Section 2.1 - Scatterplots

Statistics 104

Autumn 2004

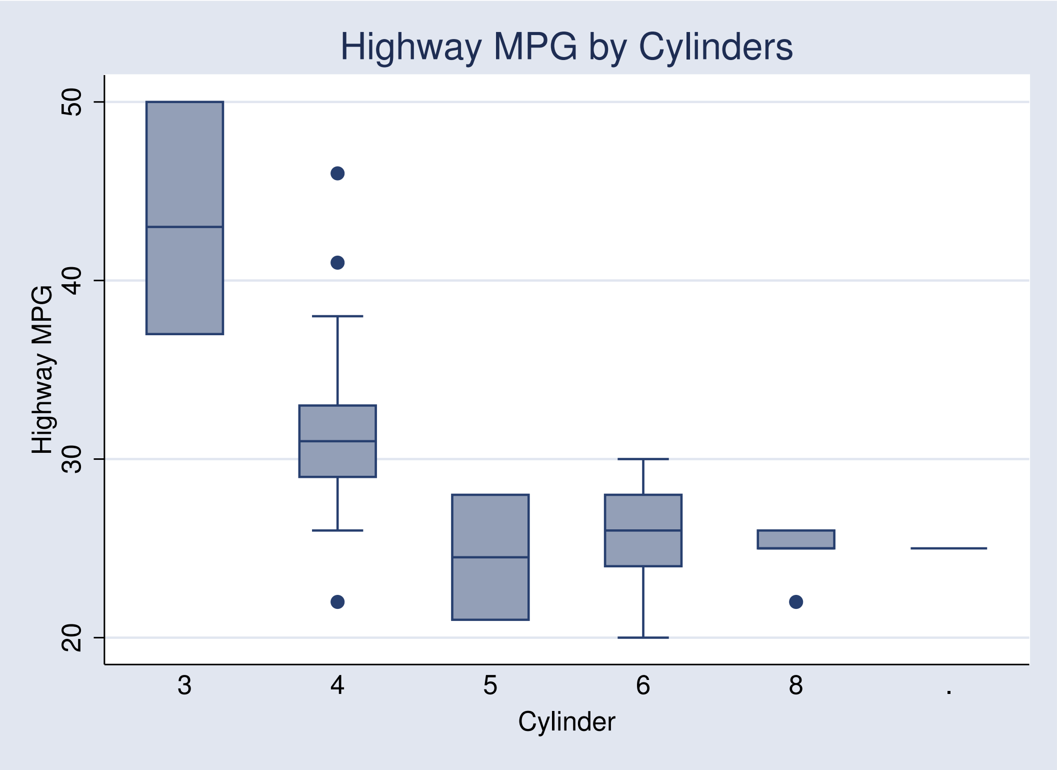
Looking at Data: Relationships

Examples:

- Temperature and number of O-ring failures in Shuttle launches
- AZT doseage and preventing onset of AIDS
- Smoking and lung cancer
- Engine size and gas milage
- Crude oil prices and GNP

Terminology

Association Two variables measured on the same individuals are associated in some values of one variable tend to occur more often with some values of the second variable than with other values of that variable.



Response variable Outcome measure of a study

Explanatory variable A variable which explains or causes changes in the response variables.

Examples:

# O-ring failures	Launch temperature
Response	Explanatory
Engine size	Highway MPG
Explanatory	Response
Twin 1 height	Twin 2 height
???	???

Alternative terminology

Explanatory variable:

- Predictor variable
- Independent variable
- Exogenous variable (Economics)

Response variable:

- Dependent variable
- Endogenous variable (Economics)

As independent and dependent have other meanings in statistics and probability, we won't use these terms in this class.

Scatterplots

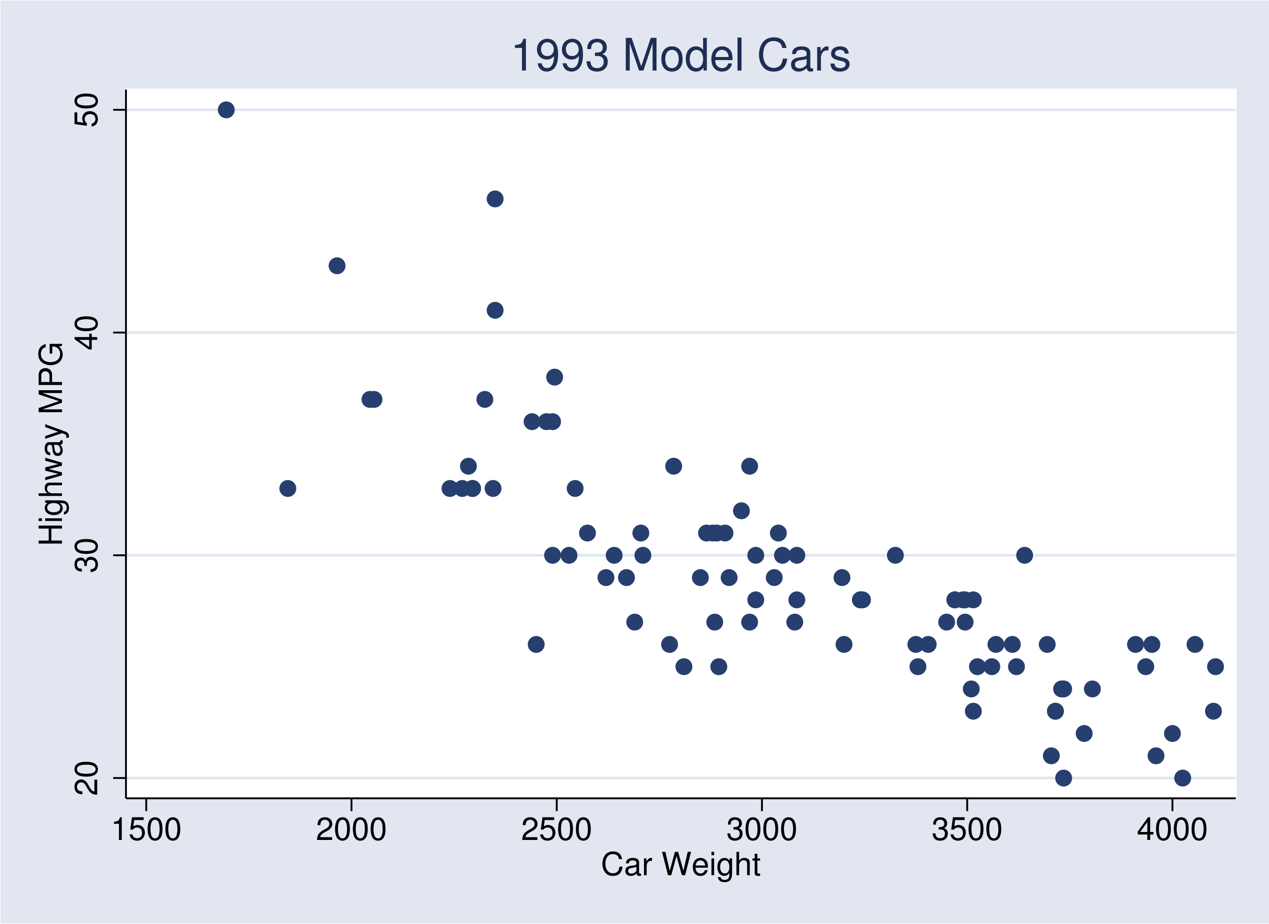
Useful for comparing 2 quantitative variables.

Scatterplot conventions

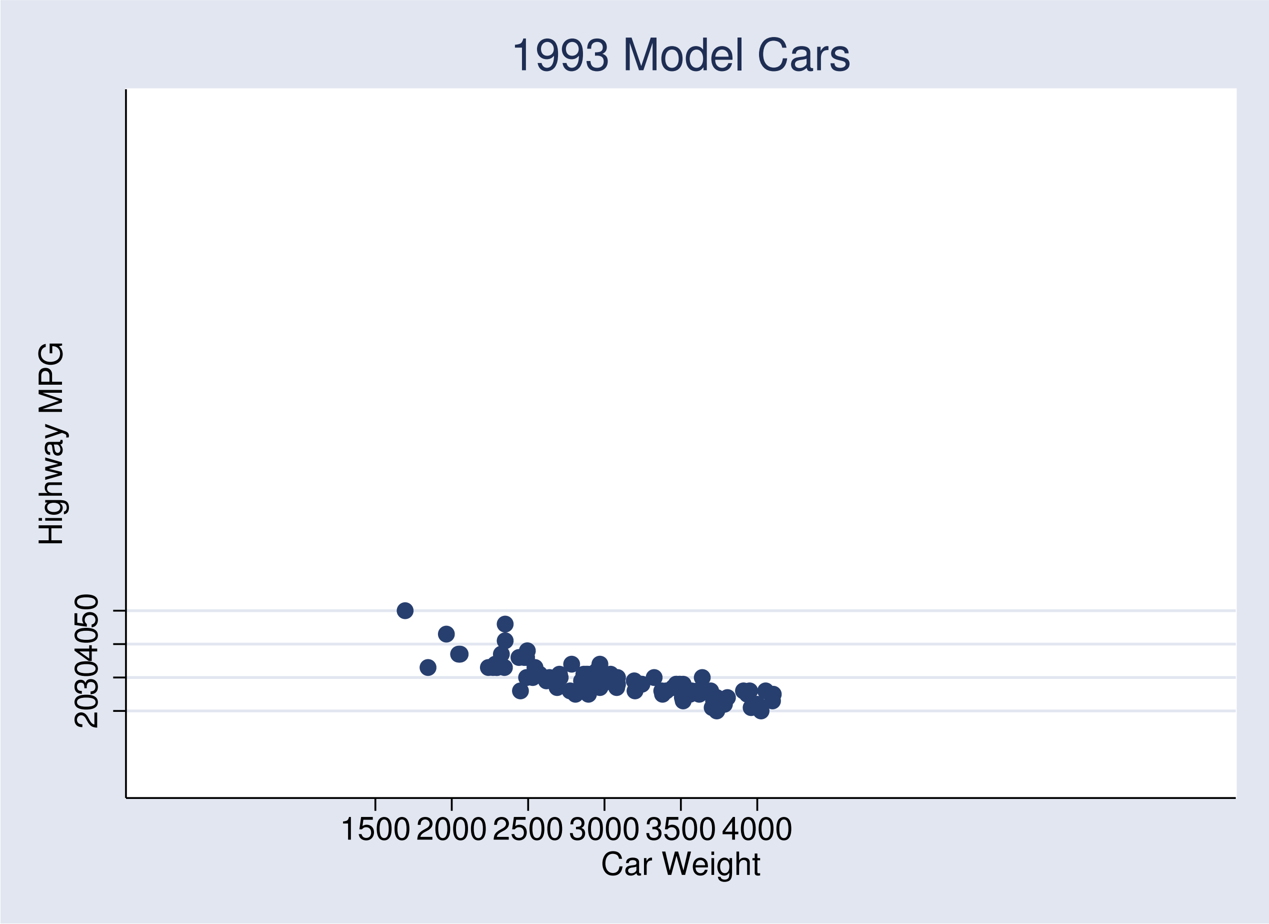
1. x-axis (horizontal): explanatory variable
   y-axis (vertical): response variable

2. Pick ranges for axes that match the data

Example: Car Weight vs Highway MPG



Stata Default



Adjusted ranges

For scatterplots, you must have paired (matched) data

Example: Comparing to brands of gasoline

1. 40 cars - 20 cars get Hess, 20 cars get Getty

2. 20 cars - 20 cars get Hess on 1st day, 20 cars get Getty on 2nd day

In the first case, the data aren't paired. How do you decide which observation on Getty gas gets plotted with which observation on Hess.

A study like this would be better examined with with side by side box plots.

In the second case, the data is paired and it is clear on how to do the plot.

Note: while the pairing is a good idea, this particular study design isn't very good to answer a question like this. We'll see why when we get to chapter 3.

Interpreting scatterplots

Look for overall pattern and striking deviations

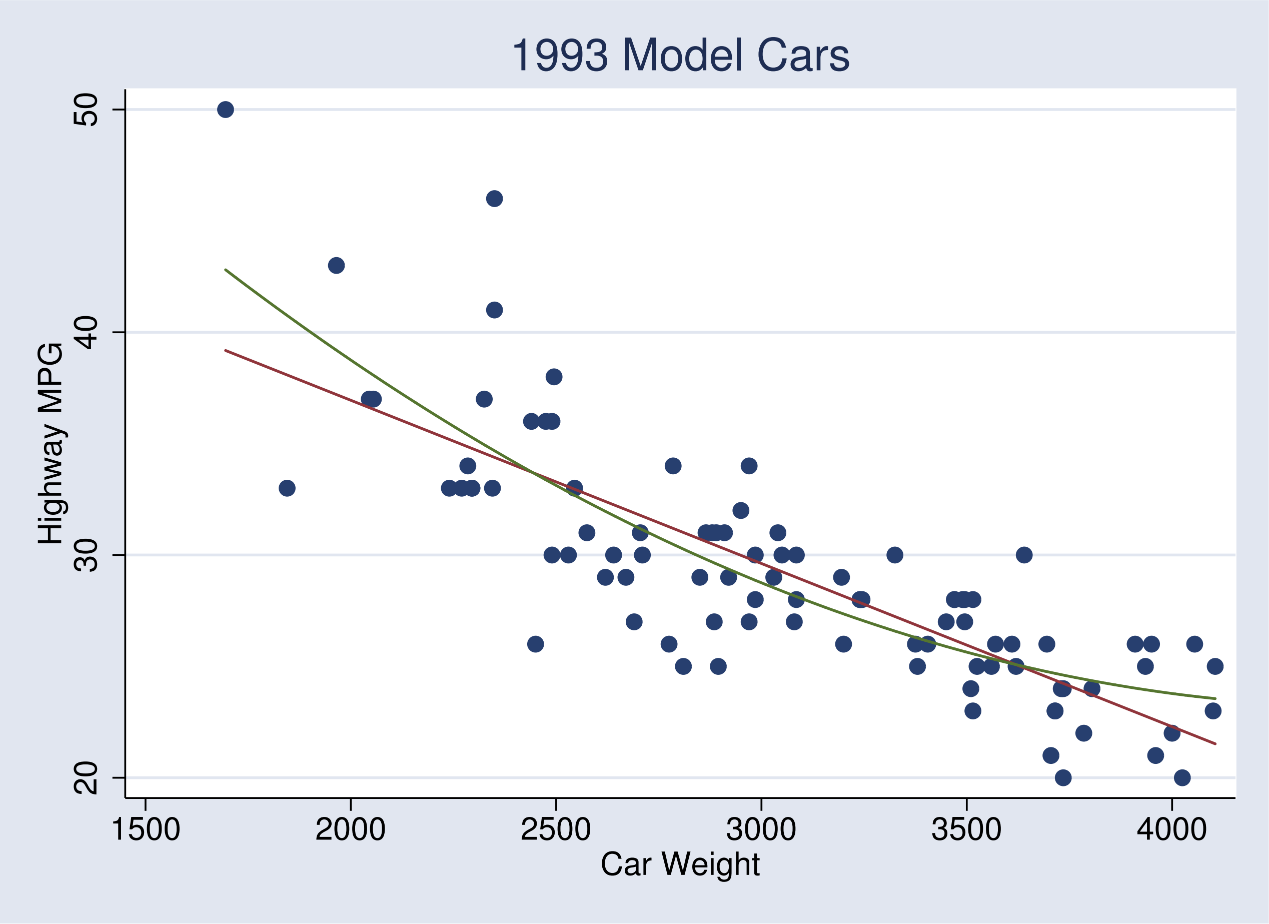
Form, direction, and strength

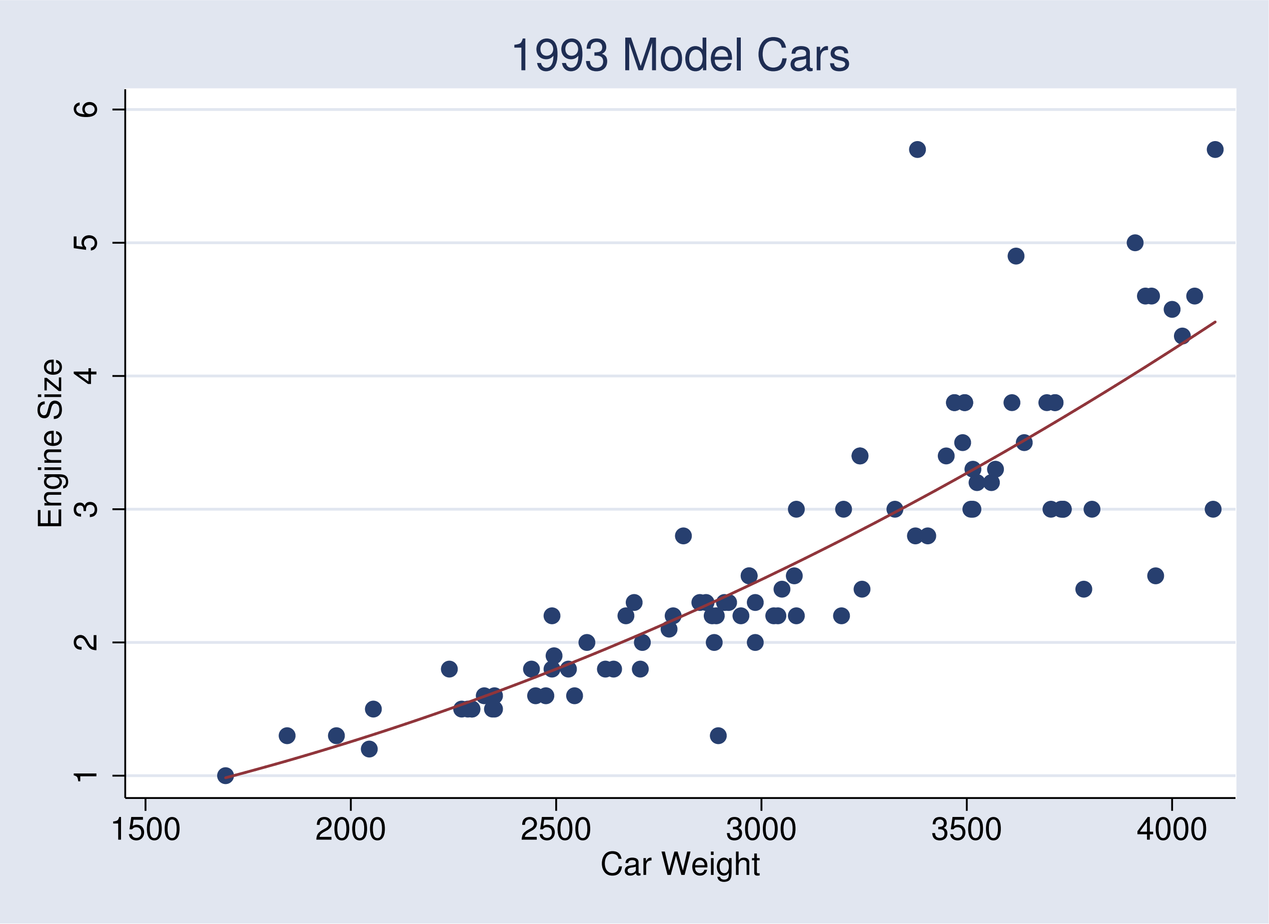
- Form: linear or curved

- Direction:

  Increasing - larger x's tend to be with larger y's (Positive association)

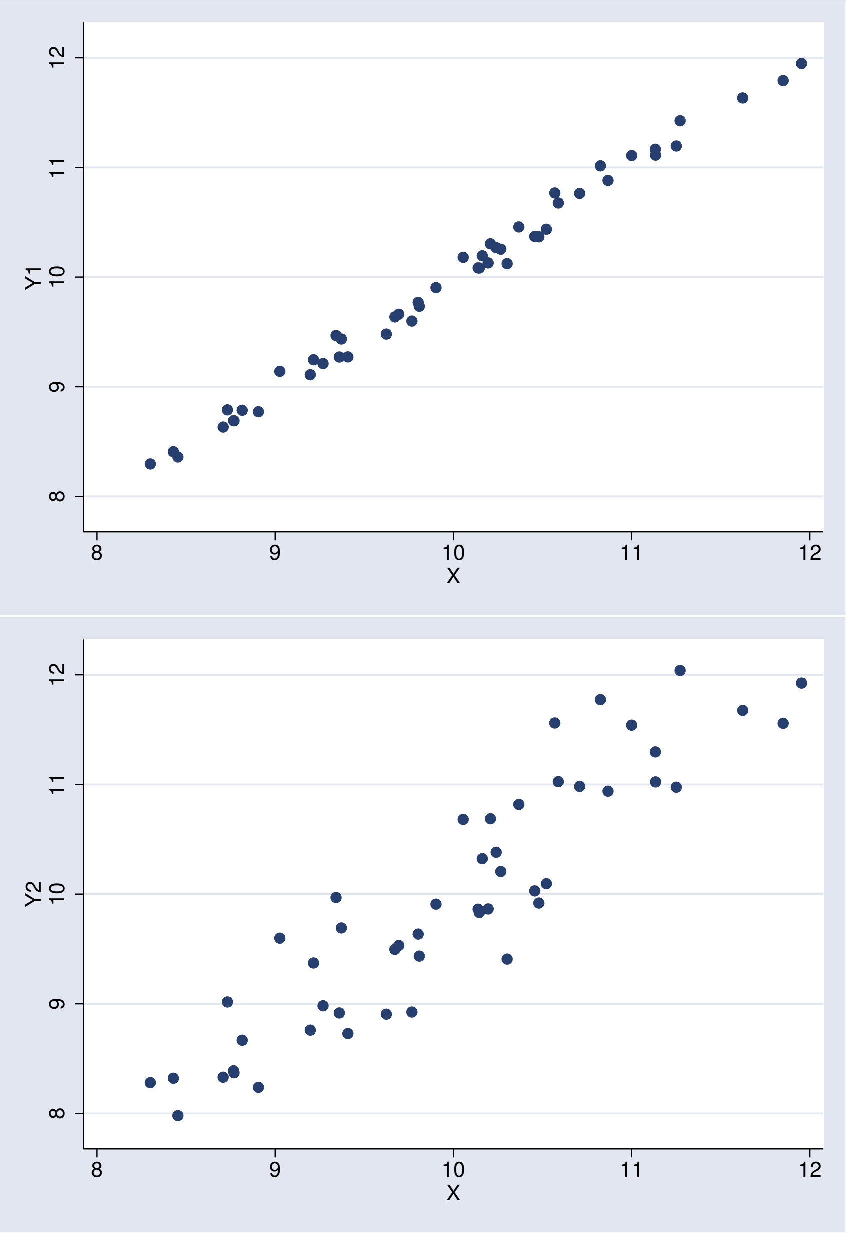
  Decreasing - larger x's tend to be with smaller y's (Negative association)


1993 Model Cars
Highway MPG
20
30
40
50
1500
2000
2500
3000
3500
4000
Car Weight



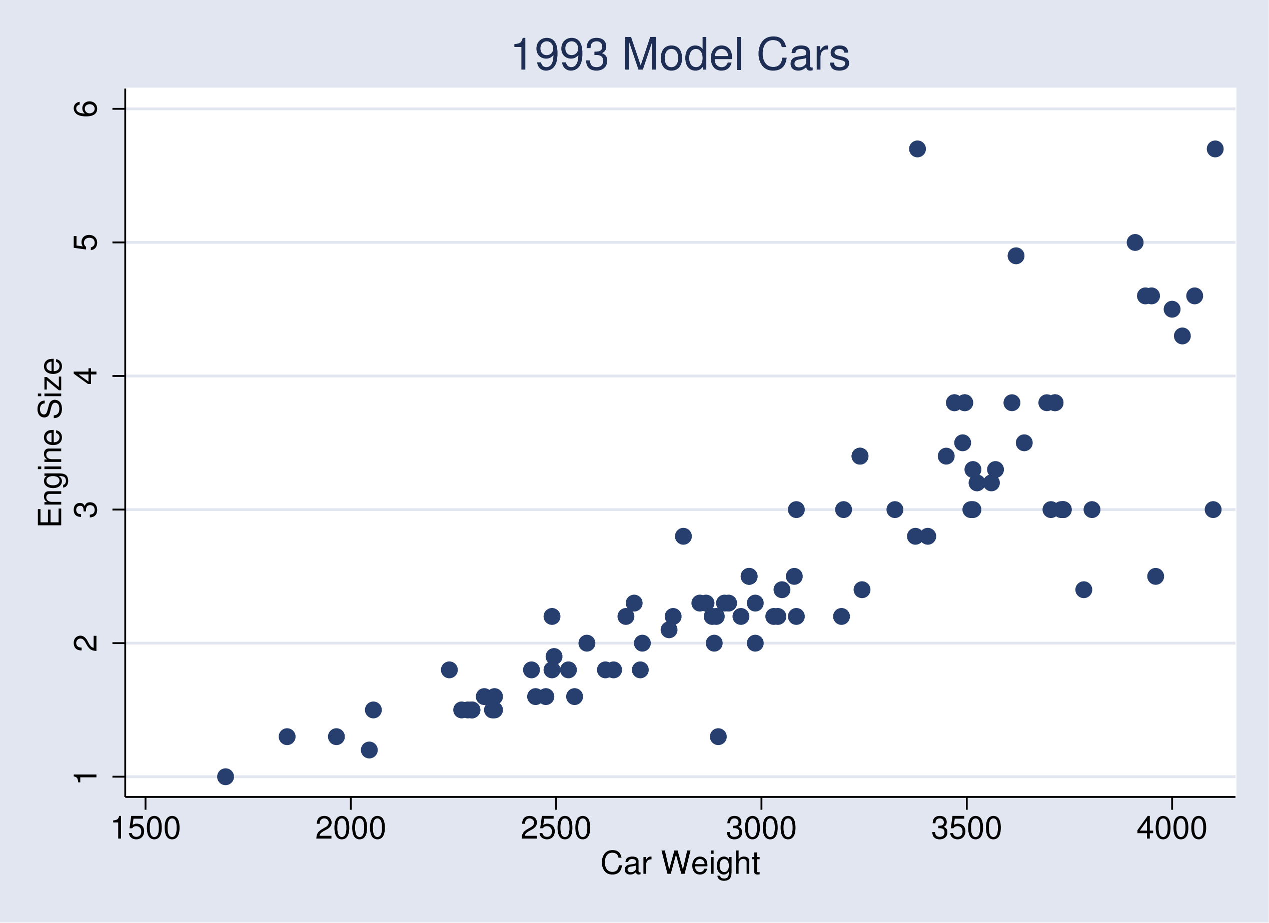
- Strength: Are points tightly clustered around the basic pattern or are they more spread out.

  Is the spread around the pattern the same across the range of the x's.


Y1
8
9
10
11
12
X
8
9
10
11
12
Y2
8
9
10
11
12
X
8
9
10
11
12

- Outliers (Deviations from the basic pattern)

  Look at values that have the same (or similar) x values.



For example, lets look at cars weighing around 3500 pounds.

Most have engine sizes from 2.8 to 3.8 litres (approximately). However one car has an engine size close to 6 litres.

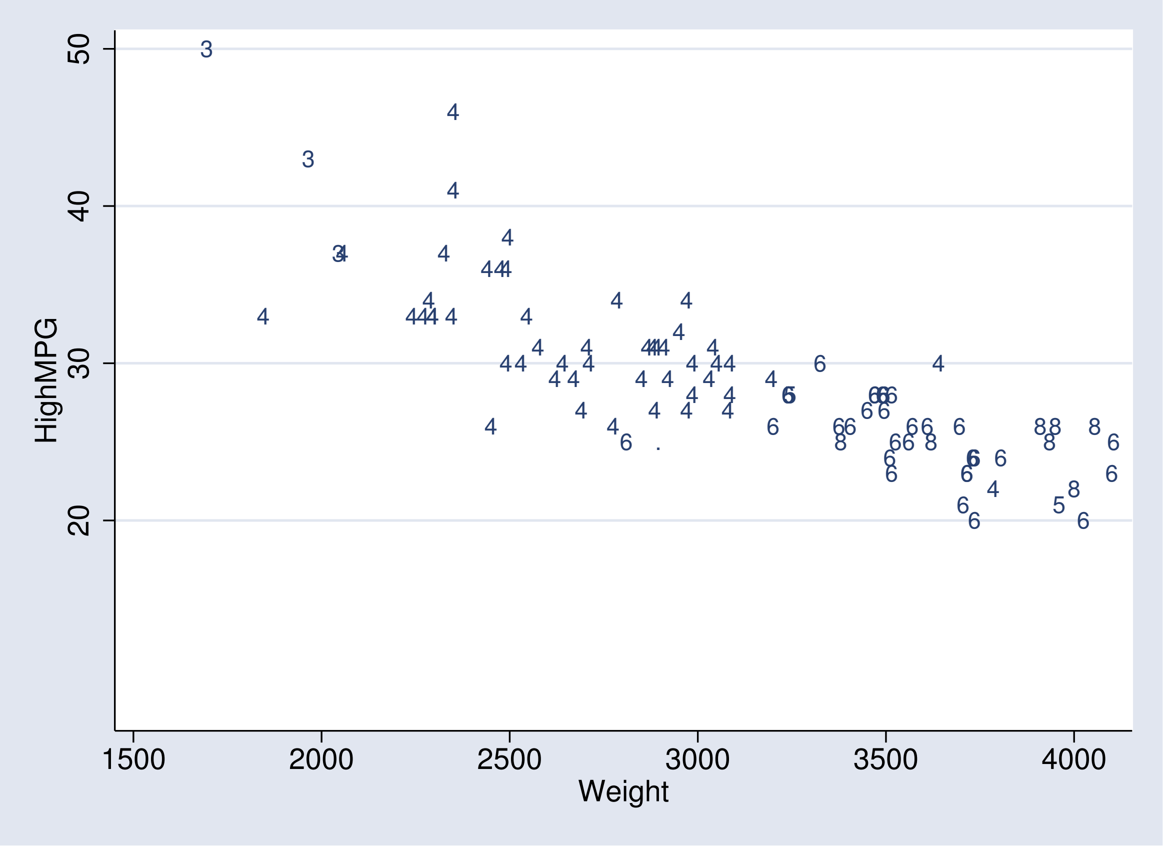
It happens to be the Corvette (weight = 3380 lbs, engine size = 5.7 litres).

Some other cars with similar weights are the Ford Taurus (3325 lbs), the Pontiac Bonneville (3495 lbs), and the Audi 90 (3375 lbs)

The Corvette is the only Sporty cars with a weight between 3300 and 3500.

Adding categorical variable information

Give each level of a categorical variable a different label



```
twoway (scatter HighMPG Weight, msize(vtiny) mcolor(none)
mlabel(Cylinder) mlabposition(0)), yscale(range(7.8 12.2))
```

Scatterplot smoothing

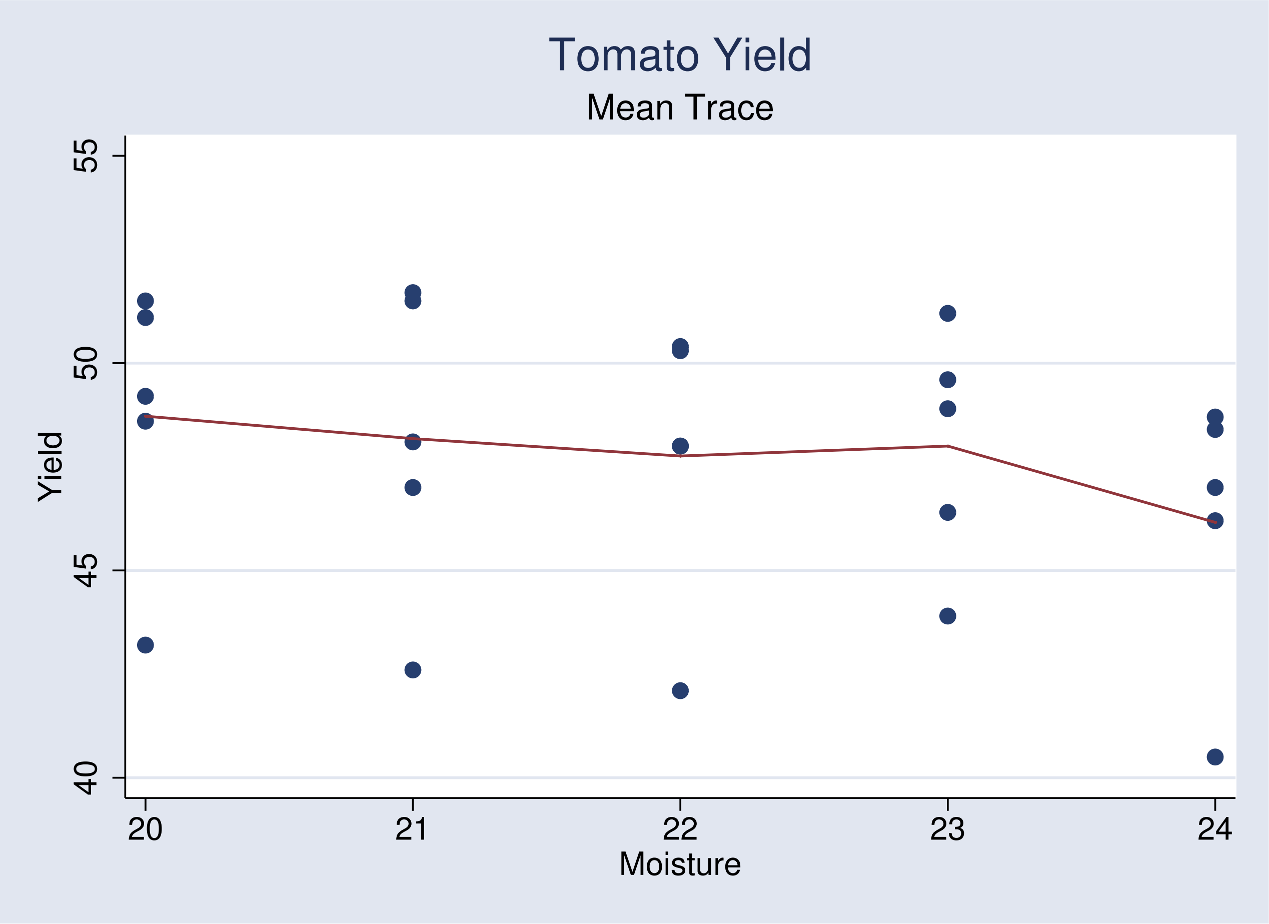
An approach for trying to describe the pattern in a scatter plot

- Mean trace (Like Figure 2.4 in Example 2.5)

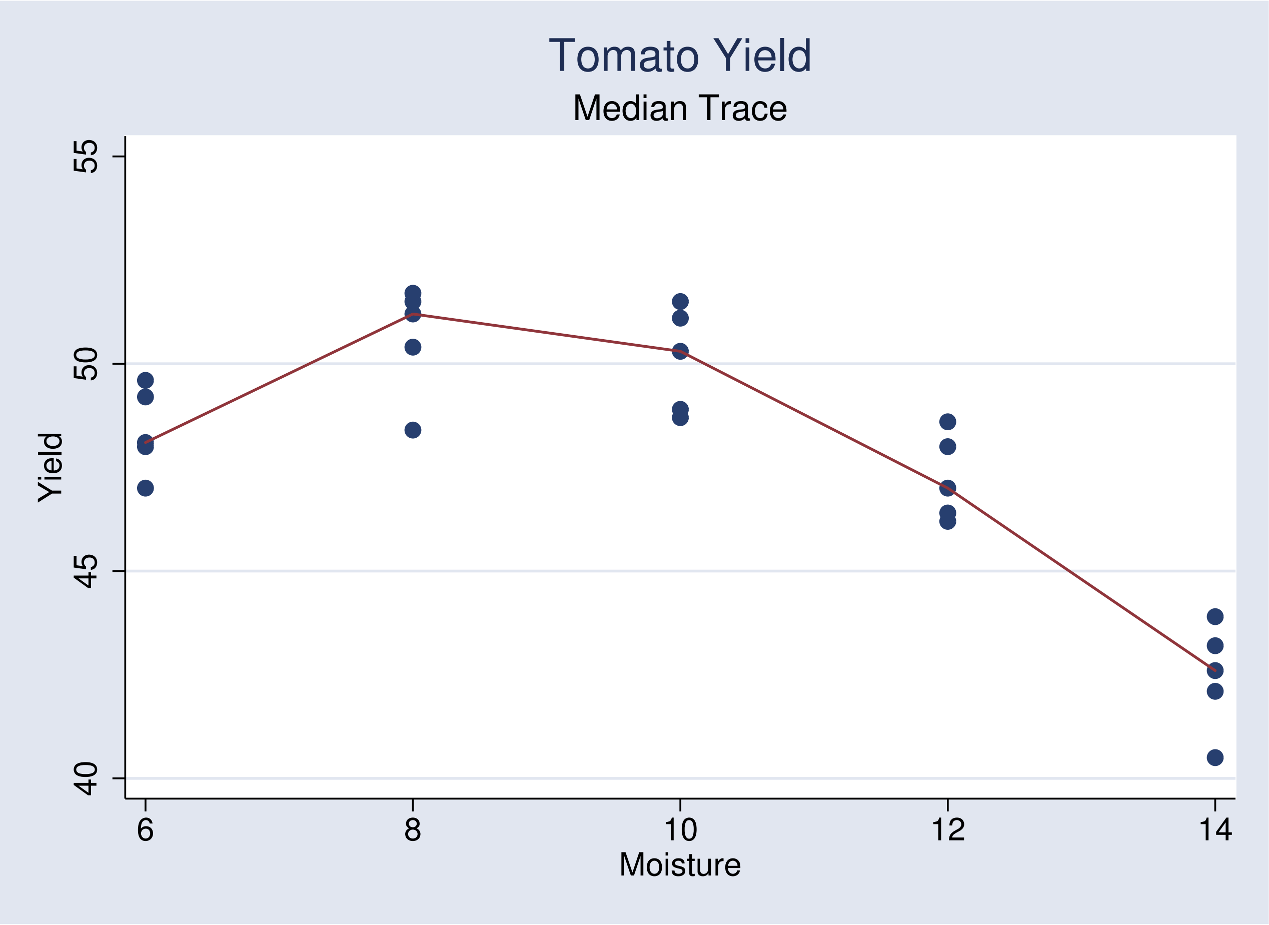
  Average all the y-values with the same x value. Plot each of the averages against its corresponding x value.

  Example: Tomato yield

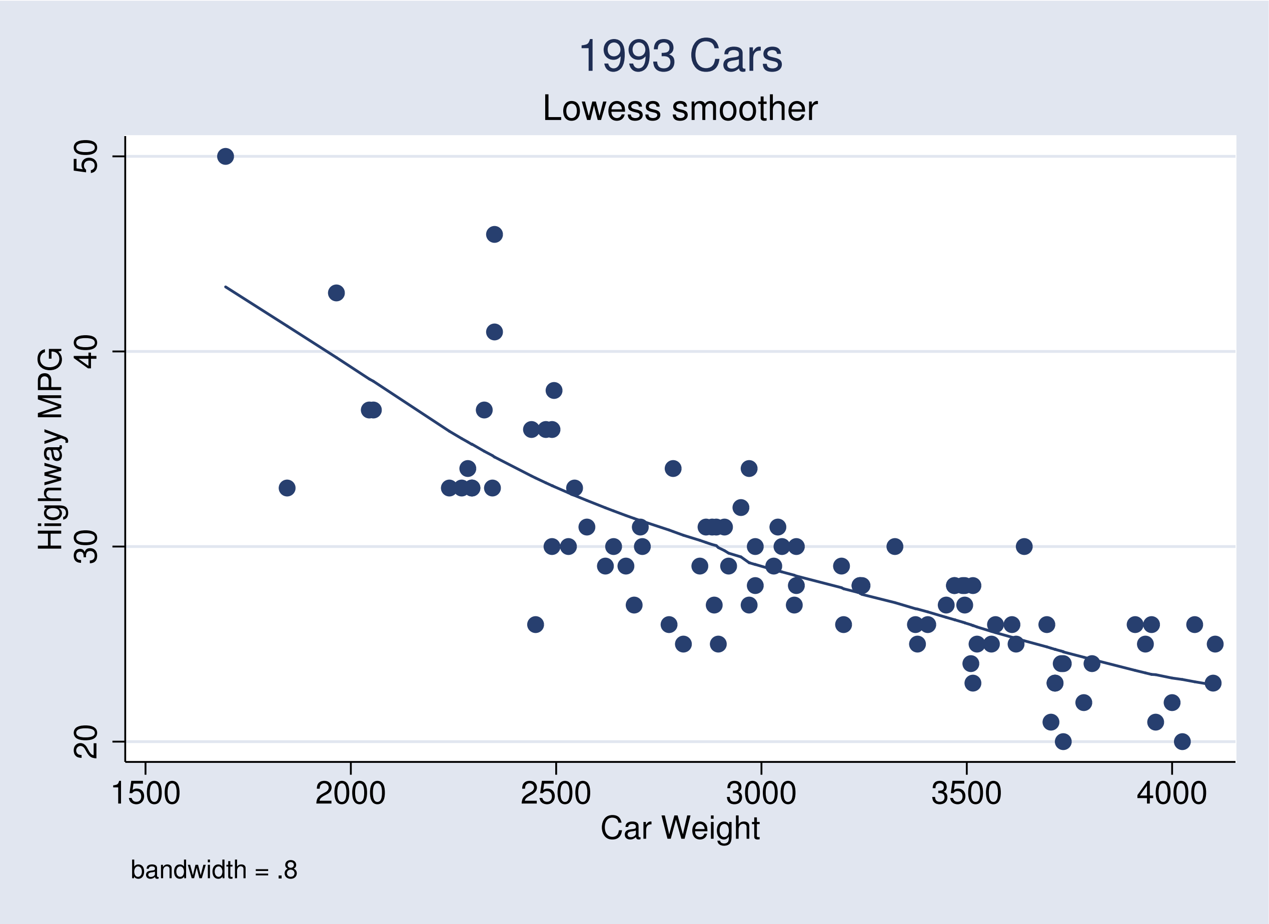
  Experiment looking at the effect of Temperature (20-24°C) and Moisture (6, 8, 10, 12, 14 in) on Tomato yield. One observation per temperature/moisture combination.


Tomato Yield
Mean Trace
Yield
55
50
45
40
20
21
22
23
24
Moisture

- Median trace - use medians instead of means

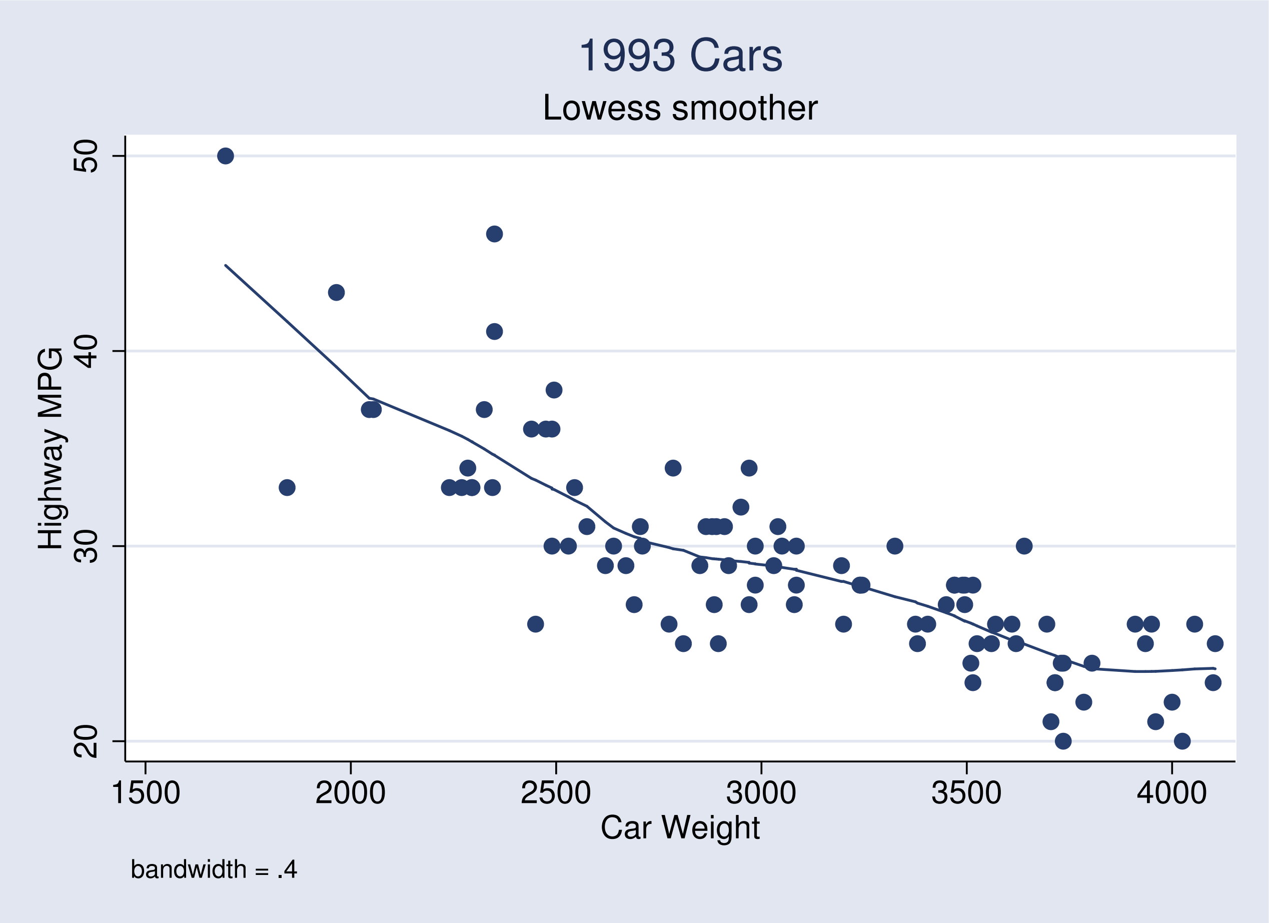


- Lowess smoother



```
lowess HighMPG Weight, ytitle(Highway MPG) xtitle(Car Weight)
title(1993 Cars) subtitle(Lowess smoother)
```

The smoothness of the smoother can be adjusted by the bandwidth



```
lowess HighMPG Weight, bwidth(0.4) ytitle(Highway MPG)
xtitle(Car Weight) title(1993 Cars) subtitle(Lowess smoother)
```

There are many other types of smoothers

- Smoothing splines
- Kernel smoother
- Nearest neighbour