

# Statistical graphics: The good, the bad, and the ugly

Peter Flom

National Development and Research  
Institutes, Inc

[www.peterflom.com](http://www.peterflom.com)

# Questions to ask

- What to graph?
- Why do you want to graph it?
- Who will see it?
- How are they going to look at it?
- When do you need it done?

# Principles of good graphics (Tufte)

- Show the data
- Induce the viewer to think about the substance
- Avoid distorting the data
- Present many numbers in a small space

# Principles of good graphics (Tufte)

- Make large data sets coherent
- Encourage the eye to look at different pieces of the data
- Reveal the data at several levels of detail
- Serve a clear purpose

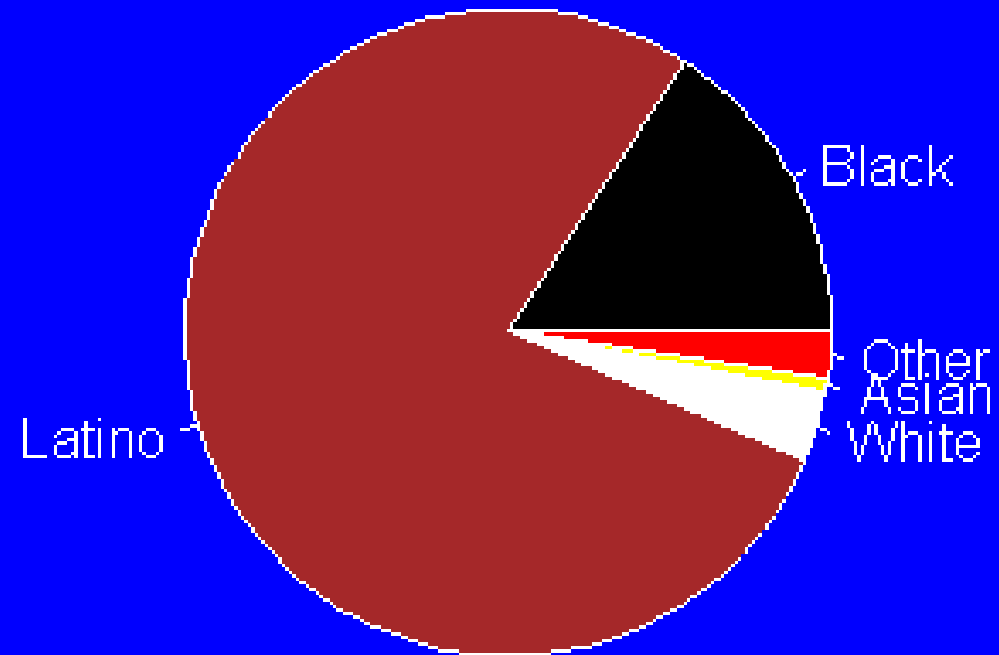
# Principles of good graphics (Cleveland)

- Clear vision
- Clear understanding
- Banking
- Axes, scales, and tick marks

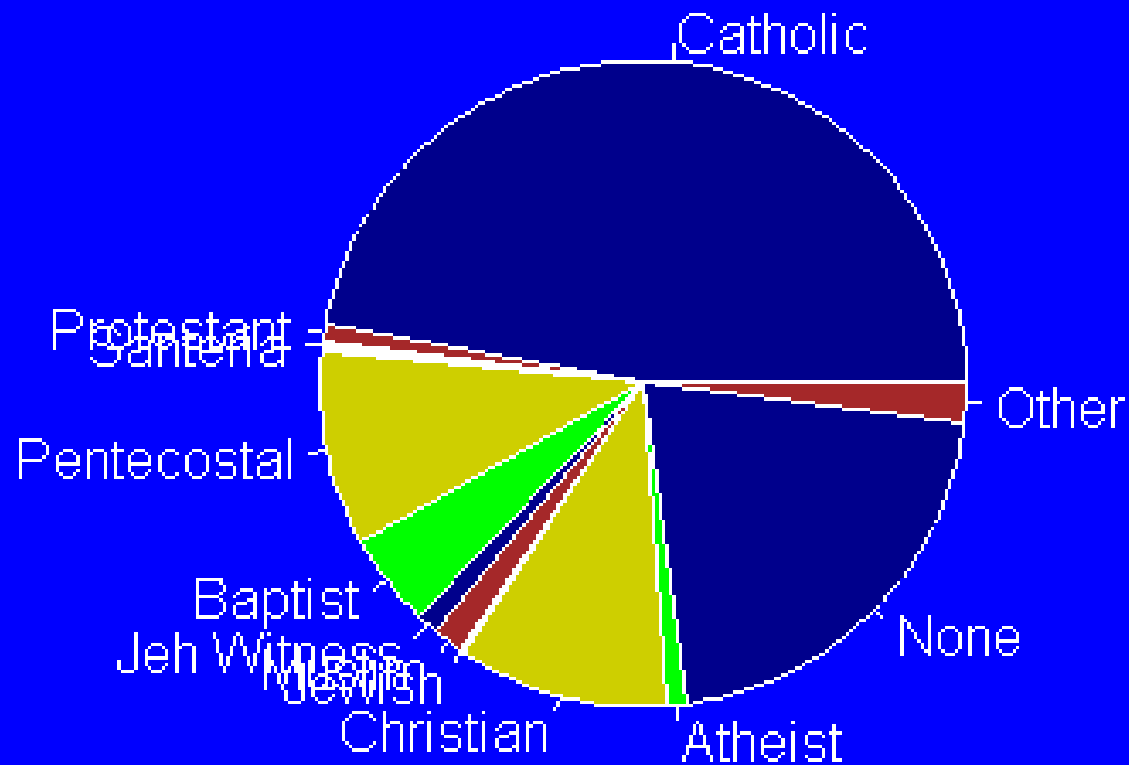
# Univariate categorical data

- With few categories, a table is often best
- Pie charts are to be avoided
- Dot charts are often good

## Pie chart of race/ethnicity in Bushwick

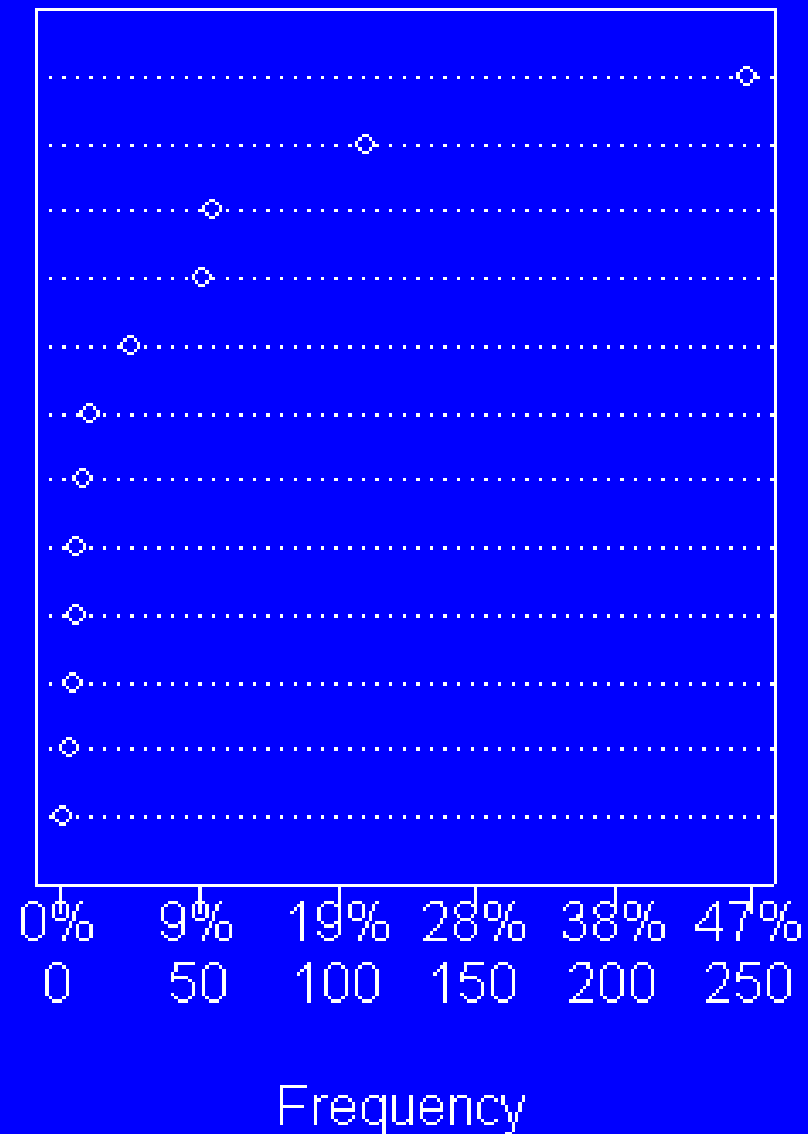


## Pie chart of religions in Bushwick



## Religions in DUHRAY

Catholic  
None  
Christian  
Pentecostal  
Baptist  
Other  
Muslim  
Atheist  
Jeh Witness  
Protestant  
Santeria  
Jewish



# Univariate continuous data

- A histogram is the most common choice, but not ideal. As Cleveland put it, longevity and ubiquity are not guarantees of utility
- A better choice for general purposes is a density plot, with multiple smooths

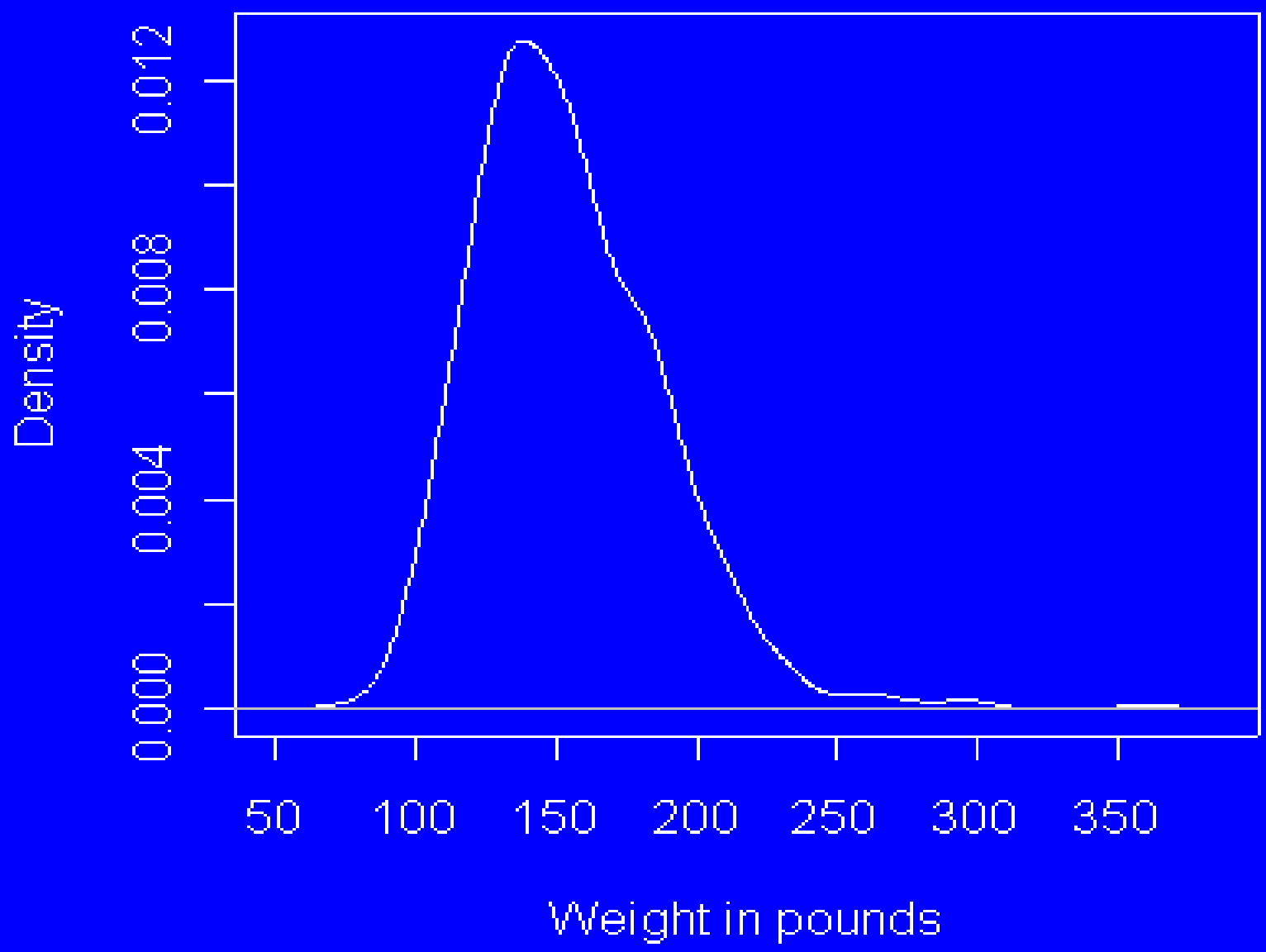
# Problems with histograms

- Dependent on bin width and limits
- Wasteful of space
- Categorizes continuous data
- 2 dimensions for 1 dimensional data

# Basic density plot

- Really just a smoothed histogram
- Still dependent on smoothing
- Doesn't categorize data
- One dimension

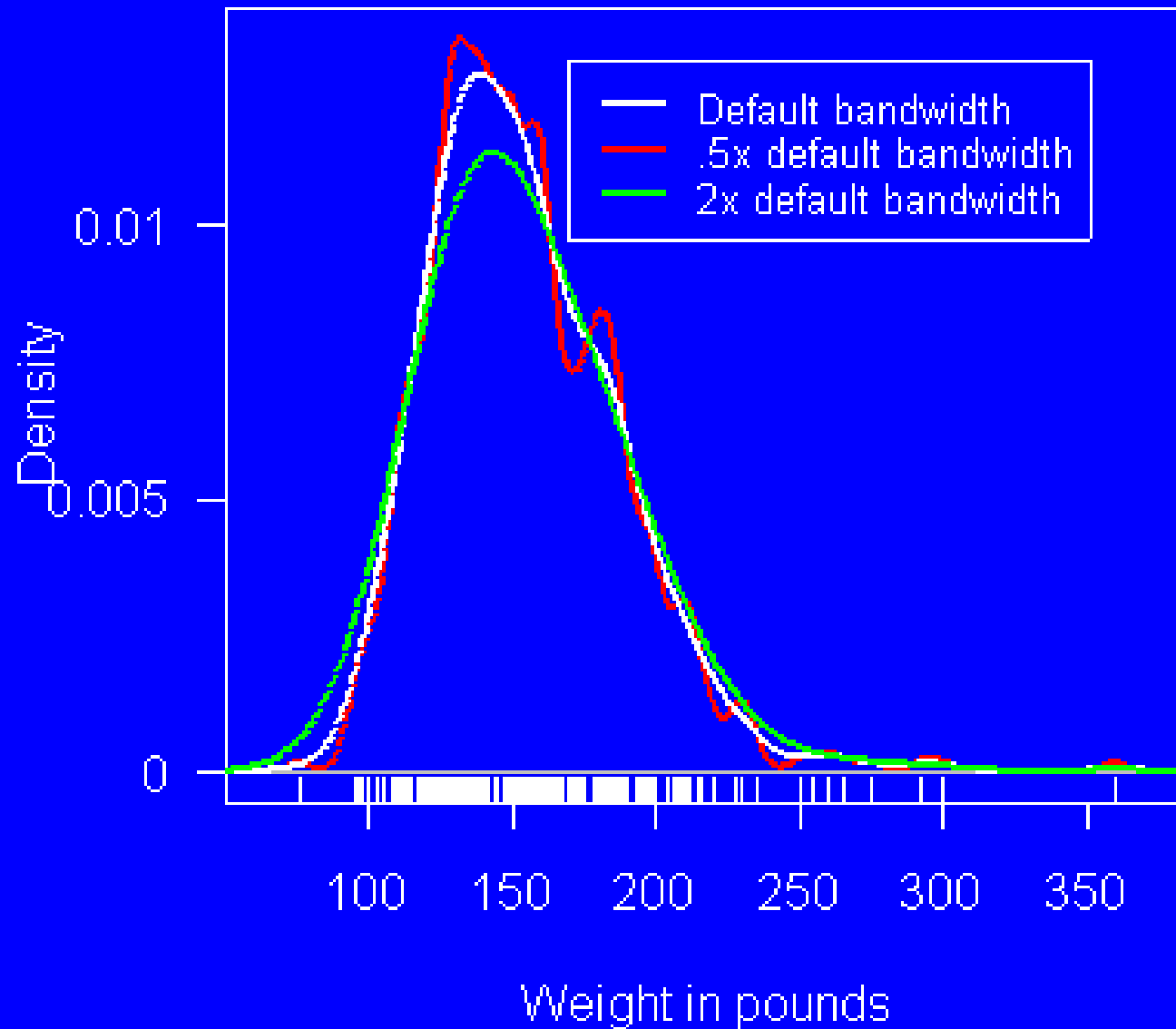
## Density plot of weight, DUHRAY



# Improvements to density plot

- Tick marks horizontal
- Multiple smooths
- Add more information –rug plot

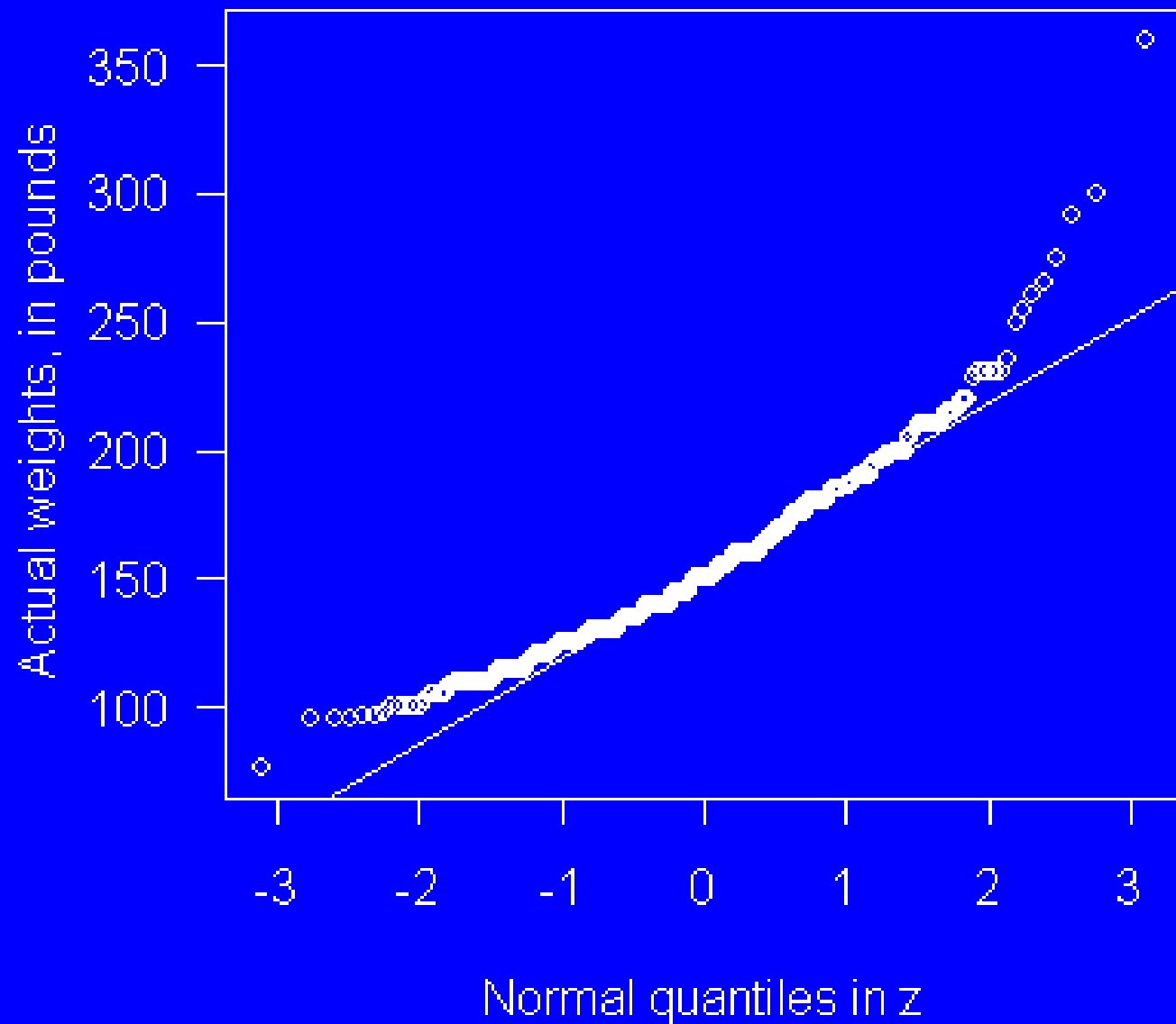
## Density plot of weight, DUHRAY



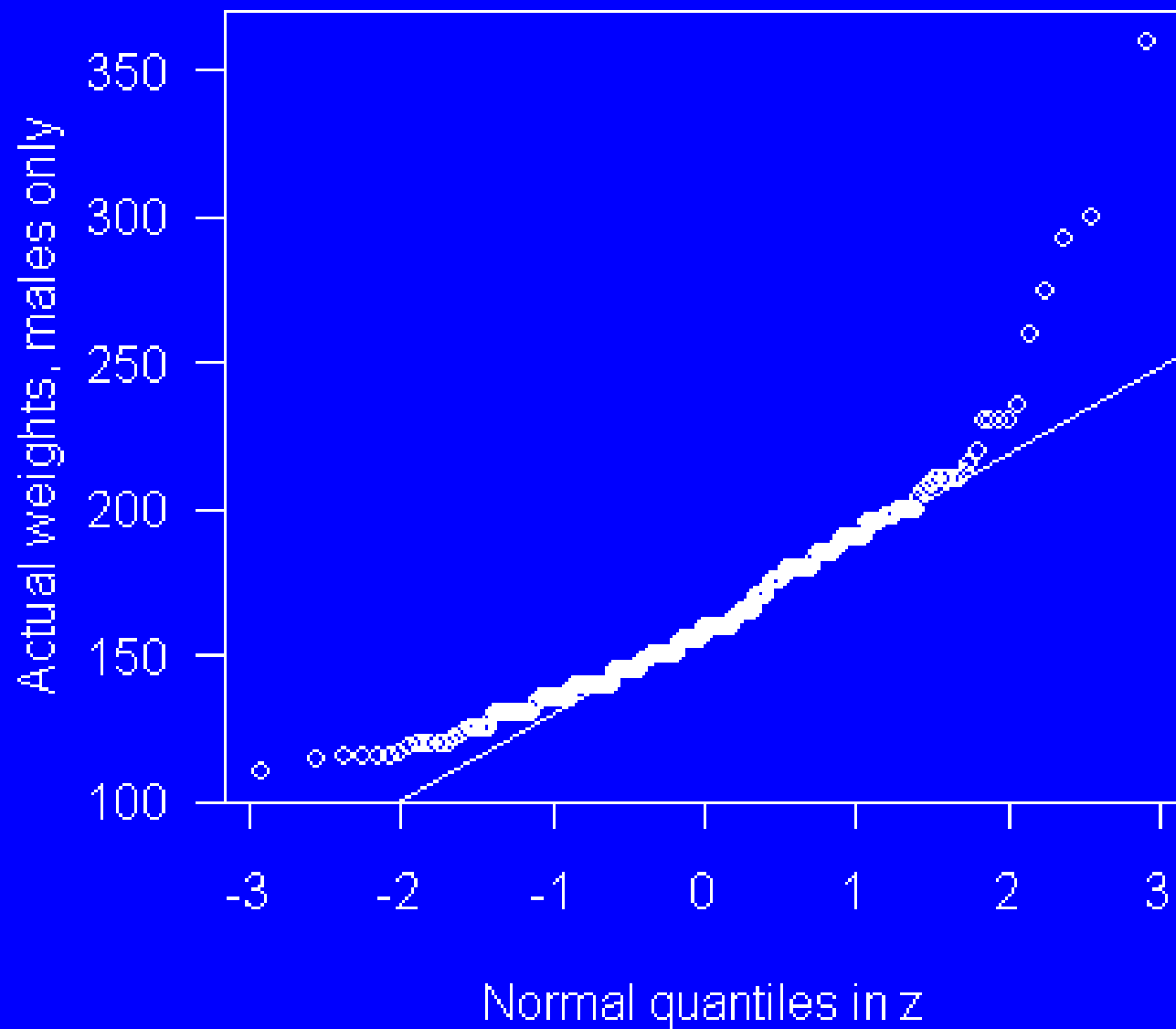
# Univariate continuous data – checking assumptions

- Often you want to check whether some variable is distributed according to some distribution – usually normal
- Here, the best plot is a quantile plot
- Weight had mean = 154, sd = 34, skew = 1.13, kurtosis = 3.19

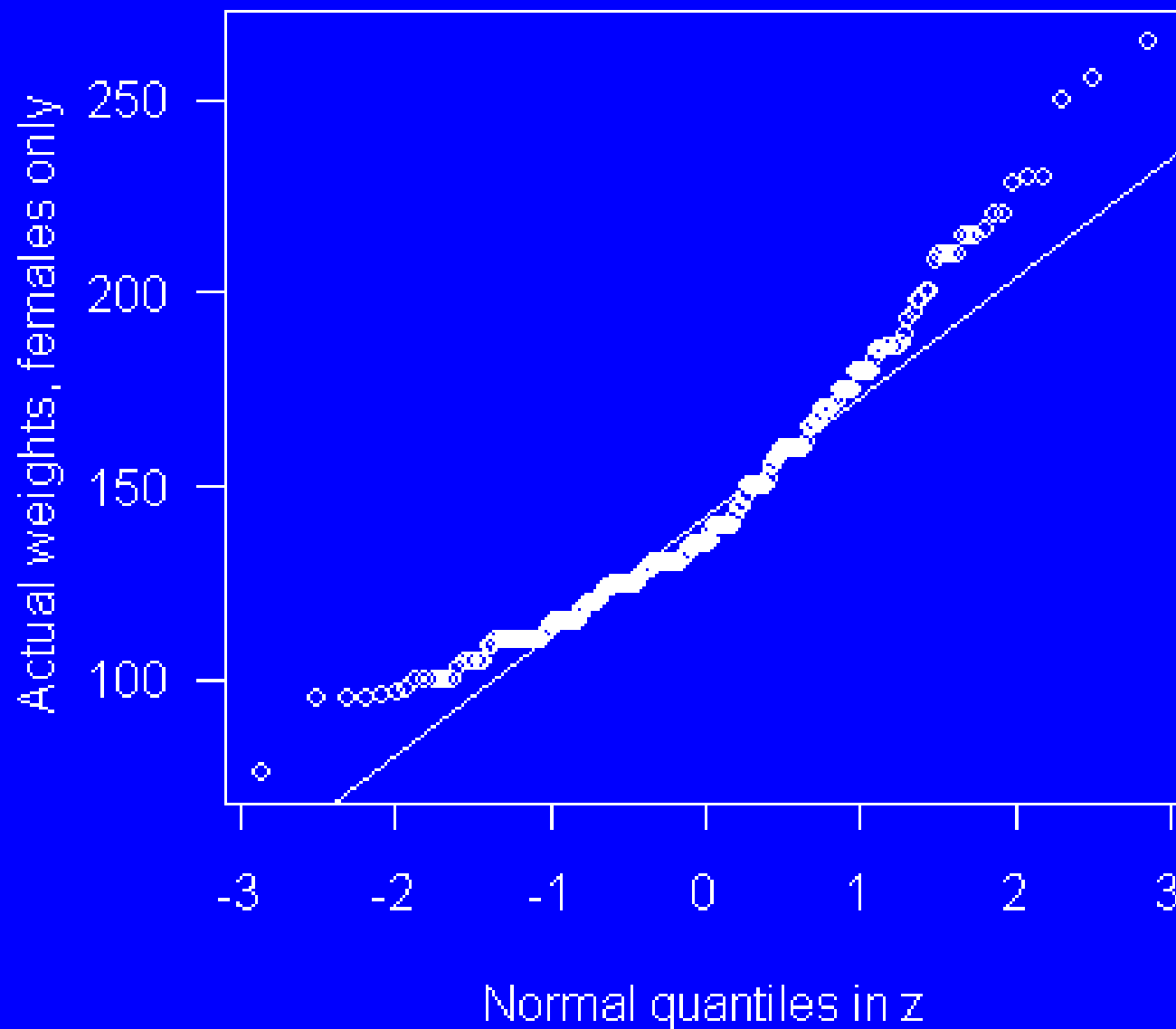
## Normal quantile plot of weight, DUHRAY



## Normal quantile of weight, males, DUHRA



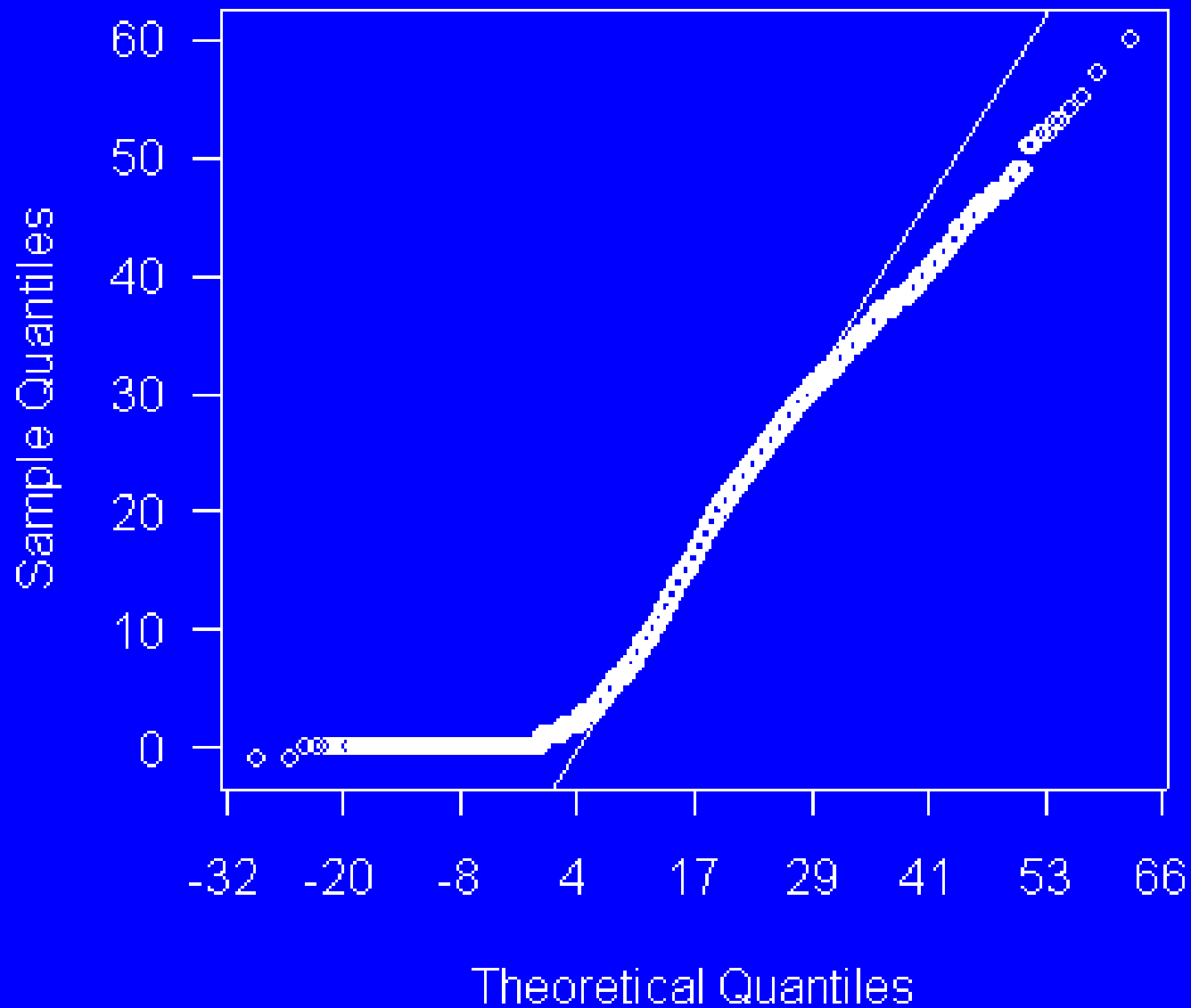
## Normal quantile of weight, females, DUHRAY



## Checking normality (cont)

- Sometimes, descriptive statistics give a misleading picture
- For number of years injecting, mean = 16, median = 16, sd = 12, skew = 0.3, kurtosis = -0.9

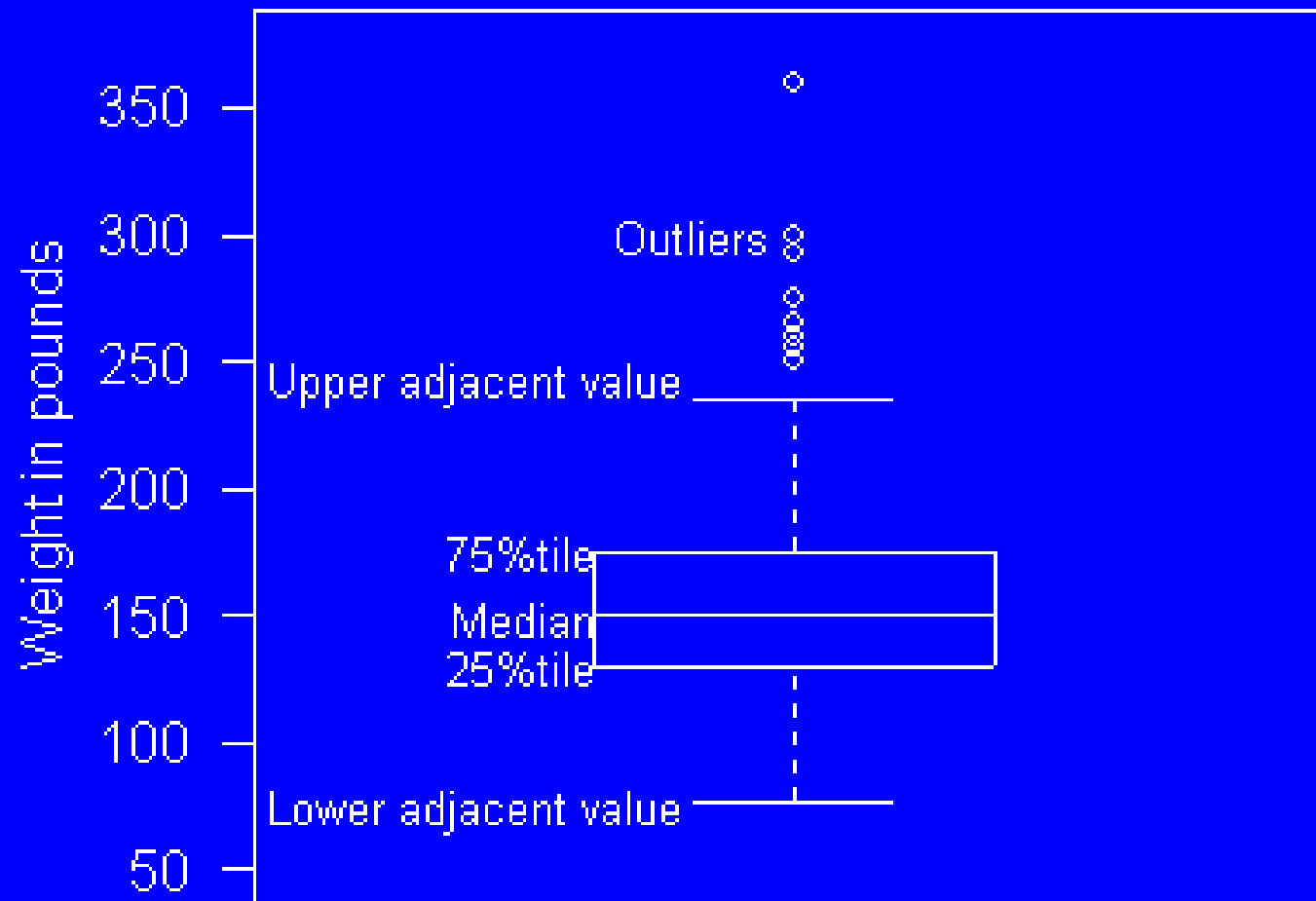
## Normal quantile plot number of years injecting



# Boxplots

- Another simple choice is the boxplot
- Rather wasteful of space
- There are some interesting options

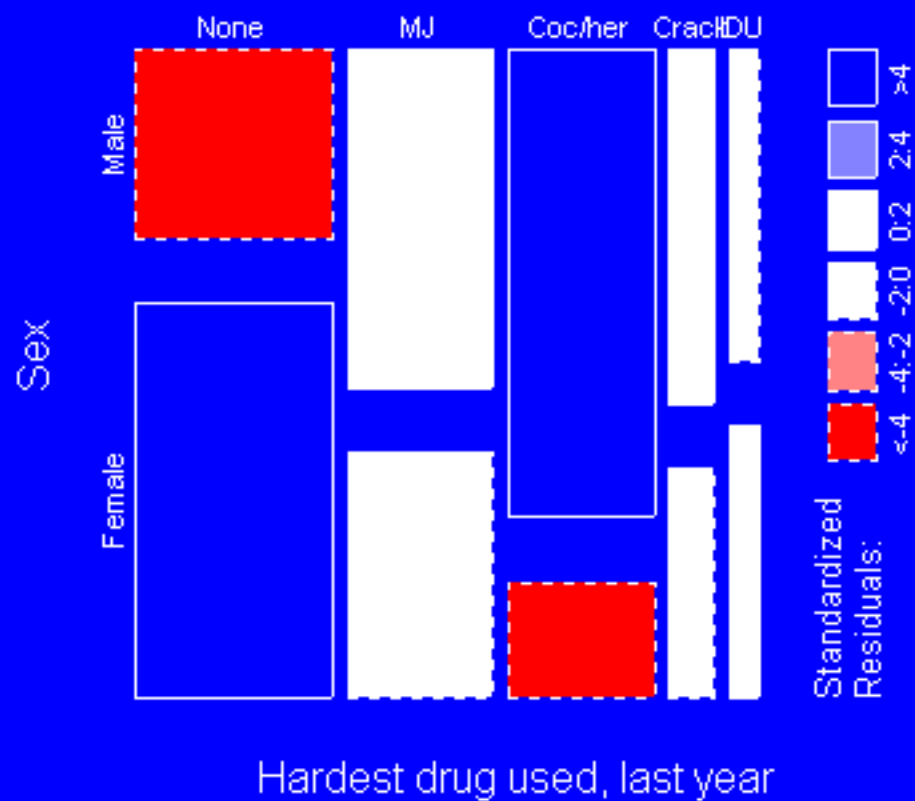
## Box plot of weight, DUHRAY



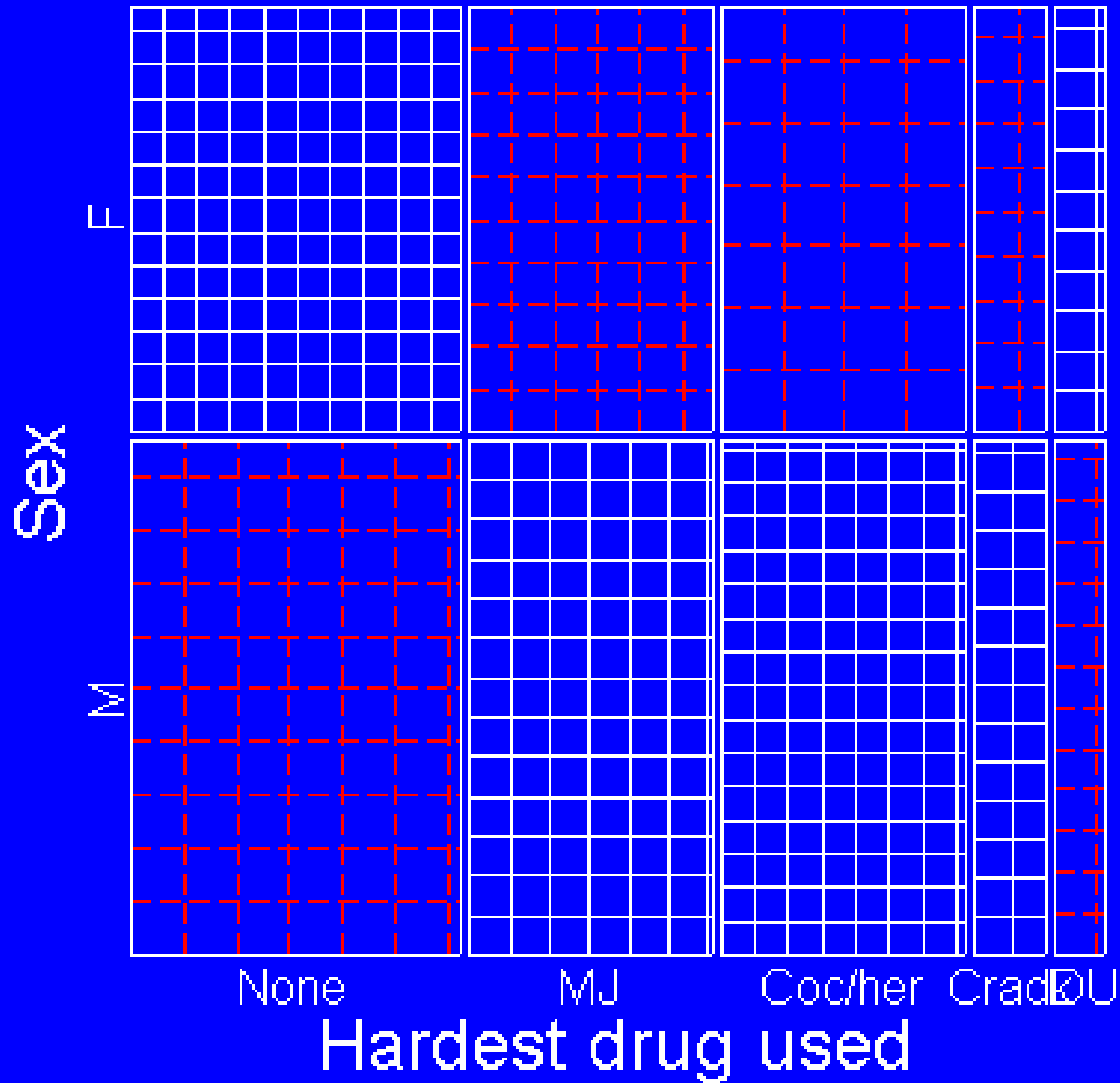
# Bivariate data

- When both are categorical a table is usually best, but mosaic plots and sieve plots can be useful too

## Drug use and sex, DUHRAY



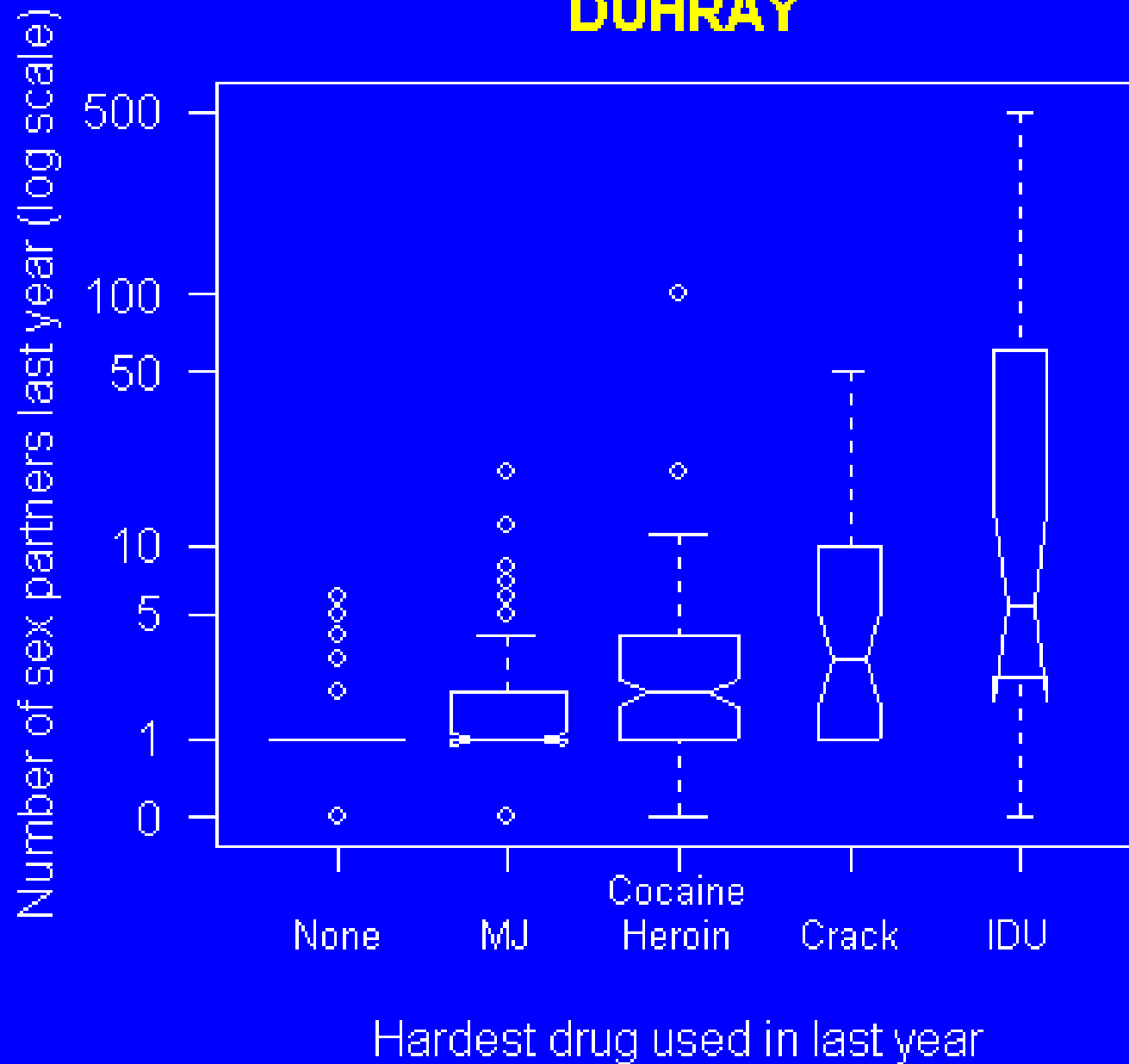
# Sieve plot of sex and drug use DUHRAY



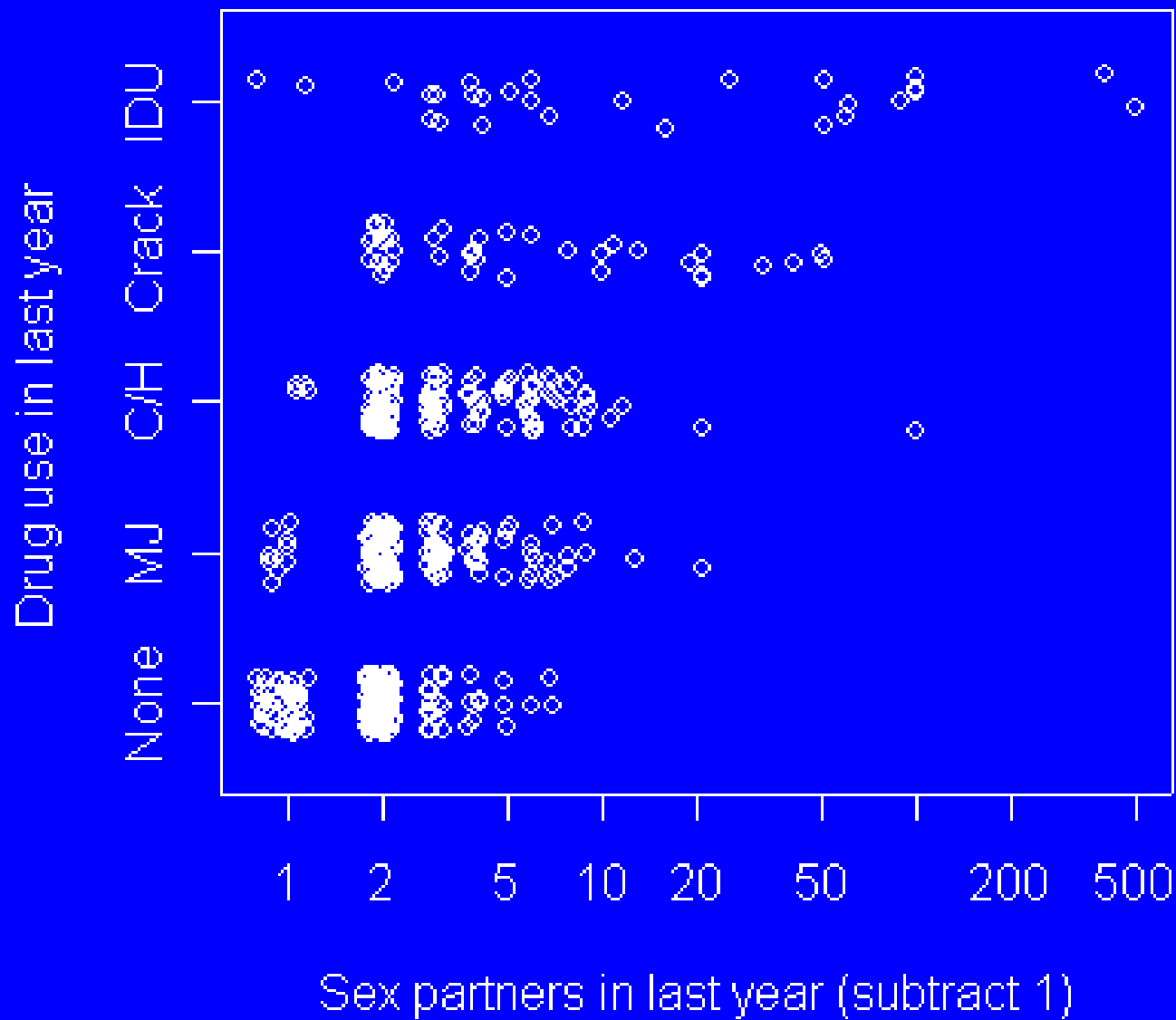
## Bivariate data (cont)

- When there is one categorical and one continuous variable, the best option depends on number of categories
- With few categories, parallel boxplots can be good
- Other choices are a strip plot or quantile quantile plot

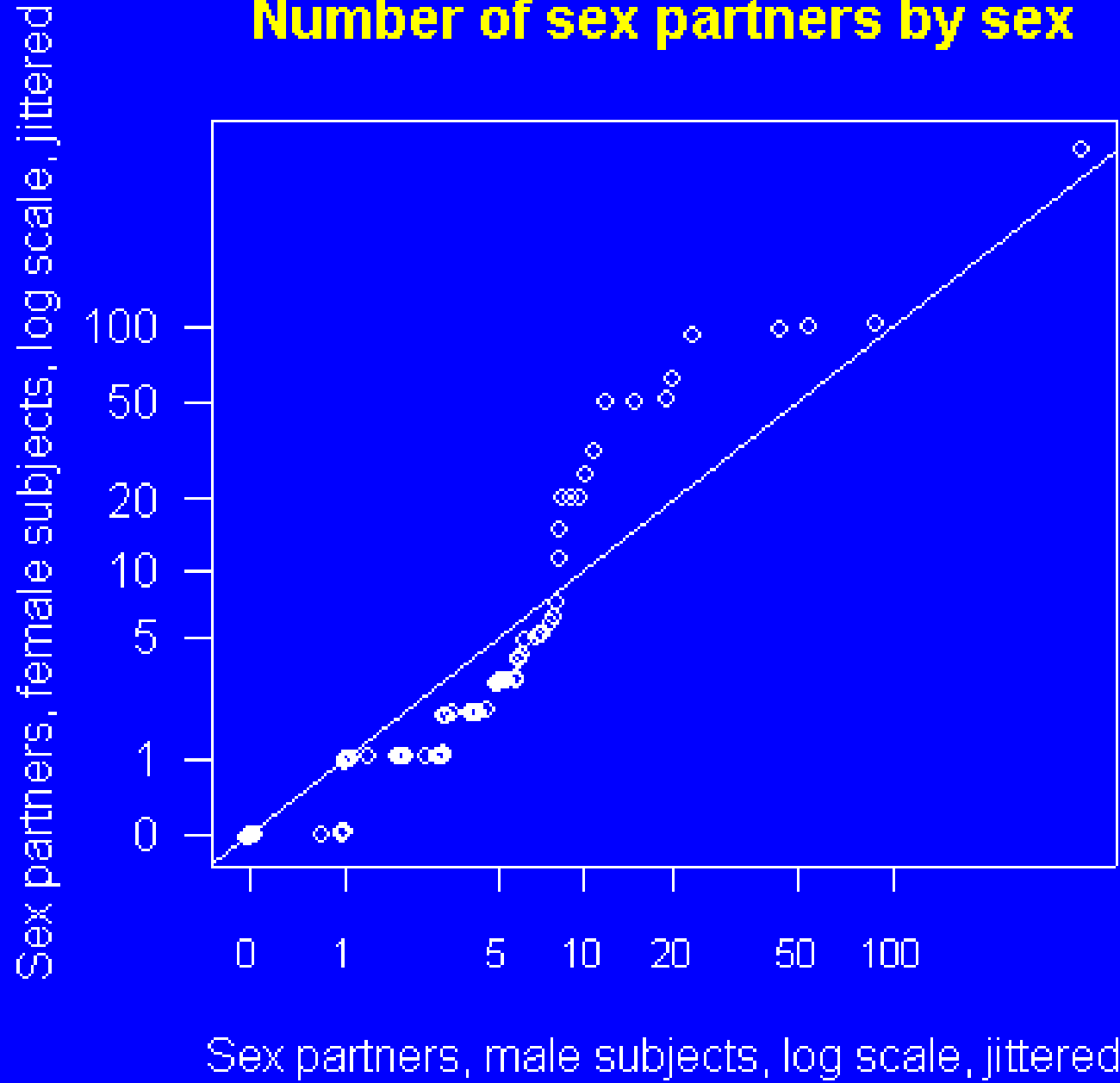
## Drug use and number of sex partners, DUHRAY



# Strip chart of sex partners and drug use DUHRA



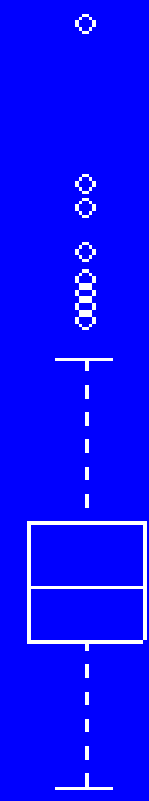
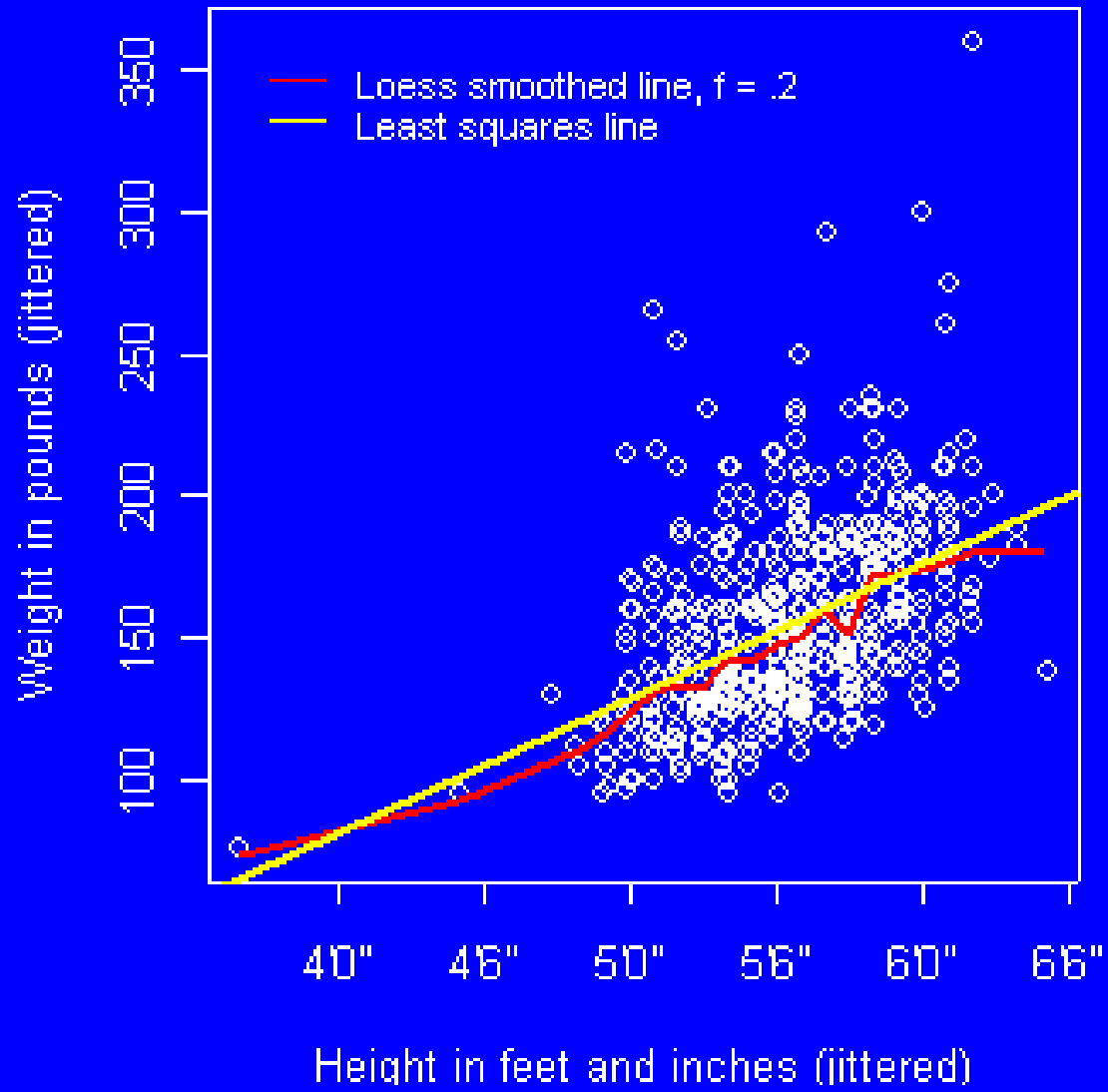
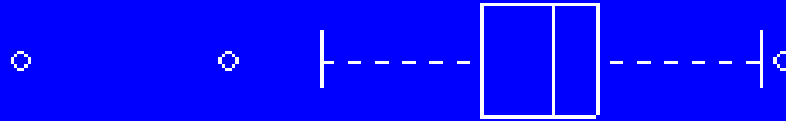
## Number of sex partners by sex



## Bivariate data (cont)

- When both are continuous, the standard method is a scatterplot
- But many improvements can be made

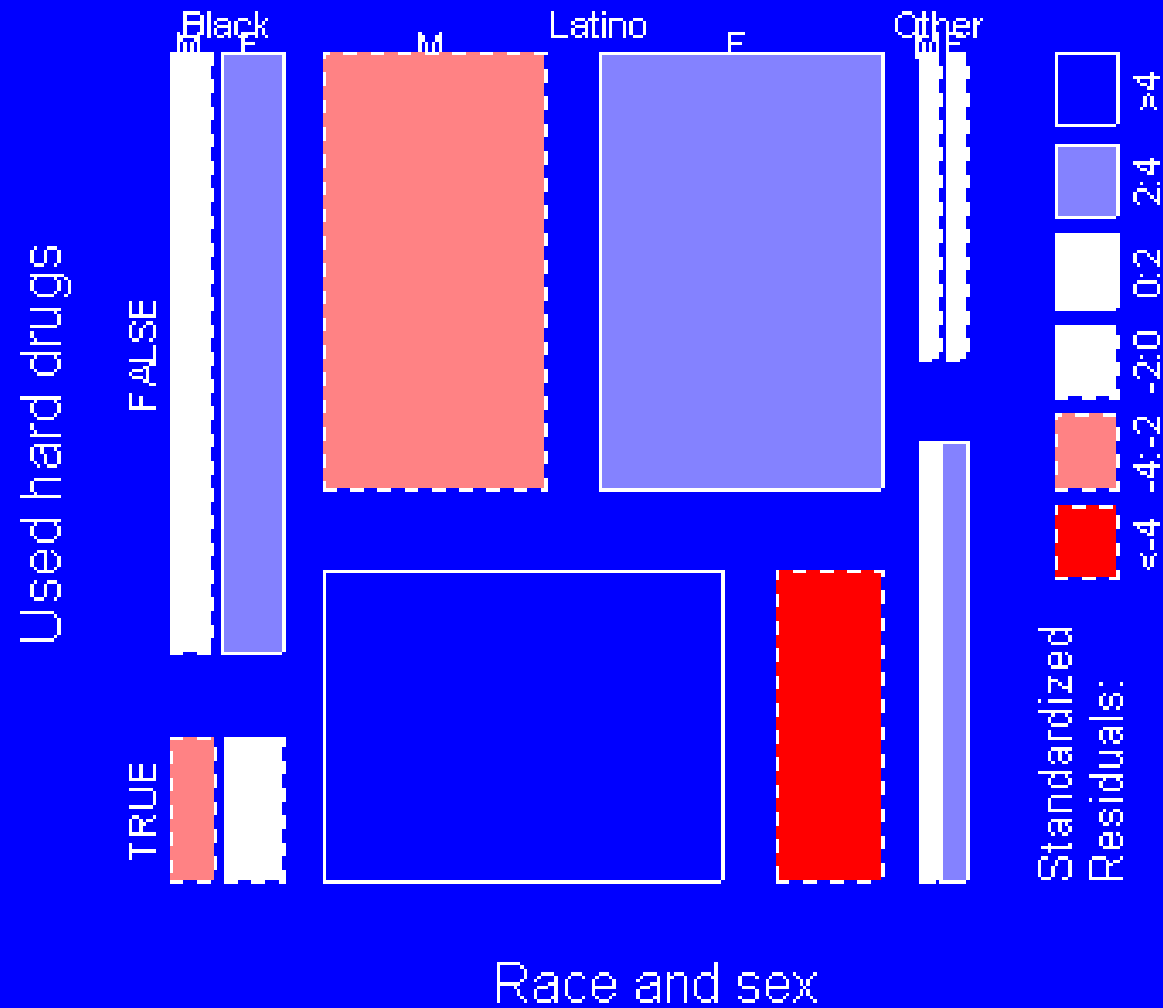
# Scatterplot of height and weight, DUHRAY



# Trivariate data: All categorical

- A mosaic plot can be useful, as can a mosaic matrix

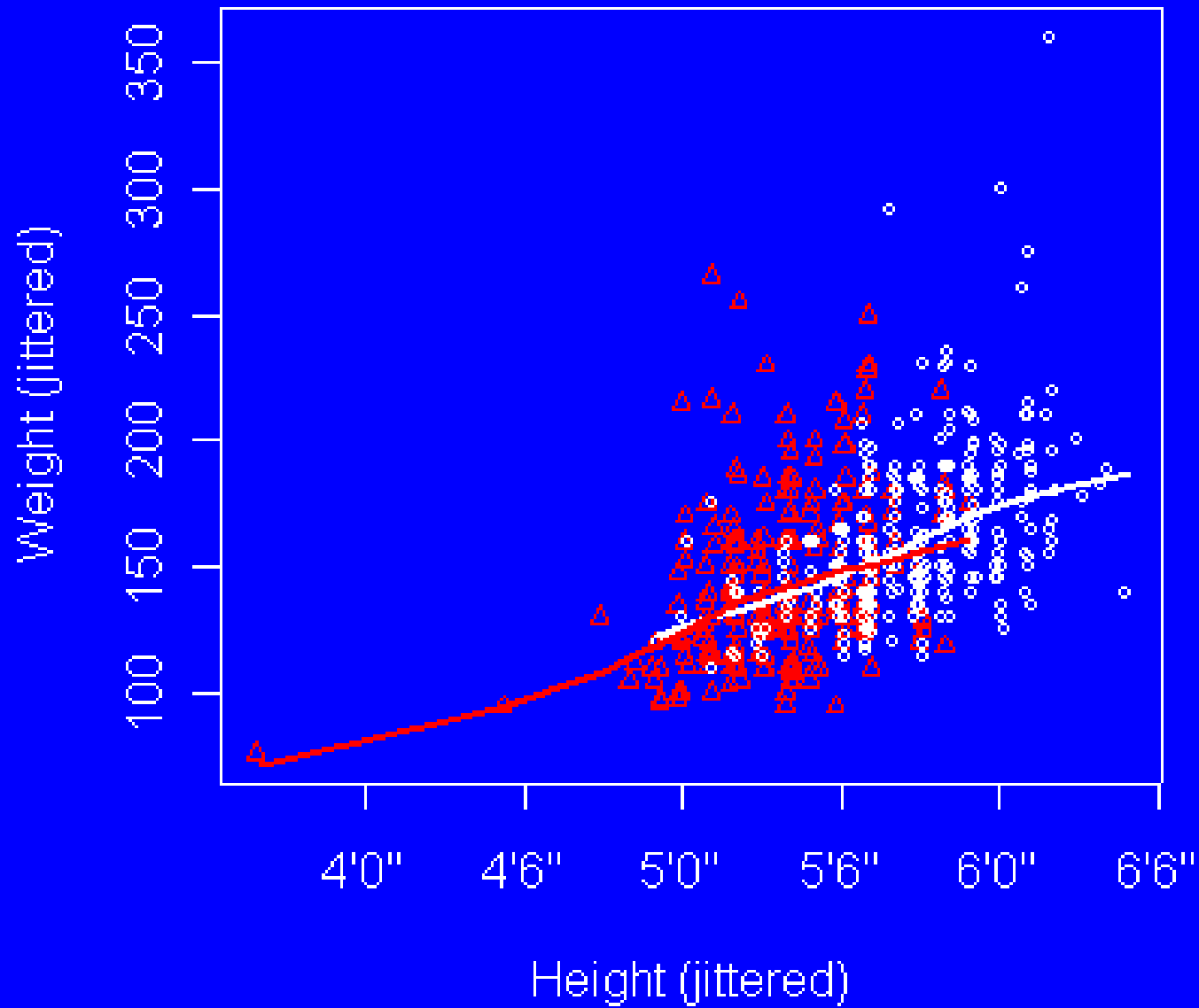
# Sex, race, and hard drug use



# Trivariate data: Two continuous, one categorical

- If there are few categories, you can use a variation in plotting symbol to represent them

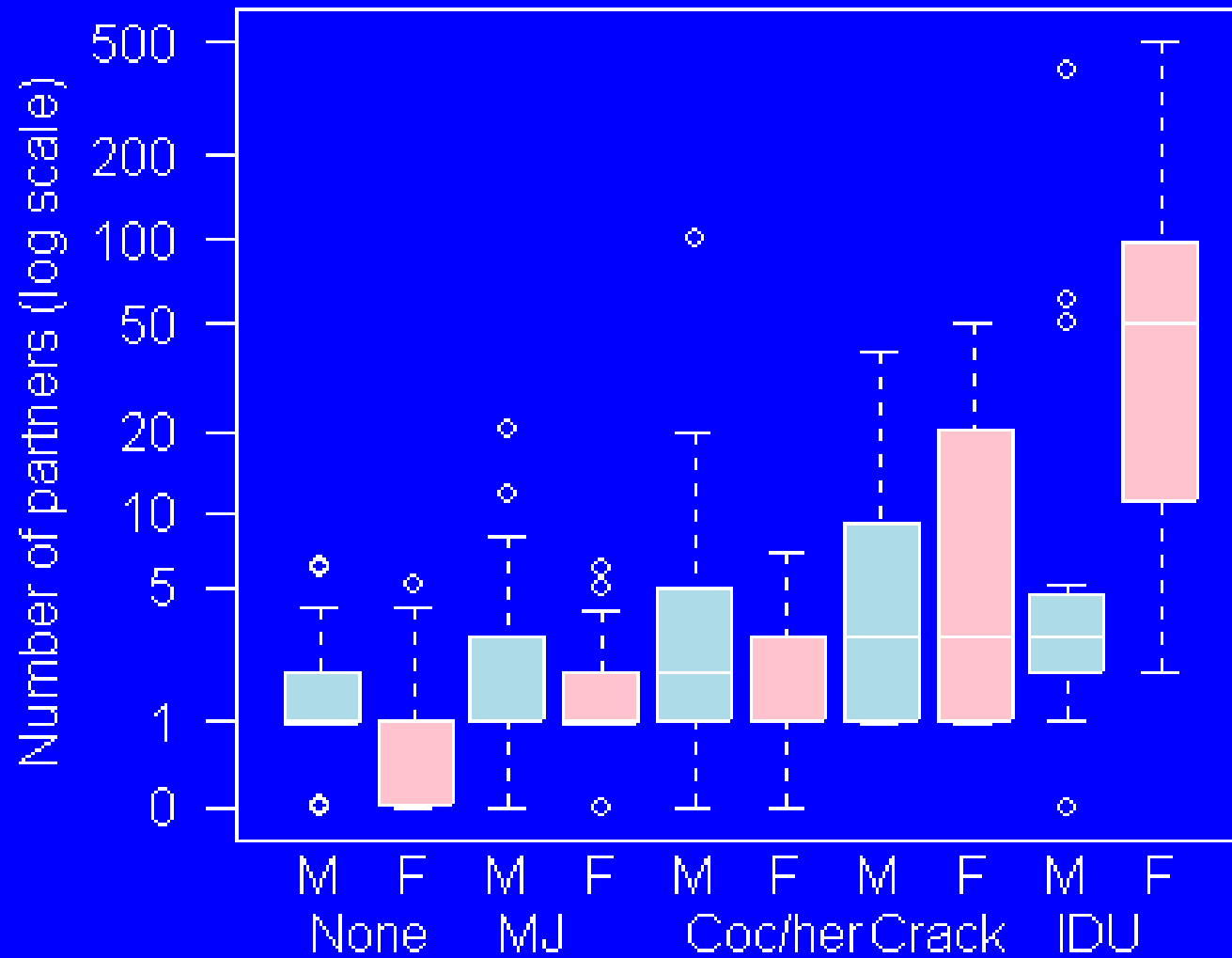
## Height and weight, by sex



# Trivariate data: Two categorical, one continuous

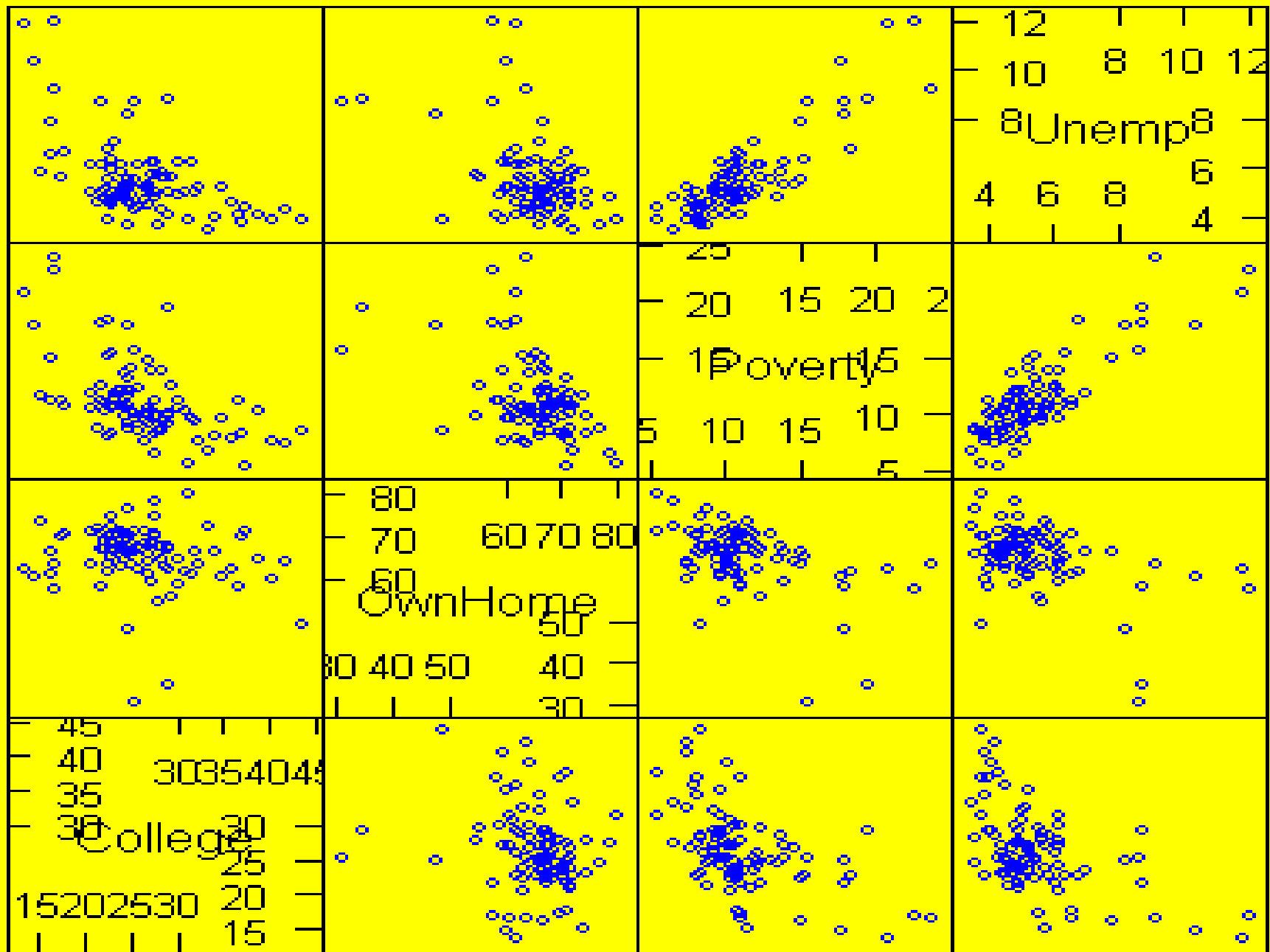
- In this case, parallel boxplots can be very useful

## Sex partners by sex and drug use



# Multivariate data

- One tool is the scatterplot matrix
- With categorical data, mosaic plots can be useful

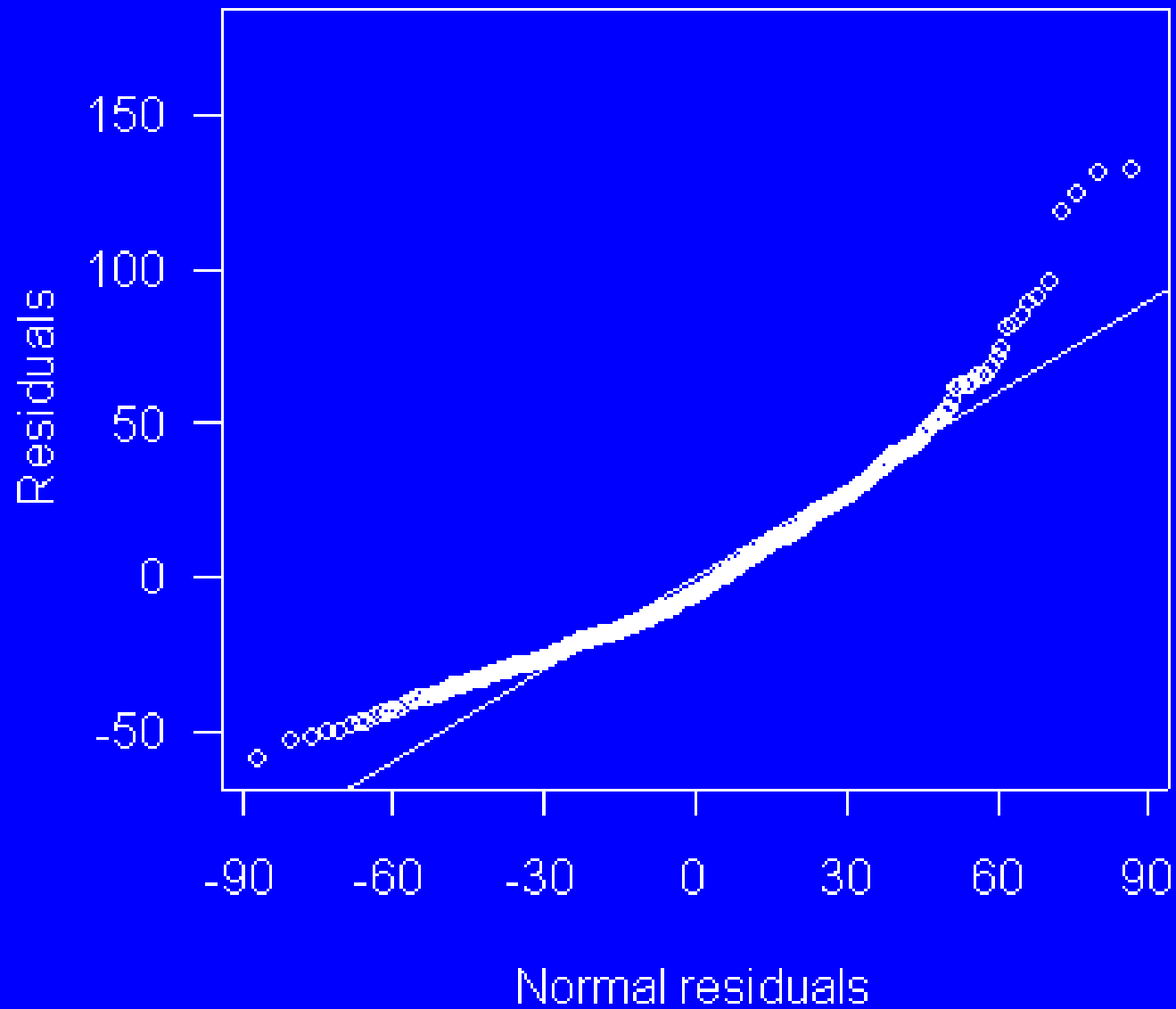


Scatter Plot Matrix

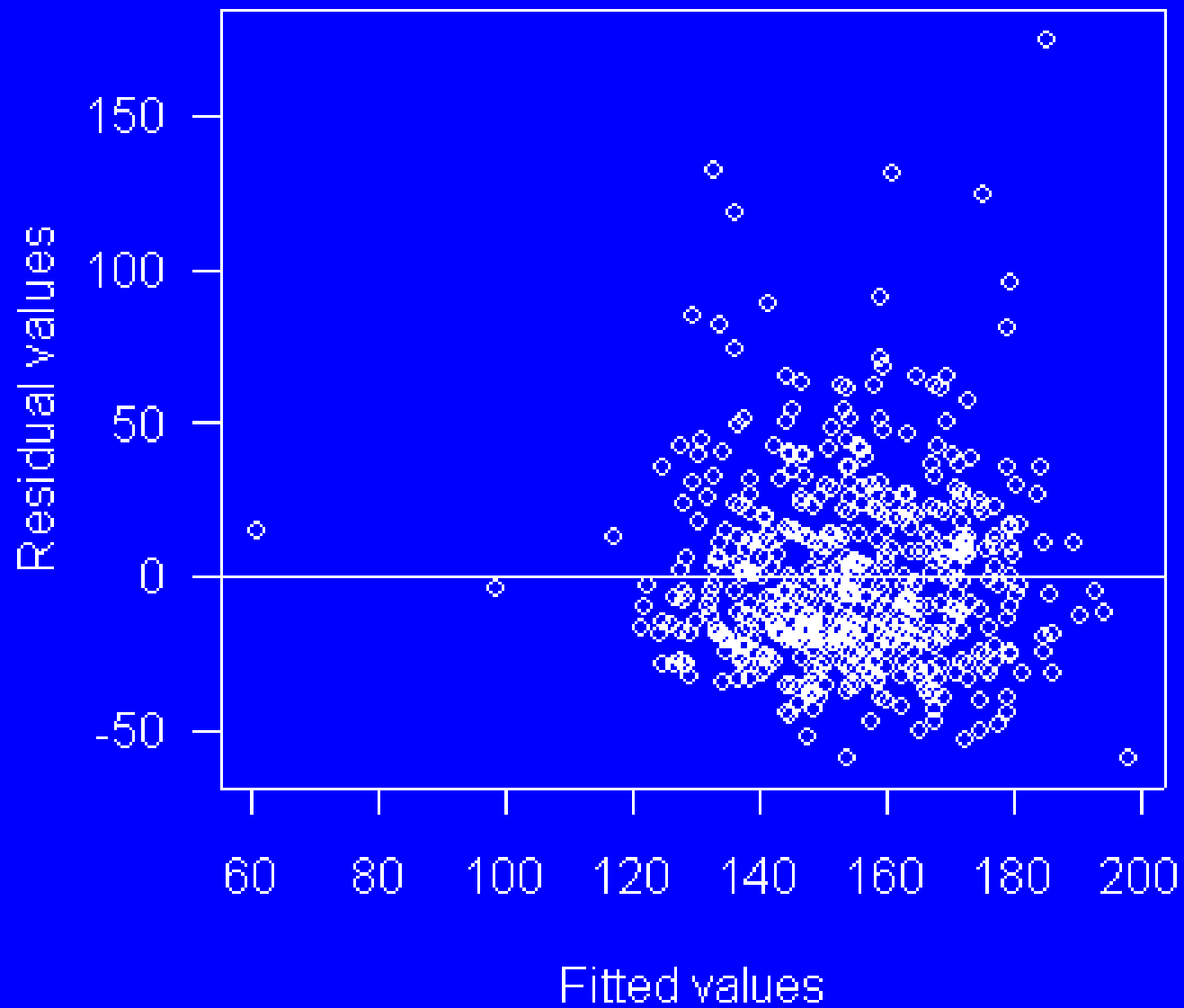
# OLS regression: Model checking

- Normality of residuals –finding outliers
- Residuals vs. fitted – heteroscedasticity
- Model here is weight  $\sim$  height, sex, and age

## Quantile plot of residuals



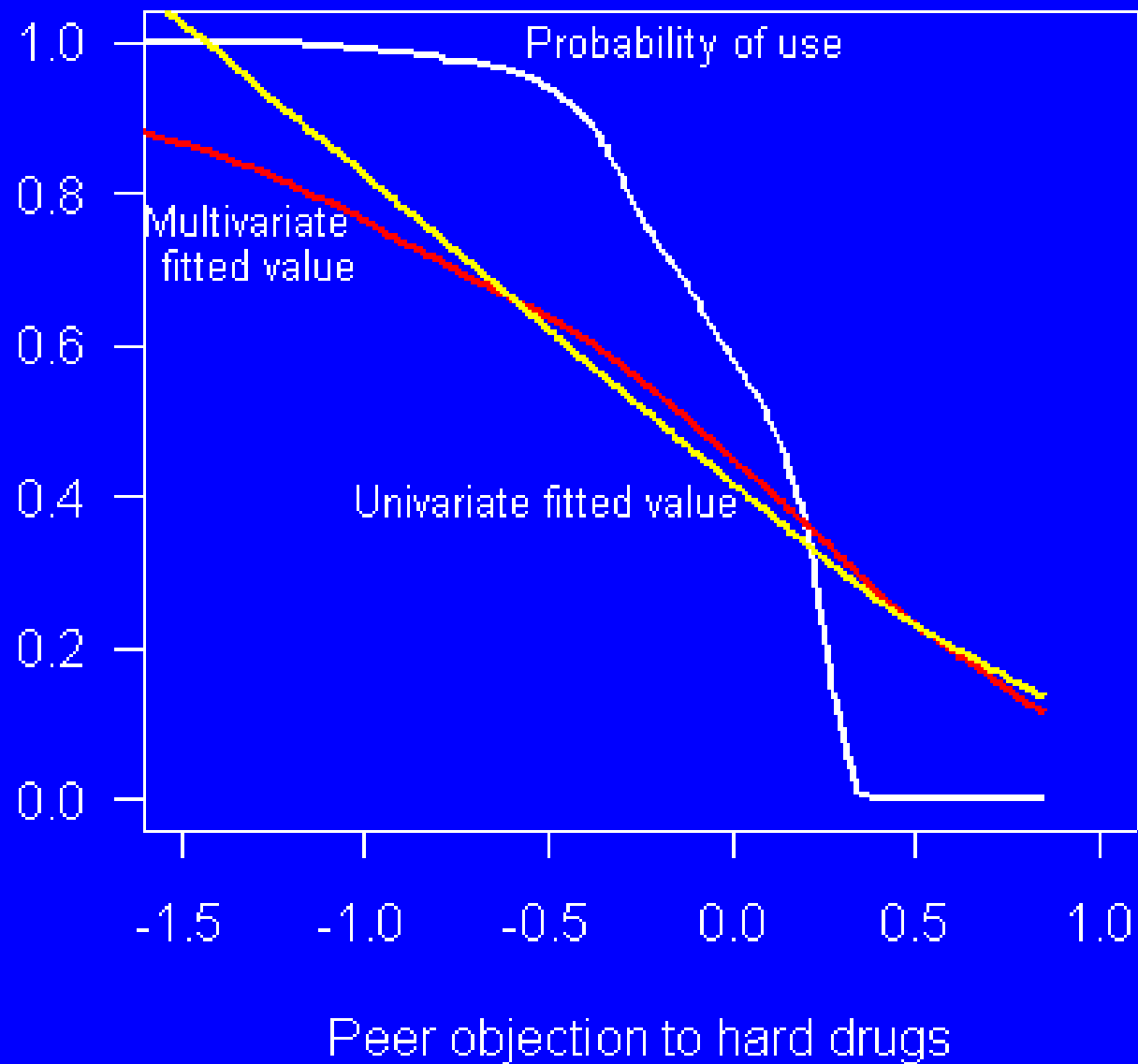
## Residuals vs. fitted



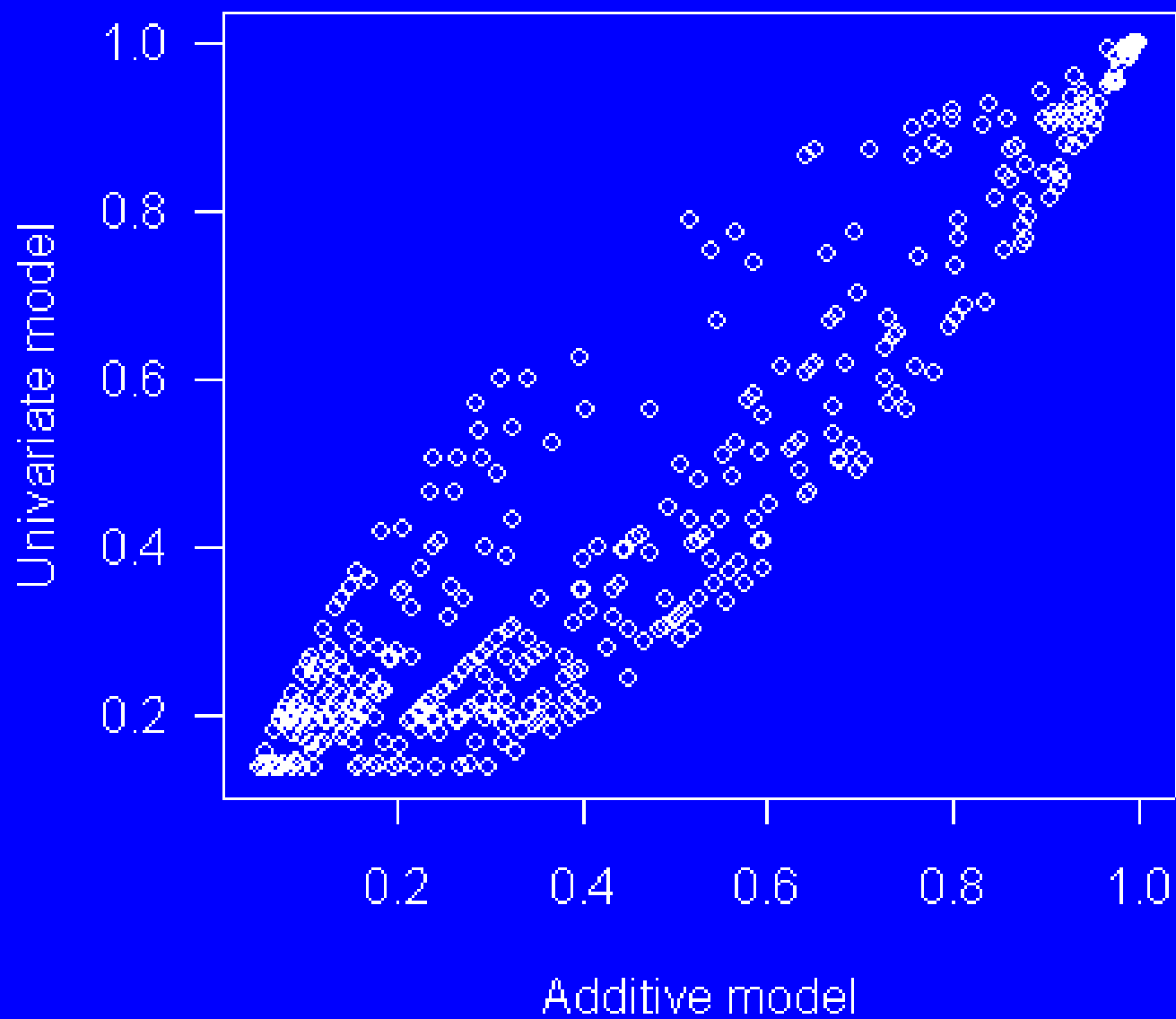
# Logistic regression

- Smoothed plots of probability vs. values of IVs
- Comparisons of fitted values

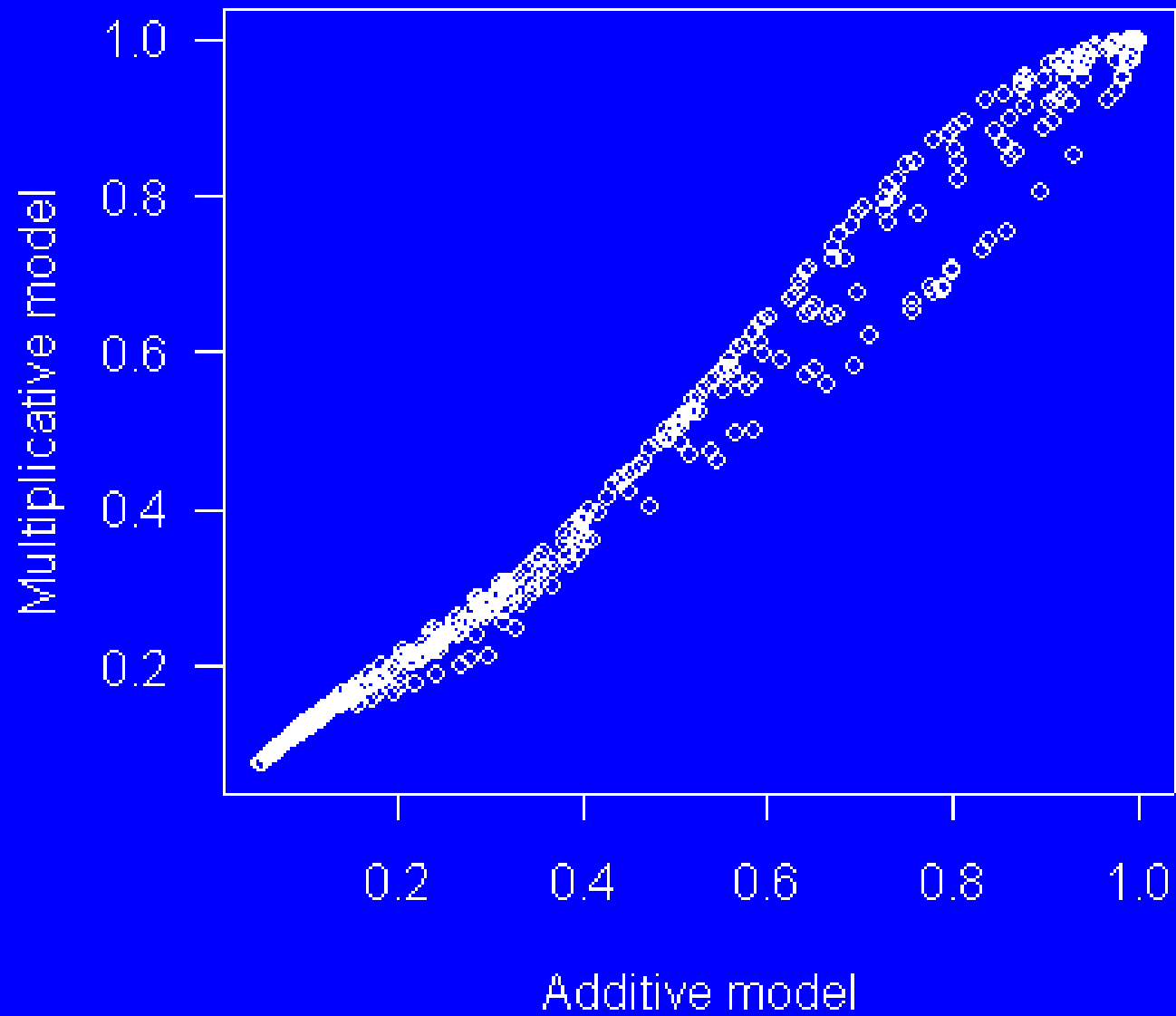
## Smoothed scatterplot of norm and...



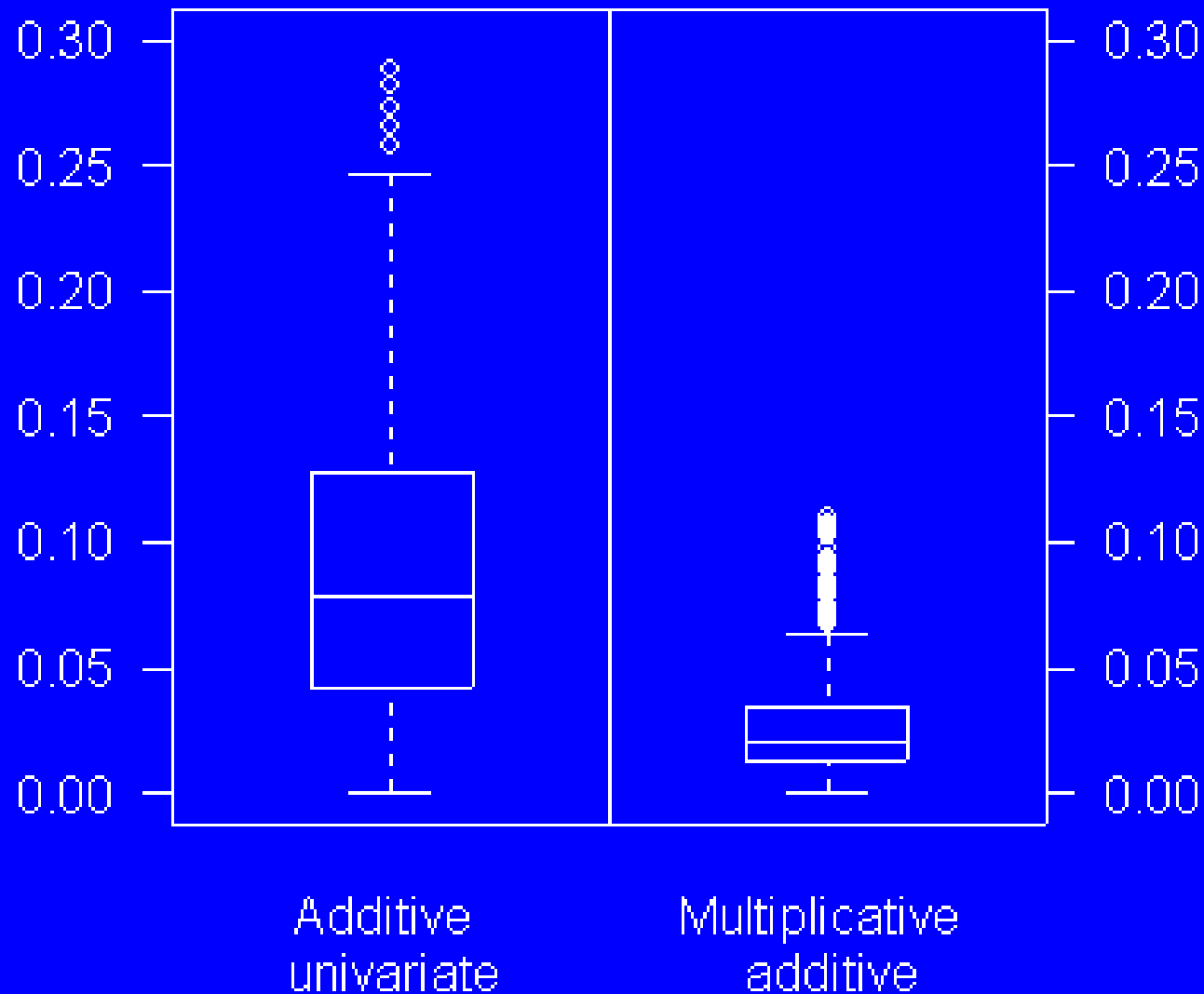
## Scatterplot of fitted values



## Scatterplot of fitted values



Boxplots of absolute differences in fitted values



# References and further reading

- Cleveland: The elements of graphing data
- Cleveland: Visualizing data
- Tufte: The visual display of quantitative information
- Friendly: Visualizing categorical data
- Dalgaard: Introductory statistics with R
- Maindonald: Data analysis and graphics using R