IQR Rule for Outliers

1. Arrange data in order.

2. Calculate first quartile (Q1), third quartile (Q3) and the interquartile range (IQR=Q3-Q1). CO2 emissions example: Q1=0.9, Q3=6.05, IQR=5.15.

3. Compute Q1−1.5 × IQR (=−6.825) Compute Q3+1.5 × IQR (=13.775) Anything outside this range is an outlier.

So by this criterion, US at 19.7 is an outlier, Russia at 9.8 is not.

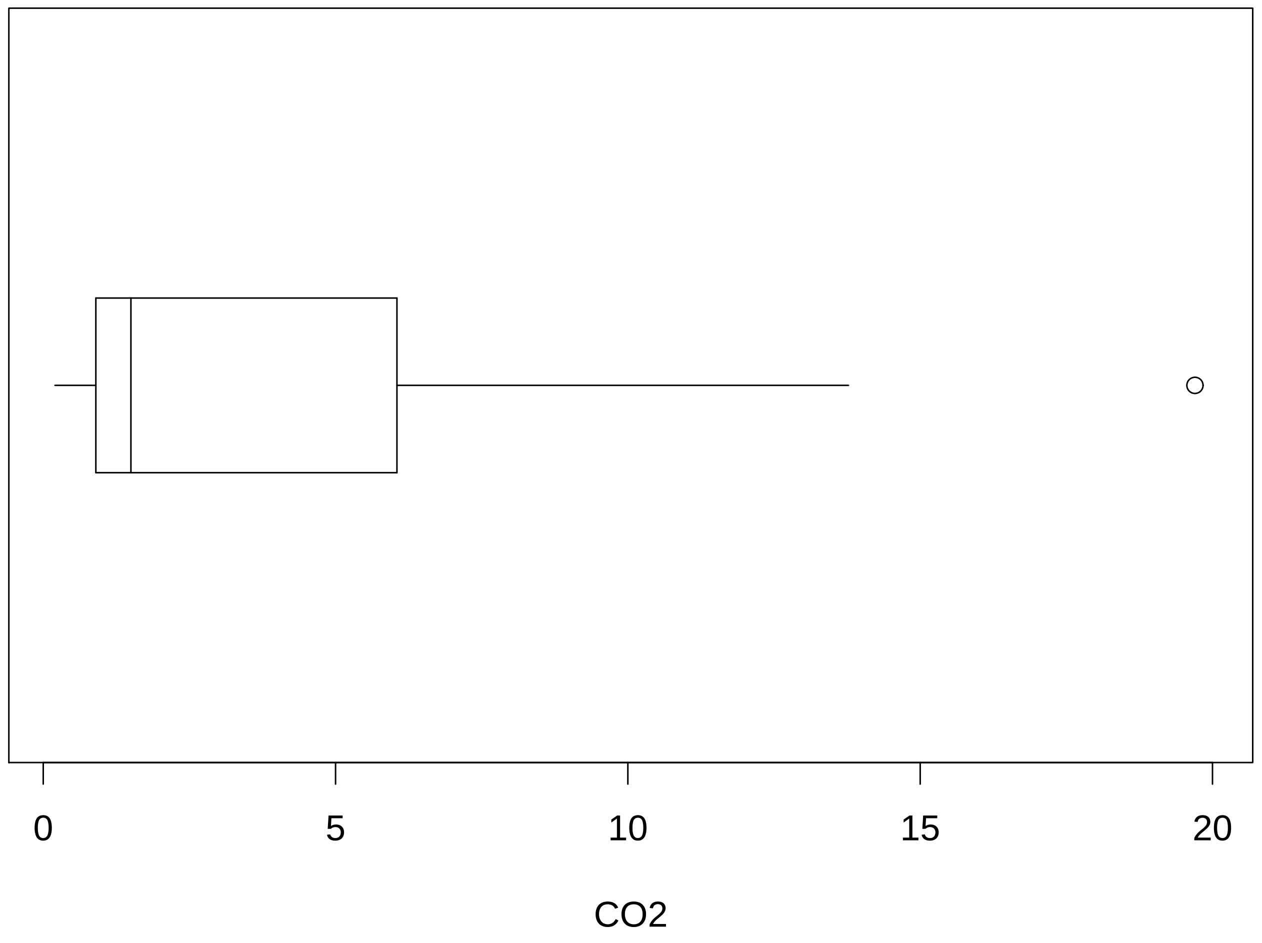
Exercise: Are there any outliers in the datasets of class heights? (Q1=63, Q3=68.5, min and max observations are 60 and 77)

The Boxplot

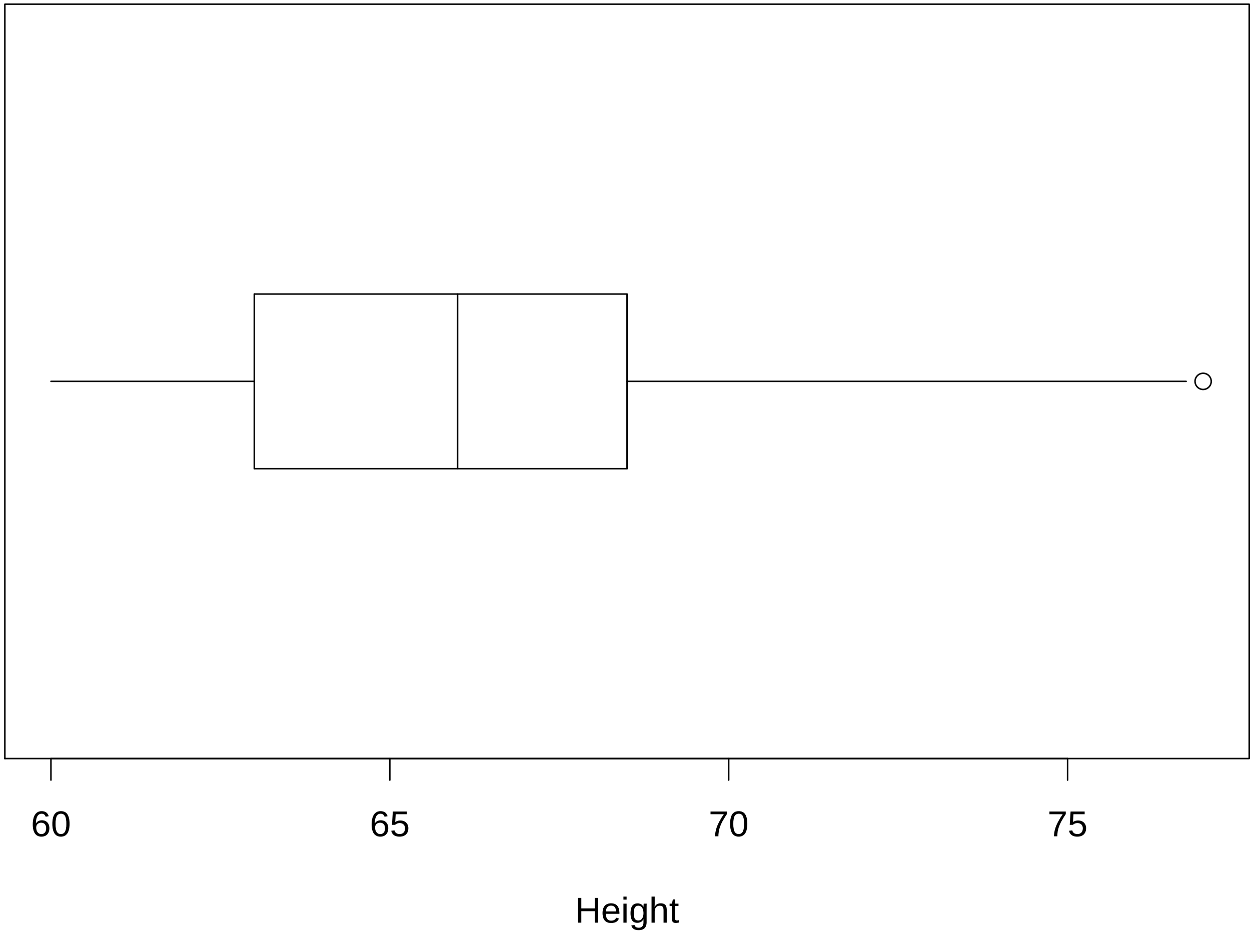
Purpose: a simple graphical device to display the overall shape of a distribution, including the outliers.

1. Calculate Q1, median, Q3 and the 1.5 IQR outlier limits.
2. Draw a “box” from Q1 to Q3 with bars at Q1, Q3 and the median. (In these examples the box is horizontal, but it could also be vertical.)
3. Draw a straight line from Q3 to *either* the largest observation *or* the Q3+1.5 IQR upper outlier bound, whichever is smaller.
4. Draw a straight line from Q1 to *either* the smallest observation *or* the Q1-1.5 IQR lower outlier bound, whichever is larger.
5. Any remaining observations (the outliers) are shown as individual points on the plot.

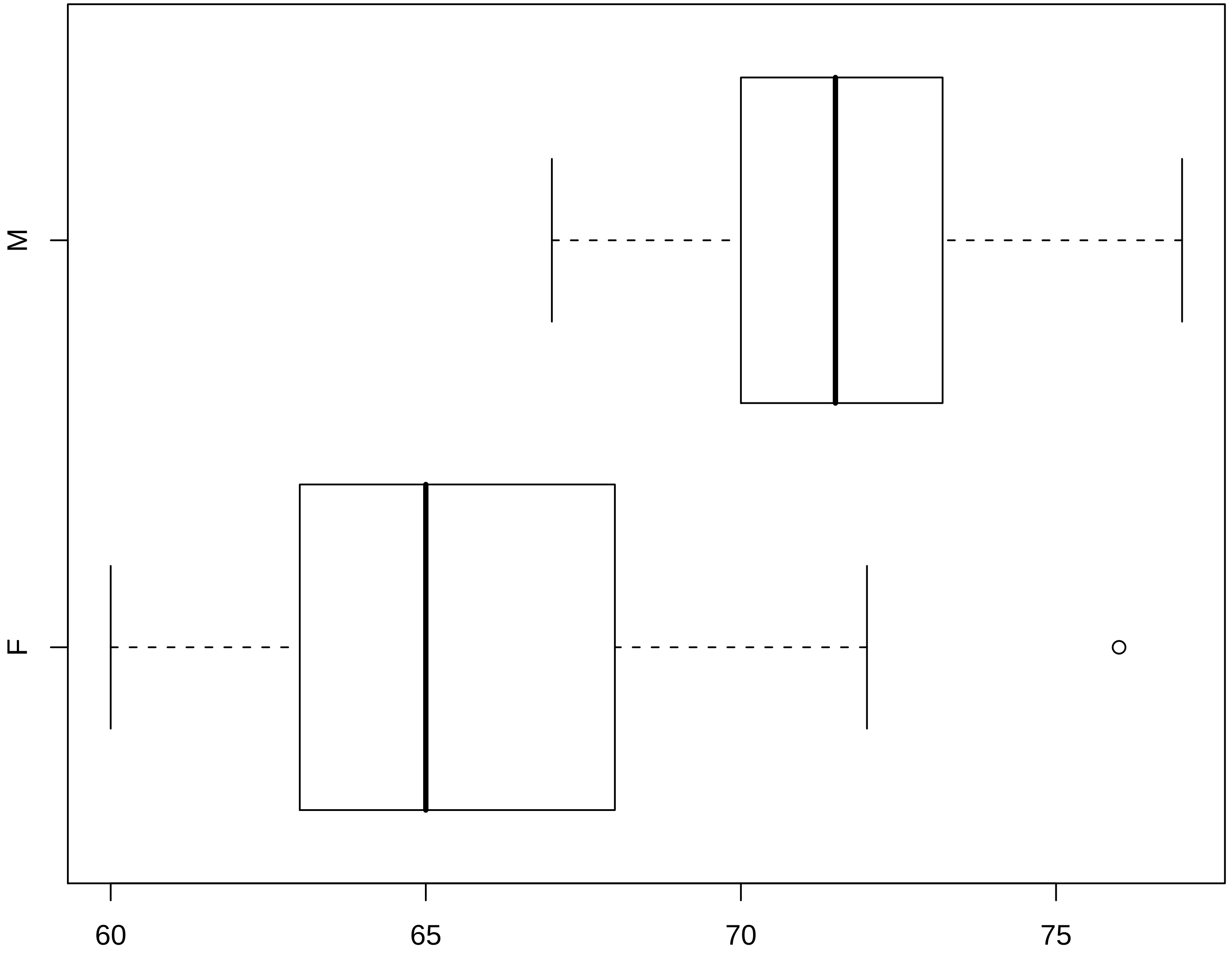
Box plot of CO2 data



Box plot of student heights



Side by side boxplots for M/F (thanks to Vangelis)



Chapter 3:
Association, Correlation and Regression

The **response variable** is the outcome variable on which comparisons are made.

The **explanatory variable** defines the groups to be compared with respect to values of the response variable.

Association means that the values of the response in some way depend on the explanatory variable. At this level of discussion, talking about association does not imply that there is an actual causal effect, because the association may be spurious (example of mortality rates in British women, grouped into smokers and non-smokers)

Contingency Tables

Used when we want to look at associations among two categorical variables.

Each entry or **cell** of the table contains the **frequency** of a particular combination of the two variables.

Note: Frequency is a count, not a proportion. We'll talk next about converting counts into proportions.

Example Based on Political Affiliation by Gender

Party	Female	Male	Total
Democrat	30	4	34
Republican	17	4	21
Independent	10	2	12
Total	57	10	67

Converting Frequencies to Proportions

The key point is that there are different ways to do this.

Unconditional proportions: express everything as proportion of the grand total (67).

Party	Female	Male	Total
Democrat	.448	.060	.507
Republican	.254	.060	.313
Independent	.149	.030	.179
Total	.851	.149	1.000

Conditional proportions: if we're interested in comparing party affiliation by gender, divide each column by the total for that column.

Party	Female	Male	Total
Democrat	.526	.400	.507
Republican	.298	.400	.313
Independent	.175	.200	.179
Total	1.000	1.000	1.000

We could also standardize by row instead of by column. In this example, it is arguable that knowing the proportion of women among Democrats is less interesting than knowing the proportion of Democrats among women (especially when the distribution of men/women in the sample is very far from 50:50). However, as a statistical operation, either form of standardization is valid.

Associations of Categorical Variables

The question arising from all this is, when is there an association?

Two variables are associated if the conditional proportions of the response variable depend on the explanatory variable.

Note that this definition does not settle how large the samples need to be for the differences to be “significant”.

Associations of Quantitative Variables

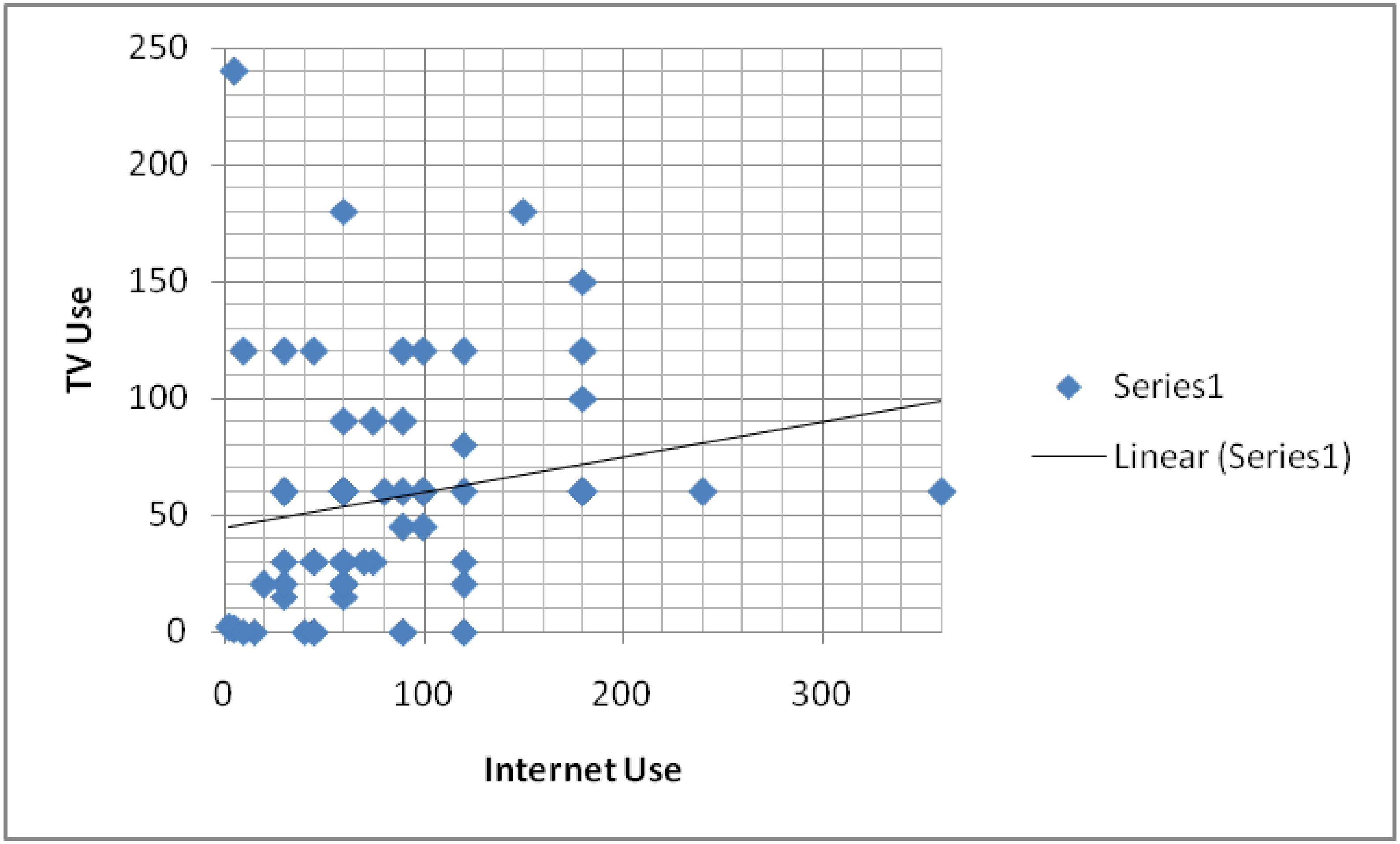
Different tools — leading role play by **scatterplots**.

Different uses for a scatterplot:

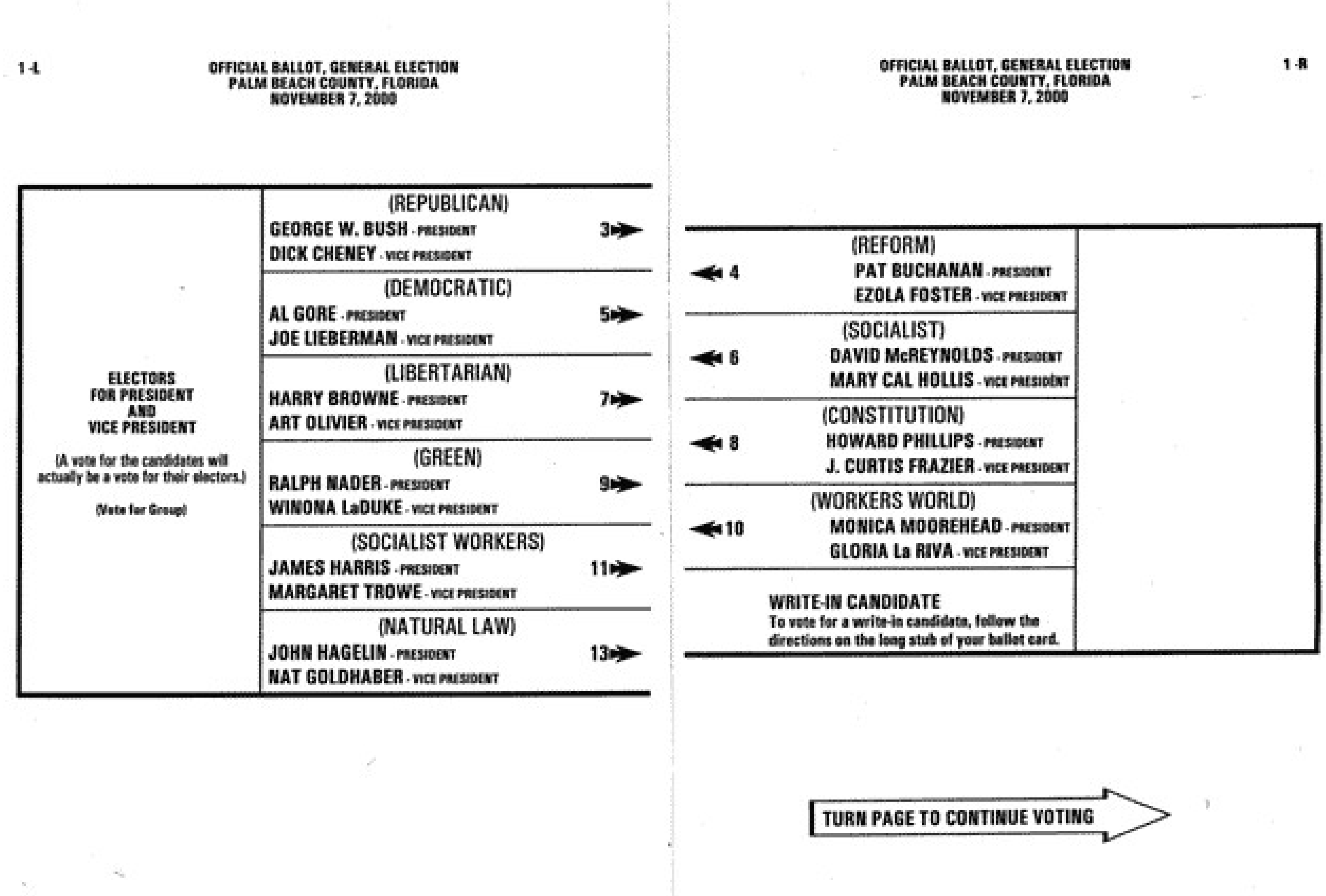
- Look for general associations, e.g. by plotting as trendline (option in Excel)

- A scatterplot can also be useful for detecting other features of the data, e.g. outliers.

Scatterplot of TV use against internet use



The “butterfly ballot”



1-L

OFFICIAL BALLOT, GENERAL ELECTION
PALM BEACH COUNTY, FLORIDA
NOVEMBER 7, 2000

ELECTORS
FOR PRESIDENT
AND
VICE PRESIDENT

(A vote for the candidates will actually be a vote for their electors.)

(Vote for Group)

(REPUBLICAN)
GEORGE W. BUSH - PRESIDENT
DICK CHENEY - VICE PRESIDENT
3➤

(DEMOCRATIC)
AL GORE - PRESIDENT
JOE LIEBERMAN - VICE PRESIDENT
5➤

(LIBERTARIAN)
HARRY BROWNE - PRESIDENT
ART OLIVIER - VICE PRESIDENT
7➤

(GREEN)
RALPH NADER - PRESIDENT
WINONA LaDUKE - VICE PRESIDENT
9➤

(SOCIALIST WORKERS)
JAMES HARRIS - PRESIDENT
MARGARET TROWE - VICE PRESIDENT
11➤

(NATURAL LAW)
JOHN HAGELIN - PRESIDENT
NAT GOLDHABER - VICE PRESIDENT
13➤

OFFICIAL BALLOT, GENERAL ELECTION
PALM BEACH COUNTY, FLORIDA
NOVEMBER 7, 2000

1-R

◄ 4
(REFORM)
PAT BUCHANAN - PRESIDENT
EZOLA FOSTER - VICE PRESIDENT

◄ 6
(SOCIALIST)
DAVID McREYNOLDS - PRESIDENT
MARY CAL HOLLIS - VICE PRESIDENT

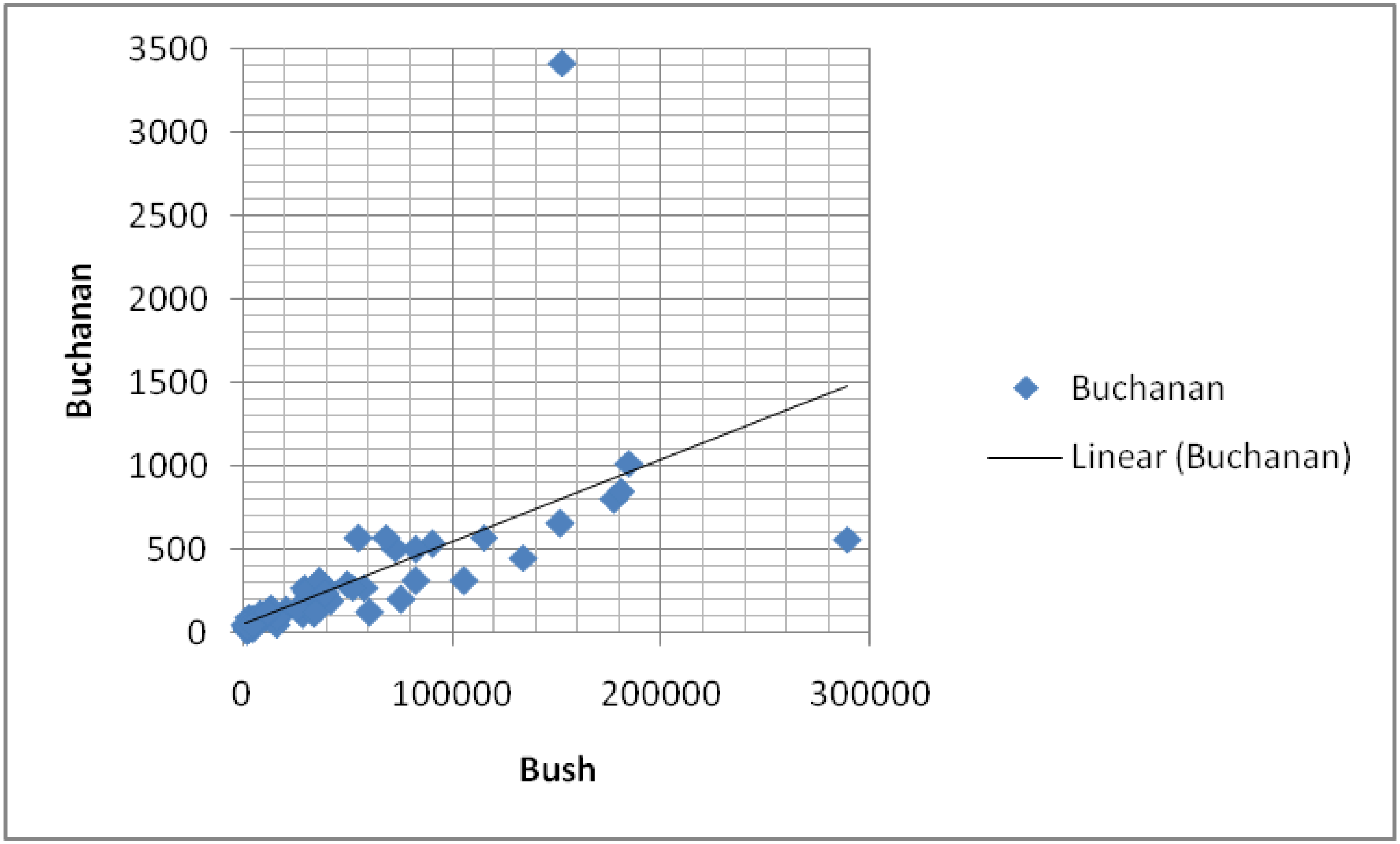
◄ 8
(CONSTITUTION)
HOWARD PHILLIPS - PRESIDENT
J. CURTIS FRAZIER - VICE PRESIDENT

◄ 10
(WORKERS WORLD)
MONICA MOOREHEAD - PRESIDENT
GLORIA La RIVA - VICE PRESIDENT

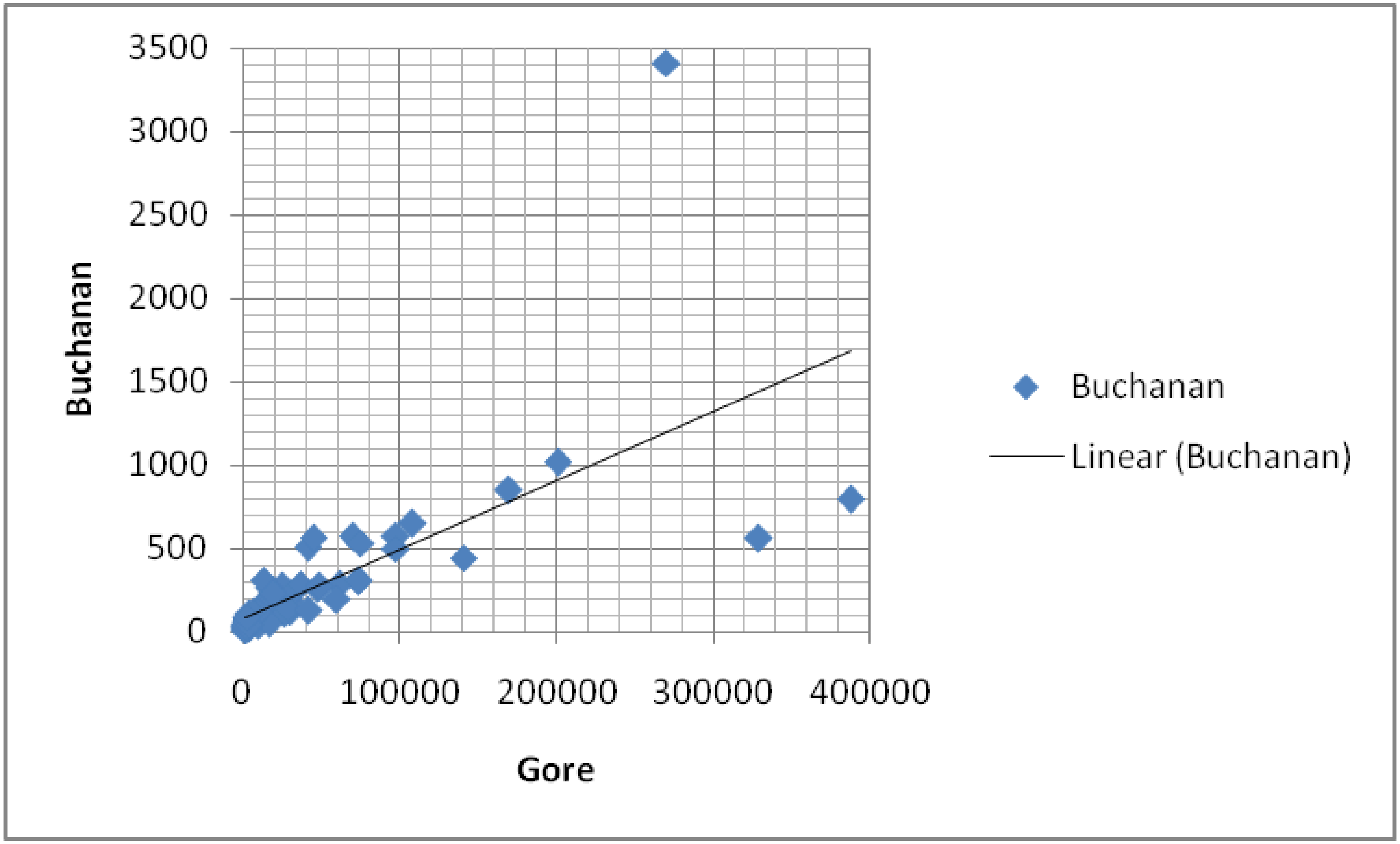
WRITE-IN CANDIDATE
To vote for a write-in candidate, follow the directions on the long stub of your ballot card.

TURN PAGE TO CONTINUE VOTING

Scatterplot of Buchanan vote against Bush vote in Florida 2000



Scatterplot of Buchanan vote against Gore vote in Florida 2000


Buchanan
Gore
◆ Buchanan
Linear (Buchanan)