Linear Regression

Kin 304W

Week 7: February 19, 2013

Think about the courses you have taken and work experiences you have had. What types of prediction equations have you seen or used before?

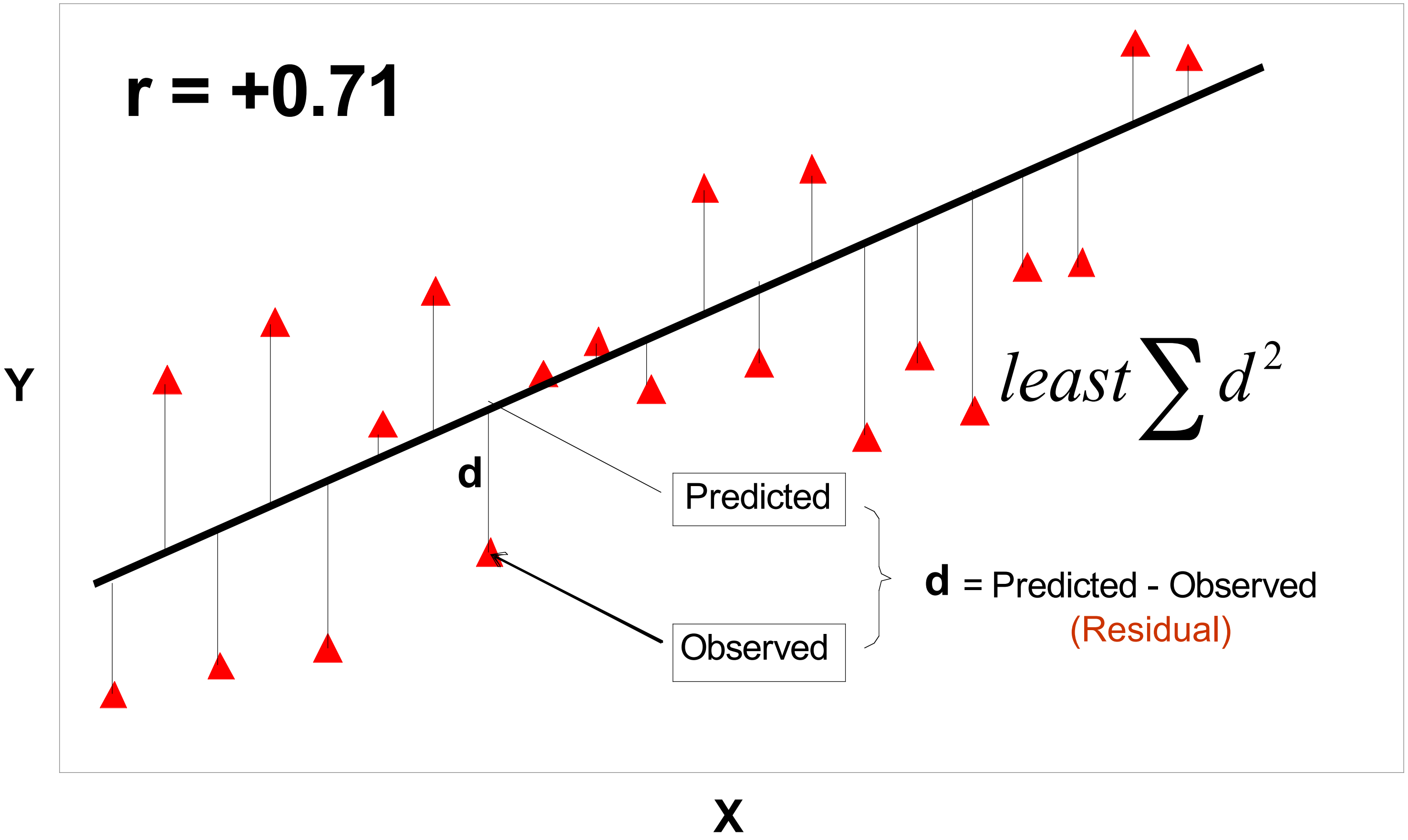
Linear Regression Allows Us To Do Prediction

- Research questions are sometimes framed as, “can we predict one variable from another?”
- Linear regression analysis
  - Fits a line, with a specific equation, to data
  - Software searches for the best fitting line
- **Y = mX + c**

m = slope (“coefficient)

c = intercept (“constant”)

Least Sum of Squares Curve Fitting



Output from Linear Regression

- Correlation Coefficient (*r*)
  - how well the line fits
- Standard Error of Estimate (S.E.E.)
  - how well the line predicts

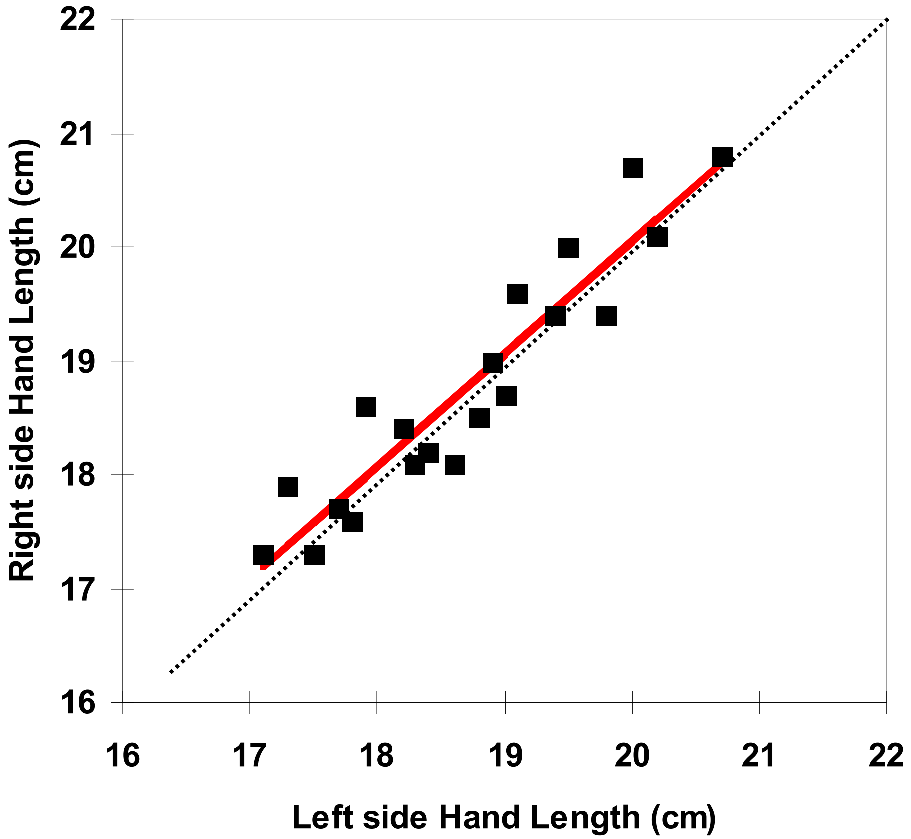
Standard Error of Estimate

- Measure of how well the equation predicts Y
- Has units of Y
- 68.26% of time (~2/3 times) the true score is within plus or minus 1 S.E.E. of predicted score
- 95% of time the true score is within plus or minus 1.96 S.E.E.s
- The S.E.E. is the standard deviation of the normal distribution of residuals

Linear Regression Example 1

Equation: Right Hand Length = 0.99 x Left Hand Length + 0.25

r = 0.94 SEE = 0.38 cm



Linear Regression Example 2

- Resting metabolic rate (RMR) is the body’s resting energy expenditure.
- Participants have to be fasting and rest for 30 minutes immediately before measurement. Then RMR is measured via indirect calorimetry with a gas exchange hood for 40 minutes. Expensive + time-consuming.
- Fortunately, there are prediction equations for RMR:
  - RMR for men (kcal/day) = (13.75 x WT in kg) + (5 x HT in cm) - (6.76 x AGE in years) + 66
  - RMR for women (kcal/day) = (9.6 x WT in kg) + (1.8 x HT in cm) - (4.7 x AGE in years) + 655

Harris-Benedict Equations

How Good Is My Equation?

- Regression equations are sample specific.

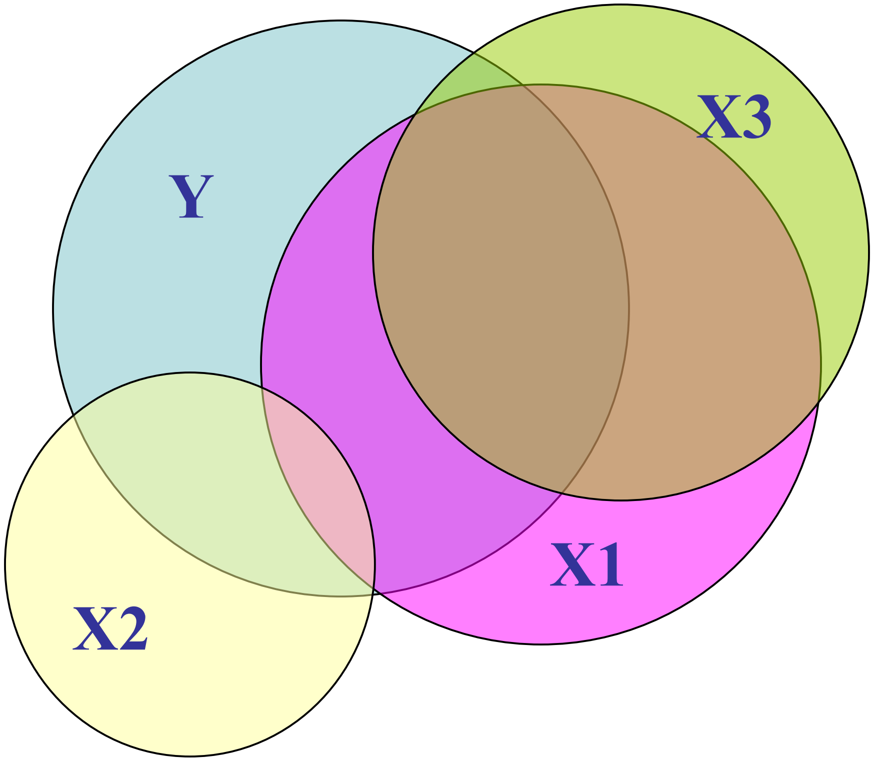
- Cross-validation studies
  - Develop your regression equation using one sample.
  - Test your equation on a different sample. Is the S.E.E. different?

- Split sample studies
  - Take a 50% random sample and develop your equation then test it on the other 50% of the sample.

Multiple Linear Regression

- Simple linear regression has one independent variable
- Multiple linear regression, more than one independent variable
- $Y = m_1X_1 + m_2X_2 + m_3X_3 \ldots\ldots + c$
- We call the m values "coefficients" rather than slopes.
- We call the c value a "constant" or "intercept."
- Same meaning for *r*, and S.E.E., just more measures are used to predict Y
- There are different ways to build a multiple regression equation:
  - Enter method (SPSS default) - force all variables into one equation
  - Stepwise regression - variables are entered into the equation based upon their relative importance

Building a Multiple Regression Equation



Goal is to explain the most variance in Y.

X1 has the highest correlation with Y; therefore it would be the first variable included in the equation.

X3 has a higher correlation with Y than X2.

However, X2 would be a better choice than X3 to include in an equation with X1 to predict Y.

X2 has a low correlation with X1 and explains some of the variance that X1 does not.

Standardized Linear Regression

- In regular linear regression, the numerical value of m_n is dependent upon the size (units) of the independent variable
  - $Y = m_1X_1 + m_2X_2 + m_3X_3 \ldots\ldots + c$
  - If X_1 was height in meters and $m_1 = 1.0$, what would happen to m_1 if height was in centimeters?
- In standardized regression, variables are transformed into standard z scores (using internal norms) before regression analysis; therefore mean and standard deviation of all independent variables are 0 and 1 respectively.
- The numerical value of zm_n now represents the relative importance of that independent variable to the prediction
  - $Y = zm_1X_1 + zm_2X_2 + zm_3X_3 \ldots\ldots + c$