STAT 500

Correlation and its Connection to
Simple Linear Regression

Population Correlation Coefficient

- Measure of linear relationship between two quantitative variables (X and Y) in population.

- Denoted as ρ.

- Defined as

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - E(X))(Y - E(Y))]}{\sigma_X \sigma_Y}$$

Properties of ρ

- $-1 \leq \rho \leq 1$
  - Perfect linear relationship $\rho = -1$ or $\rho = 1$
  - No linear relationship: $\rho = 0$
- Sign of ρ indicates direction of relationship
  - Negative linear relationship between X and Y: $\rho < 0$
  - Positive linear relationship between X and Y: $\rho > 0$
- Strength of relationship indicated by $|\rho|$

Properties of ρ

- ρ is invariant to the choice of scale for X and/or Y.
  - X = height, Y = weight
  - Same ρ whether X is measured in in. or cm.
  - Same ρ whether Y is measured in lbs. or kg.

Sample Correlation Coefficient

- Estimate ρ by taking a sample from population and calculating r.

$$r = \frac{1}{n-1}\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{S_X S_Y}\right)$$

- r has the same properties as ρ.

Hypothesis Test for ρ

To determine whether two variables X and Y have a linear relationship, we can also use r to conduct a hypothesis test for ρ: $H_0 : \rho = 0$ vs. $H_a : \rho \neq 0$.

- We can do a t-test for this null hypothesis.
- Note that b_1 is a function of r

$$b_1 = r\left(\frac{S_Y}{S_X}\right)$$

- b_1 and r have same sign
- Inference for β_1 and ρ produce same test statistic, distribution, p-value, decision, and conclusion.

Differences between Correlation and Slope

- Correlation
  - Focus is relationship between X and Y
  - Use when there is not a clear response variable
- Slope
  - Focus is explaining change in values of Y with x
  - Use when there is a clear response variable

r and R^2

- r is a function of R^2

$$r = \pm\sqrt{R^2} \qquad r^2 = R^2$$

- r is a numerical summary of the direction and strength of the linear relationship between X and Y.

- R^2 is a numerical summary of the percentage of variability in Y that can be explained by the linear regression with x.

STAT 500

Model Diagnostics for
Simple Linear Regression Models

SLR Model and Assumptions

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \text{ where } \epsilon_i \text{ i.i.d. } N(0, \sigma^2)$$

- Values of Y_i are independent (independent random errors).
- Values of x_i are fixed.
- $\mu_{Y|x_i}$ is a linear function of x_i
- Homogeneous error variance: $Var(\epsilon_i) = \sigma^2$
- Normally distributed errors: ϵ_i i.i.d. $N(0, \sigma^2)$

Independence

- Check independence of observations through details of data collection.
- Beware of
  - Observations over time
  - Clustering of observations
  - Spatial elements to observations
- Crucial assumption - must use other methods if violated.

Fixed Values of x

- Assume x is measured without error.
- Check through variable definition and through details of data collection.
- If violated, model the error in x using a random effect.

Linearity

- Scatterplot - Plot of Y_i versus x_i: linear pattern
- Residual Plot - Plot of residuals e_i versus x_i: no pattern
- Violations of linearity
  - Transform Y_i values so that relationship with x_i is linear.
  - Common transformations: log transformation and power transformation(Y^2, Y^3, $\sqrt{Y}$, etc.)
  - Conduct analysis with transformed Y values.
  - Undo transformation in drawing conclusions.

Normality and Homogeneous Variance Residuals

- Residuals are approximations for random errors:

$$e_i = Y_i - \widehat{Y}_i = Y_i - (b_o + b_1 x_i) \quad \text{for } i = 1, 2, ..., n$$

- Important properties of residuals
  - $\sum_i e_i = 0$
  - $\sum_i x_i e_i = \sum_i \widehat{Y}_i e_i = 0$
  - Residuals are negatively correlated

$$e_i \sim N\left(0, \sigma^2\left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right)\right)$$

Regression Analysis - Residuals

- Residuals do not have homogeneous variances

$$e_i \sim N\left(0, \sigma^2\left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right)\right)$$

- Sometimes use

$$r_i = e_i \Big/ \sqrt{MSE\left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right)}$$

(known as studentized residuals)

Residual Plots

- Plot residuals versus predicted values
  - detect nonconstant variance
    - Look for changes in variability around the horizontal line at 0.
    - Megaphone shaped pattern: variability of e_i increases or decreases as x_i increases.
  - detect nonlinearity
  - detect outliers

Residual Plots

- Plot residuals versus X
  - In simple linear regression, this is the same as above.
  - In multiple regression, it will be useful.
- Plot residuals versus other possible predictors (e.g., time)
  - Detect important lurking variable
- Plot residuals vs lagged residuals
  - Detect correlated errors
- Normal probability plot of residuals
  - detect nonnormality

Remedies for Model Violations

- Transformation of Y
- Adding/modifying predictors
- More sophisticated models and/or estimation procedures
  - Weighted least squares for nonhomogeneous variance
  - Time series models for correlated errors
  - Robust regression methods for nonnormality
- These will be described more fully under multiple regression

Case Diagnostics

- Leverage
- Outliers
- Influential Points

Case Diagnostics - Leverage

- Extreme values of x are called high leverage cases because they exert a large "pull" on SLR
- Measure of "potential" influence of observation on SLR
- Leverage of the ith observation is:

$$h_i = \left(\frac{1}{n-1}\right)\left(\frac{x_i - \bar{x}}{s_x}\right)^2 + \frac{1}{n}$$

- Properties of h_i
  - $1/n \leq h_i \leq 1$
  - $\sum_{i=1}^{n} h_i = 2 \qquad \bar{h} = 2/n$

Case Diagnostics - Leverage

- Often use $4/n$ or $6/n$ as a guide for determining large h_i
- In addition to an absolute cutoff, look for large h_i by examining the distribution of h_i values across observations

Case Diagnostics - Outliers

- Extreme Y_i value for a given x_i
- Three assessment methods
  - Residuals
  - Internally studentized residuals
  - Externally studentized residuals

Case Diagnostics - Residuals

- Residuals

$$e_i = Y_i - \widehat{Y}_i$$

- $\text{Var}(e_i) = \sigma^2(1 - h_i)$
- Observations with higher leverage will have residuals with smaller variability.

Case Diagnostics - Residuals

- Internally studentized residuals

$$r_i = \frac{e_i}{\sqrt{MSE(1-h_i)}}$$

- r_i will have mean zero and approximately equal variance
- Outliers will inflate MSE
- r_i is called STUDENT in SAS

Case Diagnostics - Residuals

- Externally studentized residuals

$$t_i = \frac{e_i}{\sqrt{MSE_{(-i)}(1 - h_i)}}$$

  where $\text{MSE}_{(-i)}$ is MSE without the ith observation

- t_i will have mean zero and approximately equal variance
- t_i is called RSTUDENT in SAS

Case Diagnostics - Outliers

Studentized residual values with absolute value

- Less than 2 are fine
- Between 2 and 3 indicate potential outliers
- Greater than 3 indicate outliers

Case Diagnostics - Outliers

- Outliers inflate value of $\hat{\sigma}^2$

- Will lower values of test statistics t and F

- Will inflate widths of confidence intervals for parameters and prediction intervals

Case Diagnostics - Influence

- Concerned about unusual cases that have a big influence on both:
  - $\widehat{Y}_i$ for some x_i
  - estimated slope $\widehat{\beta}_1$
- Could delete the case, refit model and examine the change

Case Diagnostics - Influence

- COOK'S D - effect of deleting the i-th case on the least squares regression model

$$D_i = \left(\frac{r_i^2}{2}\right)\left(\frac{h_i}{1-h_i}\right)$$

- D_i is large when r_i is large and h_i is large
- $D_i > 2 * \sqrt{2/n}$ indicates substantial influence