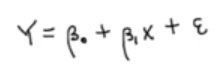
Final Review Sheet

# **Linear Regression**

* **Four assumptions of linear regression model**
  1. **Linear relationship between X and Y**



* B0 + B1 : Mean of Y for given X=x ---> The average value of Y for given X=x (value of interest)
  + - * B0 = intercept
      * B1 = slope
      * E : Error (individual effect) ---> part that can't be explained by X

1. **Conditionally Normal Distribution of Y at each given X=x**
2. **Homoscedasticity (equal variance assumption on Y at each given X=x**
3. **Conditional independence: for given X=x, Y's are independent of each other**

* **Model a linear relationship, not a causal effect**
  + If B1 = 0, then X and Y have no linear relationship
  + **If B1 != 0, then X and Y have a significant linear relationship (but does not imply causal effect)**
    - Ex. Ice Cream Sales vs. Canyon Lake Deaths
* **Exploratory analysis with scatter plot (visual) or Pearson/ Spearman correlation (quantitative)**

1. **Scatterplot** (Y vs. X) with:
   * + The **population correlation coefficient p (rho)** measures the **strength** (absolute value) and **direction** (sign) of the association between the variables
2. **Correlation** 
   * + Always between -1 and +1 (no unit; standardized measure)
     + p = 0 implies **no linear** relationship (may still have a relationship that is not linear)
       1. When p is close to -1 or +1, there is a strong linear relationship (**p measures linear relationship**)
     + Pearson Correlation --> ***sensitive*** to outliers
     + **Spearman** Correlation --> ***robust*** to outliers
       1. Still measures for a linear relationship and is robust to outliers in the data

* **Multicollinearity issue**

**Text

Description automatically generated**

* + **Detect multicollinearity among predictors through scatter plot matrix or VIF**

**Scatterplot**

**A picture containing diagram

Description automatically generated** Graphical user interface, text, application

Description automatically generated

**VIF**

* + - Remove each VIF one by one and re-run (remove the predictor with the largest VIF first and chip away until all VIFs are below 10)
  + **What happens if highly correlated predictors are in the model? Check simulation studies in Week10 materials**

Text

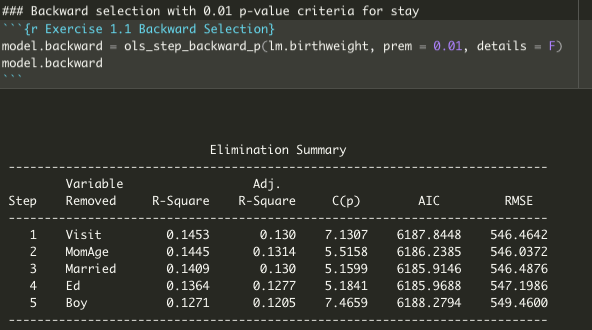
Description automatically generated Text

Description automatically generated

* **Model** **selection**
  + **Automatic selection (forward, backward or stepwise selection)**
    - **Automatic Selection** is based on the p-value (significance of terms)
      * **Backward Elimination**
      * **Forward Selection**
      * **Stepwise Selection**

**Text

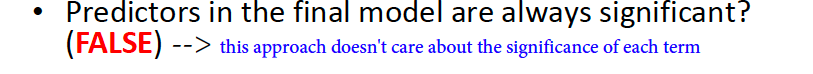
Description automatically generated**

****

**Text

Description automatically generated**

* + **Best subset approach**
    - **Balance between good prediction and simple model e.g., Adjusted R^2, AIC, BIC, Mallow’s Cp**

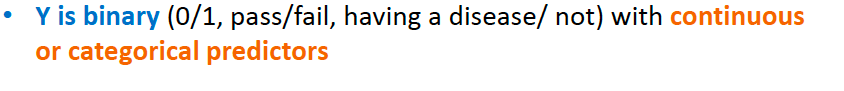
****

**Best Subset Approach** w/ goodness of fit criteria

* + - Simple model = less # of predictors (we want to achieve a good prediction with least number of predictors)
    - Uses:
      * **Adjusted R-Squared**
        + The bigger the better
      * **AIC , BIC, SBC**
        + The smaller the better
      * **Mallows Cp Criteria**
        + Cp <= P
* **Interpretation of R output**

* **Interpretation of R^2**
* **Determine significance of parameters and interpret the model**
* **Specify estimated regression line based on estimated parameters**
  + **When all predictors are continuous**
  + **When model includes categorical predictor**
  + **When interaction between categorical and continuous predictors is included**
* **Assess and remedy diagnostic issues. Guess which assumption is violated based on diagnostics**
  + **residual plots**
  + **histogram and QQ-plot**
  + **Cook’s distance**
* **Transformation of Y**

# Logistic Regression (binary logit)

* **Response variable is binary**
  + ****
* **Important to set an event of interest (decide what is our event of interest)**
* **Goal of logistic regression**
  + **Text

    Description automatically generated**
* **Odds and Odd Ratio**
* **Choose the best model (model selection)**
* **Specify estimated logistic regression line based on estimated parameters**
* **Assess diagnostic issues (e.g. residual plots, cook’s-d measure)**
* **Determine significance of parameters and interpret the model**
* **Interpretation of estimated coefficients with Odds Ratio (specifically)**
* **Guess what kind of observation would have the highest chance to have an event**
* **Howser-Lemeshow test (Goodness of fit test)**
* **Classification tool based on estimated probabilities**
  + **How we can classify each observation based on given cut-off**
  + **Example of cut-off’s: 0.5, sample proportion..**