STA 321 Spring 2014

Lecture 7 *Tuesday, Feb 11th*

- **Summarizing Bivariate Data**

  - **Two quantitative variables**

    Scatter Diagram
    Regression Line
    Correlation Coefficient, Coefficient of Determination, Slope and Intercept of Regression Line

Describing the Relationship Between Two Quantitative Variables

Scatter Diagram

- In applications where one variable depends to some degree on the other variables, we label the dependent variable Y and the independent variable X
- Sometimes, this choice is ambiguous

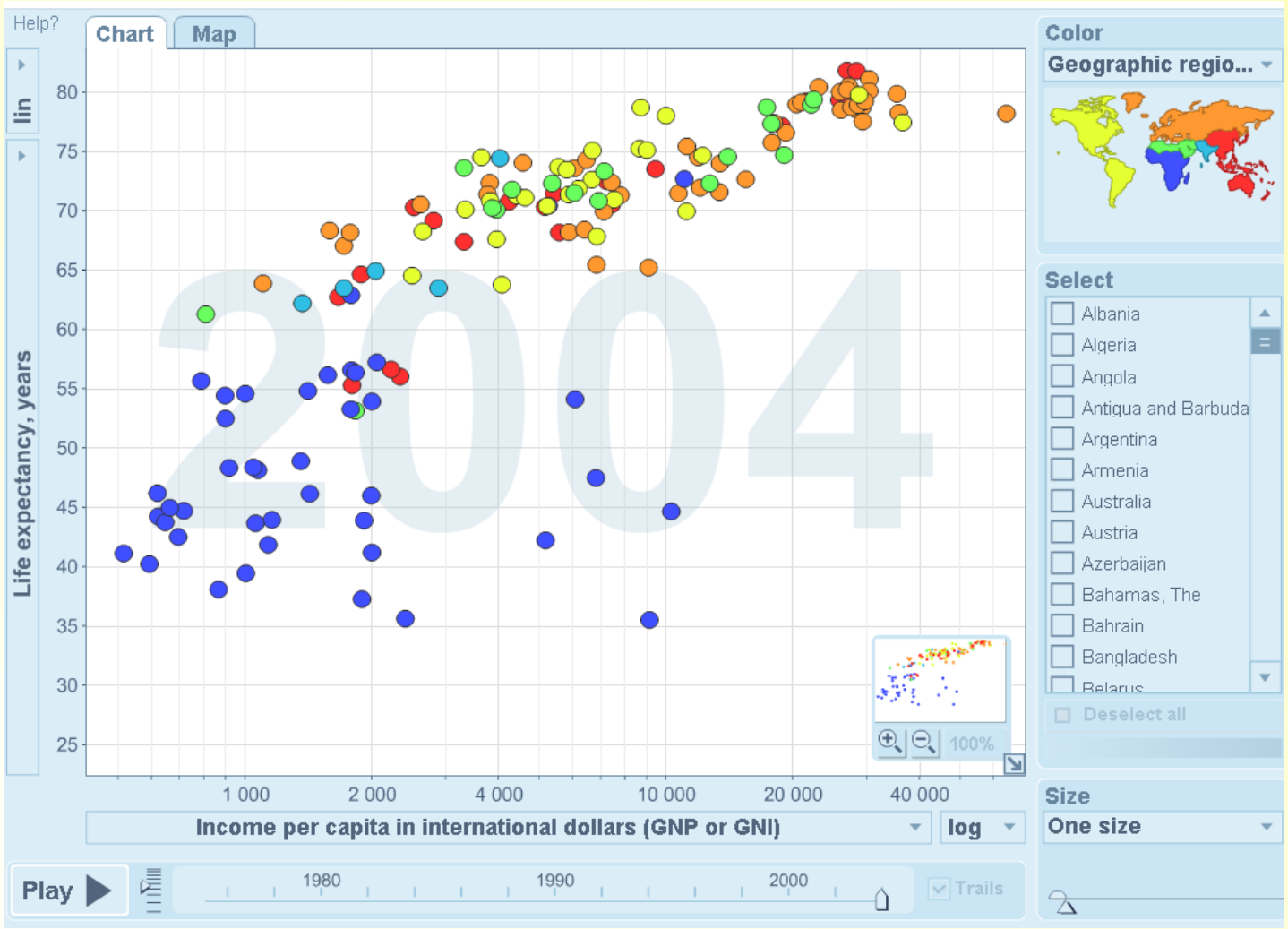
  Example:

  Income = X

  Life expectancy = Y

- Each point in the scatter diagram corresponds to one observation (one country)

Scatter Diagram of Life Expectancy (Y) and Income (X) for Several Countries



Source: gapminder.org

Analyzing Linear Relationships Between Two Quantitative Variables

- Is there an association between the two variables?
- Positive or negative?
- How strong is the association?
- Notation
  - Response variable: *Y*
  - Explanatory variable: *X*

Recall: Association / Independence

- ***Association***: The distribution of the response variable changes in some way as the value of the explanatory variable changes
- "*No association*" is called ***Independence***.
- In practice there is rarely data with perfect independence, and in samples, there is sampling variation
- A measure of association is a statistic that summarizes the strength of the statistical dependence between two variables

Sample Measures of Linear Relationship

- Sample Covariance:

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{1}{n-1}\left(\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i \right)$$

- Sample Correlation Coefficient:

$$r = \frac{s_{xy}}{s_x s_y}$$

- Here, s_x and s_y are the standard deviations of the x and y variables
- Population measures: Divide by N instead of n-1

Properties of the Correlation I

- The value of *r* does not depend on the units (e.g., changing from inches to centimeters)
- *r* is standardized
- *r* is always between –1 and 1, whereas the covariance can take *any* number
- *r* measures the ***strength and direction of the linear association*** between *X* and *Y*
- r>0 positive linear association
- r<0 negative linear association

Properties of the Correlation II

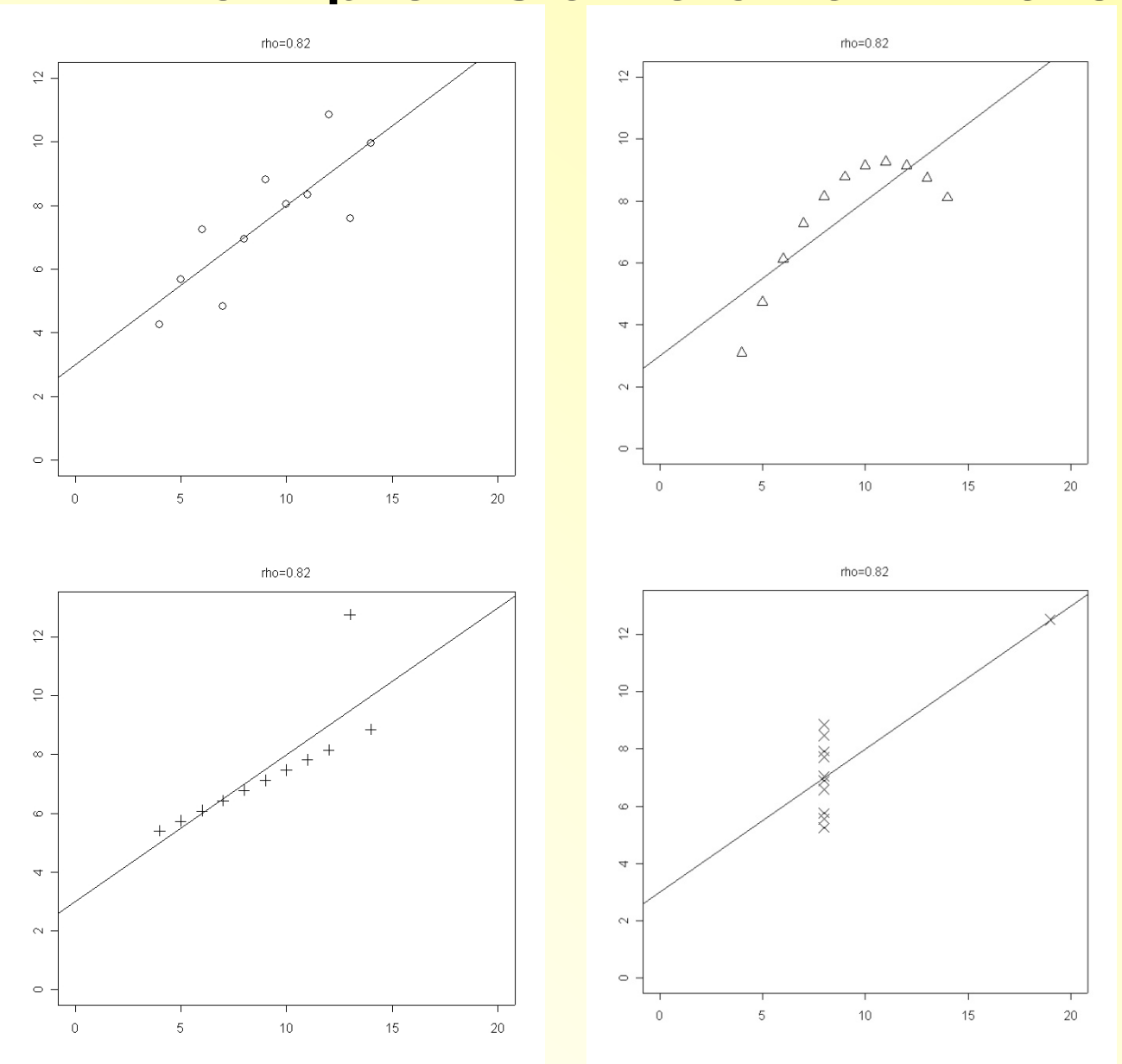
- $r = 1$ when all sample points fall exactly on a line with positive slope *(perfect positive association)*
- $r = -1$ when all sample points fall exactly on a line with negative slope *(perfect negative association)*
- The larger the absolute value of r, the stronger is the degree of linear association

Properties of the Correlation III

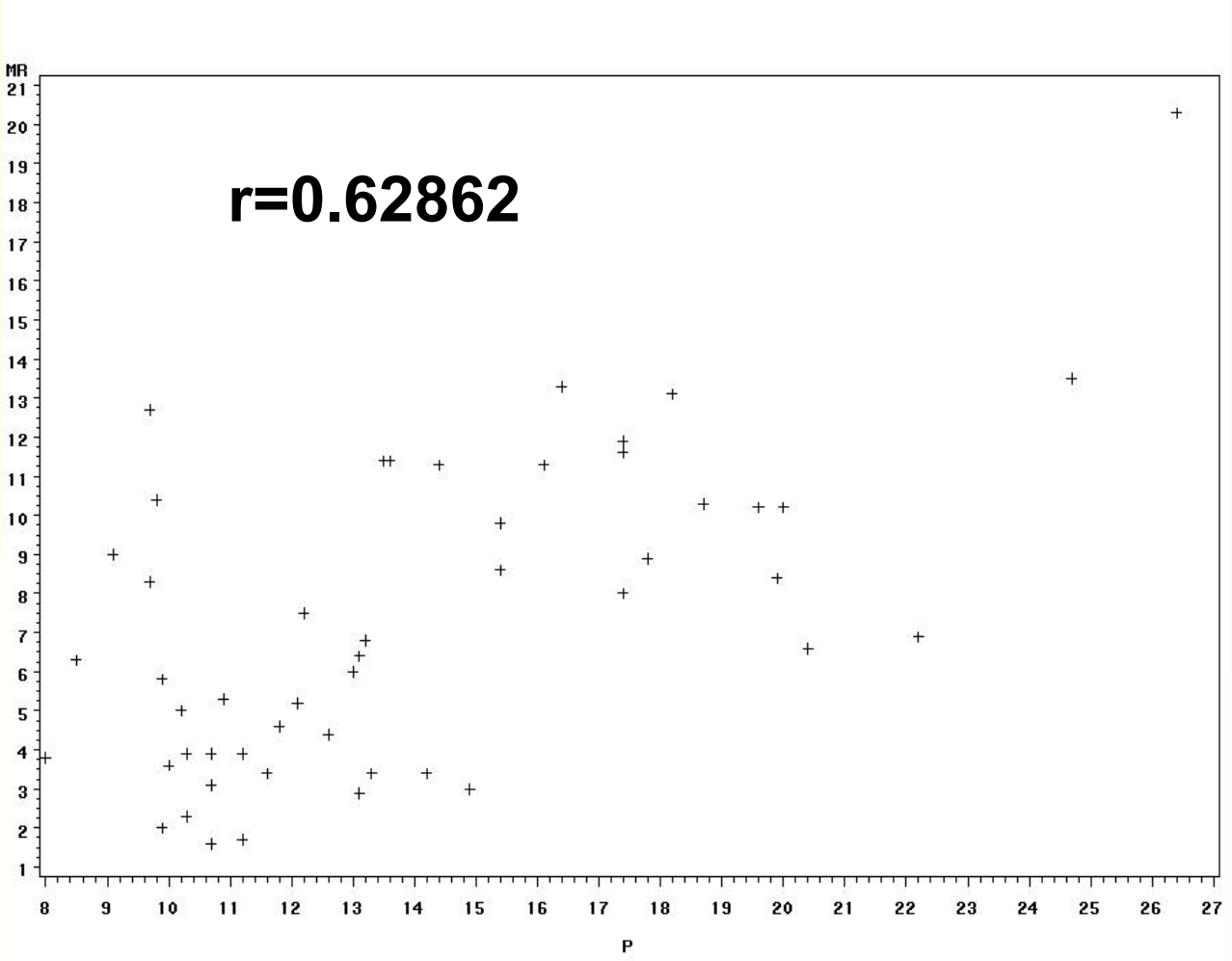
- If *r* is close to 0, this does not necessarily mean that the variables are not associated
- It only means that they are not *linearly associated*

- The correlation treats *X* and *Y* symmetrically
- That is, it does not matter which variable is explanatory (*X*) and which one is response (*Y*), the correlation remains the same

Example: Correlation = 0.82



Scatter Diagram of Murder Rate (Y) and Poverty Rate (X) for the 50 States



Correlation and Scatterplot Applet 1

Simple Regression Analysis Tool

Model Assumptions and Violations

- **Factors Influencing the Correlation**
- The sample correlation depends on the range of *X*-values sampled
- When a sample has a much narrower range of variation in *X* than the population, the sample correlation tends to underestimate the population correlation
- The sample (*X*, *Y*) values should be a random sample of the population
- It should be representative of the *X* population values as well as the *Y* values

Correlation and Scatterplot Applet

Correlation = Linear Association

- *r* measures the ***strength and direction of the linear association*** between *X* and *Y*
- In other words: *r* measures how well the scatter plot of the data can be approximated by a straight line

- *Which straight line is best?*

Correlation: Example

- For a sample of 100 people, the correlation coefficient between X = hours of statistics instruction and Y = annual income (in dollars) equals 0.70

a) Suppose instead that X refers to minutes instead of hours, and Y refers to monthly income converted into Euro. State the correlation.

b) Suppose that Y is treated as the explanatory variable and X is treated as the response variable. Will the correlation coefficient change in value?

Method of Least Squares

- There are many possible ways to choose a straight line through the data
- Goal: Make the vertical distances between the observations and the straight line as small as possible
- Vertical distances: residuals
- The sum of the residuals should be zero
- There are many possible ways to choose a straight line through the data such that the sum of the residuals is zero

Method of Least Squares

- Better Goal: Minimize the sum of the squared residuals $\sum (y_i - \hat{y}_i)^2$
- The squared residuals are the squared vertical distances between the straight line and the data
- Correlation by Eye Applet Minimize the MSE (mean square error)

- This method is called the ***method of least squares*** (Gauss)

Method of Least Squares

- This leads to the ***prediction equation*** or ***least squares equation***

$$\hat{y} = b_0 + b_1 \cdot x$$

- With the following coefficients

$$\text{slope} \qquad b_1 = \frac{s_{xy}}{s_x^{\,2}} = \frac{\sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

$$\text{intercept} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$

- In practice, the calculations for slope and intercept are done using the computer, not by hand

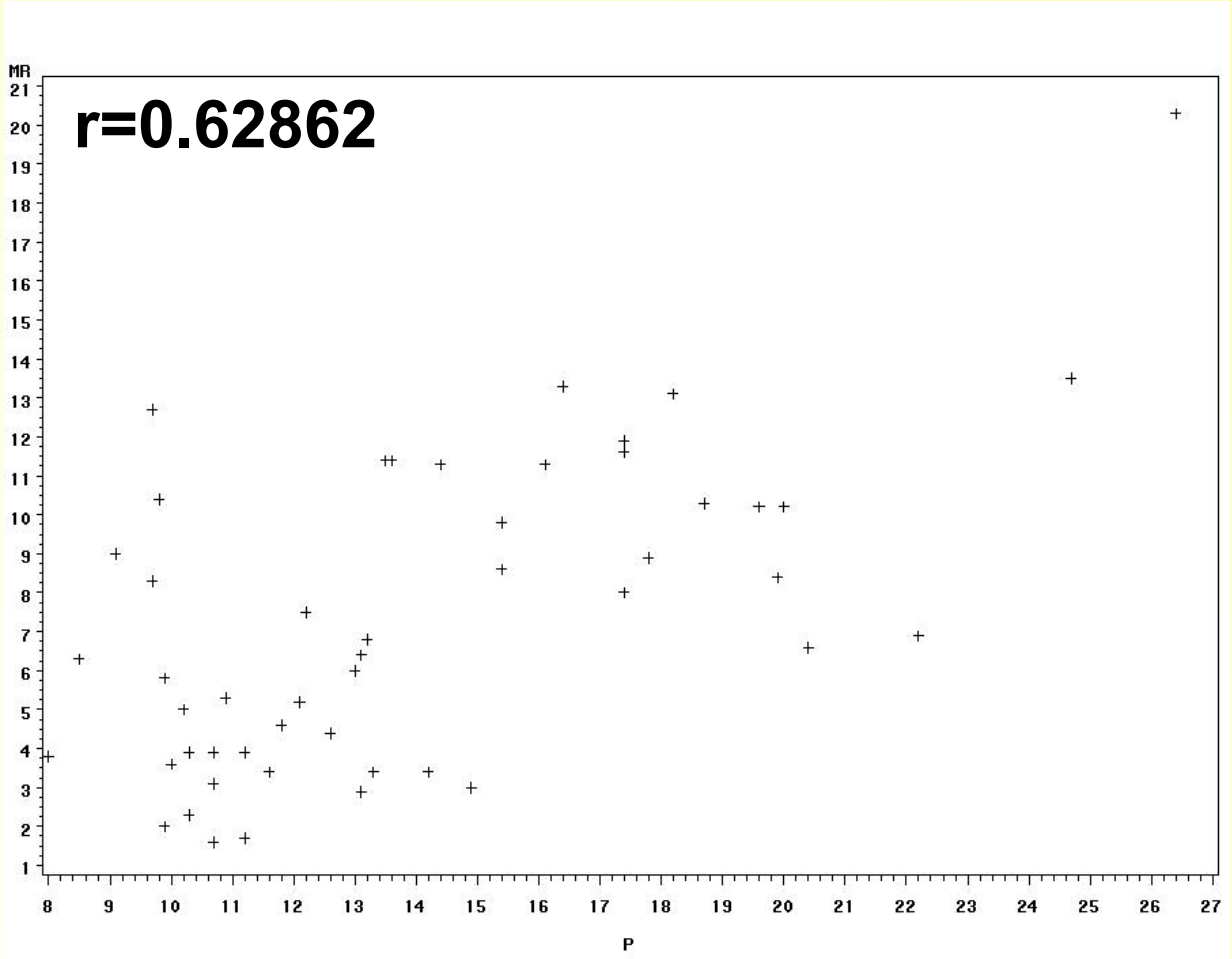
Simple Regression Analysis Tool

Descriptive Statistics Applet for Bivariate Quantitative Samples

Method of Least Squares

- The least squares estimates b_0 and b_1 in the prediction equation are the values that make the sum of squared residuals minimal
- The equation is called prediction equation because it can be used to predict values of the response variable when knowledge about the explanatory variable is available

Scatter Diagram of Murder Rate (y) and Poverty Rate (x) for the 50 States



- Any other linear equation will lead to a larger sum of squared residuals
- The observed data points fall closer to this line than to any other line

$$\hat{y} = b_0 + b_1 \cdot x = -0.8567 + 0.5842 \cdot x$$

Interpretation of Slope and Intercept

- Slope: rise/run
  - Change in *y* (rise) for a one-unit increase in *x* (run)

- Intercept
  - Intersection of the straight line with the (vertical) *y*-axis
  - (Hypothetical) predicted value of *y* for *x=0*
  - Often, the intercept has little practical meaning because the data does not have observations with *x=0*

Example: HW vs. Midterm Scores

- Someone claims that the prediction equation

$$\hat{y} = 53.4 + 13.7x$$

approximates the relationship between y=midterm exam score (between 0 and 100) and x=homework average (x=10 equals 100%)

a) Is this realistic? Why or why not?

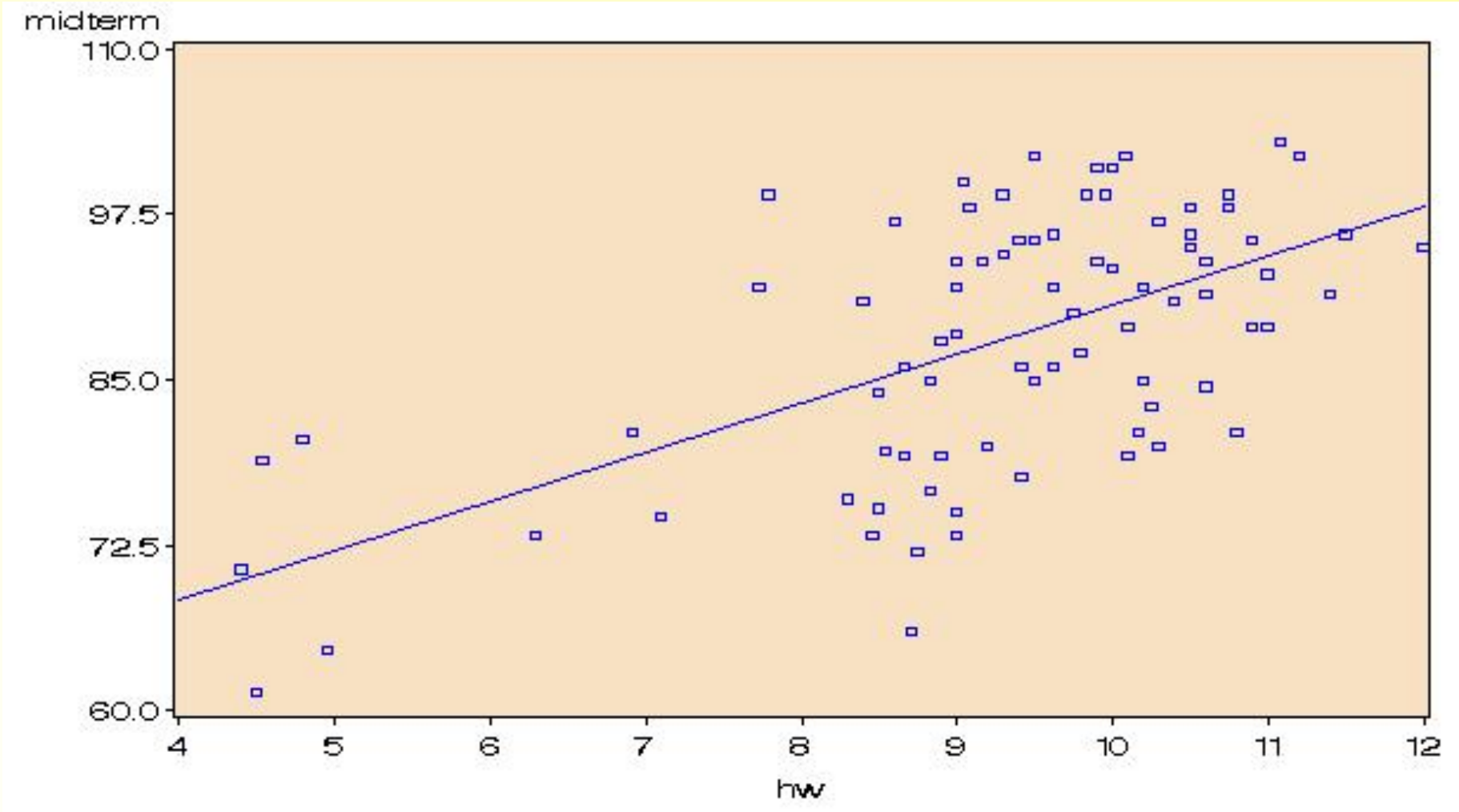
b) Suppose that the prediction equation is actually

$$\hat{y} = 53.4 + 3.7x$$

Interpret the slope.

c) Using the prediction equation in b), find the predicted midterm exam score for a student having a homework percentage of (i) 9 points = 90% and (ii) 10 points = 100%

Example: HW vs. Midterm Scores



$$\hat{y} = 53.4 + 3.7x$$

Example (Data from the 50 States)

- y = violent crime rate
- x = poverty rate
- The relation can be approximated by the straight line y = 210 + 25 x
- Interpretation of y-intercept and slope?
- y = violent crime rate
- x = percentage of high school graduates
- Approximated by y = 1756 - 16 x
- Interpretation of y-intercept and slope?

Linear Function, Slope

- When $b_1 = 0$, the value of x has no influence on the value of y
- $b_1 > 0$: positive relationship between the variables
- $b_1 < 0$: negative relationship

Slope (b_1) vs. Correlation (r)

- The slope b_1 of the prediction equation tells us the direction of the association between the two variables
  - Positive b_1 : Slope upward
  - Negative b_1 : Slope downward
- The slope does *not* tell us the *strength* of the association
  - It depends on the units and can be made arbitrarily small or large by choice of units
  - It does ***not*** treat *X* and *Y* symmetrically
- A measure of the strength of the linear association is the correlation coefficient *r*

Correlation and Slope

- The value of r does not depend on the units – it is a standardized regression coefficient
- r is always between –1 and 1
- Recall that b_0 and b_1 could take *any* value
- r measures the ***strength of the linear association*** between X and Y
- r has the same sign as the slope b_1
- r is symmetric in x and y

Correlation Coefficient and Slope

- The correlation coefficient r is a standardized version of the slope b_1 of the prediction equation

$$r = \frac{s_x}{s_y} \cdot b_1$$

where b_1 is the slope in the equation

$$\hat{y} = b_0 + b_1 \cdot x$$

- Advanced Interpretation:
  - Slope: If x increases by **one unit**, then y is expected to increase by $\boldsymbol{b_1}$ **units**
  - Correlation coefficient: If x increases by **one standard deviation**, then y is expected to increase by $\boldsymbol{r}$ **standard deviations** (does <u>not</u> depend on units!)

Regression Toward the Mean

- Sir Francis Galton (1880s): correlation between *x*=father's height and *y*=son's height is about 0.5
- Interpretation: If a father has height one standard deviation below average, then the predicted height of the son is 0.5 standard deviations below average
- More Interpretation: If a father has height two standard deviations above average, then the predicted height of the son is 0.5 x 2 = 1 standard deviation above average
- Tall parents tend to have tall children, but not ***so*** tall
- This is called "regression toward the mean"

  → statistical term "regression"

The regression effect is often misunderstood: Please read the article on the Regression Fallacy by Gerard E. Dallal.

Effect of Outliers

- Outliers can have a substantial effect on the (estimated) prediction equation
- In the murder rate vs. poverty rate example, DC is an outlier
- Prediction equation with DC:

  y-hat = -10.13 + 1.32 x
- Prediction equation without DC:

  y-hat = -0.86 + 0.58 x

Effect of Outliers

- Removing the outlier would cause a large change in the results
- Observations whose removal causes substantial changes in the prediction equation, are called ***influential***
- It may be better not to use one single prediction equation for the 50 states and DC
- In reporting the results, it has to be noted that the outlier DC has been removed
- Correlation and Regression Applet

Prediction

- The prediction equation y-*hat* = $b_0 + b_1 x$ is used for predictions about the response variable *y* for different values of the explanatory variable *x*
- For example, based on the poverty rate, the predicted murder rate for Arizona is

$b_0 + b_1 x = -0.8567 + 0.5842 \times 20 = 10.83$

Dependent Variable	Predicted Value	Residual	
10.2	10.8281	-0.6281	*(Arizona)*
6.6	11.0618	-4.4618	*(Kentucky)*

Residuals

- The difference between observed and predicted values of the response variable (*y* – *y-hat*) is called a ***residual***
- The residual is negative when the observed value is smaller than the predicted value
- The smaller the absolute value of the residual, the better is the prediction
- The sum of all residuals is zero

Scatterplot

- Is linear regression/correlation always appropriate for two quantitative variables?
- How to decide whether a linear function may be used?
- *Always* ***plot the data*** *first*
- Recall: A ***scatterplot*** is a plot of the values (*x*,*y*) of the two variables
- Each subject is represented by a point in the plot
- If the plot reveals a non-linear relation, then linear regression is not appropriate, and the (Pearson) correlation coefficient is not informative