Health Studies 315

- **Alicia Y. Toledano, Sc.D.**
- **Assistant Professor, Department of Anesthesia and Critical Care, Department of Health Studies, and Cancer Research Center**
- **Donna (DACC) - 702-6378**
- **Penny (DHS) - 702-2453**
- **toledano@dacc-41.bsd.uchicago.edu**

Regression, Prediction and Adjustment

- **Advanced Therapy in Thoracic Surgery, Ch. 59, Statistical Techniques and Analysis in Thoracic Surgery, with Jemi Olak, MPH, MD (in press)**

Relationships Between Variables

- **1. correlation**
- **2. linear regression**
- **3. multiple linear regression**
- **4. logistic regression**
- **regression: one or several variables that explain or predict a response variable.**

Example: Hypothetical Data

- **relationships of length of postoperative stay and incidence of wound infection to patient age, sex, diagnosis, type of surgery, and operation duration.**
- **n=74**
- **analyses performed in Stata (StataCorp. 1997. Stata Statistical Software: Release 5.0 College Station, TX: Stata Corporation).**

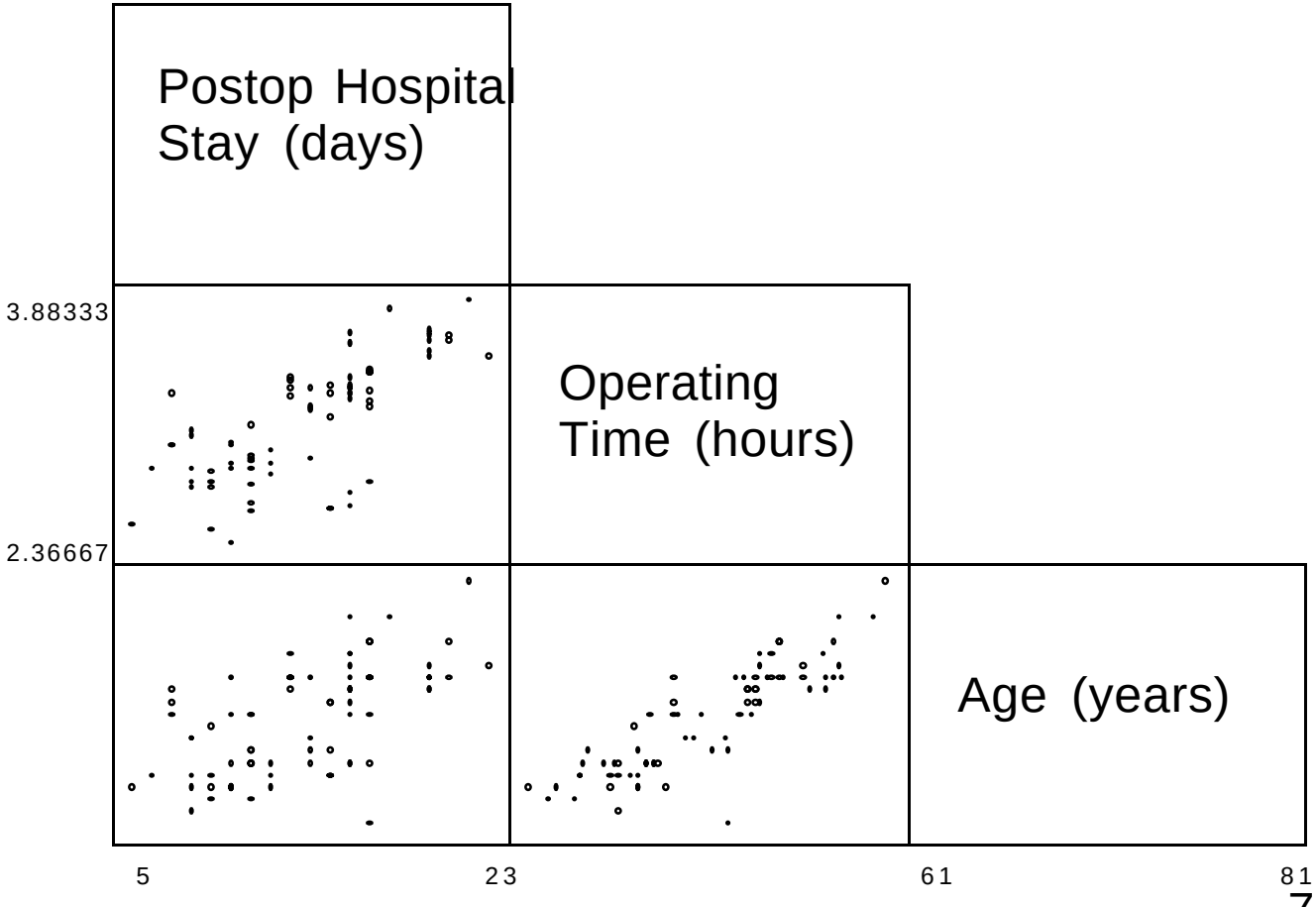
1. Correlation

- **describes strength of linear association between two variables.**
- **Two issues:**
- **linearity**
- **distinction between association and causation**

Example: Linear Association

- relationship between any pair of variables appears to be linear
- observed pairs of values cluster around a straight line
- stronger association seen as tighter clustering around line
- relationship between age and operating time strongest (next slide)

Example: Correlation



Linear Association, cont.

- **correlation will not reflect the strength of a curvilinear association.**
- **even if curvilinear association is strong, correlation may be low**
- **reflects only the linear part**

Association Does Not Imply Causation

- **no matter how strong association is**
- **correlation describes strength of linear association; you must determine whether it is a causal association or not**
- **e.g., consistent in several situations, cause before effect, biologically plausible, dose-response, experimental evidence**

Measuring Correlation

- **correlation coefficient.**
- **values from -1 to +1.**
- **closer to +1 or to -1 indicate stronger association.**
- **between 0 and +1 implies direct relationship: values for both variables increase together**
- **between -1 and 0 implies inverse relationship: as values for one variable increase, values for other variable decrease.**
- **relationships shown in Figure 1 all direct.**

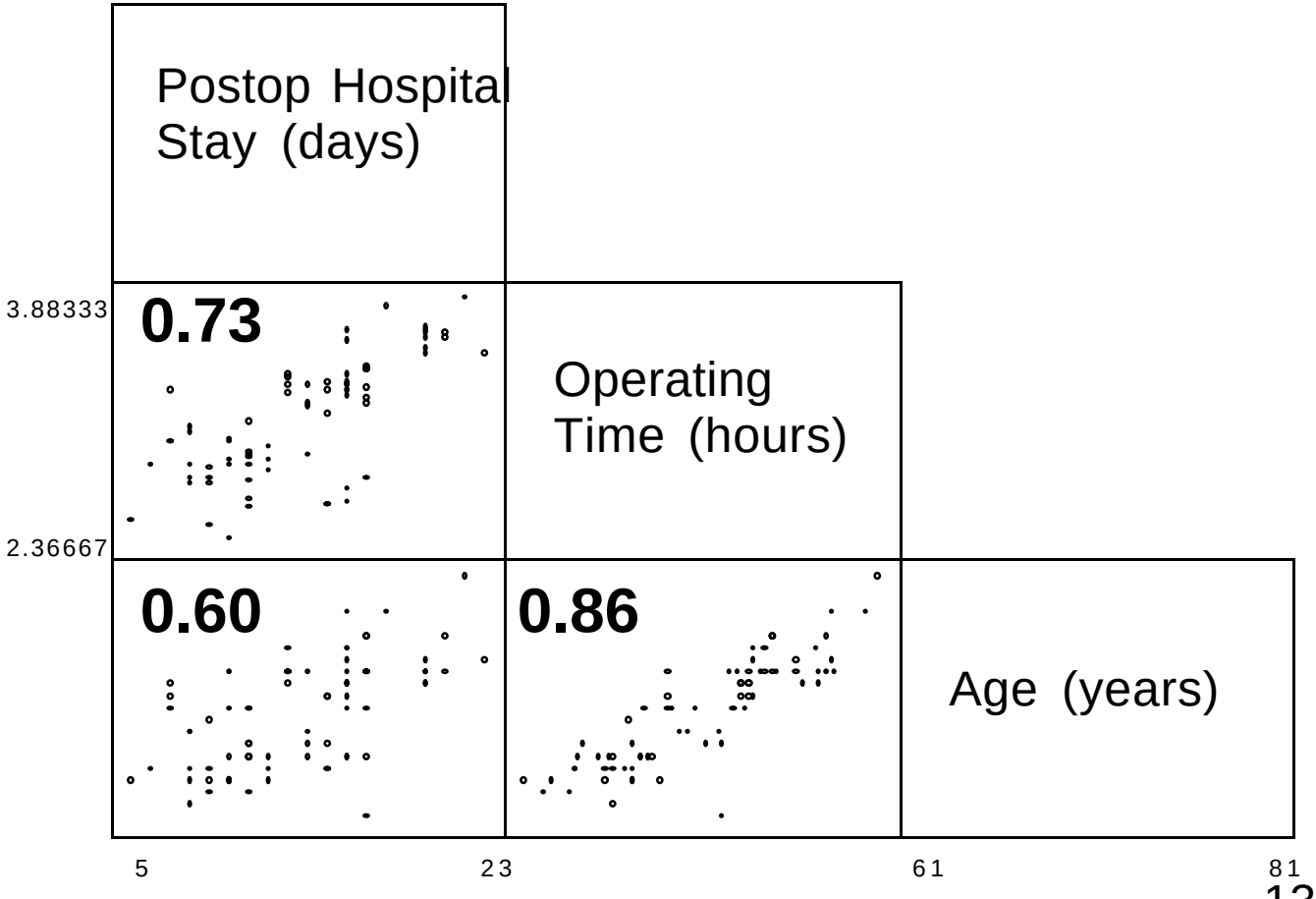
Measuring Correlation, cont.

- **Pearson Product-Moment correlation for continuous variables from normally distributed data**
- **e.g., length of postoperative stay, operating time, and age**
- **Spearman Rank Order correlation used for other types of data**
- **special measures of association used for contingency tables**

Estimated Pearson Correlation Coefficients

Variables	Estimate	95% CI
postop stay (days) & age (years)	0.60	(0.41,0.79)
postop stay (days) & operating time (hours)	0.73	(0.57, 0.89)
operating time (hours) & age (years)	0.86	(0.75, 0.98)

Example: Correlation



Unexpected Results

- **graph can provide clues to why correlation smaller than expected.**
- **large errors in measurement of variables cause points to spread in large cloud**
- **association not linear.**
- **association, no matter how strong, does not imply causation - only you can be the judge of that.**

Regression

- **describes relationship between two or more variables.**
- **can also be used to predict the value of an outcome (dependent variable Y) based on the values of covariates (independent variables X_1, X_2, X_3, ...)**
- **Y should take values along a continuum**
- **each X can take values along a continuum or may be dichotomous or categorical**

2. Linear Regression (Single)

- **mathematical model used to fit a line to pairs of data points Y and X**
- **X should take values along a continuum.**
- **Example: relationship between postoperative stay (Y) and operating time (X).**

Regression Assumptions

- values of Y must be independent of each other
- measure single outcome once on each patient
- independent information
- not just many measurements on same patient
- special tools when measure outcome several times on each patient, or several outcomes on each patient
- paper should indicate assumptions checked

Regression Assumptions, cont.

- **for each distinct value of X, the subset of Y follows a normal (bell-curve) distribution**
- **means of normal distributions for Y variables related to values of X in linear fashion**
- **variances of normal distributions for Y variables similar for all values of independent variable X**

Regression Assumptions, cont.

- **make a graph of the (X,Y) pairs (as above)**
- **sometimes need to transform data to make linear model appropriate**
- **legitimate if one-to-one, e..g., log(X) or log(Y)**
- **check using computer package (Òregression diagnosticsÓ)**

Regression Equation

- $Y = b_0 + b_1 X + e$
- b_0 is the mean of Y when X = 0
- b_1 is the change in the mean of Y for a one unit increase in the value of X
- e is random error.
- R^2: proportion of variation in Y that is explained by variation in X
- square of correlation coefficient

Example: Length of Stay

- **Postoperative stay (days) =**
- **-12.47 + 8.37 * Operating time in hours**
- **R^2 = 0.53, i.e., 53% of variation in length of stay explained by variation in operating time**
- **for each one hour increase in operating time, length of stay increased an average of 8.37 days.**

Example, cont.

- **95% confidence interval (6.51, 10.24)**
- **increase in average length of stay per additional hour of operating time could be from 6.51 to 10.24 days**
- **T-statistic = 8.97, p < 0.001**
- **predict: average length of stay for shortest operating time (2.37 hours) is**
- **-12.47 + (8.37 * 2.37) = 7.37 days**
- **only predict in range of observed data**

Clinical vs Statistical Significance

- clinical significance indicated by large (positive or negative) value for b_1
- observe real differences in the value of Y as the value of X changes.
- statistical significance measured by comparing $b_1 / SE(b_1)$ to critical value from Student's t-distribution.

Clinical vs Statistical Significance

- **if a lot of variability in data, clinically significant b_1 may not be statistically significantly different from zero**
- **if data are measured precisely, but only small changes in Y occur for unit changes in X, b_1 may be statistically significantly different from zero**
- **but not useful in clinical practice.**

3. Multiple Linear Regression

- **dependence of Y on each of several X variables X_1, X_2, X_3, ...**
- **some of these variable may take values along a continuum, and others may be dichotomous or categorical.**
- **Y values must be continuous**
- **Example: how postoperative stay (Y) depends on type of operation, operating time, age, diagnosis, and sex**

Regression Assumptions, cont.

- **draw graph of (X,Y) pairs for each X variable**
- **as above**
- **data transformations may need to be examined for some of the X variables**
- **also methods to check these assumptions in statistical packages.**

Multiple Regression Equation

- $Y = b_0 + b_1 X_1 + b_2 X_2 + ... + e,$
- b_0 overall mean of Y when $X_1=X_2= . . . = 0$
- b_k change in mean of Y for one unit increase in value of X_k
- e is random error
- F-test determines whether H_0: Òall b_k's are zeroÓ is true.
- if H_0 rejected, examine b_k's to determine which of them are informative and which are not.

Example: Multiple Regression

Independent Variables	Estimate	95% CI
Pneumonect. (yes/no)(a)	-1.07	(-2.75,0.62)
Resection (yes/no)(a)	-1.33	(-3.46,0.80)
Thoracotomy (yes/no)(a)	-1.66	(-3.42,0.10)
Operating time (hours)	7.60	(3.79,11.54)
Age (years)(a)	-0.16	(-0.47,0.15)
Met. carcinoma (yes/no)	3.51	(0.66,6.35)
Benign disease (yes/no)	-0.63	(-2.49,1.24)
Sex (female)	2.85	(1.23, 4.47)
b_0		0.25

(a) overall, F(4,65) = 1.28, P-value = 0.288

Model Building

- **clinical and statistical considerations.**
- **if no firm theoretical requirement for including a variable, can drop from model if not statistically significant.**
- **distinguish between clinical and statistical significance**
- **confounders**
- **effect modifiers**

Example: Model Building

- reasonable to believe that operation time and type of operation are closely correlated, such that only one is needed in the model.
- age is closely related to operating time in this hypothetical data so that it, too, may be dropped

Example: Final Model

n=74 patients, overall F(4,69) = 31.12, P < 0.0001

Independent Variables	Estimate	95% CI
Operating time (hours)	6.30	(3.90, 8.83)
Met. carcinoma (yes/no)	3.22	(0.49, 5.96)
Benign disease (yes/no)	-0.37	(-2.15, 1.42)
Sex (female)	3.14	(1.56, 4.73)
b_0		-7.59

R^2 (adjusted) = 0.6227

was 0.6827 with full model

Example: Interpretation

- **62% of variability in length of stay explained by variation in operating time, diagnosis, and sex**
- **for two patients with same diagnosis and sex, additional hour of operating time associated with an average increased length of stay of 6.36 days.**
- **On average, patients with metastases stayed 3.22 days longer than those without**
- **and women stayed 3.14 days longer than men**

4. Logistic Regression

- special regression model used when dependent variable Y is dichotomous
- e.g., whether wound infection occurs.
- X variables still independent variables
- model probability Y is “yes”given X values
- linear regression: relationship between mean value of Y and X values
- probability called π, should always be estimated between 0 and 1

Logistic Regression Equation

- **logit(π) = $b_0 + b_1 X_1 + b_2 X_2 + ... + e$.**
- **logit(π) = log[π / (1-π)]**
- **e^b: odds ratios for occurrence of the event (e.g., wound infection) per unit change in the value of X_k (e.g., additional hour of surgery)**
- **"pseudo-R^2"**

Example: Wound Infection

(n=74 patients, overall Chi-square(6) = 53.70, P < 0.0001)

Independent Variables	Estimated OR	95% CI
Not lobectomy (yes/no)	18.80	(1.60, 220.61)
Postoperative stay (days)	1.41	(0.99, 2.02)
Operating time (hours)	1.80	(0.02, 164.76)
Age (years)	0.54	(0.33, 0.88)
Not benign disease (yes/no)	20.11	(1.96, 206.85)
Sex (male)	29.21	(1.23, 693.67)

pseudo-R^2 = 0.5962

Example: Interpretation

- **odds ratio for each independent variable is adjusted for all the other independent variables.**
- **odds of patients with malignant disease getting wound infections are approximately 20 times the odds of patients with benign disease getting wound infections.**
- **odds of infection are approximately halved for each one year increase in age.**

Example: Interpretation

- **effects of**
- **diagnosis,**
- **age,**
- **type of surgery (lobectomy or not), and**
- **sex,**
- **both clinically and statistically significant.**

Summary

- **correlation coefficient measures strength of linear association**
- **association does not imply causation**
- **regression models linear relationship of mean of Y to one or more predictors X**
- **logistic regression models relationship of probability that Y is ÒyesÓ to one or more predictors X**

Summary, cont.

- **paper should provide evidence that regression is appropriate**
- **diagnostic plots**
- **report coefficients, not just p-values**
- **coefficients tell you size and direction of effects of X variables on the outcome Y**
- **multiple (logistic) regression can be used to adjust for effects of other X variables**