Relationships Between Two Variables: Regression and Prediction

- **Example 1: Could we predict or explain a state's Democratic vote % in the 2008 election based on its Democratic vote % in the 2004 election?**
- **Scatterplot shows a positive linear association between 2004 Democratic vote percentage and 2008 Democratic vote percentage.**
- **Example 2: Long term study of families measures two variables for each family: the father's cholesterol level at age 50, and the son's cholesterol level at age 50.**
- **Could you use the observed relationship between the two variables to predict a young man's cholesterol level at age 50, based on his father's age-50 level?**

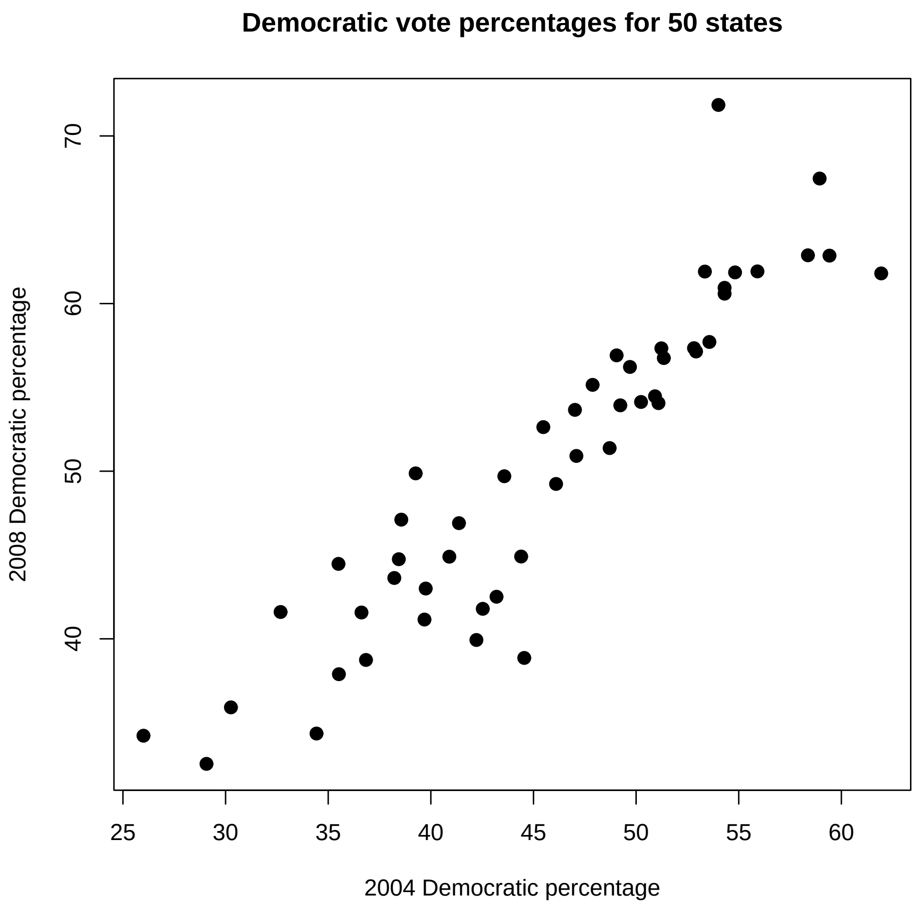

Figure 1: Scatterplot: 2004 Democratic percentage on horizontal axis, 2008 Democratic percentage on vertical axis.

Regression Lines

- ***Regression Analysis*** **is a statistical procedure that describes the relationship between two variables with a mathematical function.**
- **In regression, one variable is called the *explanatory variable* (denoted X) and the other is the *response variable* (denoted Y).**
- **In *linear regression*, we use a *straight line* to approximate the relationship between Y and X.**
- **Example 1: For a state that voted 45% Democratic in 2004, what is the predicted Democratic percentage in 2008?**

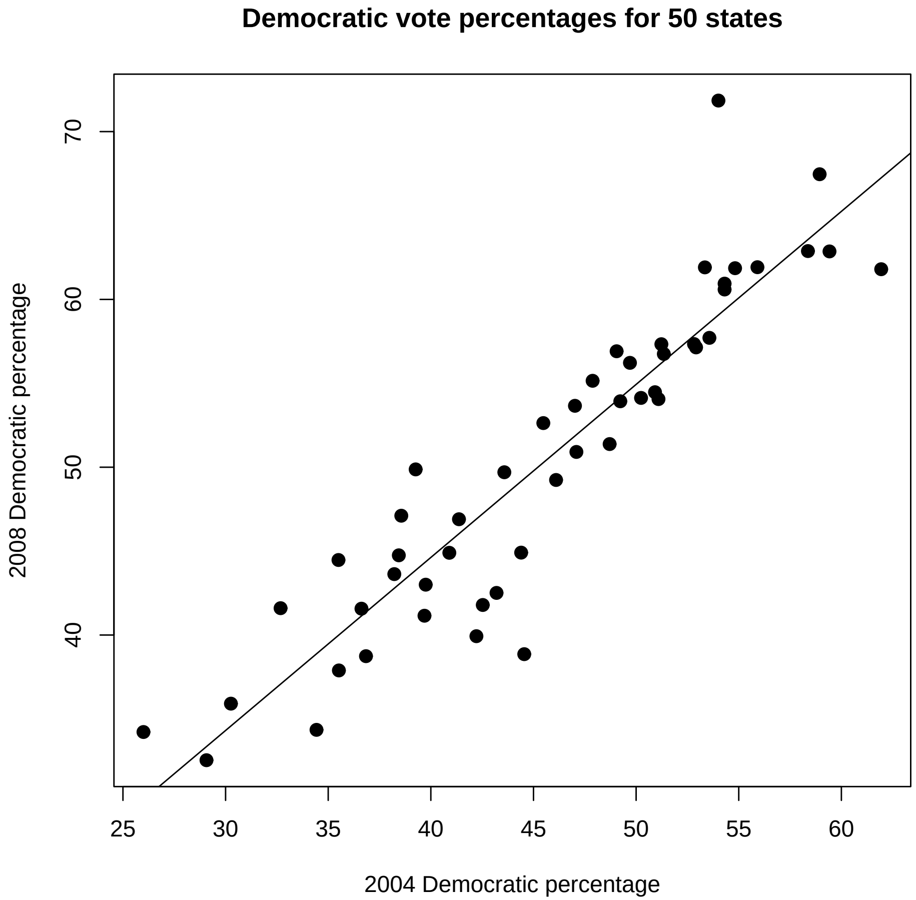

Figure 2: Scatterplot: 2004 Democratic percentage on horizontal axis, 2008 Democratic percentage on vertical axis, with regression line included.

Least Squares Regression Line

- **Suppose we have a scatterplot showing a data set with two variables measured on each observation.**
- **If the variables appear linearly associated, we could draw a line through the points to approximate the relationship between the variables.**
- **How do we figure out exactly *which* line would be best?**
- ***Least Squares Method:* Pick the line that makes the squared vertical distances from the data points to the line add up to the smallest number possible (see Fig. 15.3 picture).**
- **Statistical software can give us the equation of the *least squares regression line*.**

Using the Regression Line Equation

- **The equation of the regression line for the Democratic vote data is:**
  $Y = 3.377 + 1.031X$

- **In this equation, *Y* represents the response variable (2008 Democratic percentage, here) and *X* represents the explanatory variable (2004 Democratic percentage, here).**

- ***Note:* Sometimes statisticians use $\hat{Y}$ in the regression formula rather than *Y*, to emphasize that the regression equation gives a *predicted* *Y* value, not an observed *Y* value.**

- **The District of Columbia (not part of the data set) had 89.18% Democratic votes in 2004.**

- **What is the predicted 2008 Democratic percentage for DC?**

Using the Regression Line Equation

- **The equation of the regression line for the Democratic vote data is: $Y = 3.377 + 1.031X$**
- **The District of Columbia (not part of the data set) had 89.18% Democratic votes in 2004.**
- **What is the predicted 2008 Democratic percentage for the District of Columbia?**
- **Predicted 2008 Democratic % for DC is $3.377 + 1.031 \times 89.18 = 95.32$.**
- **In fact, the true 2008 Democratic % for DC was 92.46% (fairly close to predicted value).**

Clicker Quiz 1

Consider the Democratic vote percentage regression line. If State A had a 2004 Democratic percentage of 45%, and State B had a 2004 Democratic percentage of 50%, which is true?

A. State A will have a higher predicted 2008 Democratic percentage than State B will.

B. State A will have a lower predicted 2008 Democratic percentage than State B will.

C. State A and State B will have an equal predicted 2008 Democratic percentage.

D. It is impossible to compare the predicted 2008 Democratic percentages of State A and State B.

More about the Regression Line Equation

- **Recall the equation of the regression line for the Democratic vote data is: $Y = 3.377 + 1.031X$**
- **The first number (3.377 here) is called the *intercept* of the regression line.**
- **The number that is multiplied by X (1.031 here) is called the *slope* of the regression line.**
- **The intercept represents the predicted Y-value when $X = 0$ (if an X-value of 0 makes sense in the data set!)**

Still More about the Regression Line Equation

- **The slope is the *rate of change*: How much will the predicted *Y* change when *X* increases by one unit?**
- **The slope is *positive* when the two variables have a positive linear association.**
- **The slope is *negative* when the two variables have a negative linear association.**

Clicker Quiz 2

Consider the following regression equation where *X* = age-50 LDL cholesterol level of the father and *Y* = age-50 LDL cholesterol level of the son: $Y = 4 + 1.1X$. For a father with age-50 LDL level of 100, what is the predicted age-50 level of his son?

A. 110

B. 106

C. 15

D. 114

Warnings about Prediction

- **The *linear regression model* assumes that the relationship between the two variables is roughly *linear*.**
- **If a scatterplot of the two variables shows a *curved* association, a different form of regression model should be used.**
- **Predictions are more *precise* when the association between the two variables is *strong* rather than *weak*.**
- **Beware of *extrapolation*! It is risky to predict *Y* for an *X*-value that is much smaller or larger than the *X*-values of your sample observations. (recall DC Democratic vote prediction!)**
- **Linear trend seen in sample data may not be true for much smaller or larger *X* values. (Example: Congressional Budget predictions)**

Other Facts about Regression

- **The least-squares regression line may be greatly affected by outliers (see precipitation example).**
- **The square of the correlation (denoted r^2) is the proportion of variation in the *Y* values that may be explained by the linear association between *Y* and *X*.**
- **The r^2 for the Democratic vote regression is about 0.85.**
- **So about 85% of the variability in the states' 2008 Democratic percentages may be explained by their linear association with the 2004 Democratic percentages.**

Clicker Quiz 3

Recall the r^2 for the Democratic vote regression is about 0.85. What is the correlation coefficient between 2008 Democratic vote percentage and 2004 Democratic vote percentage in this sample?

A. $-\sqrt{0.85}$ **= -0.92**

B. -0.85

C. $\sqrt{0.85}$ **= 0.92**

D. 0.85

Clicker Quiz 4

A study has shown the r^2 for a regression of SAT score and college GPA is about 0.27. What is a correct conclusion?

A. About 27% of students who take the SAT go to college.

B. About 27% of the variation in college GPA may be explained by its linear association with SAT score.

C. The correlation between SAT score and college GPA is 0.27.

D. There is a 27% chance that there is a relationship between SAT score and college GPA.

Causation

- **A strong relationship between two variables does not mean that changing one variable will cause changes in the other.**
- ***Lurking variables* may account for the association between the two variables (television & life expectancy example)**
- ***Example:* Study of obesity in 9- to 12-year-old girls measured each girl's body mass index (BMI), along with mother's BMI, and other variables such as physical activity level, diet, television.**
- **Strongest correlation was between girl's BMI and mother's BMI (r = 0.506)**
- **Was heredity the main cause of girls' weights?**

Causation Example (continued)

- ***Example:*** **Study of obesity in 9- to 12-year-old girls measured each girl's BMI, along with other variables.**
- **Strongest correlation was between girl's BMI and mother's BMI (r = 0.506)**
- **Was heredity the main cause of girls' weights?**
- **Effect of heredity could be *confounded* with the effect of environment.**
- **The environment the family lives in (exercise, diet, TV habits) may affect both the girl's and the mother's BMI.**
- **Also, the example the mother sets may have as much of an effect on the girl's BMI as the mother's genetic background does.**

More about Causation

- The relationship we see between two variables may be due to (1) direct causation, (2) common response, or (3) confounding (see Figure 15.5)
- Without a well-designed experiment, it is difficult to determine which of these is the precise reason for the association.
- We can still use the relationship for prediction, even if we can't establish direct causation between the two variables.

When can We Conclude Causation from an Observational Study?

When all of the following are true:

- **When the association is *strong* and *consistent*.**
- **When extreme values of the “cause variable” are associated with extreme values of the “effect variable.”**
- **When the alleged cause *precedes* the effect chronologically, and when the alleged cause is *plausible*.**