36-350: Data Mining

Handout 20
November 3, 2003

Linear regression and visualization with categorical predictors

Indicator coding—A predictor with C categories is turned into $C - 1$ binary indicators. In a linear model, this gives each category (after the first) its own coefficient. But for visualization, other methods are needed.

Average monthly temperature in Arizona:

Month	Place	Temp
July	Flagstaff	65.2
Aug	Flagstaff	63.4
Sept	Flagstaff	57.0
Oct	Flagstaff	46.1
Nov	Flagstaff	35.8
Dec	Flagstaff	28.4
Jan	Flagstaff	25.3
July	Phoenix	90.1
Aug	Phoenix	88.3
Sept	Phoenix	82.7
Oct	Phoenix	70.8
Nov	Phoenix	58.4
Dec	Phoenix	52.1
Jan	Phoenix	49.7
July	Yuma	94.6
Aug	Yuma	93.7
Sept	Yuma	88.3
Oct	Yuma	76.4
Nov	Yuma	64.2
Dec	Yuma	57.1
Jan	Yuma	55.3

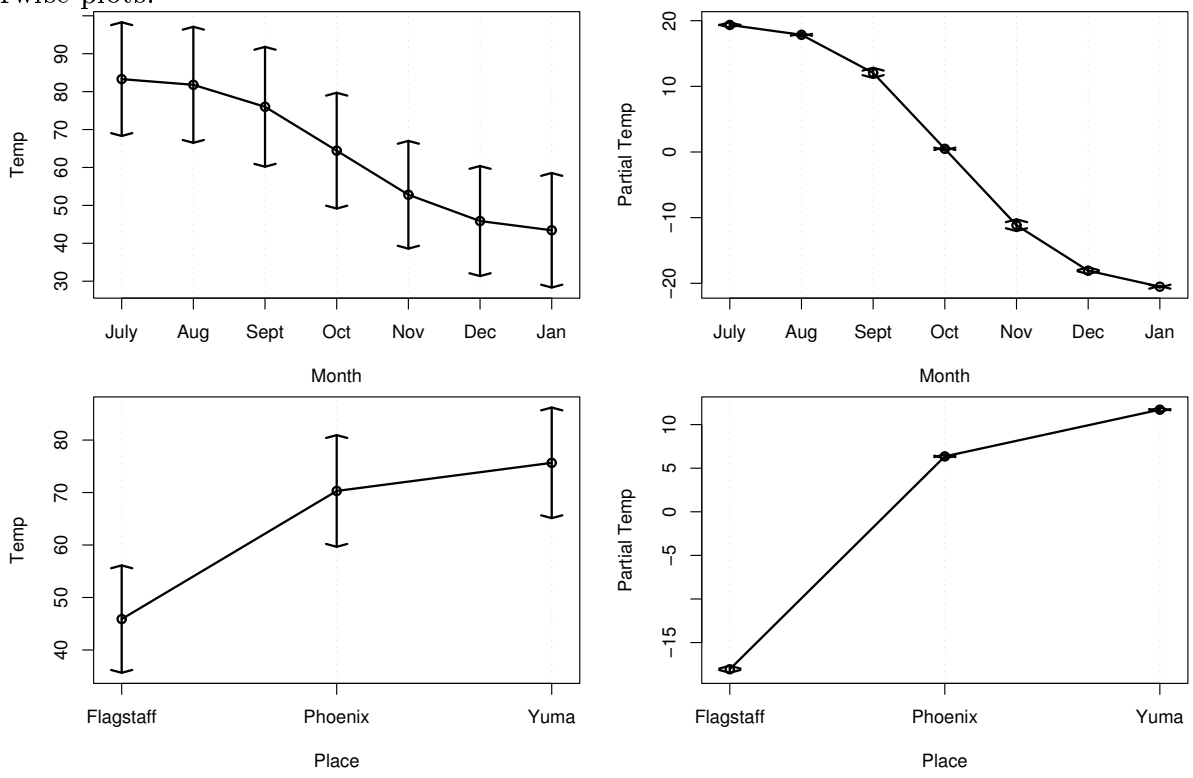
As a table:

```
Temp
      Place
Month  Flagstaff Phoenix Yuma
  July      65.2    90.1 94.6
  Aug       63.4    88.3 93.7
  Sept      57.0    82.7 88.3
  Oct       46.1    70.8 76.4
  Nov       35.8    58.4 64.2
  Dec       28.4    52.1 57.1
  Jan       25.3    49.7 55.3
```

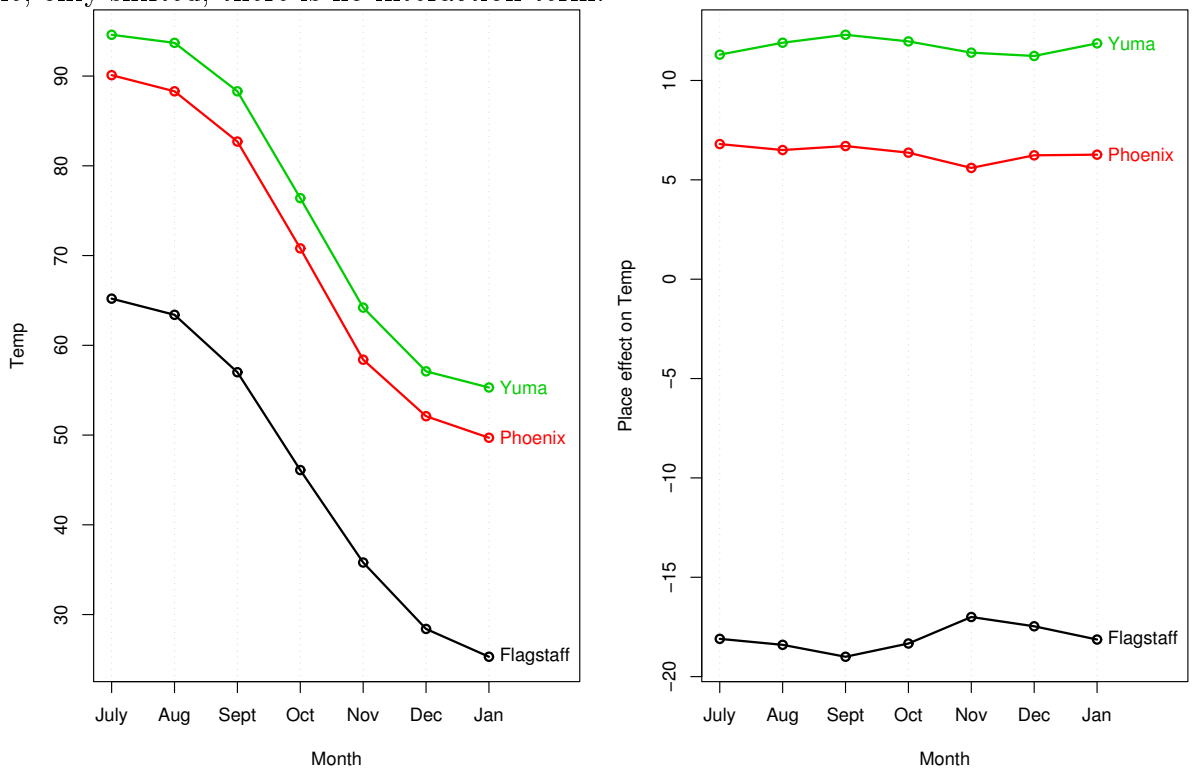
A linear model to predict temperature:

```
Coefficients:
   (Intercept)        MonthAug       MonthSept
        95.010          -1.500          -7.300
      MonthOct        MonthNov        MonthDec
       -18.867         -30.500         -37.433
      MonthJan  PlaceFlagstaff    PlacePhoenix
       -39.867         -29.771          -5.357
```

Pairwise plots:

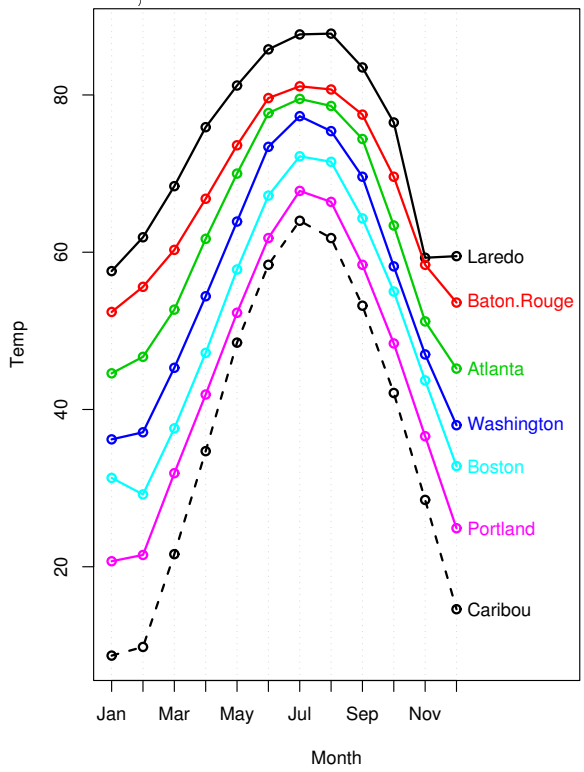


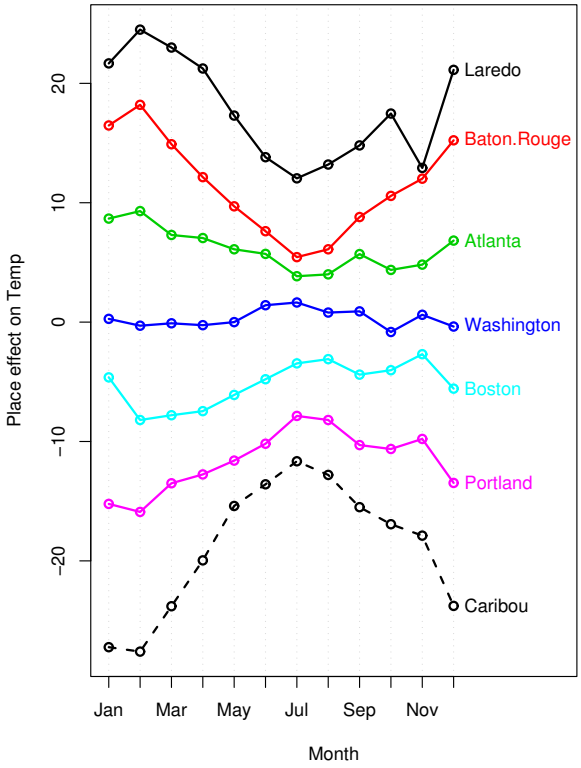
Partial residuals show that both variables are important. But is an interaction term needed? Contour plots won't work. Use a line chart, the equivalent of a slice plot. If the curves are the same, only shifted, there is no interaction term.

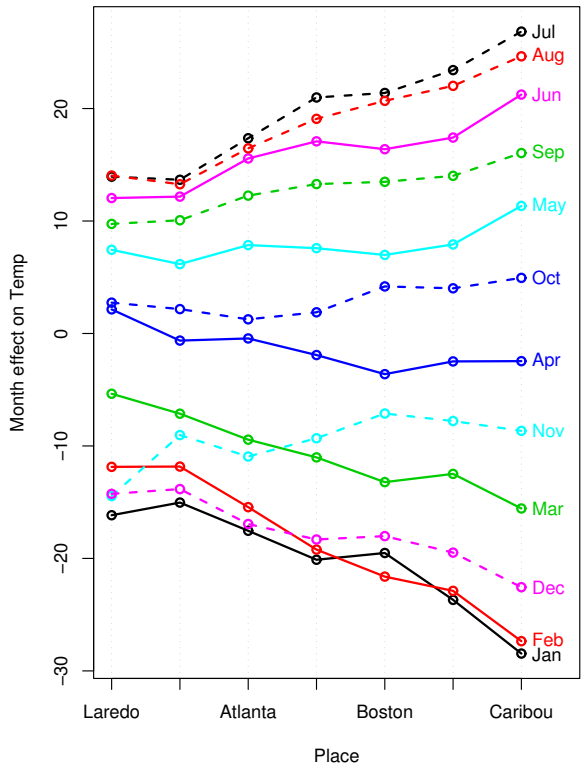


In the right plot, the mean of each `Month` is subtracted, which should make the curves flat if there is no interaction term. The "Place effect" is the change from the mean, and measures the importance of `Place`. "No interaction term" is equivalent to "Place effect is constant," i.e. the temperature difference between places is the same every month. Equivalently, the temperature difference between months is the same in each place.

More months, more cities:







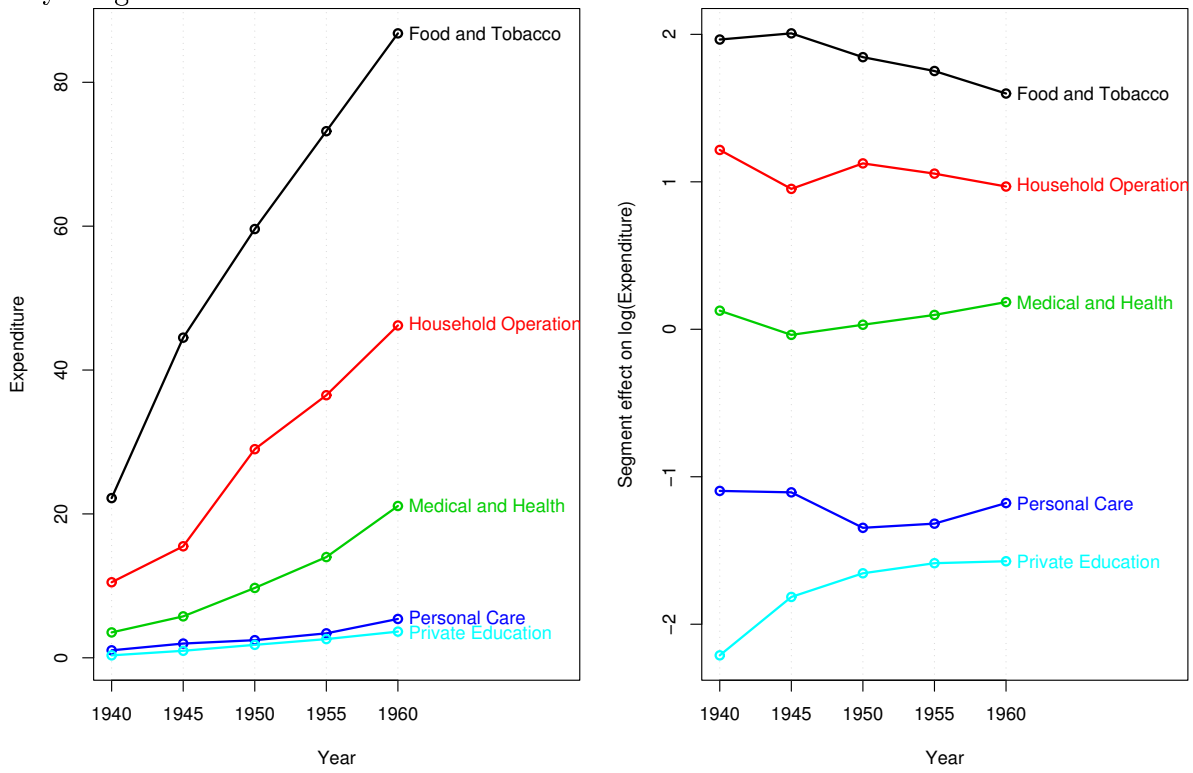
To describe the interaction: `Place` has the largest effect in winter, and `Month` has the largest effect in the North.

U.S. personal expenditures, by decade, in billions of dollars:

Expenditure

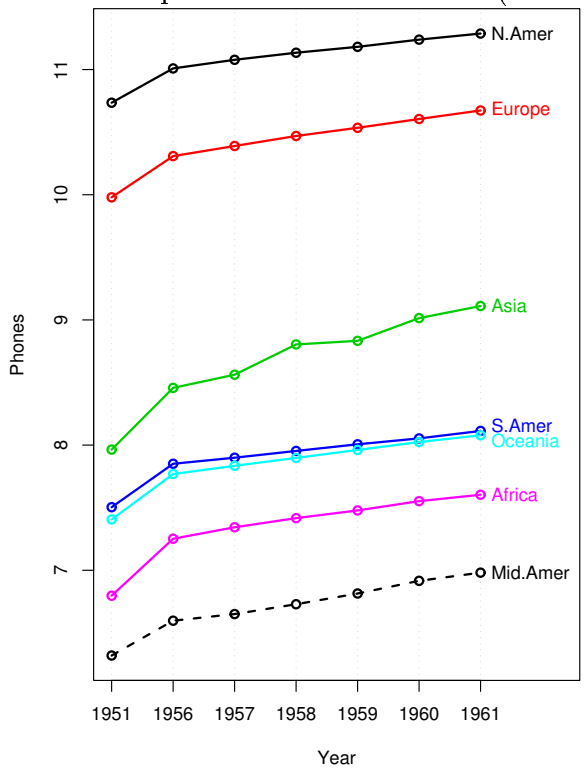
Segment	Year 1940	1945	1950	1955	1960
Food and Tobacco	22.200	44.500	59.60	73.2	86.80
Household Operation	10.500	15.500	29.00	36.5	46.20
Medical and Health	3.530	5.760	9.71	14.0	21.10
Personal Care	1.040	1.980	2.45	3.4	5.40
Private Education	0.341	0.974	1.80	2.6	3.64

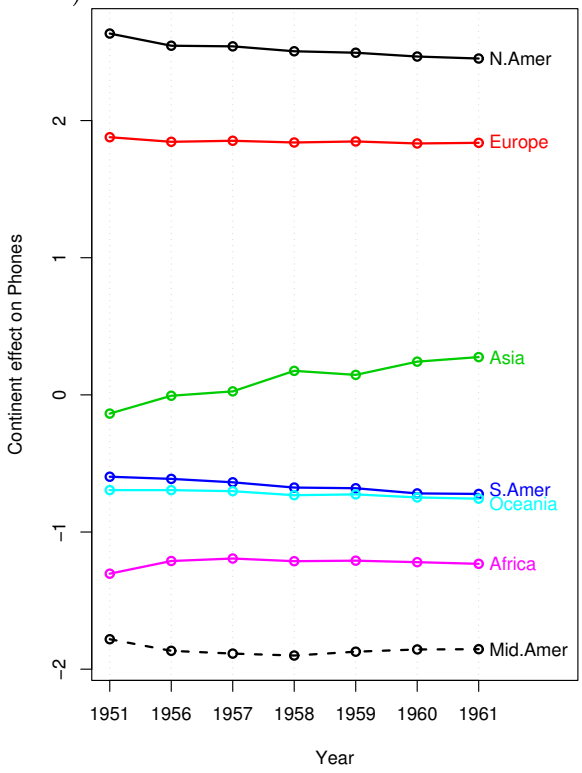
See anything unusual?

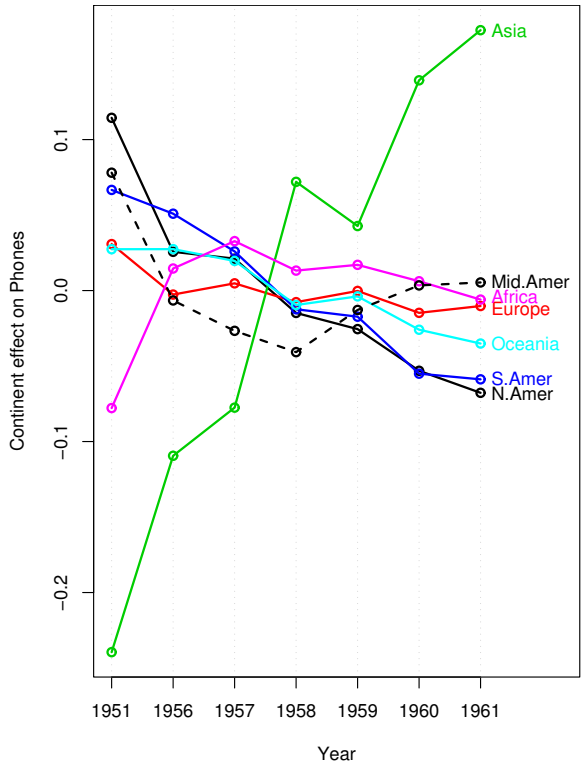


No interaction term means the ratio of expenditures between segments is constant over time.

Number of telephones across the world (in thousands):

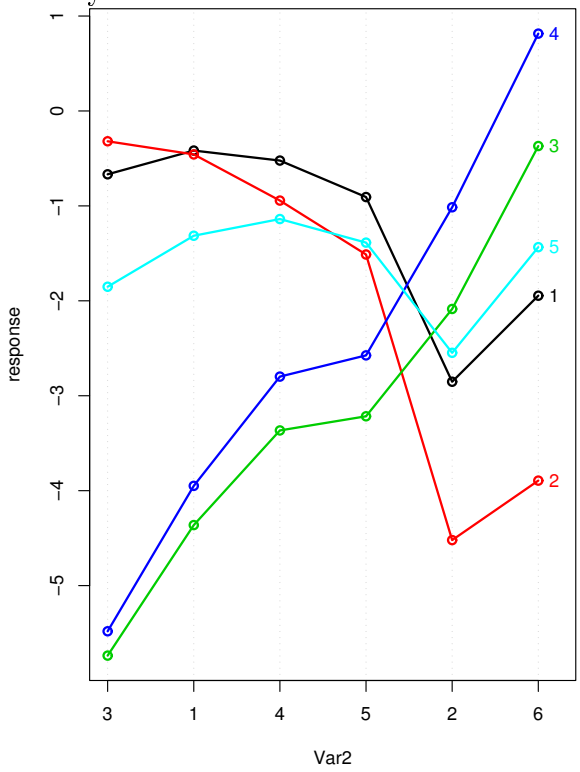


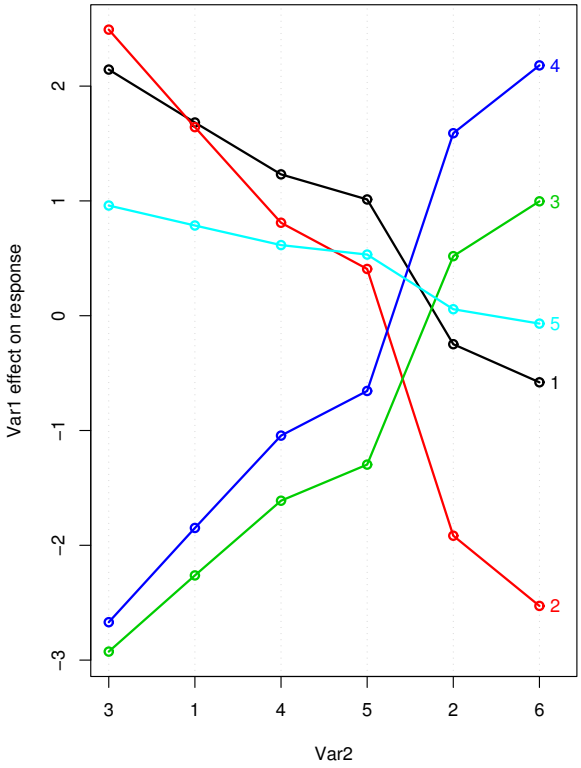


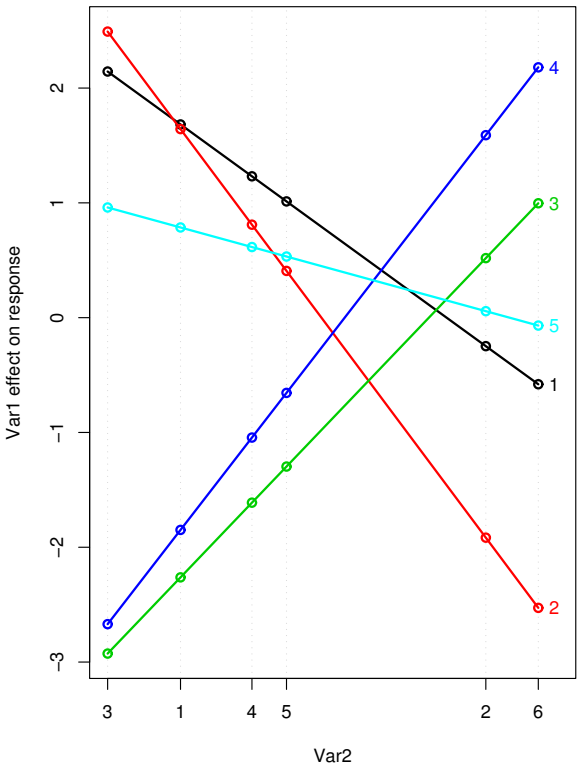


Residuals clarify the interaction.

A perfectly bilinear interaction term:



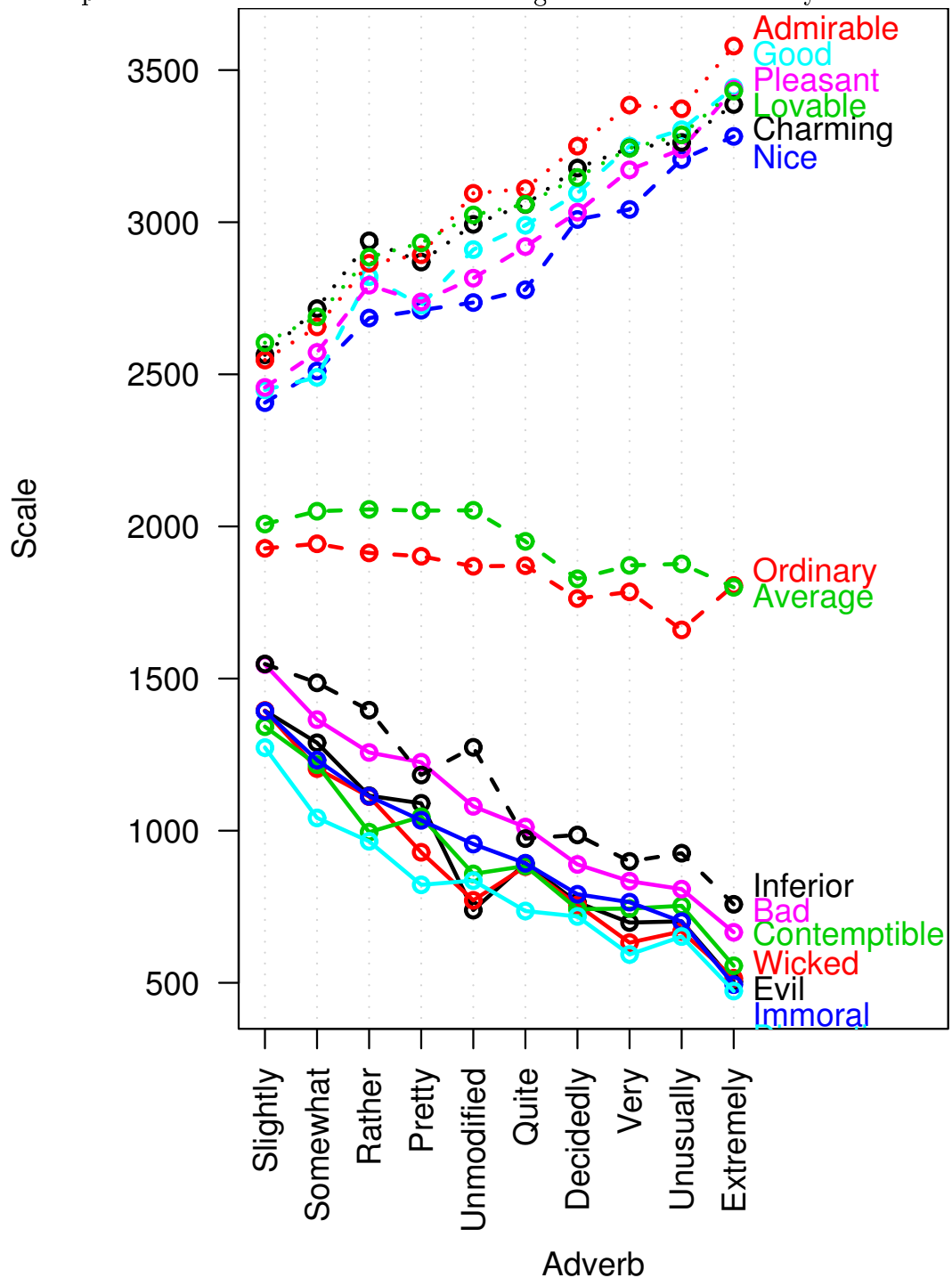




Sorting and spacing the categories helps reveal bilinearity.

$$y = m + a_i + b_j + u_i v_j$$

Adverb-adjective pairs. Is it better to be "rather average" or "rather ordinary"?



Sorting and spacing the categories to make the lines straight also puts the adverbs in a natural order.