3 Types of Analytics

1. Descriptive - find human interpretable patterns that describe the data
    a. Large scale summaries, local patterns - finding fraud
2. Predictive - use variables to predict unknown or future values of other variables
    a. Regression, classification, ranking/recommendation
3. Prescriptive (may need causal reasoning)

Analysis is often retrospective (not prospective) data was not collected in a methodical way that is tailored for the analytical task

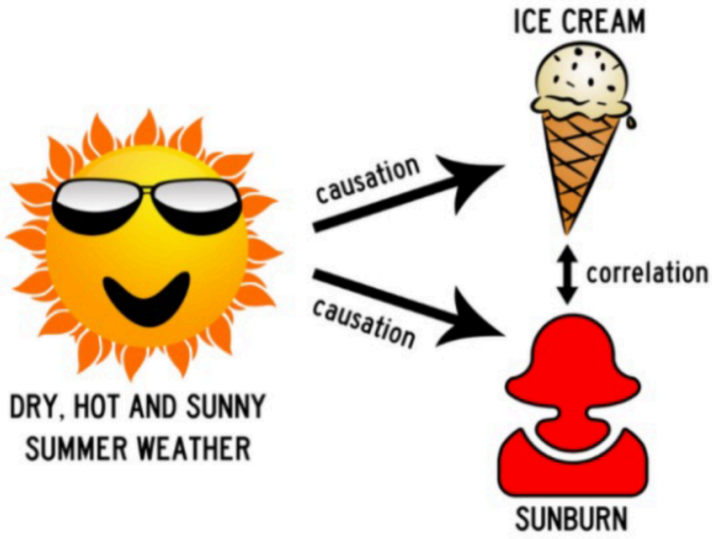
- Data may be collected in the past without a specific analytical task in mind, which can lead to challenges because the data was not gathered with a methodology to ensure relevance or completeness for the analytics to be done - more risk for biases that are not controlled

No Free Lunch - there is no universally best model, so tradeoffs need to be understood

- Deep nets and complex models are very general and powerful (few to no assumptions about the nature of relationships) but require lots of data, hyperparameter optimization, compute, little statistical or human insights, and the solution may not be robust

Sanity checks are crucial to make sure that model's assumptions are correct, preventing overfitting, ensuring integrity of the model

Confounders



- Here, the confounding variable is temperature: high temperatures cause people to both eat more ice cream and spend more time outdoors under the sun, resulting in more sunburns.
- Confounders are correlated with the independent variable (may be causation) and **causally** related to the dependent variable

- Confounders bias results and prevent you from seeing actual relationships

Bayes Rule

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)},$$

- Prior probability = p(Y)
    - Probability of case without information
- Posterior probability = p(Y|X)
    - Probability of case taking into account new information

Independence

- If p(X,Y) = p(x)*p(y) or p(Y|X) = p(Y)

Likelihood Function

- Maximum likelihood is an approach for determining the parameters in a probability distribution using an observed data set. Approach is to find **parameter values that maximize the likelihood function**
- We maximize the log of a function because it's more convenient - more mathematically simple and because the product of a large number of small probabilities can underflow the numerical precision of the computer but the sum of log probabilities wouldn't
- First maximize mean and then variance
- **Over a sufficient data set the mean should equal the true mean but the variance estimate will underestimate by a factor of (N-1)/N - this is due to us maximizing the mean and then the variance, creating bias**
    - Becomes less severe with higher N, but more severe with more parameters
- Negative log likelihood can be interpreted as a cost function similar to MSE for estimating the parameters of the distribution
- **Likelihood refers to when you are trying to find the optimal value for the mean or standard deviation for a distribution given a bunch of observed measurements**
    - **Probability is the likelihood of observing measurements given a distribution**

Marginal distribution

- Probability distribution of a subset from a larger set of variables
    - Represents probabilities of the subset while ignoring or summing over the probabilities of the other variables

Conditional Distribution

- Distribution of one variable give something true about the other variable
    - Usually in % - % of people who have a gpa Y **given** they are age 20

Covariance

- Statistical measure that indicates the degree to which 2 random variables change together, does not provide information on the strength of the relationship (not normalized like correlation)

Multiple Linear Regression - MLR

- Statistical format =

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

  - Coefficients are betas, dependent variable is Y, estimates are indicated by hats
  - Learned by Ordinary Least Squares OLS - hyperplane is fitted by minimizing the mean squared error MSE
- Alternative statistical format

$$Y|X_1 \ldots X_p \sim N(\beta_0 + \beta_1 X_1 \ldots + \beta_p X_p, \sigma^2)$$

  -
  - Meaning: Y|Dependents ~ Error terms
    - Y|Dependents: Conditional mean of Y is linear in the dependent variables
    - Error term:
      - Normally distributed
      - Independent from each other
      - Identically distributed
  - Minimizing MSE on the training data yields the maximum likelihood estimate (MLE) solution of the assumed generative model
- ML Notation (Generalized linear regression)
  - Generalized linear regression is a linear combination of basis functions

$$y(x, \mathbf{w}) = \sum_{i=0}^{M} w_i \phi_i(x) = \mathbf{w}^\top \boldsymbol{\phi}(x)$$

    - y(x,w) is the output (dependent)
    - W is the vector of weights of coefficients
    - $\phi_i(x)$ are the basis functions which transform the input x
    - M represents the number of basis functions
  - Parametric model - Assume a predetermined structure and a set of parameters(weights) to estimate using data
  - Steps for learning a parametric model:
    - Determine the functional form of the model (polynomials, neural networks, linear functions) which are the mathematical structure that describes how the inputs relate to the outputs
    - Learn the parameters (weights) using the training data. In linear models, parameters would be the coefficients/betas

    - **Special cases**
        - Linear regression - most basic case where the basis function $\phi_i(x)$ is simply the input variables themselves - no transformation
        - Polynomial regression with scalar x - in this case the input x has polynomial terms (x, x^2, x^3) so the MLR model uses $\phi_i(x) = x^i$ which allows the model to fit curves to the data
- Interpreting betas

$$\beta_j = \frac{\partial E[Y|X_1, \ldots, X_p]}{\partial X_j}$$

    - Beta is the partial derivative of the expected value of Y with respect to the predictor X_j, holding all other predictors constant - B_j is how much Y is expected to change for a one-unit change in X_j all other variables fixed

Collinearity

- Collinearity is a situation where 2 or more predictor variables in a multiple regression model are highly correlated
- Mostly effects interpretability, does not affect MSE or R^2
    - When there is collinearity the predictors are not independent, which causes the coefficients to be less interpretable - changes in one variable are associated with the other which makes it hard to isolate the effect of individual predictors
- Increases uncertainty in coefficient estimates - small changes in the data can then lead to large fluctuations in the estimated values of the coefficients
- Dummy variables/one hot encoding
    - Must always do 1 less than the total amount
    - If you do create dummies for each category then you have **perfect multicollinearity** - sum of all dummies will always equal 1 for each observation. This **linear dependency makes the model matrix singular**, therefore the regression algorithm cannot invert the matrix to estimate the parameters.

Ordinary Least Squares (OLS)

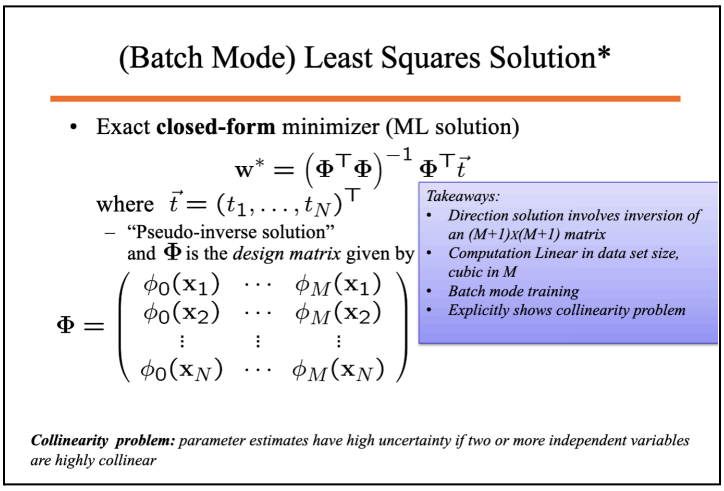
- Method used to estimate the parameters/weights of a regression model by minimizing the sum of squared errors (SSE)

- Loss function E(w) is the loss function OLS tries to minimize

$$E(w) = \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{w}^T\phi(x_n) - t_n\right)^2$$

    - ½ is for mathematical simplicity because you take the derivative and the squared cancels it out
- Minimizing MSE/SSE on the training data yields the Maximum likelihood estimate solution
    - Under the assumptions that the errors are normally distributed with constant variance

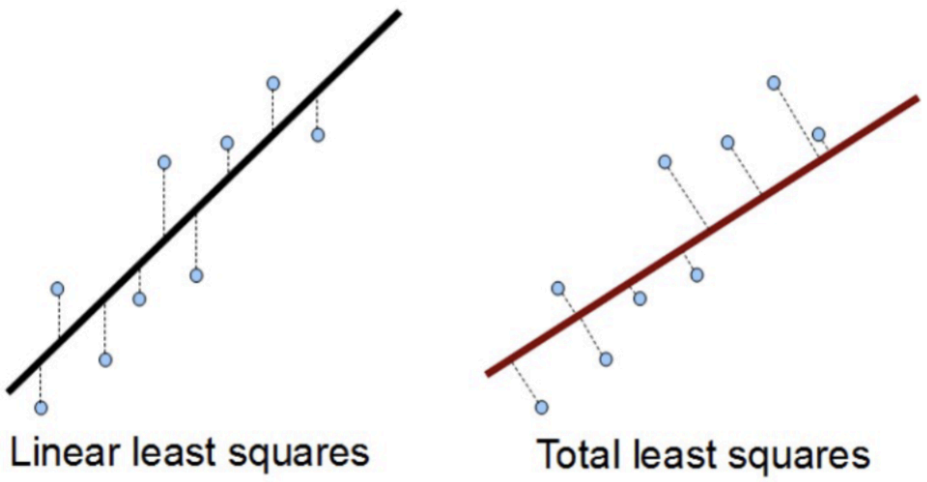
(Batch Mode) Least Squares Solution



- Method for finding the optimal parameters (w) in a linear regression by using an **exact closed form solution**
- Uses matrix inversion on an (M+1) x (M+1) matrix where M is the number of basis functions (features). $\Phi_0 x = 1$ which corresponds to the intercepts, where w_0 is the coefficient for the intercept
    - Can become computationally demanding as M increases
- Computation is linear in the size of the dataset (N) but cubic in the number of features (M) - more manageable to scale with more data points than with more features
- Method is **batch mode** which means it process the entire dataset at once to calculate the solution which is efficient for small or medium datasets but may not be for larger ones
- Collinearity problem is explicit because collinear variables makes the matrix difficult to invert, leading to high uncertainties in the parameter estimates

Linear Least Squares vs Total Least Squares

Total Least Squares (Aside)



Which one is better?
Which one should you choose?

- Linear least squares is used by OLS and minimizes the vertical distances between the observed data and regression line - only accounts for errors in the dependent variable and assumes the predictors are measured without error
    - Best for when error or noise is only found in the dependent variable
- Total least squares minimizes perpendicular distances from the data points to the regression line, which accounts for errors in both dependent and independent variables
    - Best for when there are measurement errors in both independent and dependent

Model Complexity and Overfitting

- High complexity (more features) increases risk of overfitting, high N (more data points) reduces risk of overfitting with more complex models
- Regularization - impose penalties on less desirable solutions

$$\text{Cost} = \text{MSE} + \lambda \text{ Penalty (f)}$$

    - Lambda is alpha in SKlearn
    - Regularization penalty (Penalty(f)) is a functional (maps the function of the regularization onto a number) which quantifies how complex or undesirable the solution is
    - Intercept is not penalized

    - Variables should be standardized so all features are on the same scale/magnitude otherwise features with large scale would dominate the penalty term
- Popular penalties
    - Ridge Regression - penalty is **sum of squared weights**
        - Discourages large weight values and shrinks them toward zero (never exactly zero) - also known as shrinkage or weight decay. Useful when we want to keep all features
        - Ridge regression has an exact closed form solution
    - Lasso Regression - adds penalty on the **sum of the absolute values of the weights**.
        - Can push some weights to exactly zero which performs feature selection
    - Elastic Net - combines both ridge and lasso penalties by shrinking weights and setting some weights to zero
    - Number of non-zero weights - leads weights to be driven to zero
    - Smoothness of function - penalize models that are not smooth (punishes I'm fluctuation)

Kaggle data set issue

- Test set may be different if you sample the data again, meaning other models may perform better on other test sets - want to find how good the data would be on an infinite test set (true generalization error)

Adjusted R^2

- Adjusts for the difference between your training data and true data - so as the amount of training data increases, the difference can be expected to shrink