2. Linear Models

Regression

Some Concepts

- Linear regression describes a functional linear relationship between two variables
- Correlation gives a value as (0-1) as to how related two variables are
- Regression gives a line that best describes the relationship between two variables – suggests a hypothesis
- Lurking Variables are variables that are not among studied variables, but may still influence interpretation of results
- Regression is not always causal – significant regression =/= causal r'ship

Reading Equation of a Line

- Y = a + bX
  - a is the intercept of Y when X is o
  - b is the slop of the line

Calculate Equation of a Line

1. Start with graphical display of data
   - No point fitting model if r'ship does not look linear
2. Find the line of best fit

- Line of best fist aims to minimize the residuals (vertical distances between data points and the line)
- Calculate the average values of X and Y
  - Sum up all X values and divide by number of points
  - Sum up all Y values and divide by number of points

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

- Calculate the average slope
  - Subtract X average from X value of each point
  - Subtract Y average from Y value of each point
  - Multiply two values
  - Add output of each data point
  - Divide by sum of each X average subtracted from X value

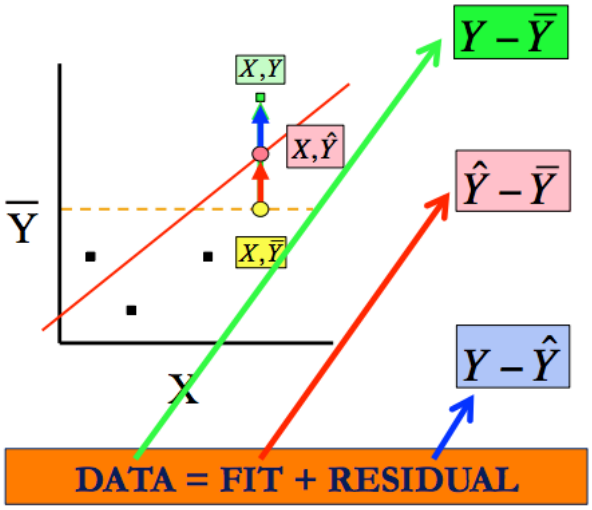
$$b = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})}$$

- Create Linear equation
  - Y is average Y values
  - X is average X values
  - b is average slope

$$a = \bar{Y} - b\bar{X}$$

Testing the Significance of Slope

- Testing the line is statistics describing line differ significantly to 0 (= no r'ship between X and Y)
- *Is the sample taken from a population where the average slope (β) is zero?*
- The further away from 0 the value of Y', the stronger the relationship
- Components
    - Total Variation (distance from line where slope = 0 to data point)
    - Fitted Variation - distance explained by regression line (distance from line where slope = 0 to fitted line)
    - Residual Variation – unexplained variation, error (distance from fitted line to data point)



- F Ratio is how much of the variation can be accounted for by the fitted regression model
    - IE: compares ratio of fitted and residual variation within total variation
    - Calculate MS regression
        - Subtract average Y value from each Y value on the line
        - Square each value
        - Add all values
        - Divide by number of degrees of freedom (always 1)

$$\text{MS regression} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{1}$$

    - Calculate MS residual
        - Subtract Y value on the line from each Y data point
        - Square each value
        - Add all values
        - Divide by number of degrees of freedom (n-2)

$$\text{MS residual} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$$

    - Calculate F Ratio
        - Divide MS regression by MS residual
        - F = 1 means fitted and residual variation is equal
        - F<1 means residual variation is LARGER than fitted variation (higher probability F is from population where variables have no relationship)
        - LARGER F means more variation is accounted for by model (higher chance that F explains the relationship within the population)

$$F_{1,n-2} = \frac{\text{MS regression}}{\text{MS residual}}$$

- R^2 Value is the proportion of the total variation explained by the regression
    - Sum of squares (fitted variation) divided by sum of squares (total variation)
    - If $r^2 = 1$, we can predict X and Y values perfectly
    - X can be predicted from Y by

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

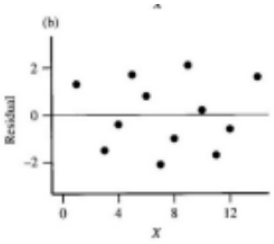
- BUT don't make predictions beyond the values of X

$$X_i = \frac{Y_i - a}{b}$$

Assumptions

Normality and homogeneity of variance

- Y values are assumed to be from a population that is normal and evenly distributed about the regression line
- ASSESS by checking residual plot – residuals should be spread evenly around line



residuals

Little/No Measurement Error

- Independent variable might be hard to measure accurately eg: temperature of violently erupting magma

Dependent Variable is Determined by the Independent Variable

Relationship Between X and Y is Linear

- ASSESS by checking scatter plot for departures
- Can transform variables to give more linear relationship
- NOTE when transforming…

$$Y = aX^b$$

$$\log(Y) = \log(aX^b)$$

$$\log(Y) = \log(a) + b\log(X)$$