

UCLA STAT 13

Introduction to Statistical Methods for the Life and Health Sciences

- **Instructor:** Ivo Dinov,
Asst. Prof. of Statistics and Neurology
- **Teaching Assistants:** Chris Barr & Ming Zheng

University of California, Los Angeles, Fall 2004
http://www.stat.ucla.edu/~dinov/courses_students.html

STAT 13, UCLA, Ivo Dinov Slide 1

Chapter 3: Exploratory Tools for Relationships

Tools for assessing relationships between

- Two quantitative variables
- A quantitative and a qualitative variable
- Two qualitative variables

STAT 13, UCLA, Ivo Dinov Slide 2

Use scatter plots to explore relationships between quantitative variables

Figure 3.1.1 Scatter plot of SYSVOL versus DIAVOL for the heart-attack data in Table 2.1.1.

From Chance Encounters by C.J. Wild and G.A.F. Seber. © John Wiley & Sons, 2000.
STAT 13, UCLA, Ivo Dinov Slide 3

Example: Deaths and radiation in milk after Chernobyl Accident

Region	Peak radioactivity in milk (picocuries/L)	Percentage increase in death rate
Middle Atlantic	23	2.2
South Atlantic	20	2.4
New England	22	1.9
East North-Central	29	3.9
West North-Central	32	3.6
East Southern	21	2.6
Central Southern	16	0
Mountain	37	4.2
Pacific	44	5

Figure 3.1.2 Chernobyl data.

From Chance Encounters by C.J. Wild and G.A.F. Seber. © John Wiley & Sons, 2000.

Example: Computer timings data

Number of terminals:	40	50	60	45	40	10	30	20
Time Per Task (secs):	9.9	17.8	18.4	16.5	11.9	5.5	11	8.1
Number of terminals:	50	30	65	40	65	65		
Time Per Task (secs):	15.1	13.3	21.8	13.8	18.6	19.8		

Figure 3.1.3 Computer timings data.

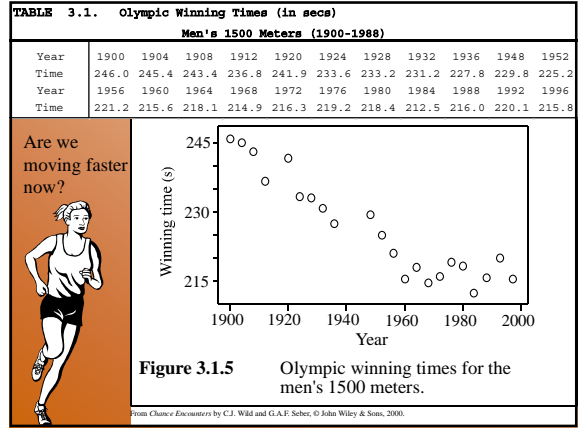
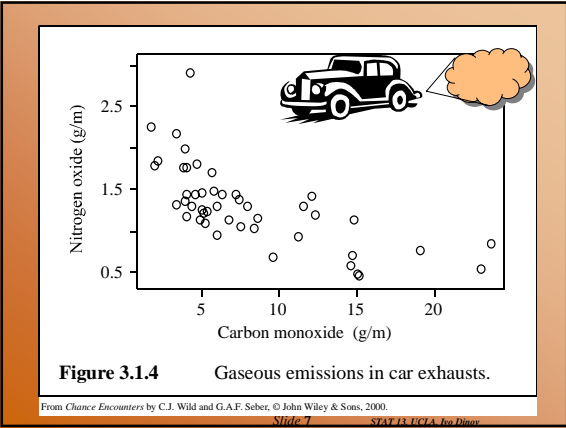
STAT 13, UCLA, Ivo Dinov Slide 5

Example: Car emissions

Car	HC	CO	NOX	Car	HC	CO	NOX	Car	HC	CO	NOX
1	0.50	5.01	1.28	17	0.83	15.13	0.49	32	0.52	4.29	2.94
2	0.65	14.67	0.72	18	0.57	5.04	1.49	33	0.56	5.36	1.26
3	0.46	8.60	1.17	19	0.34	3.95	1.38	34	0.70	14.83	1.16
4	0.41	4.42	1.31	20	0.41	3.38	1.33	35	0.51	5.69	1.73
5	0.41	4.95	1.16	21	0.37	4.12	1.20	36	0.52	6.35	1.45
6	0.39	7.24	1.45	22	1.02	23.53	0.86	37	0.57	6.02	1.31
7	0.44	7.51	1.08	23	0.87	19.00	0.78	38	0.51	5.79	1.51
8	0.55	12.30	1.22	24	1.10	22.92	0.57	39	0.36	2.03	1.80
9	0.72	14.59	0.60	25	0.65	11.20	0.95	40	0.48	4.62	1.47
10	0.64	7.98	1.32	26	0.43	3.81	1.79	41	0.52	6.78	1.15
11	0.83	11.53	1.32	27	0.48	3.45	2.20	42	0.61	8.43	1.06
12	0.38	4.10	1.47	28	0.41	1.85	2.27	43	0.58	6.02	0.97
13	0.38	5.21	1.24	29	0.51	4.10	1.78	44	0.46	3.99	2.01
14	0.50	12.10	1.44	30	0.41	2.26	1.87	45	0.47	5.22	1.12
15	0.60	9.62	0.71	31	0.47	4.74	1.83	46	0.55	7.47	1.39
16	0.73	14.97	0.51								

Source: Lorenzen [198]
 HC = hydrocarbons; CO=carbon monoxide; NOX = nitrogen oxides; grams/mile measurements; 46 identical vehicles tested.

STAT 13, UCLA, Ivo Dinov Slide 6



Review

- What is a **quantitative** variable?
- What basic **tool** is used for **exploring relationships** between quantitative variables?
- What is a **controlled** variable? (variables whose values are determined in the experimental design, as opposed to random variables who are evaluated once the experiments are conducted (e.g., number of terminals vs. task completion time))
- What is the difference between a random and a nonrandom variable? (variables whose values are not to be observed as random events during the experiment, i.e., these are controlled, deterministic or predictable variables, e.g., **year** for the Running Time experiment).

Slide 9 STAT 13, UCLA, Jon Dineer

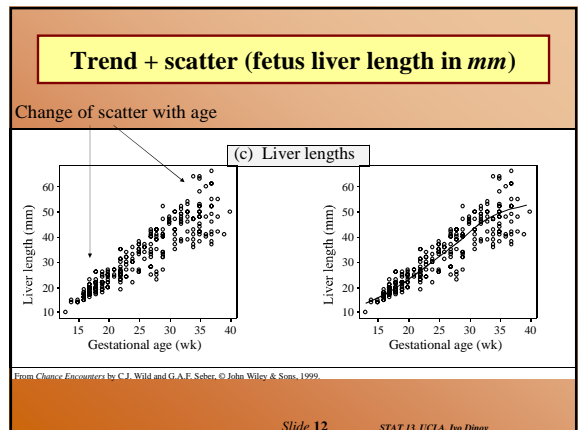
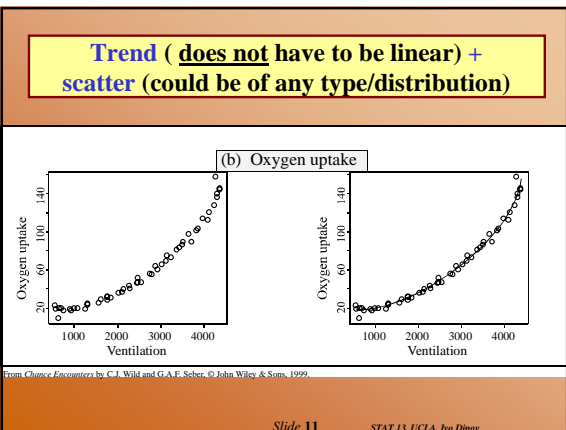
Regression relationship = trend + residual scatter

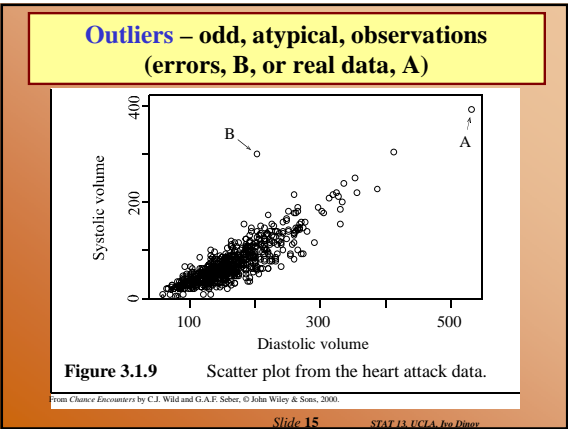
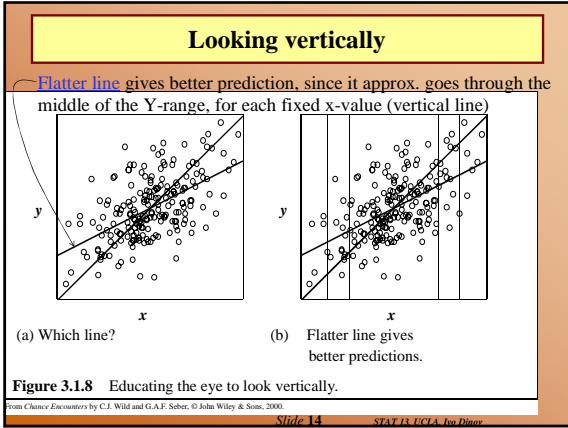
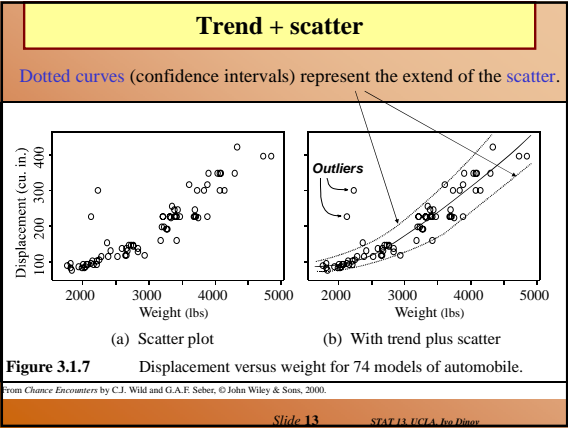
(a) Sales/income

Two scatter plots showing the relationship between Disposable income (\$) on the x-axis and Retail sales (\$) on the y-axis. The x-axis ranges from 9000 to 12000, and the y-axis ranges from 5400 to 6600. The left plot shows the raw data points, and the right plot shows the same data points with a regression line fitted to them, illustrating the trend and residual scatter.

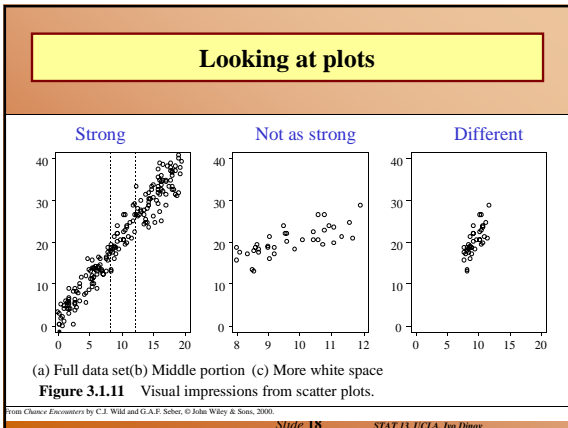
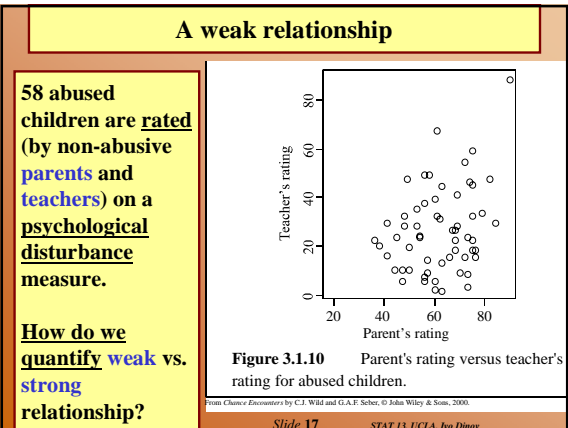
- **Regression** is a way of **studying relationships** between variables (random/nonrandom) for predicting or explaining behavior of 1 variable (**response**) in **terms of** others (**explanatory variables** or **predictors**).

Slide 10 STAT 13, UCLA, Jon Dineer





- ### Review, Fri., Oct. 05, 2001
- Finished univariate data presentation, plotting, summarizing, histogramming, stem-and-leaf plots.
 - What are the differences between the three major measures of central tendency (mean, median, mode).
 - Multivariate relationships and tools for data exploration (scatter plot, regression relationships, trend + scatter).
 - Examples of linear vs. non-linear trends.
- Slide 16 STAT 13, UCLA, Jon Dinger*



Strong and weak relationships

- Plotting a strong relationship only on a small X-range will make the relationship much weaker (So, be ware sample size and sample representativeness do matter).
- The x-range scale and y-range scale need to be taken into account when investigating strong/weak relationships (extending or compressing any of the axes could significantly change the relation trend).

Slide 19 STAT 13, UCLA, Jon Dinger

Questions ...

- When people talk about plotting Y versus X , which variable is conventionally represented on the horizontal axis and which on the vertical axis?
- What are the roles of the response variable and the explanatory variable in regression?
- On a scatter plot, which axis is conventionally used for the explanatory variable and which for the response?

Slide 20 STAT 13, UCLA, Jon Dinger

Questions ...

- What are the two main components of a regression relationship?
- What do we call observations that are further from the trend curve than expected when compared with the usual level of scatter?
- Should outliers simply be discarded when analyzing data?

Slide 21 STAT 13, UCLA, Jon Dinger

Questions ...

- What should you immediately do when you identify an outlier?
- What makes some relationships look weak and others look strong?
- Under what circumstances can a strong relationship look weak in a scatter plot?
- What do we mean by association between two variables? (scatter plot trend that can not be explained by chance alone, implies the two variables are associated) A positive association? (If y and x are associated and y increases with x). A negative association? (If y and x are associated and y decreases with x).

Slide 22 STAT 13, UCLA, Jon Dinger

The prediction problem – can not predict a new response form a weak relationship

Prediction Interval

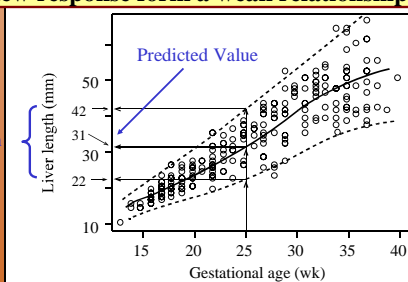


Figure 3.1.12 The prediction problem for liver-length data.

From Chance Encounters by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Slide 23 STAT 13, UCLA, Jon Dinger

Problems with prediction ...

Study of bacteria colony growth in urine samples

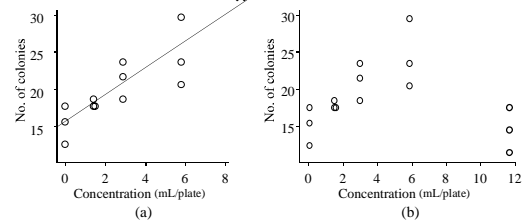


Figure 3.1.13 The dangers of predicting outside the range of the data. (Plotted from data in Margolin [1988]).

From Chance Encounters by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

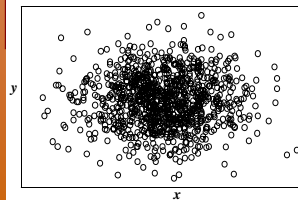
Slide 24 STAT 13, UCLA, Jon Dinger

Be very cautious when (extrapolating) predicting outside the range of the data.

Slide 25 STAT 13, UCLA, Jon Dinger

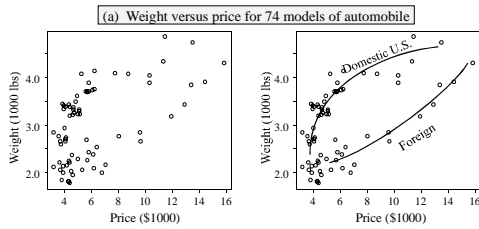
Questions ...

- Why can we **not predict** with any precision from a **weak relationship**?
- Under what **circumstances** can **prediction** be used with any confidence?
- Why is **prediction outside** of the range of the data **dangerous**?



STAT 13, UCLA, Jon Dinger

Other patterns

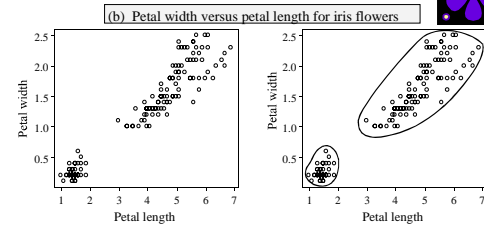


From *Chance Encounters* by C.J. Wild and G.A.F. Seber. © John Wiley & Sons, 2000.

Does **mean**, **regression** or **prediction** make sense in these situations?
How to fix the **problem**? These are some of the issues we'll address later.

Slide 27 STAT 13, UCLA, Jon Dinger

Other patterns - clusters



From *Chance Encounters* by C.J. Wild and G.A.F. Seber. © John Wiley & Sons, 2000.

Slide 28 STAT 13, UCLA, Jon Dinger