Regression and Correlation methods

Learning Objectives

1. Describe the Linear Regression Model
2. State the Regression Modeling Steps
3. Explain Ordinary Least Squares
4. Compute Regression Coefficients
5. Understand and check model assumptions

What is a Math/Stats Model?

1. Often Describe Relationship between Variables

2. Types
   - Deterministic Models (no randomness)

   - Probabilistic Models (with randomness)

Deterministic Models

1. Hypothesize Exact Relationships
2. Suitable When Prediction Error is Negligible
3. Example: Body mass index (BMI) is measure of body fat based

- Metric Formula: $BMI = \dfrac{\text{Weight in Kilograms}}{(\text{Height in Meters})^2}$

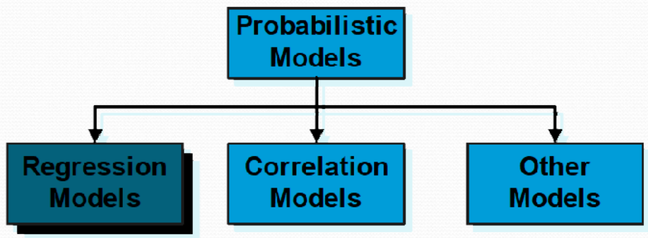
- Non-metric Formula: $BMI = \dfrac{\text{Weight (pounds)} \times 703}{(\text{Height in inches})^2}$

Probabilistic Models

1. Hypothesize 2 Components
   - Deterministic
   - Random Error
2. Example: Systolic blood pressure of newborns Is 6 Times the Age in days + Random Error
   - *SBP* = 6xage(d) + ε
   - Random Error May Be Due to Factors Other Than age in days (e.g. Birthweight)

Regression Models

Types of Probabilistic Models



Regression Models

- Relationship between one dependent variable and explanatory variable(s)
- Use equation to set up relationship
  - Numerical Dependent (Response) Variable
  - One or More Numerical or Categorical Independent (Explanatory) Variables
- Used Mainly for Prediction & Estimation

Regression Modeling Steps

- 1. Hypothesize Deterministic Component
  - Estimate Unknown Parameters
- 2. Specify Probability Distribution of Random Error Term
  - Estimate Standard Deviation of Error
- 3. Evaluate the fitted Model
- 4. Use Model for Prediction & Estimation

Model Specification

Specifying the deterministic component

- 1. Define the dependent variable and independent variable

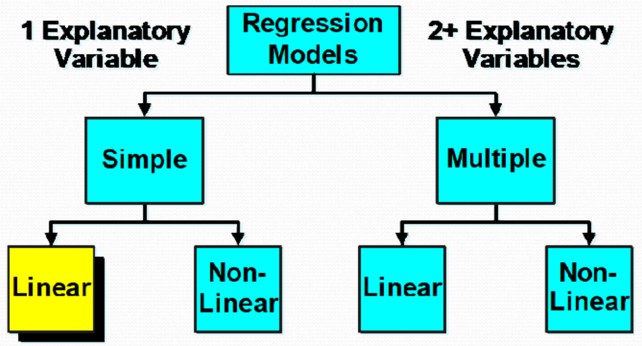
- 2. Hypothesize Nature of Relationship
  - Expected Effects (i.e., Coefficients' Signs)
  - Functional Form (Linear or Non-Linear)
  - Interactions

Model Specification is Based on Theory

- 1. Theory of Field (e.g., Epidemiology)
- 2. Mathematical Theory
- 3. Previous Research
- 4. 'Common Sense'

Types of Regression Models

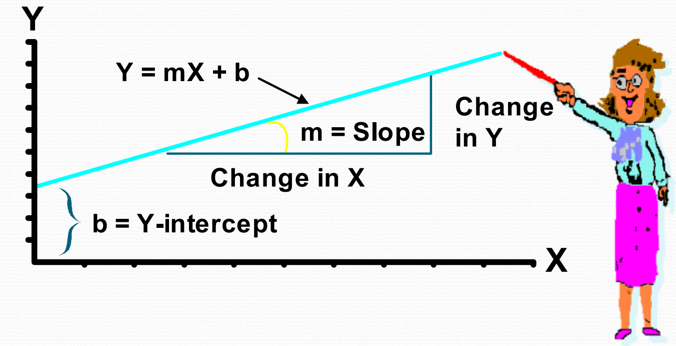
Types of Regression Models



Assumptions

- Normality of response variable

Linear Equations



Least Squares

- 'Best Fit' Means Difference Between Actual Y Values & Predicted Y Values is a Minimum. *But* Positive Differences Off-Set Negative ones. So square errors!

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

  - LS Minimizes the Sum of the Squared Differences (errors) (SSE)

Coefficient Equations

- Prediction equation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Sample slope

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

- Sample Y - intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Coefficient of determination

- To measure the strength of the linear relationship we use the coefficient of determination.

$$R^2 = \frac{\left[\sum (x_i - \bar{x})(y_i - \bar{y})\right]^2}{s_x^2 s_y^2}$$

$$or \quad R^2 = 1 - \frac{SSE}{\sum (y_i - \bar{y})^2}$$

Note that the coefficient of determination is r^2

Regression Diagnostics - I

- The three conditions required for the validity of the regression analysis are:
  - the error variable is normally distributed.
  - the error variance is constant for all values of x.
  - The errors are independent of each other.
- How can we diagnose violations of these conditions?

BASIS FOR COMPARISON	CORRELATION	REGRESSION
Meaning	Correlation is a statistical measure which determines co-relationship or association of two variables.	Regression describes how an independent variable is numerically related to the dependent variable.
Usage	To represent linear relationship between two variables.	To fit a best line and estimate one variable on the basis of another variable.
Dependent and Independent variables	No difference	Both variables are different.
Indicates	Correlation coefficient indicates the extent to which two variables move together.	Regression indicates the impact of a unit change in the known variable (x) on the estimated variable (y).
Objective	To find a numerical value expressing the relationship between variables.	To estimate values of random variable on the basis of the values of fixed variable.

Regularization

Ridge Regression

$$RSS_{\text{ridge}} = \sum_{i=1}^{n}\left(y_i - x_i^T\beta\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

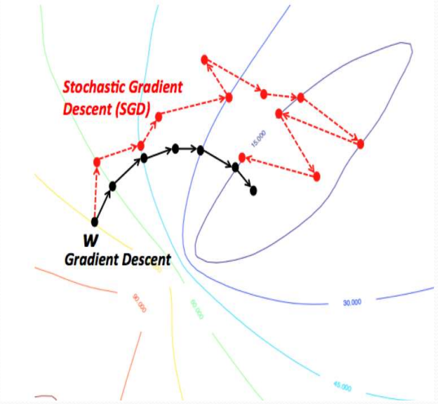
LASSO Regression

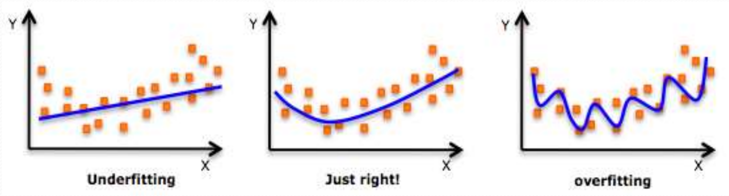
Least Absolute Shrinkage and Selection Operator (LASSO)

$$RSS_{\text{lasso}} = \sum_{i=1}^{n}\left(y_i - x_i^T\beta\right)^2 + \lambda\sum_{j=1}^{p}\beta_j$$

Parameter fine tuning

- Gradient descent
  - Batch gradient descent
  - Stochastic Gradient Descent (SGD)

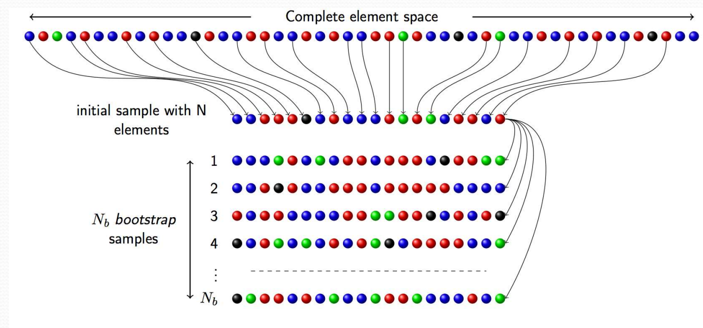


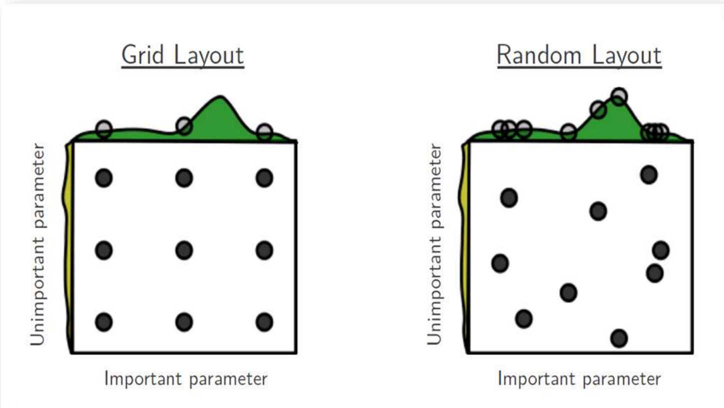


K-Fold Cross-Validation

- Primary method for estimating a tuning parameter (such as subset size)
- Divide the data into K roughly equal parts (typically K=5 or 10)

1	2	3	4	5
Train	Train	Test	Train	Train




Grid Layout
Random Layout
Unimportant parameter
Important parameter
Unimportant parameter
Important parameter