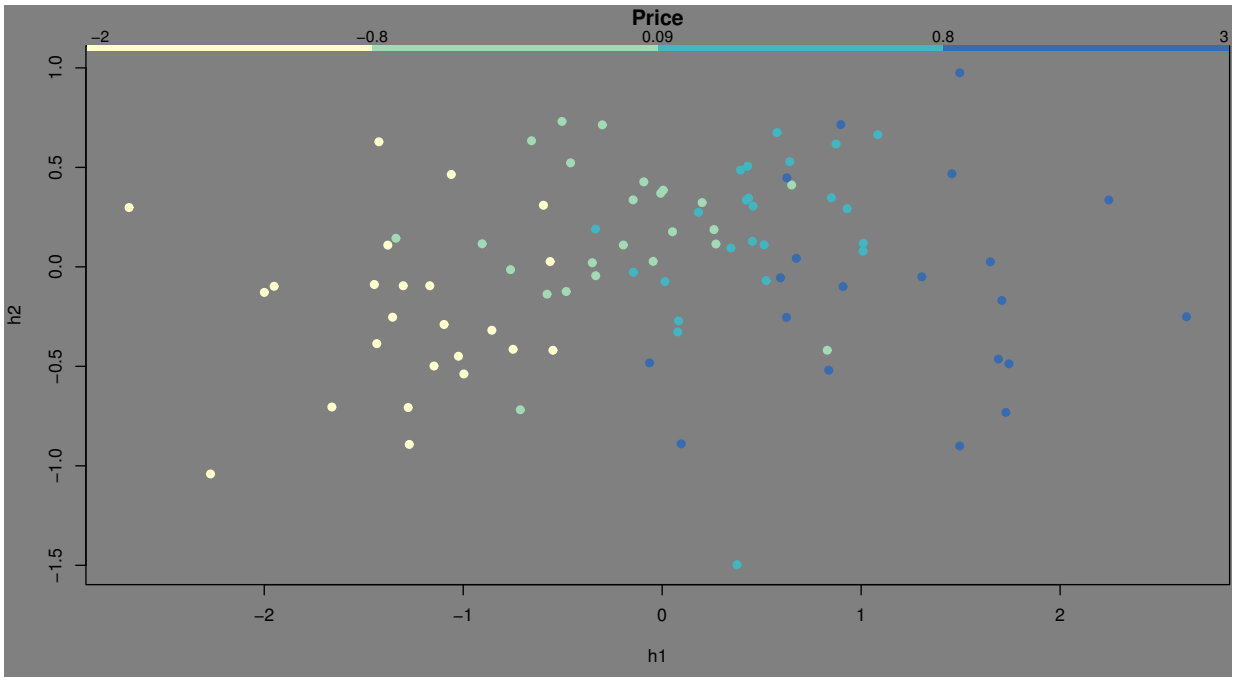
Most informative pair turns out to be `(Horsepower,Wheelbase)`, in spite of the negative interaction.

Regression projection—A projection which is informative about the response variable. To compute it, turn the response into a subgroup variable, then apply mv-projection.

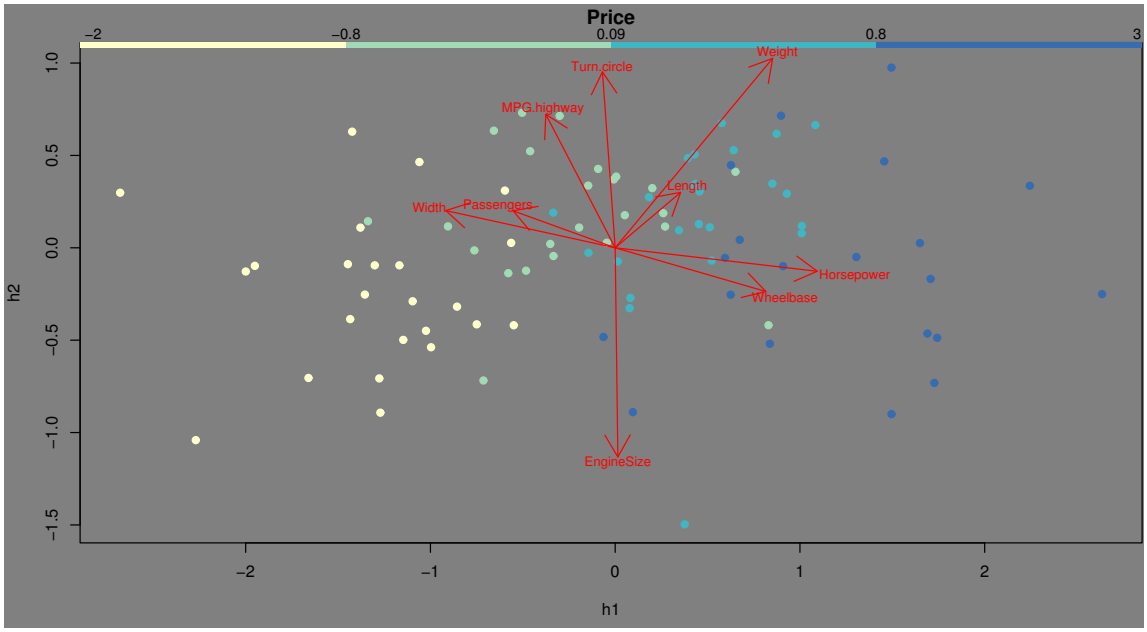
PCA projection:



Regression projection:



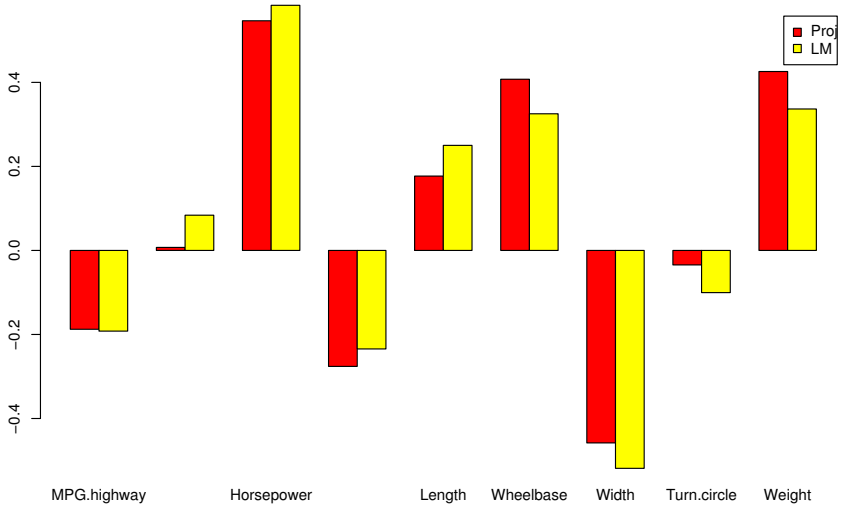
Mazda RX-7 was removed since it is a `Price` outlier.

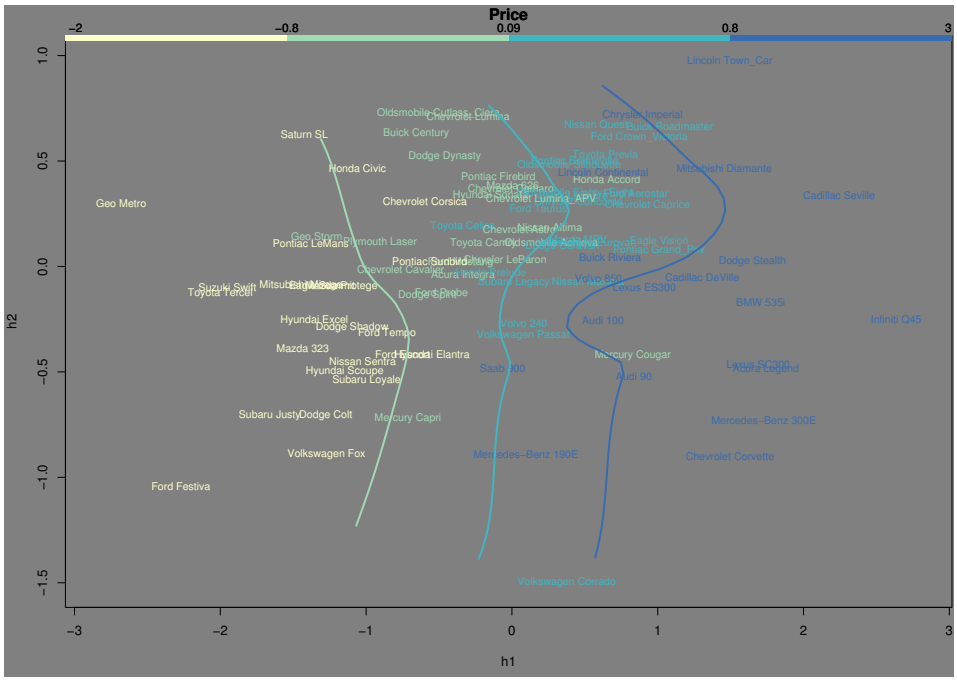


	h1	h2
MPG.highway	-0.19	0.36
EngineSize	0.01	-0.57
Horsepower	0.55	-0.06
Passengers	-0.28	0.10
Length	0.18	0.15
Wheelbase	0.41	-0.12
Width	-0.46	0.10
Turn.circle	-0.03	0.48
Weight	0.43	0.51

`h1` is a linear combination of the attributes which is most informative about the response. Thus, if the relationship is linear, regression projection is the same as fitting a linear model.

Coefficients for `h1` versus the coefficients of a linear model (via least squares):





Some outliers are apparent. We can inspect them using parallel plots.