

# Statistics

Chi-square test, Relationship among variables:  
scatterplots, correlation, simple linear regression,  
multiple regression and coefficient of determination

February 10, 2010

Jobayer Hossain, Ph.D. & Tim Bunnell, Ph.D.

*Nemours Bioinformatics Core Facility*



# Chi-square Test

- USE
  - Testing the population variance  $\sigma^2 = \sigma_0^2$ .
  - Testing the goodness of fit.
  - Testing the independence/ association of attributes
- Assumptions
  - Sample observations should be independent.
  - Cell frequencies should be  $\geq 5$ .
  - Total observed and expected frequencies are equal

# Chi-square Test

- Formula: If  $x_i$  ( $i=1,2,\dots,n$ ) are independent and normally distributed with mean  $\mu$  and standard deviation  $\sigma$ , then,

$$\sum_{i=1}^n \left( \frac{x_i - \mu}{\sigma} \right)^2 \text{ is a } \chi^2 \text{ distribution with } n \text{ d.f.}$$

- If we don't know  $\mu$ , then we estimate it using a sample mean and then,  $\sum_{i=1}^n \left( \frac{x_i - \bar{x}}{\sigma} \right)^2$  is a  $\chi^2$  distribution with  $(n - 1)$  d.f.

# Chi-square Test

- For a contingency table, we use the following chi- square test statistic,

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}, \text{ distributed as } \chi^2 \text{ with } (n - 1) \text{ d.f.}$$

$O_i$  = Observed Frequency

$E_i$  = Expected Frequency

## Chi-square Test

	Male O(E)	Female O(E)	Total
Group 1	9 (10)	11 (10)	20
Group 2	8 (10)	12 (10)	20
Group 3	11 (10)	9(10)	20
	30	30	60

# Chi-square Test– calculation of expected frequency

- To obtain the expected frequency for any cell, use:
- Corresponding (row total X column total) / grand total
- E.g: cell for group 1 and female, substituting:  $(30 \times 20 / 60) = 10$

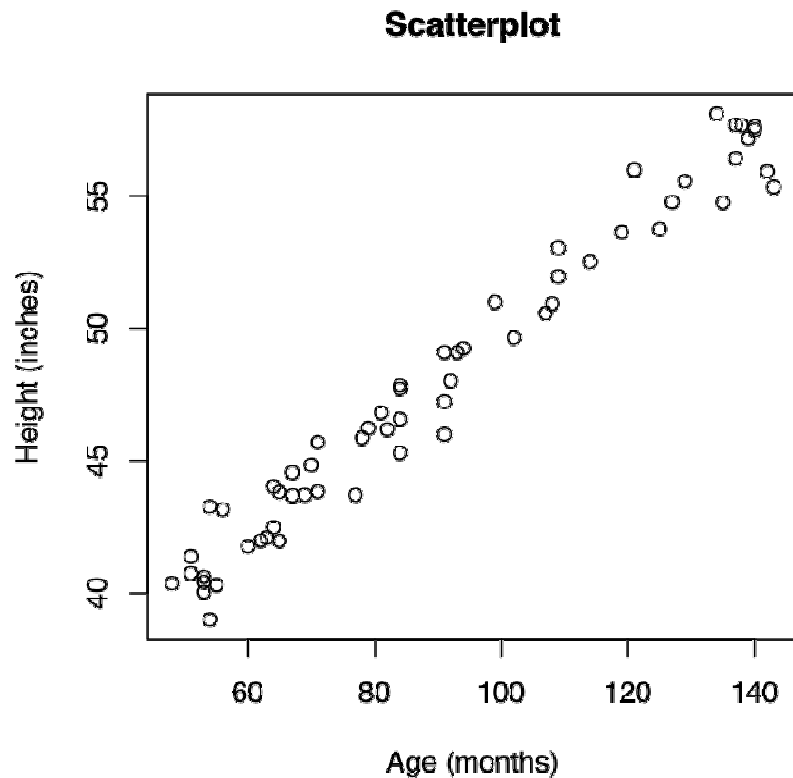
# Chi-square Test: SPSS demonstration

- Analyze->Descriptive statistics -> Crosstabs -> Pick row and column variables, select other options and click ok

# Relationships among variables

- **Response (dependent) variable(s)** - measure the outcome of a study.
- **Explanatory (Independent) variable(s)** - explain or influence the changes in a response variable
- **Outlier** - an observation that falls outside the overall pattern of the relationship.
- **Positive Association** - An *increase* in an independent variable is associated with an *increase* in a dependent variable.
- **Negative Association** - An *increase* in an independent variable is associated with a *decrease* in a dependent variable.

# Scatterplots

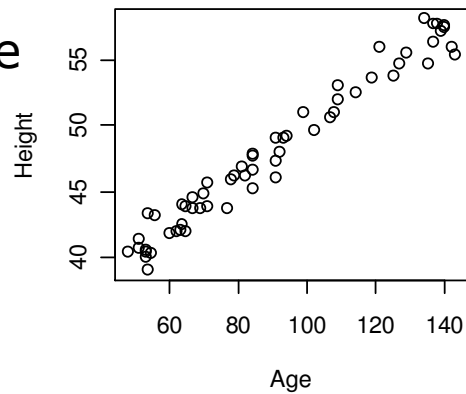


- Shows relationship between two variables (age and height in this case).
- Reveals **form**, **direction**, and **strength** of the relationship.

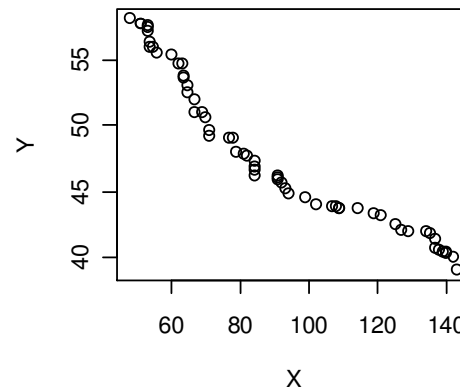
# Scatterplots

Strong positive  
association

Scatterplot of Age vs Height

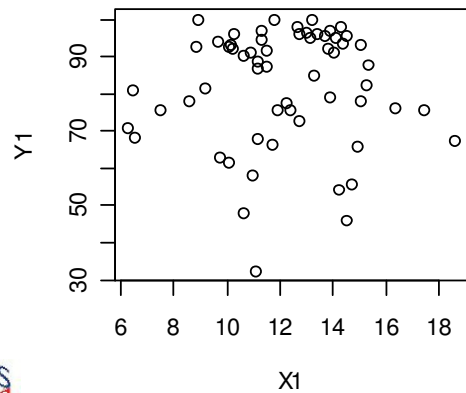


Scatterplot of variables X and Y



Strong negative  
association

Scatterplot of variables X1 and Y1



Points are scattered with a poor  
association

## Scatterplots - SPSS Demo

- Graphs->Legacy Dialogues-> Scatter/Dot->select Simple scatter and click on Define. In the new window select variables for x axis (independent) and y axis (response), write titles and labels and then click ok.

# Correlation

- Correlation measures the degree to which two variables are associated.
- Two commonly used correlation coefficient:
  - Pearson Correlation Coefficient
  - Spearman Rank Correlation Coefficient

# Correlation

- **Pearson Correlation Coefficient:** measures the direction and strength of the relationship between two quantitative variables. Suppose that we have data on variables  $x$  and  $y$  for  $n$  individuals. Then the correlation coefficient  $r$  between  $x$  and  $y$  is defined as,

$$r = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

Where,  $s_x$  and  $s_y$  are the standard deviations of  $x$  and  $y$ .

- $r$  is always a number between -1 and 1. Values of  $r$  near 0 indicate little or no linear relationship. Values of  $r$  near -1 or 1 indicate a very strong linear relationship.
- The extreme values  $r=1$  or  $r=-1$  occur only in the case of a perfect linear relationship, when the points lie exactly along a straight line.

# Correlation

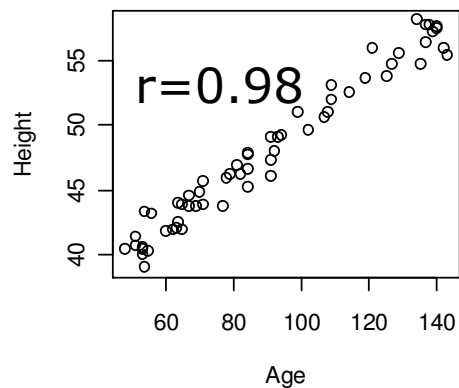
- Positive  $r$  indicates positive association i.e. association between two variables in the same direction, and negative  $r$  indicates negative association.
- Scatterplot of Height and Age shows that these two variables possess a strong, positive linear relationship. The correlation coefficient of these two variables is 0.9829632, which is very close to 1.

# Correlation

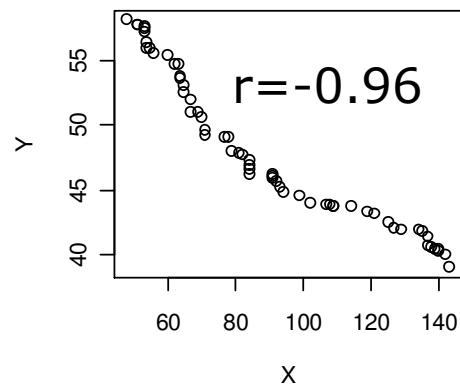
- Spearman Rank Correlation Coefficient:
  - This is non-parametric measure of correlation between two variables
  - This is basically a pearson correlation coefficient of the ranks of data of two variables instead of data itself.

# Correlation

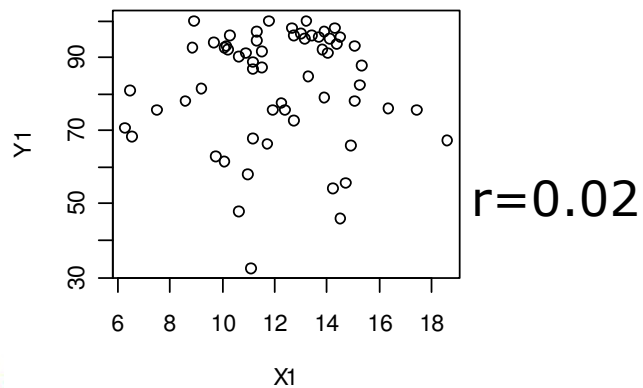
Scatterplot of Age vs Height



Scatterplot of variables X and Y

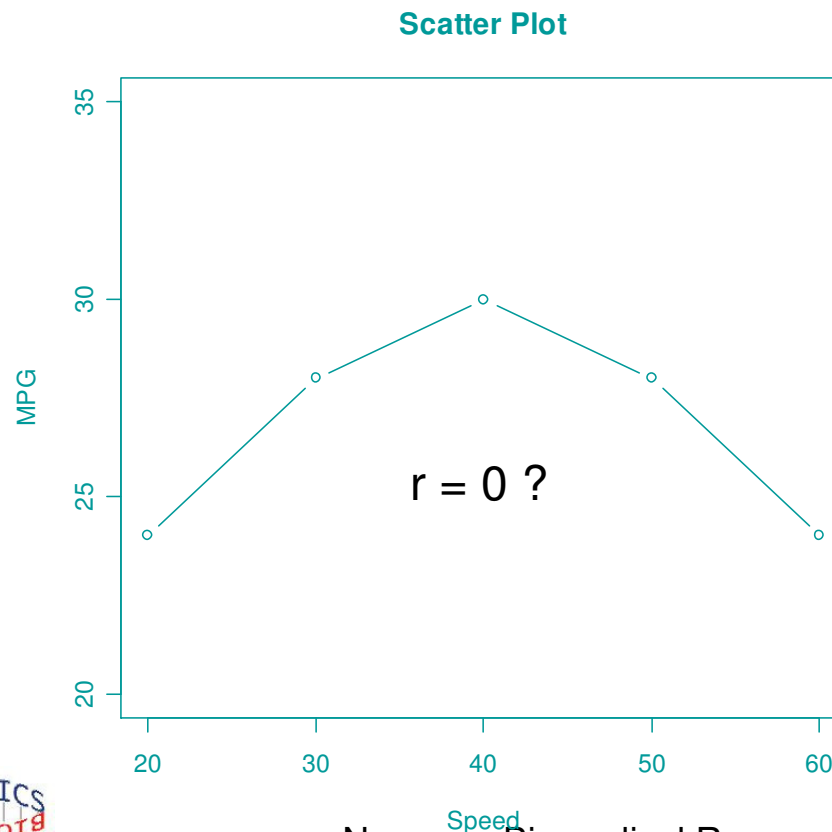


Scatterplot of variables X1 and Y1



# Correlation

**Strong association but no correlation:** Gas mileage of an auto mobile first increases than decreases as the speed increases like the following data:



Speed	20	30	40	50	60
MPG	24	28	30	28	24

Scatter plot shows an strong association. But calculated,  $r = 0$ , why?

It's because the relationship is not linear and  $r$  measures the linear relationship between two variables.

# Correlation

- **Influence of an outlier**

- Consider the following data set of two variables X and Y:

X	20	30	40	50	60	80
Y	24	28	30	34	37	15

$$r = -0.237$$

- After dropping the last pair,

X	20	30	40	50	60
Y	24	28	30	34	37

$$r = 0.996$$

# Correlation: SPSS demonstration

- Analyze-> Correlate -> Bivariate and then select variables for correlations

# Simple Linear Regression

- Regression refers to the value of a response variable as a function of the value of an explanatory variable.
- A regression model is a function that describes the relationship between response and explanatory variables.
- A simple linear regression has one explanatory variable and the regression line is straight.
- The response variable is quantitative and independent variable (s) can be both quantitative and categorical.
- Categorical variables are handled by creating dummy variable (s).

# Simple Linear Regression

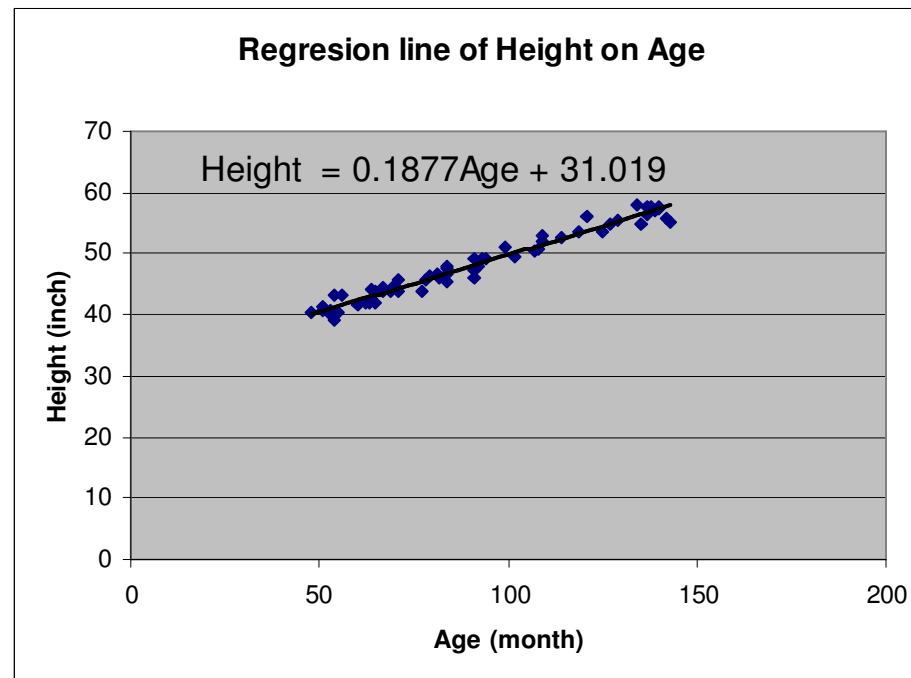
- The linear relationship of variables Y and X can be written as in the following regression model form

$$Y = b_0 + b_1X + e$$

where, 'Y' is the response variable, 'X' is the explanatory variable, 'e' is the residual (error), and  $b_0$  and  $b_1$  are two parameters. Basically,  $b_0$  is the intercept and  $b_1$  is the slope of a straight line  $y = b_0 + b_1X$ .

# Simple Linear Regression

A simple regression line is fitted for height on age. The intercept is 31.019 and the slope (regression coefficient) is .1877.



# Simple Linear Regression

## Assumptions:

- o Response variable is normally distributed.
- o Relationship between the two variables is linear.
- o Observations of response variable are independent.
- o Residual error is normally distributed with mean 0 and constant standard deviation.

# Simple Linear Regression

- **Estimating Parameters  $b_0$  and  $b_1$** 
  - Least Square method estimates  $b_0$  and  $b_1$  by fitting a straight line through the data points so that it minimizes the sum of square of the deviation from each data point.
  - Formula:

$$\hat{b}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}$$

# Simple Linear Regression

- **Fitted Least Square Regression line**

- Fitted Line:  $\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$

- Where  $\hat{Y}_i$  is the fitted / predicted value of  $i^{th}$  observation ( $Y_i$ ) of the response variable.

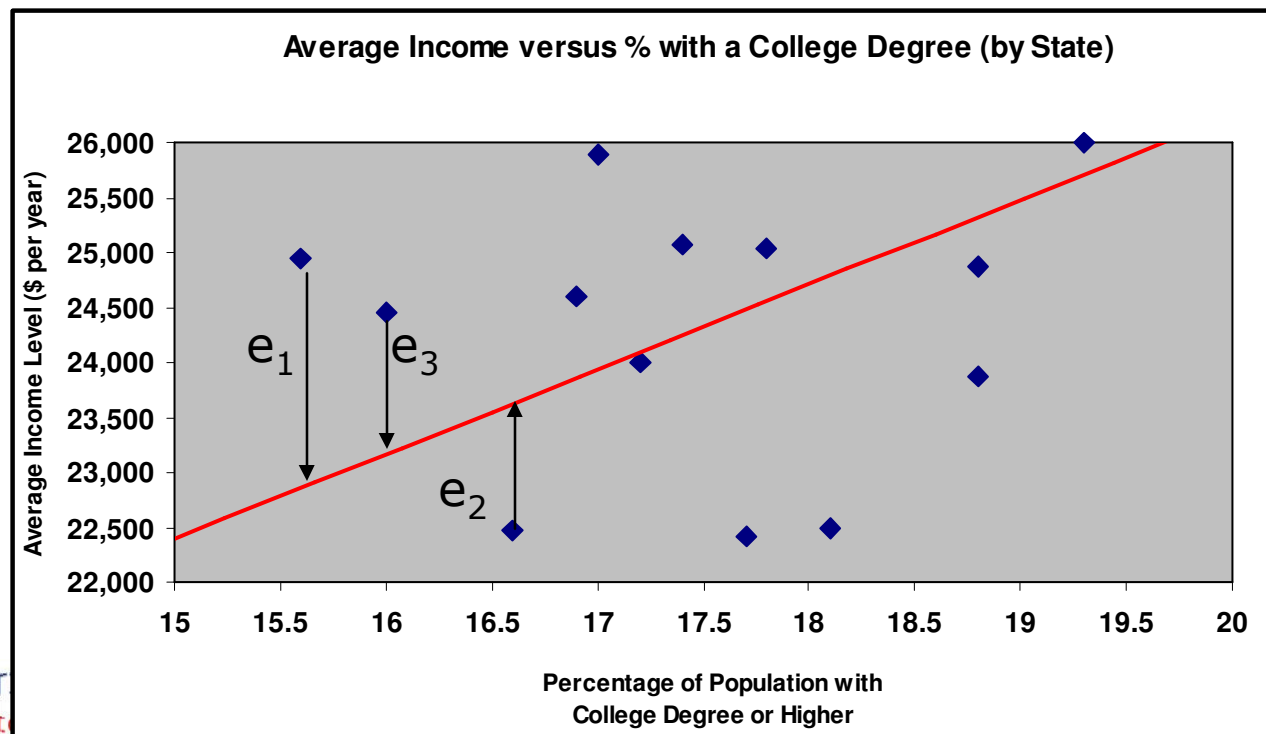
- Estimated Residual:  $\hat{e}_i = Y_i - \hat{Y}_i$

- Least square method estimates  $b_0$  and  $b_1$  to minimize the summed error:  $\sum_{i=1}^n \hat{e}_i^2$

# Simple Linear Regression

## □ Fitted Least Square Regression line

In this example, a regression line (red line) has been fitted to a series of observations (blue diamonds) and residuals are shown for a few observations (arrows).



# Simple Linear Regression

- Interpretation of the Regression Coefficient and Intercept
  - Regression coefficient ( $b_1$ ) reflects the average change in the response variable  $Y$  for a unit change in the explanatory variable  $X$ . That is, the **slope** of the regression line. E.g.
  - Intercept ( $b_0$ ) estimates the average value of the response variable  $Y$  without the influence of the explanatory variable  $X$ . That is, when the explanatory variable = 0.0.

# Simple linear regression: SPSS demonstration

- Analyze ->Regression->Linear->select a dependent (e.g. height) and an independent variable (age) and other output options.

# Multiple Regression

- Two or more independent variables to predict a single dependent variable.
- Multiple regression model of Y on p number of explanatory variables can be written as,

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_pX_p + e$$

where  $b_i$  ( $i=1,2, \dots, p$ ) is the regression coefficient of  $X_i$

# Multiple Regression

- Fitted Y is given by,

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \dots \hat{b}_p X_p$$

Where,  $\hat{b}_i$  is the estimate of  $b_i$

- The estimated residual error is the same as that in the simple linear regression,

$$\hat{e}_i = Y_i - \hat{Y}_i$$

# Multiple Regression: SPSS demonstration

- Analyze ->Regression->Linear->select a dependent variable (e.g. PLUC.pre) and more than one independent variables (e.g. age and LWAS) and other output options.

## Coefficient of Determination (Multiple R-squared)

- Total variation in the response variable Y is due to (i) regression of all variables in the model (ii) residual (error).
- Total variation of y,  $SS(y) = SS(\text{Regression}) + SS(\text{Residual})$
- The Coefficient of Determination is,

$$R^2 = \frac{SS(\text{Regression})}{\text{Total } SS(Y)} = 1 - \frac{SS(\text{Residual})}{\text{Total } SS(Y)}$$

## Coefficient of Determination (Multiple R-squared)

- $R^2$  lies between 0 and 1.
- $R^2 = 0.8$  implies that 80% of the total variation in the response variable  $Y$  is due to the contribution of all explanatory variables in the model. That is, the fitted regression model explains 80% of the variance in the response variable.

## Coefficient of Determination (Multiple R-squared)

- A  $R^2$  always increases with an increasing number of variables in the model, without consideration of sample size. This increase of  $R^2$  may be due to chance variation.
- An **Adjusted  $R^2$**  accounts for sample size and number of variables are being used in the model and reduce the possibility of chance variation.

# Coefficient of Determination (Multiple R-squared): SPSS demonstration

- It's in the output of Multiple regression.
- For the previous example, Coefficient of determination is 0.039.

# Thank you



Nemours Biomedical Research