

SLHS 1302

Chapter 8

Visualizing relationships among variables

Key terms

- ▣ **time plot: A picture of a sample of time series data, where time is plotted along one axis**
(typically the x-axis) and the paired observations are plotted along the other axis.
- ▣ **scatterplot: A graph in which two *linked variables are plotted against each other*. Frequently,**
the predictor or *independent variable is plotted on the x-axis, and the dependent or predicted value is plotted on the y-axis*.
- ▣ **vowel plot: A special kind of scatterplot, which plots the first and second formants of one or**
more vowels from one or more talkers. F1 is plotted in *descending (greatest to least) order* along the y-axis, and F2 is plotted in *descending order along the x-axis*.

-
- ▣ **linear relationship: can be described by the slope intercept formula. It means you can draw a straight line describing the trajectory of the data along 2 axes.**
 - ▣ **positive relationship: as x increases, so does y.**
 - ▣ **negative relationship: as x increases, y decreases.**
 - ▣ **“slope intercept” formula: $y=mx+b$. which includes:**
 - y-intercept: Where the line goes through the y-axis (x,0), represented by b in the formula.**
 - slope: The change in y accounted for by each unit change in x , represented by the variable m .**

-
- ▣ **scatter:** The variation in x and y . The dispersion along one dimension may be called “spread.” When describing dispersion along two dimensions, as in a scatterplot, the term we use is “scatter.”
 - ▣ **exponential function:** A sharply curving increase or decrease in y , such that it changes exponentially for each unit of x . Exponential growth, such as happens with bacteria or an investment, is an example of an exponential function.

Background knowledge

- The basic objective of data visualization is to provide an efficient graphical display for summarizing and reasoning about quantitative information.
- Data visualization approaches can be classified into several groups, starting from creating informative charts and diagrams (statistical graphics and infographics) and ending with advanced statistical methods for visualizing multidimensional tables containing both quantitative and qualitative information.
- In this chapter, we will mainly look at scatter plots.

Scatter Plot Defined

- A **scatter plot** or **scattergraph** is a type of mathematical diagram using Cartesian coordinates to display values for two variables for a set of data.
- The data is displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis.

Scatter plot

- A scatter plot is used when a variable exists that is under the control of the experimenter. If a parameter exists that is systematically incremented and/or decremented by the other, it is called the *control parameter* or independent variable and is customarily plotted along the horizontal axis.
- The measured or dependent variable is customarily plotted along the vertical axis. If no dependent variable exists, either type of variable can be plotted on either axis and a scatter plot will illustrate only the degree of correlation (not causation) between two variables.

Example

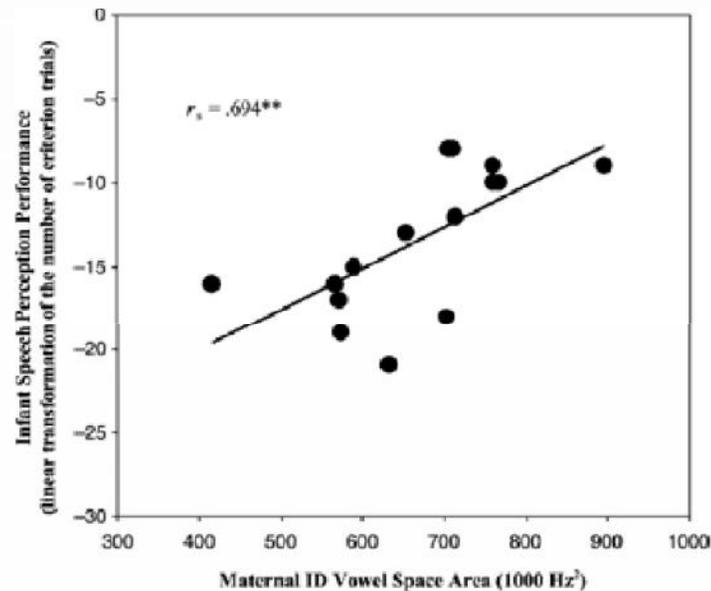


Figure 4. Scatter plot showing the relationship between performance in a test of speech perception by individual 10–12-month-old infants and their mother's speech clarity (see text for details). The results show that mothers with clearer speech have infants with better speech perception performance.

Research Question

▣ **Can you tell how old a person is from his voice?**

▣ Pitch

▣ ...

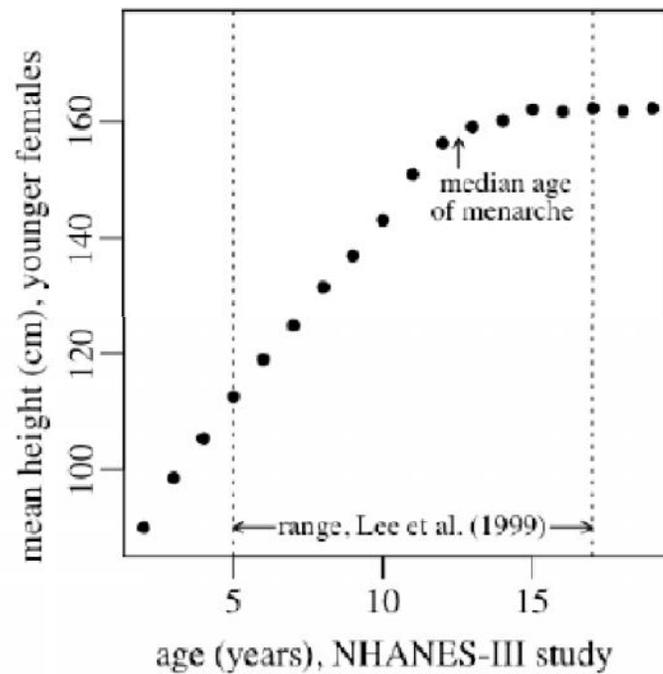


Figure 8.3. Relationship between age and mean height for girls and teenage women in the NHANES-III study. N for plot is 5,539. Upward pointing arrow at median age of menarche (Anderson, Dallal, & Must, 2003).

Interpreting scatter plot data

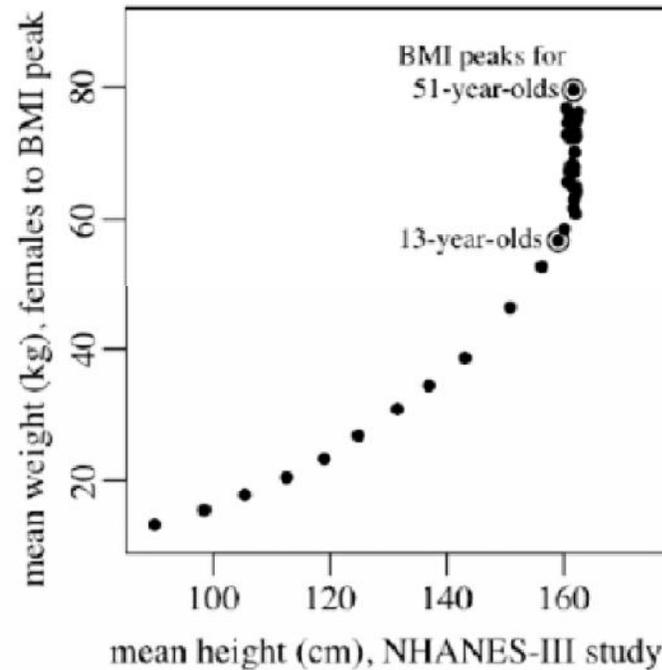


Figure 8.5. Relationship between mean weight and mean height for the 10,838 female participants who were between 2 and 51 years in the NHANES-III study.

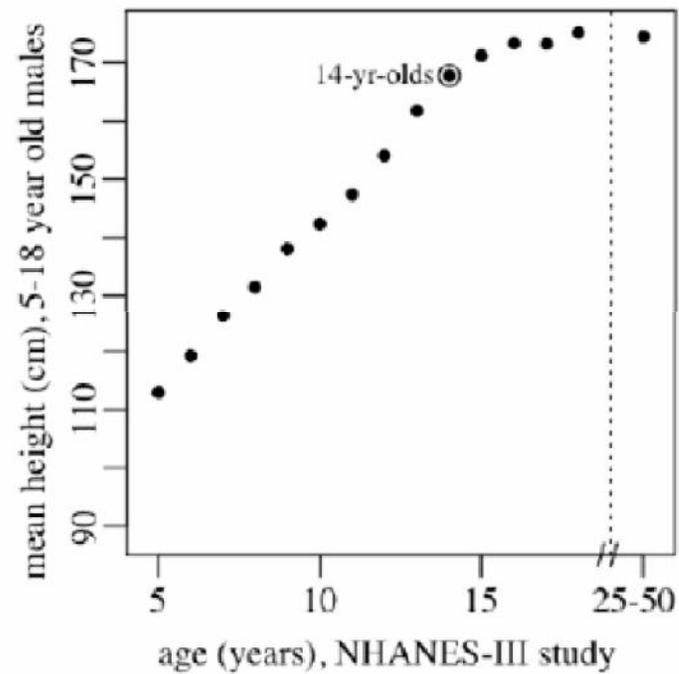


Figure 8.7. Relationship between age and mean height for the 3,471 boys in the NHANES-III study who were between 5 and 18 years old. For reference, we also show the mean for 25 to 50 year old men.

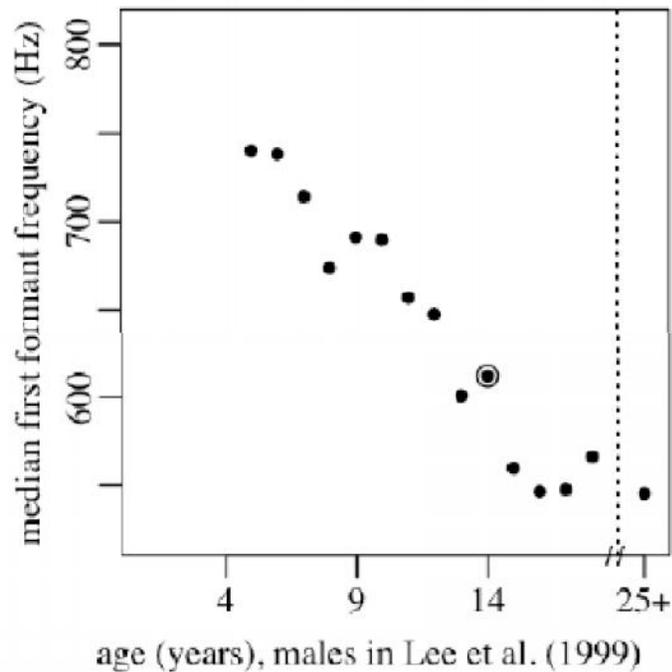


Figure 8.9. Relationship between age and median first formant frequency measured at the vowel steady state for the 258 male participants in the Lee, Potamianos, and Narayanan (1999) study.

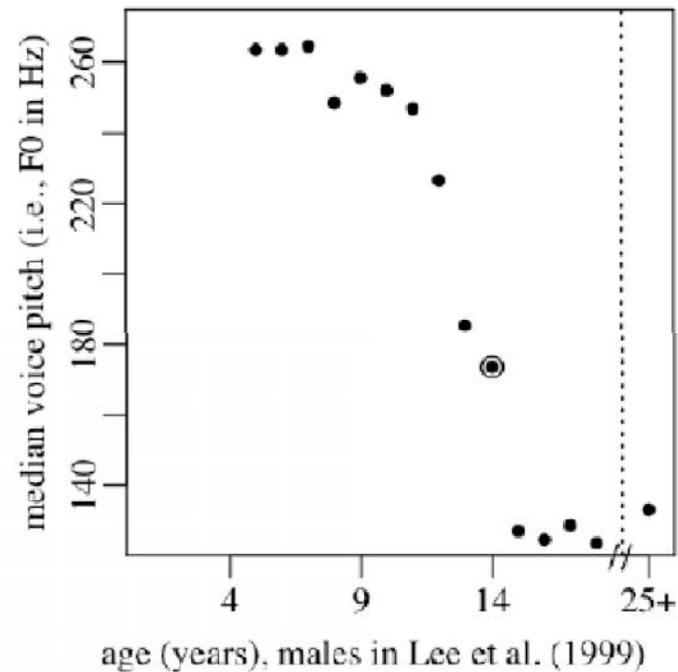


Figure 8.10. Relationship between age and median voice pitch (fundamental frequency, F0) measured at the same vowel mid-point as F1 for the 258 male participants in the Lee et al. (1999) study.

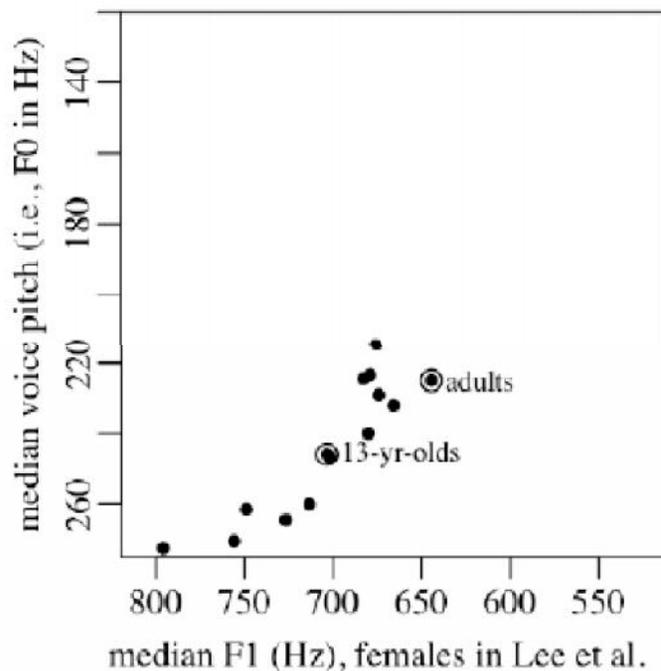


Figure 8.12. Relationship between median voice pitch (fundamental frequency, F0) and median F1 (the y-axis variable in Figure 8.1) for the 234 female participants in the Lee et al. (1999) study.

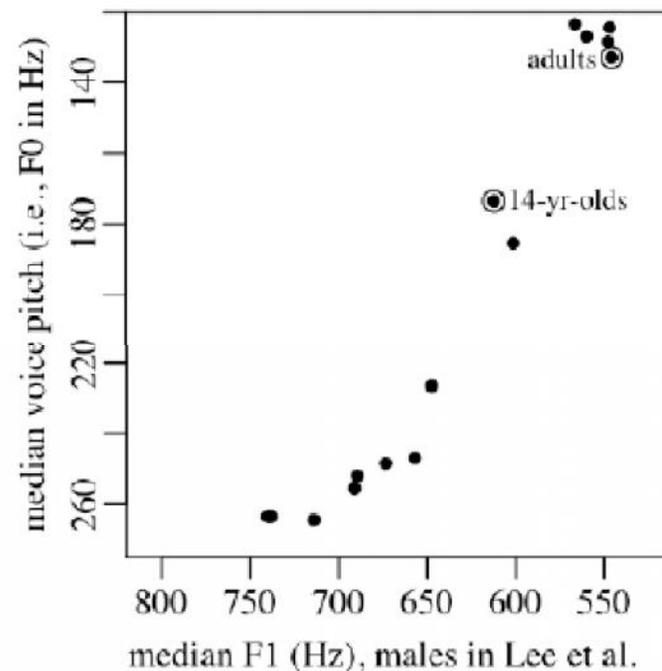


Figure 8.13. Relationship between median F0 (the y-axis variable in Figure 8.10) and median F1 (y-axis variable in Figure 8.9) for the 258 male participants in the Lee et al. (1999) study.

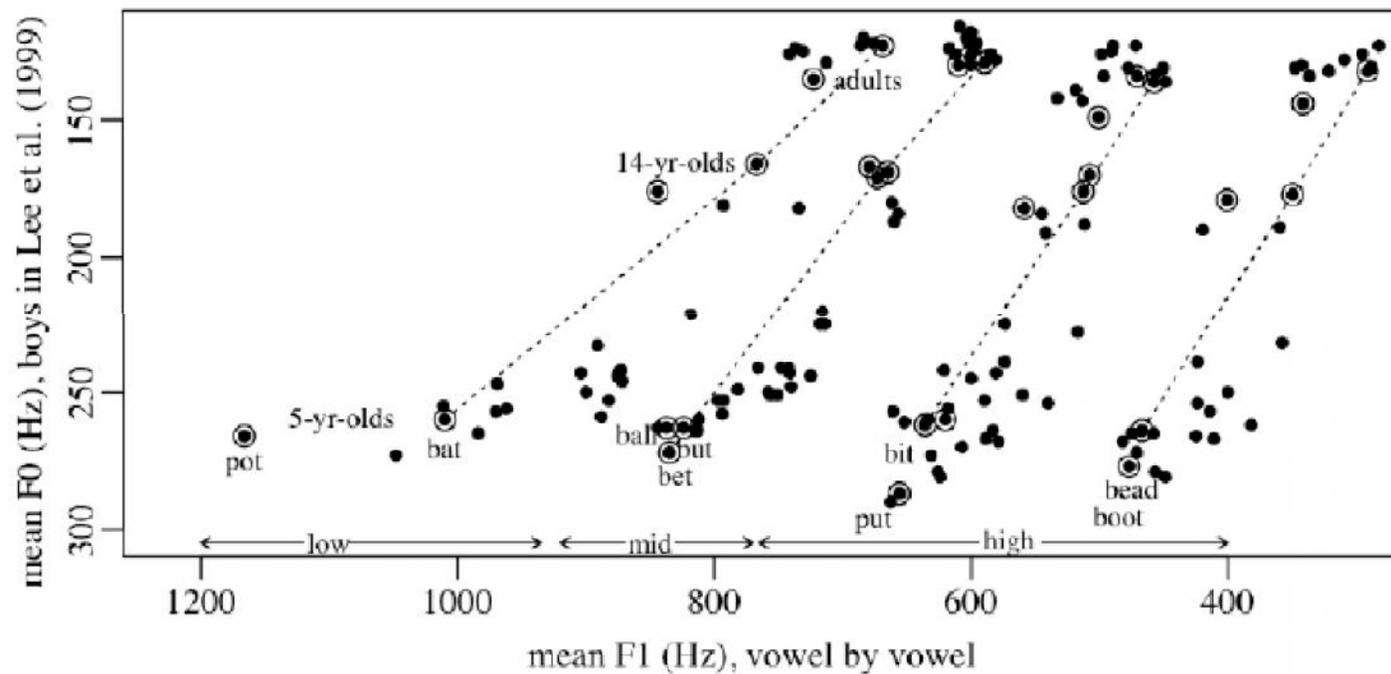


Figure 8.16. Same data from Lee et al. (1999) as in Figure 8.13, but with average values calculated separately vowel type by vowel type for each age group.

What can scatter plots tell us?

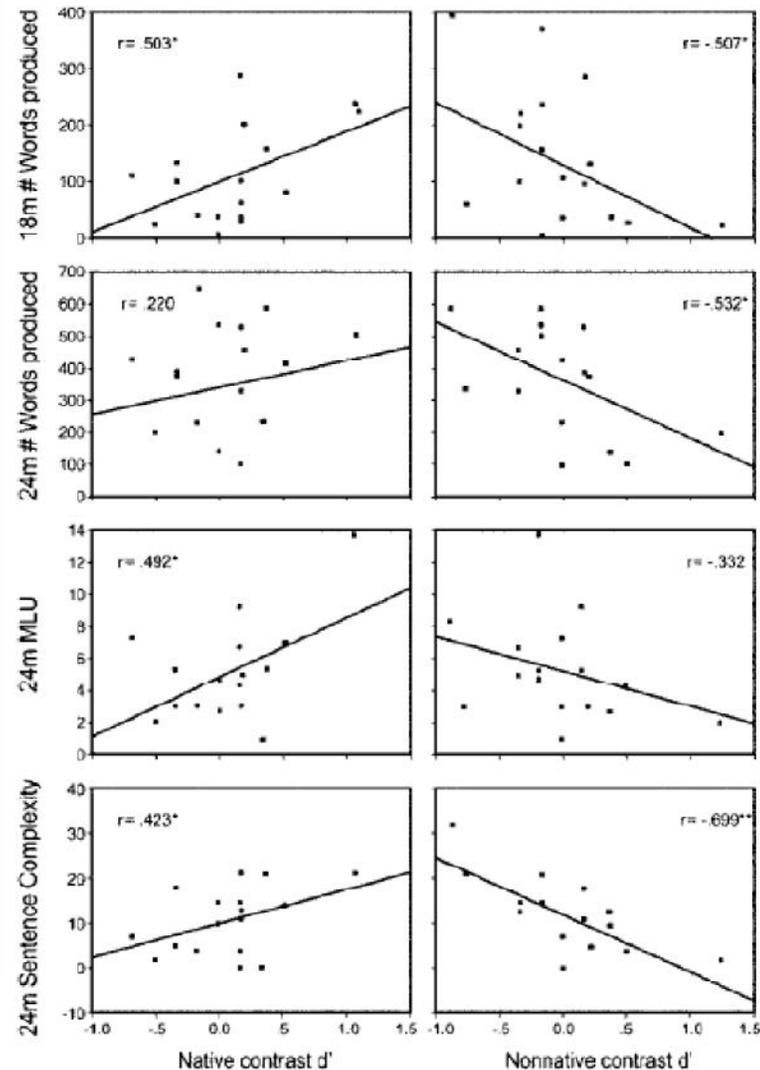
- Time series plot is a special kind of scatter plot
 - Effects of time (age) – time series data
 - Current trends and predictions of future development
- Interpolation of missing data (or gaps)
- Visualization of relationship among two variables of interest

How to describe the relationship between two variables using a scatter plot?

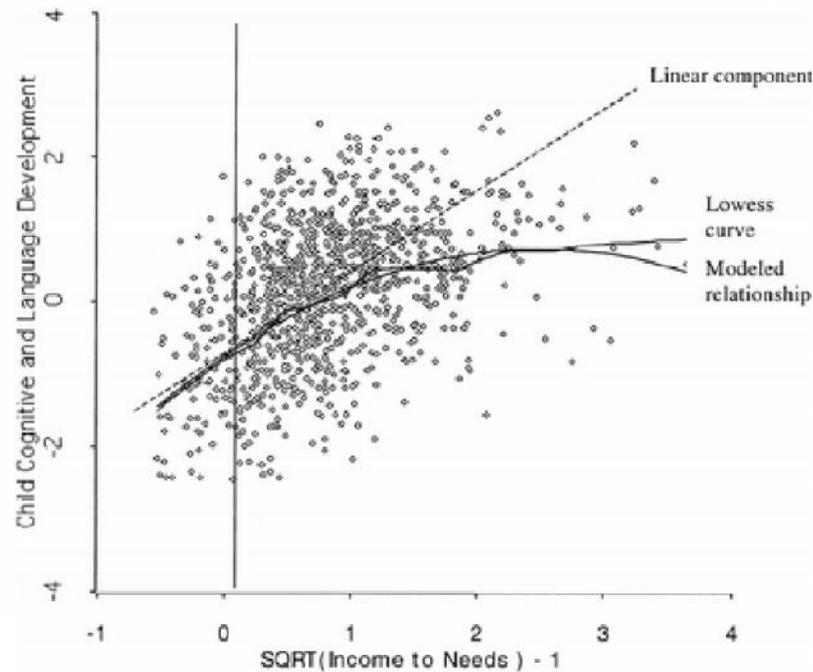
■ Linear function

- $y = mx + b$
- Slope m : positive or negative
- Y-intercept: baseline

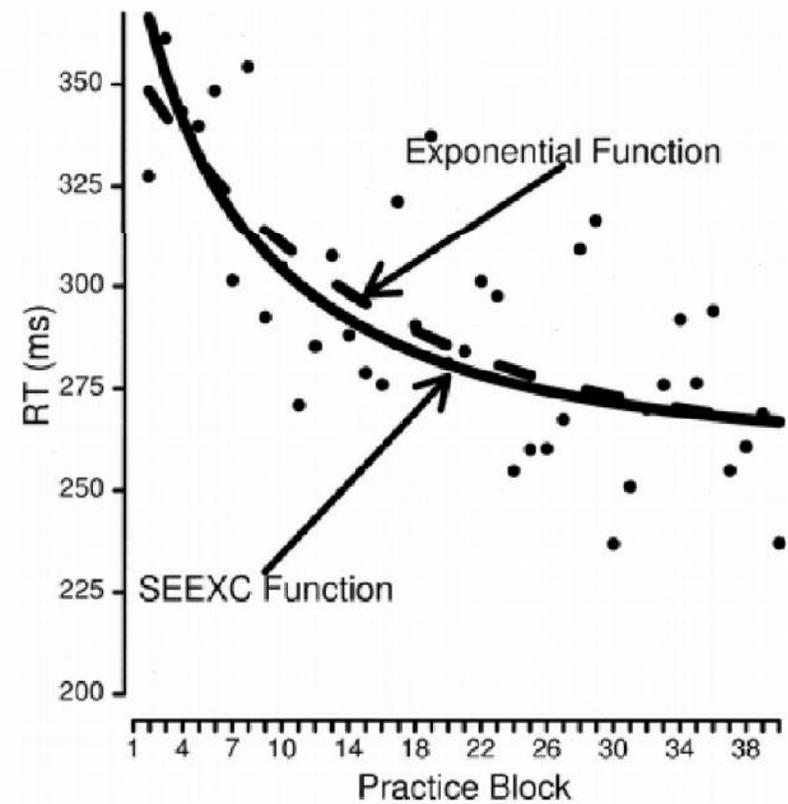
■ Example



- Exponential function: $y = x^m + b$
- Degrees of scatter



■ Exponential function



Answering research question

- Voice pitch changes as a function of age and height due to physiological changes in the vocal structure.
- The onset of puberty can be a landmark for making the distinction of vocal pitches by male and female speakers.
- It remains to be tested how one can predict gender correctly based on pitch measurement and what the statistical test results can tell us about the confidence degree of the prediction.

R codes

- ▣ Set the directory to where the data are
 - ▣ `setwd('xxx')`
- ▣ Read in the data as a tab-delimited table with a header
 - ▣ `nchs=read.table("NCHSwh.txt",header=TRUE,sep="\t")`
- ▣ Show names of the header
 - ▣ `names(nchs)`
“height” “weight” “age” “sex”

- Exclude infants in the data plot

- `notbaby=subset(nchs,!is.na(height))`

- Make a 3x3" plotting window, , with nice tight margins. Use `quartz()` for mac, replace it with `windows()` for windows, and `x11()` for linux.

- `windows(width=3.1,height=3.1,pointsize=12)`

- `par(family="serif",oma=rep(0,4),mar=c(2.8,2.7,0.1,0.1),mgp=c(1.7,0.5,0))`

-
- Plot all the children's and adults' height and weight, using the plotting formula for predicting y from x (remember $y \sim x$), and `pch=20` for filled in circles, and setting up the y -axis so that it matches the next plot.
 - `plot(notbaby$weight~notbaby$height,pch=20,ylim=c(10,175),xlab="mean height of the children (cm)",ylab="mean weight (kg), NHANES-III study")`
 - ...practice all the R codes in the chapter.