Introduction to Statistical Learning

1 Research / Elevator Speech

- **Background and Research**
  - I started my PhD in statistics and I have worked on high-dimensional regression and covariance estimation with noisy and missing data, topic modeling and low rank estimation for matrix- and tensor-variate data, and distributed spectral clustering algorithms. The range of projects I lead at the University of Michigan forced me to cope with challenges that are both theoretical and practical in nature.

    I have mostly used Python for running numerical experiments for my methods and writing up a software for my collaborators at the Cell and Development Biology department to trace neurons in Brainbow images. The joint work with the Brainbow group allowed me to accurately communicate my findings, constantly discuss the solutions to tracing problems, and update the deliverables for collaboration. This dynamic research environment naturally translated to the experience during my military service at the Republic of Korea army headquarters.

    There were few problems that I lead at the ROKA headquarters, and the most successful project was developing an automatic security system which required me to constantly work with Bigrandier and Major generals, soldiers who are actively in surveillance duty, and policymakers at the headquarters. The communication with a wide range of groups was essential for our project as we need to meet the requirements to deploy our system, convince the generals of the benefits of automatic security systems, and ensure the practicality of actually deploying our project.

    I believe the combination of the communication and technical skills that I gained during my PhD studies and the military service is a natural fit to being a scientist at Uber, as there are various business models at the company. The creativity and the persistence to solve unique problems to provide reliable transportation system in various fields excites me as a scientist and I hope to Data Science team at Uber and positively contribute to solving Problems at Uber.

- **Why Uber?**
  - Uber mission: Reliable Transportation, everywhere for everyone.
  - Fitness for the role
  - **Uber Works**
  - Reliable transportation is a really ambitious mission to accomplish, and it is already a technical feat that Uber provides a consistent riding experience. As a user, as much as I love the ride experience, I made a good usage of Uber eats app to surprise my friends and loved ones in US while I was in Korea for my military service. The whole experience of getting the right recommendation of food and reliable delivery system was more than enough to connect with my loved ones while I was in another country.

    Besides, the variety of problems arising from the open market place at Uber makes Uber one of the most exciting place to work. The problems of providing incentives for drivers, recommendation systems, and driver / rider matching system all have a common goal of maintaining a healthy marketplace. With the success of Uber eats, Uber is now confronted with additional dimension for improving the marketplace. After I digged deeper on why Uber was working the problems they work on, the intersection of the projects at Uber motivates me and I would be thrilled to work as part of the team.

- **Data Science**: The process of analyzing data to achieve actionable insights. Reliable Transportation. My job would be to determine correct data sets and variables and collect structured and unstructured data from varying sources to find useful insights and create opportunities.

- **Graph Neural Network**

Case Study

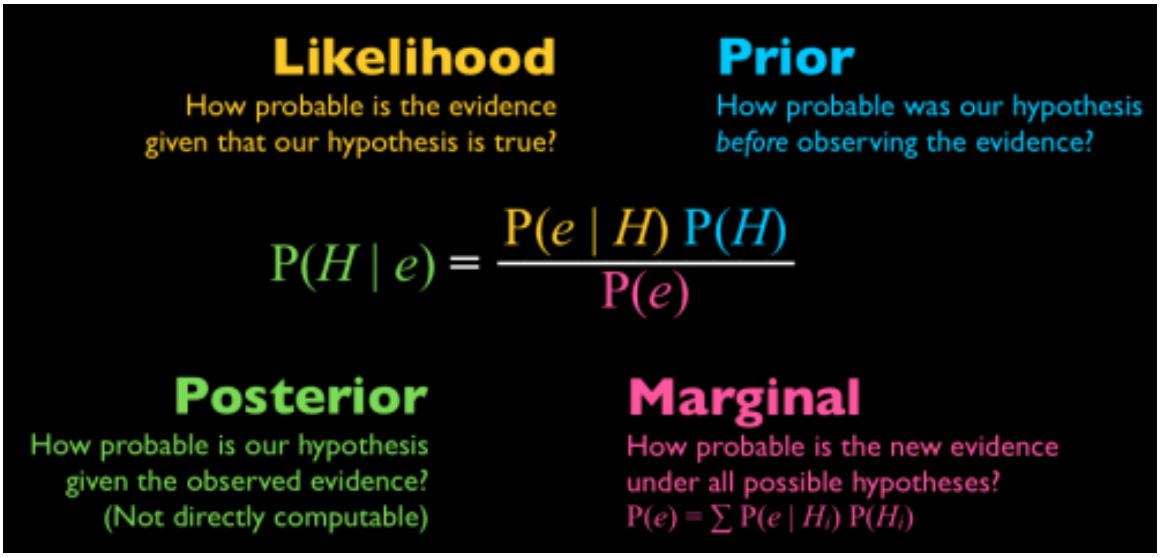
Diagnose a problem with a **target metric**. **Think about marketplace and how the company earns avenue from their business models**.

1. Clarify the scenario and the metric
   - Check outliers and set a quantile threshold
   - If extreme values, investigate values. Data collected correctly?
   - Look at other algorithms or problems that is involved with the question.
2. How does the company work in business perspective vs. user experience?
3. Look at the time factor
   - Sudden change vs progressive change
4. Check other products by the same company
   - If they experience the same change
   - Bigger problem
5. Segment users by dimensions to isolate the issue
6. Decompose the metric
7. Summarize the overall approacy
   - Systematic approach
     - Structure
     - Comprehensiveness
     - Feasibility
   - Forecasting, Dispatch, Experimentation

1. **Goal**
2. **Data**
3. **Computational Constraint**
4. **Formulaized**
5. **Method Consider**
6. **Model Evaluation**

2 Sample Questions

- Data
  - **Steps for Data wrangling and cleaning**
    Discover, structuring, cleaning, enriching, validating
- **Neural Network**
  - Why neural network and why is it a blooming field?
  - Explain deep learning model to customers.
- **Probability**
  - Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event. A good exmaple would be if the risk of developing health problems is known to increase with age, Bayes theorem allows the risk to an individual of a known age to be assessed more accurately than simply assuming that the individual is typical of the population as a whole



  - **Likelihood**: How probable is the evidence given the hypothesis
  - **Prior**: How probable was hypothesis before observing the evidence
  - **Posterior**: How probable is hypothesis given the observed evidence
  - **Priori (Marginal)**: How probable is the new evidence under all hypothesis

- **Central Limit Theorem**: If a large sample is taken from any distribution, regardless of the distribution is discrete or continuous, the sample means will be approximately normal distribution. The mean of the sample means is μ, the standard deviation of the sample means is $\frac{\sigma}{\sqrt{n}}$.
  - **Standard deviation**: $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$
  - **Standard Error**: Statistical inference based on sampling distribution. Accuracy of a sample mean by measuring sample-to-sample variability of the sample means.

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

- **Statistical Power**
  - The probability that a test will correctly reject a false null hypothesis. Statistical power has relevance only when the null is false.
  - Power = 1 - Type II Error
  - P(True Positive) = 1 - P(False Negative)
- **Classification Measure**
  - Precision = (True Positive)/(Total Predicted Positive) = 1 - FDR, where FDR = TP/(Predicted Positive).
    - Good when the costs of False Positive is High
  - Recall = (True Positive)/(Total Actual Positive) = 1 - FNR, where FNR = FN/P
    - When high cost associated with False Negative
  - **F1-score** = (Precision * Recall) / (Precision + Recall)
    - Seek balance between precision and recall
    - Uneven class distribution
  - **ROC**: TPR vs FPR.
    - TPR = Sensitivity or Recall
    - FPR = 1 - TNR
- **P-value**
  - **P-value** = P(Type I Error) is the probability of observed results of a test assuming that the null hypothesis is correct.

- **Overfitting**
  - Use dropout, ℓ_1, ℓ_2 regularization. Reduce complexity, CV
- Lasso shrinks to exactly zero as we reduce by constant factor, Ridge shrinks by proportion.
- **bias-variance**
  - high bias = underfitting, high variance = overfitting
- Logistic regression vs SVM
  - $n \gg 0$ use LR or SVM without a kernel
  - $n < p$, use SVM with Gaussian Kernel
  - $n \ll p$, create more features than LR or SVM without Kernel
- **Gradient Boost**
  - **Boosting**: Converting weak learners into strong learners. Effectively reduce the variance of the model

  Trains many models in a gradual, additive, and sequential manner.
- **When naive Bayes?**
  - Categories kept simple, but if the predictors are dependent, bad
  - Works well on small datasets
  - High bias and low variance.
- **Random Forest vs Boosting**
  - Random forest put together based on randomized covariates to reduce bias in a parallel manner.
  - Boosting sequential learning where try to learn weakness of the previous modely by fitting the residuals
  - Random forest out of the box, but Boosting needs a tuning.

3 Coding

3.1 Algorithm and Complexity

1. Sorting algorithms
2. Complexity for the code

4 Statistics

Look up Data Science questions

- A/B Testing
- Causal Inference
- FDR/FNR / Imbalanced Data / Precision / Recall
- Distributions

1. Casella-Berger 510 / 511 problems
   - Bayes Rule
   - One- or two-sided confidence interval / hypothesis testing / likelihood ratio problem / Neyman-Pearson
   - Binomial variable / Multinomial Distribution
   - Imbalanced data
2. Introduction to Statistical Learning: Every Algorithm practice summarizing in 30 seconds, 2 advantages and 2 disadvantages

- Goodness of fit
- Model evaluation: ROC
- Random Forest model and explanation
- Detecting multi-colinearity assumptions
- Assumption of GLM
- Breadth. Classification models / Regression / Decision Trees / Random Forest / Gradient Boosting

3. Regression

5 Case Study

Guide them through how you are considering variables

- **Driver / Rider**
- **Geographic / Time / Weather**
- **Trips / Delivery**
- **Rating**

Precision and **Recall**

5.1 Key Points

1. **Goal / Deliverables**: Clarify and simply the question with a goal / deliverable in mind.
   - **Context**: User experience / Increase Revenue / Long- vs Short-term goal
   - **Options**: Tackle two to three problems and offer the options.
2. **Brainstorm**: Initial few minutes to organize thought and categorize three things. 1) Retrospective analysis, 2) Modeling, 3) Evaluation
   - **Exploratory**: Demographic information. Usage frequency. Do feature selection
     (a) Correlation between response and features
     (b) Dimension reduction visualization to see linear vs non-linear
     (c) **Features**: Work hours, average hr per week (past two months), driver longevity, idle vs busy time, city vs suburb. Earnings average. Rating of drivers to customer. Cancellation rate. Last trip they take, multiple look back for these values
   - **A/B Testing**
     (a) **Design**: How long to run an AB Testing? We need to determine the sample size. $Samplesize = \frac{16\hat{\sigma}^2}{\delta^2}$, where δ is difference between treatment and control. $\rightarrow$ Since we don't know δ, we use minimum detectable effect.
       – Type II error or Power
       – Significance level
       – Minimum detectable effect
     (b) If we are doing multiple testing, we need to use lower significance level. Otherwise, FDR increases!
       – **Bonferroni Correction**: Significance level / number of tests $\rightarrow$ too conservative
       – **Control FDR**:
     (c) **Novelty and Primacy Effect**: Also known as welcome the changes vs reluctant to change. Larger or small initial effects will be due to these.
       – Run tests only on first time users
       – Compare first users vs old users in the treatment group.
     (d) **Interference**: Split users randomly and users are independent
       – Assumption fails if **Social Network**(facebook, linkedin, etc) or **two-sided markets** (Uber, Lyft, Airbnb)
       – **Network Effect**: User behaviors are impacted by others and spillover the control group. The difference underestimates the treatment effect.
         * **Network clusters**: Assign clusters randomly as people interact mostly within the cluster.
         * **Ego-network randomization**: A closter composed of an ego and her alters.
       – **Two-sided markets**: Resources shared among control and treatment groups.
         * **Geo-based randomization**: Split by geolocations (NYC vs SF). Big variance since markets are unique
         * **Time-based randomization**: Split by day of week and assign all users to either treatment or control. Only when treatment effect is in short time. $\rightarrow$ *Not for long time effect.* Ex) Surge-time pricing algorithm is short-term and time based randomization is useful.
   - **Modeling**
   - **Evaluation**
   - **Summarize**

5.2 Basic Data Science Question

1. **How to deal with outliers?**
   - Use a more complex model. The outlier may come from model mismatch → ex) If the residual is increasing constantly, we can do Poisson regression.
   - Normalize the data to shift the extreme values closer
   - Adopt algorithms not affected by outliers, ex) random forest
2. **Collaborative filtering**

   (a)

5.3 AB Testing

1. **Choose and characterize metrics for sanity check and evaluation**
   - **Inavriant metrics**: Not supposed to be affected by the experiments.
   - **Evaluation metrics**: Daily active users to measure user egagement, click through rate to measure button design.
2. **Design of Experiment and define methodology**: Clearly state hypotheses
   - PICOT (Population, Intervention, Comparison, Outcome, Time)
     - **Population** under study (visitors, drivers, riders)
     - **Intervention** different services or incentives
     - **Comparison** group
     - **Outcome** what we will measure
     - **Time** when you will measure it
   - **Randomization Strategy**: Randomly assign to a Treatment/Control group. → Avoid **Confounding bias** by correct unit of randomization.
     - Randomly assign visitors to Layout A or B homepage
     - Avoid counfounding factors using **propensity score**. Potential confounders are used to build a statistical model that assigns each person a propensity score. People with high scores are more likely to have certain confounders.
   - **Practical significance boundary**: Revenue increase \$2/user could be significant enough to move on. Power of the test(0.8) and significance level(0.05) to calculate sample size.
3. **Result Measurement / Analysis**
   - **Metric**: Set before the experiments start to help you understand what kind of changes your experiment is causing.
     - **Key metric** to judge the experiment and make business decisions
     - Ex) Impression count, click-through rate, buttom hover time, time spent on page, bounce rate
     - **Monitoring metrics** to measure negative effects and existing metrics to estimate impact.
   - Metrics should pickup changes you care about and not the one you don't care.
   - **A/A** tests to compare people in the control group and check if the metric picks up difference between the two. If varies a lot, the metric of interest might be too sensitive to use in experiment. → use historical data.
   - **Exposure and duration**
     - **What's the size of the eligible population?**
     - **Imact on the user experience and business**
   - **Power Analysis** to determine the smallest sample size for detecting the effect of a given test.
4. **Product Insights / Run Experiment**
   - **Duration**: What's the best time to run it?
   - **Exposure**: What fraction of traffic to expose?
   - **Learning effect**: User behavior becomes stable over time. Run on a smaller group of users for a longer period of time.
5. **Analyze**

- **Sanity Check**: Check if invariant metrics have changed. If failed, 1) retrospective analysis or 2) look into learning effect
- **One metric not significant**: Break down into different platforms, day of the week. The change affects the new user and experienced user differently.
- **Multiple metrics at the same**: Bootstrap and run experiment again and again. Bonferroni correction.

6. **Launch Decisions Platform Team**
   - Proceed once practical AND statistical significance is observed.

5.4 Key-Metric

- **Uber Eats**:
  - **Total Sales**: On-premise sales + delivery sales + pickup sales
  - **Goal**: Order food from user and connect customers and restaurants. Provide delivery service
  - **Revenue**: Short-term (increase revenue), Long-term (Market share)
  - Uber Eats earns from commission fee and advertisement fee.
  - **Acquisition**: Customer visits websites or install app. Sign-up for account.
    - number of website visits / app downloads
    - App openings - number of customers looking for food
    - number of new accounts
  - **Activation**: Find a restaurant, find a dish, order food, receive food, give feedback
    - Number of restaurants, variety of restaurants, operation hour, time for finding restaurant. Finding before ordering, number of dishes per restaurant. Avg price per dish. Order volume. Avg distance / delivery time / fee
  - **Retention**: Revisiting customers

What geospatial distributions on a map makes the most sense for them?

- **Supply**: Number of drivers and total driving time
- **Wait Time**: Driver acceptance rate (accepted requests by drivers)/(total offers to drivers)
  - **Network Value**: Evaluate earning based on the current requested trip and future potential earning opportunities.
  - (Number of requested trips)/(number of available drivers) → Historical and real time status.
- **User Events (Marketing / Growth)**
  - **Installs**: Where are people installing the app? Serviceable in that area?
  - **Searches**: Where are users searching for a ride? Do we have drivers in those hotspots? How far are they?
  - **Bookings**: Where do bookings take place? How long after the user searches for a ride? What is the frequency?
  - **Trips**: Where do trips take place due to promotions? where do discounted trips happen?
  - **Cancellations**: Where are trips canceled by the user? Do they cancel because of the ETA?
  - **Churn**: Where do people search but don't book?Which step do they drop-off? Is it preference of cars or price?
  - **Cohorts**: Repeated customers? Valuable customers?
- **Driver Metrics**:
  - **Total Supply**: What is the total number of drivers / off-duty or full-time drivers? Total driving time?
  - **Busy / Idle**: How many are busy or idle?
  - **Time**: Where do drivers spend the most amount of idle time? How far are they from the next demand hotspot?
  - **Cancellations**: where do they cancel? Location? Time?
  - **Earnings per Hour / Incentives**: Where and when do drivers make the most amount of earnings or incentives?
- **Strategy / Growth**:

  - **Trip Start / End**: Pickup and destination? Are they repeatable? How long was the trip? Where do power users start the trip from?
  - **Origin-Dest Pair**: What are the top O-D pair in a city? What are the most profitable?
  - **Paths**: What are the most common paths that repeatable users take?
  - **Idle Spots**: What are the most common spots where drivers are idle in trip?
- **Business**:
  - **Revenue**: How is revenue distributed in a city? What are the lowest revenue but the highest driver cost areas?
  - **Driver Utilization**: How is driver utilization spend across city and how doe shtat change with different times of the day?
  - **Requests per hour**: How many bookings requests do we get per hour in different areas?
  - **Conversion Rate**: How's the conversion rate of the requests in terms of rider acceptance? Where and when do riders don't accept bookings?
  - **Completion Rate**: What's the completion rate of the trips started?

5.5 Uber Eats Metric

- **Short-term**: Increase revenue
- **Long-term**: Makret share
- **Acquisition**: Customer visit websites, download apps, sign-up for an account
- **Activation**: Find restaurants, find a dish, order food, receive food, give feedback
- **Retention**:

6 Regression

Prediction vs Interpretability

- **Restrictive**: Lasso and Linear Model
- **Middle**: Generalized additive models
- **Flexible**: bagging, boosting, support vector machines, neural networks

In some cases, a less flexible method could lead to more accurate predictions. *rightarrow* Potentially overfitting in flexible methods.

6.1 Assessing Model Accuracy

- **Quality of Fit**: Measure how well the model's predictions actually match the observed data. $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$.
  - Interested in the accuracy of the predictions on test data.
  - **No Test Data**: No guarantee the lowest training MSE leads to lowest test MSE. Try splitting the training data.
  - As model flexibility increases, training MSE decrease but not necessarily the same for test MSE. i.e. overfitting.

$$MSE = \mathbb{E}(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon) \tag{1}$$

  - **Variance**: The amount by which $\hat{f}$ would change if we estimated it using a different training data set. $\rightarrow$ More general models have high variance.
  - **Bias**: Error introduced by approximating a real-life problem.
  - **Tradeoff**: Bias increases dramatically for flexible models, and vairance increases more for restrictive models.
- **Classification**: For classification tasks, we look at the error rate $\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$.
  - **Bayes Classifier**: Assigns each observation to the most likely class gitven its predictor values. i.e.

$$\hat{y}_0 = \arg\max_j P(Y = j | X = x_0) \tag{2}$$

  - The Bayes classifier produces the lowest possible test error rate, namely *Bayes error rate.*
  - **K-Nearest Neighbors**: The Bayes classifier serves as an unattainable gold standard. KNN estimate the probability by looking at the neighbors, i.e.

$$P(Y = j | X = x_0) = \frac{1}{K}\sum_{i \in N_0} I(y_i = j) \tag{3}$$

  - KNN is a classic example of low bias high variance model.
- **Linear Regression**: For the simple linear regression problem $Y = \beta_0 + \beta_1 X$, we have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{4}$$

  - How accurate is the sample mean? We compute the standard error of $\hat{\mu}$ as following

$$Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma^2}{n} \tag{5}$$

    where σ is the standard deviation of each of the realization y_i.
  - Residual standard error (RSE) and the R^s statistics shows the quality of a linear regression.

$$RSE = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

- R^2 provides the proportion of variance explained.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{7}$$

- where $TSS = \sum(y_i - \overline{y})^2$ and $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. R^2 **measures the proportion of variability in Y that can be explained using X**.
  - R^2 near 0 means either 1) the linear model is wrong, or 2) the error variance σ^2 is too high.
  - The value of R^2 could be low in some fields. Applications in biology, etc. rough approximation to the data, as the residual errors due to other unmeasured factors.
  - The correlation

$$Cor(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{8}$$

  - In the simple regression setting, $R^2 = r^2$. **Correlation quantifies the association between a single pair of variables rather than a large number of variables.**
  - **Adjusted** R^2: Adjusted based on the number of predictors in the model.The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected. Always lower than R^2

- **Multiple Linear Regression**: We can just think of it as a projection on col span of X and we have $\beta = (X^T X)^{-1} X y$.
  - The hypothesis testing of $H_0 = \beta_1 = \beta_2 = \cdots = \beta_p = 0$ to determine if there is any relationship between the response and predictors. We use $F - statistic$,

$$F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)} \tag{9}$$

    If H_a is true, expect $F > 1$. No relationship leads to F close to 1. We need larger F for small n.
  - If we only test the last q variables, $H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$,

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n - p - 1)} \tag{10}$$

  - 1) Check multi-colinearity, 2) Calculate VIF, 3) Look at the individual $t - statistics$.
    - Variational Inflation Factor (VIF) is used for multi coliearity. Let R_i^2 come from fitting X_i on other variables. Then $VIF = \frac{1}{1 - R_i^2}$.
  - Principal Component Regression, Regularization (Ridge or Lasso), or take the average between two. Partial Least Square.
  - AIC minimize KL-divergence. BIC prefers smaller models.
  - $Cor(Y, \hat{Y})^2$. The fitted model maximizes the correlation among all possible linear models.
  - Prediction intervals larger than confidence intervals as it also takes the irreducible errors into account.
  - Confidence interval for fitted value

- **Qualitative Predictors**: Dummy variable with indicator 0/1 or -1/1. Just write out the formula.

- If we include an interaction in a model, we should also include the main effects regardless of the significance of the coefficients. If you use the interaction term, you can have different slope and intercept.

- **Non-linear Relationships**: Polynomial regression is an easy way to incorporate this by transforming the covariates. The three assumptions of regression are

  1. Linearity
  2. Additivity
  3. Homogeneous and Uncorrelated error.

  - **Non-linearity of the data**: *Residual plots vs fitted values* for identifying non-linearity. If linear, should be no pattern.
  - **Correlation of Error Terms**: If there is correlation among error terms, the estimated standard errors underestimate the true standard errors. → narrower confidence interval. Occurs in the time series data. Plot residuals over time. If positively correlated, tracking in the residuals will appear.

- **Non-constant Variance of Error Terms**: The linear regression model assumes a constant variance. $Var(\epsilon_i) = \sigma^2$. *Heteroscedasticity* (non-constnt variance in errors) shape of the residual plot. $\rightarrow$ transform the response Y. log or $\sqrt{}$ shrinks the larger response and reduction in heteroscedasticity. $\rightarrow$ If we know the variance, we can do weighted least squares with wieght proportional to inverse vairances.
- **Outliers**: RSE is used to compute confidence intervals and p-values. So removing outlier is important. Residual plots can be used to identify outliers. $\rightarrow$ studentized residuals: dividing each individual e_i by estimated standard error. Greater than 3 possible outlier.
- **High Leverage Points**: Observation with *high leverage* have an unusual value for x_i. Removing high leverage data affects the regression more than removing the outlier.
- **Collinearity**: Bigger standard error for $\hat{beta}_j$ to grow. The power of the hypothesis testing (correctly detecting a non-zero coefficient) reduced by collinearity.
- **Multi-collinearity**: Compute the variance inflation factor (VIF). Smallest possible value for VIF is 1.

$$VIF(\hat{beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}} \tag{11}$$

  1) Drop problematic variables. 2) Combine the colinear variable into a single predictor.

- Regression Question to answer
  1. Is there a relationship between sales and advertising budget? $\rightarrow$ Hypothesis testing
  2. How strong is the relationship? $\rightarrow R^2$
  3. How accurately can we predict future observation? $\rightarrow$
  4. Is the relationship linear? $\rightarrow$ residual plots
- Parametric methods tend to outperform non-parametric approaches when there is a small number of observations per predictor.
- Compare to the KNN model to explain the parametric vs non-parametric.

7 Classification

Logistic regression, linear discriminant analysis, quadratic discriminant analysis, naive Bayes, KNN. Classification can be formulated with the Bayes theorem.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{12}$$

- Binary classification with linear regression is same as linear discriminant analysis. The reason not to use regression for binary classification
  1. Regression cannot accomodate beyond binary case
  2. Not provide meaningful estimates of $P(Y|X)$
- sensitivity:
- specificity:
- **Logistic Regression**: Model $P(X) = P(Y = 1|X)$ with logit function.

$$P(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1}} \tag{13}$$

  – Use maximum likelihood approach.

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{j:y_j=0} (1 - p(x_j)) \tag{14}$$

  – $\frac{P(X)}{1-P(X)}$ is called oods, and logistic regression models the *log odds* or *logit* as a linear model.

$$\log\left(\frac{P(X)}{1 - P(X)}\right) = \beta_0 + \beta_1 X. \tag{15}$$

- **Multinomial Logistic Regression**: For $k = 1, \ldots, K$

$$P(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}} \tag{16}$$

where we select Kth class for a baseline and

$$P(Y = K|X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}} \tag{17}$$

The interpretation should be careful as it depends on the baseline class K.

Alternatively, we can use *softmax coding*

$$P(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{\sum_{l=1}^{K} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}} \tag{18}$$

  - Substantial separation between two classes $\rightarrow$ logistic regression could be unstable.
  - If we impose assumption like gaussian and the sample size is small, other method is more accurate.

- **Generative Model for Classification**
  - Let π_k be the prior on the distribution of classes and $f_k(X) = P(X|Y = k)$. Then the Bayes theorem states that

$$P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \tag{19}$$

  - This is the posterior probability of an observation $X = x$ belongs to k, but estimating $f_k(x)$ ix challenging. The Bayesian method aims to use this posterior to estimate the Bayes classifier.

- **Linear Discriminant Analysis**: Assuming f_k is Gaussian, Linear discriminant analysis (LDA) approximates the Bayes classifier by plugging estimates of $\mu_k, \hat{\sigma}^s = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$ and the prior π_k. Share same covariance matrix Σ.
  - LDA classifier results from assuming the observations within each class come from a normal distribution with a common variance.
  - Bayes classifier works by looking at $\arg\max_k p_k(X)$. To improve specificity, the criterion should be

$$P(Y = 1|X = x) > 0.5 \rightarrow P(Y = 1|X = x) > 0.2$$

    .
  - Instead of just looking at the accuracy, we want to consider precision and recall to meet the customer's needs. ex) clinical trials, credit cars, etc.

- **ROC Curve**: True Positive Rate (Sensitivity), False Positive Rate (specificity). The performance is summarized by area under the curve (AUC).

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N*	

- **Quadratic Discriminant Analysis**: Same assumption as LDA, but each has its own covariance matrix Σ_k.
  - LDA over QDA if few training observation to reduce variance. But model mismatch leads to high bias for LDA. Compared to the linear boundary of LDA, QDA has a quadratic boundary.

- **Naive Bayes**: Assumes independence of the features. Priors π_k estimated by the proportion of data belogning to the kth class. LDA and QDA can work in higher dimension with Gaussian assumption.
  - Naive Bayes assumes that the p predictors are independent within kth class. Thus the conditional distribution $f_k(X) = \prod_i f_{ki}(x_i)$

- Estimating a p-dimensional density is rough (joint distribution). $Bias \uparrow, Varaince \downarrow$
  - Quantitative
    1. $X_j|Y = k \sim N(\mu_{jk}, \sigma^2_{jk}$. Seems like QDA, but we assume diagonal Σ_k.
    2. Non-parametrically estimate $f_{kj}(x_j)$ using histogram or *kernel density estimator* (or smooth histograms).
  - Qualitative: Just count the proportion.
- Works well for n is not large enough compared to p.
- Any classifier with a linear decision boundary is a special case of naive Bayes. ex) LDA is a special case of Naive Bayes. If we have diagonal Σ for LDA, naive bayes is a special case of LDA.
- Naive Bayes gives additive fit. QDA incorporates potential interactions.

- **K-nearest neighbors**: Non-parametric approach of estimating the posterior distribution based on the neighbors. Curse of dimensionality. $Bias \downarrow, Var \uparrow$.
  - Decision boundary non-linear and moderate n, p QDA may be preferred over KNN.

7.1 Comparison of Classification Methods

- Review p.172
- Linear Boundary: Logistic Regression, LDA (break on gaussian assumption)
- Moderately non-linear: QDA or naive Bayes (break on independence)
- Complicated non-linear: KNN

7.2 Generalized Linear Models

Linear regression estimate the normal distribution based on the linear relationship, and logistic is for binary variable.

- **Poisson Regression**: Poisson pdf $P(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!}$ is adopted to estimate the distribution over the data. $\rightarrow$ model counts.
  - Poisson regression models the parameter of Poisson λ as a linear function.

$$\log(\lambda(X_1, \ldots, X_p)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{20}$$

    Thus the log-likelihood becomes

$$\ell(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} \frac{e^{\lambda(x_i)}\lambda(x_i)^{y_i}}{y_i!} \tag{21}$$

  - $\lambda = E(Y) = Var(Y)$. Thus, the variance increases with more counts, but linear regression assumes constant variance.

- **Generalized Linear Models**: Given $X_1, \ldots, X_p$, Y belongs to a certain family of distributions. We map the linear function $\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ with link function η,

$$\eta(E(Y|X_1, \ldots, X_p)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{22}$$

  - We can generalize for exponential family.

8 Resampling

Repeatedly drawing sampes from a training set and refitting model to additional information about the fitted model.

- **Model assessment**
- **Model selection**

8.1 Cross-Validation

- **Validation Set**: Only divide the data once. $Bias, Overestimate \uparrow$
- **Leave-one-out**: Leave one observation oout. $Bias, Overestimate \downarrow$. No randomness in data split. Expensive as we train n models.
  - For linear or polynomial regression, use leverage for single fit

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1-h_i}\right)^2 \tag{23}$$

    where h_i is the leverage.
- **Cross-Validation**: Gives more estimates of the test error than LOOCV. The choice of k for k-fold leads to bias-variance trade off. $k = 1 \rightarrow Bias \downarrow, Var \uparrow$, but since we use the same data, we have high correlation among split data. The correlation depends on the overlap.
  - Mean of correlated quantities $\rightarrow Var \uparrow$.
  -

8.2 Bootstrap

- Estimate the standard errors of the coefficients for estimated models.

9 Linear Model Selection and Regularization

Linear models have advantage in inference.

- **Prediction Accuracy**: Given linear model, least squares provide low bias.
  - $n \gg p \rightarrow Var \downarrow$. $n \ll p \rightarrow Var \uparrow$.
  - Shrinkage and regularize $\rightarrow Var \downarrow$ with a slight increase in bias.
- **Interpretability**: Variable selection allows to only look at strong signals.

1. **Subset Selection**: Forward selection, backward selection, Hybrid.
2. **Shrinkage**
3. **Dimension Reduction**

We can perform the following criterion to compare models that estimate test errors better.

- AIC
- BIC
- C_p Mallow
- **Adjusted R-squared**: Once all of the correct variables, adding additional noise variables lead to only a small decrease in RSS.

With advance in computational methods, we can just use cross-validation.

9.1 Shrinkage Methods

- Lasso produces a simpler and more interpretable model.
- Variance of Ridge is slightly lower.
- The choice depends on the sparse subset of β.
- When OLS has high variance, Lasso could reduce overall MSE.
- Ridge regression shrinks the coefficients equally, but Lasso decreases by same constant.

- **Ridge Regression**: ℓ_2-norm regularization.
  - Advantage of least squares for bias-variance trade-off.
- **The Lasso**: Ridge regression include all p predictions in the final model as it only shrink all of the coefficients towards zero.
  - ℓ_1 penalty force the coefficient estimates to go to zero.
- **Elastic Net**:

9.2 Dimension Reduction

- **Principal Component Regressions**: Construct principal components and use them as predictors. Standardize first, otherwise high variance.
  - PCR suffers as there is *No guarantee that the directions that best explain predictors leads to the same for predicting response.*
- **Partial Least Squares**: PCR does unsupervised way, but PLS finds direction in a supervised way. Find direction that explain both the response and the predictors.

9.3 What goes wrong in high-dimension?

When $p > nn$, OLS becomes too flexible and overfits data. We can avoid overfitting with variable selection methods.
In high-dimensional setting, never use 1) sum of squared errors, 2) p-values, 3) R-squared, or other traditional methods. $\rightarrow$ report on an independent est set.

10 Moving beyond linearity

We can do a simple non-linear modeling as following

- **Polynomial Regression**: Transform the covariates
- **Step Functions**: Fitting a piecewise constant function for K distinct regions
- **Regression Splines**: Divide range of X into K distinct gregions. Fit polynomial functions at each region. Constraints at *knots*, which are region boundaries.
- **Smoothing splines**: Minimizing RSS with smoothness penalty.
- **Local regression**
- **Generalized additive models**:

11 Generalized Additive Model (GAM)

We generalized the linear model by

$$y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i \tag{24}$$

- Fit a non-linear f_j to each X_j. No need to try transformation manually.
- Accurate prediction for the response
- Examine the effect of each X_j individually
- Smoothness summarized by the degrees of freedom
- BUT restricted to additive and no interaction. We can manually introduce $f_{jk}(X_j, X_k)$.

Of course, we can use the same modeling for classification.

12 Tree

12.1 Regression Trees

1. Divide the predictor space into J distinct and non-overlapping regions.
2. Every observation in R_j make a same prediction.

- **Splitting**: Choose a predictor and cut point that results in lowest RSS. Given $R_1(j,s) = \{X|X_j < s\}$ $R_2(j,s) = \{X|X_j \geq s\}$, minimize

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \tag{25}$$

- **Tree Pruning**: Grow a very large tree T_0 then prune the subtree. Cost complexity by

$$\sum_{m=1}^{|T|} \sum_{x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha|T| \tag{26}$$

  - For classification we use one of the following
    - **Gini index**: $G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$ to measure the total variance across all classes. Measures the node purity.
    - **Entropy**

- **Plus and minus of trees**

  \+ Easy to explain, decision-making procedure, displayed graphically and handle qualitative predictors well.

  \- High variance

12.2 Ensemble Methods

- **Bagging (Bootstrap aggregation)**: Procedure to reduce the variance of a statistical learning method. Look at the bootstrated data and take an average.
- **Random Forest**: Randomly sample m subset from p predictors on bootstrapped data to allow predictors to play a role. Effectively *decorrelating* the models.
- **Boosting**: Trees are grown sequentially and fitting the residuals of the previous tree. **Fit on a modified version of the original dataset**.

13 Support Vector Machines

13.1 Maximal Margin Classifier

- **Hyperplane** is a flat affine subspace of dimension $p-1$ defined by

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = 0 \tag{27}$$

  Switching the equality to inequality shows the either side of the plane.

- If a separating hyperplane exists, it is a natural classifier.
- The natural choice is the **maximal margin hyperplane** $\rightarrow$ the separating hyperplane that is farthest from the training observations.

14 Support Vector Classifiers

- Greater robustness to individual observations
- Better classification for most of the training observations.
- The bias-variance trade off happens for the sum of slack variables $\sum_{i=1}^{n} \epsilon_i \leq C$. $C \downarrow \rightarrow Bias \downarrow, Var \uparrow$ corresponds to less tolerant of violations to the margin.
- Decision rule is based on small subset of data, *support vectors*, and robust to the observations far from the margin.

15 Support Vector Machine

- Allows for non-linear decision boundaries.
- Enlarge the feature space to incorporate nonlinear relationships
- *kernels* allow us to use non-linear relationship.

15.1 Multiple Class

- **One vs One**: Construct $\binom{K}{2}$ classifiers and tally the number of the classes.
- **One vs All**:Fit K SVMS comparing k vs the rest. Assign where the margin is the largest. In other words, the class assignment with the highest level of confidence.

15.2 Relationship to Logistic Regression

- When classes are well separated **SVM**
- When overlapping regime, **logistic regression**.

16 Unsupervised Learning

The goal is to discover interesting things about the data without the response.

- Unsupervised learning performed as part of an *exploratory data analysis*.

16.1 Principal Component Analysis

- When faced with a set of correlated variables, PCA summarizes with a smaller number of dimensions.

16.2 Missing Values and Matrix Completion

- Using PC to impute the missing values, i.e. *matrix completion.*
- Appropriate if the missingness is random. But user rating information like Netflix missing by necessity → *recommender systems*.
  - Use imputed values for missing entries
  - Low rank estimation of the data with inputed matrix X
  - Plug in the low rank estimated entry in the missing values
  - Evaluate the completion based on the observed data
- You can check the performance by looking at the eigenvectors of the true matrix and matrix completion.

16.3 Clustering

- **K-means clustering**
  1. Randomly assign select K points
  2. Assign label to the data
  3. Recalculate the mean
- **Hierarchical Clustering**: Bottom-up aggregation to have a hierarchical clustering.

		Truth	
		H_0	H_a
Decision	Reject H_0	Type I Error	Correct
	Do Not Reject H_0	Correct	Type II Error

17 Multiple Testing

Key to conducting inference. Hypothesis tests provide framework for *yes or no* questions about data. The **power** of the hypothesis test is defined by the probability of not making Type II error, i.e. **the probability of correctly rejecting** H_0.

- **Two-sample t-statistic**: Hypothesis testing for two mean witht he following

$$T = \frac{\hat{\mu}_1 - \hat{\mu}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad and \tag{28}$$

- **Bonferroni Method**: Let A_j denote the vent that we make a Type I error for the j th null hypothesis.

$$P(\text{falsely reject at least one null hypothesis}) = P(\cup_{j=1}^{m} A_j) \leq \sum_{j=1}^{m} P(A_j) \tag{29}$$

  Thus we set the significance level as $\frac{\alpha}{m}$. BUT it is quite conservative.

17.1 False Discovery Rate

The ratio false positives/total positive is *false discovery proportion.*
Benjamini-Hochberg procedure to control FDR.

17.2 Resampling method

- When no theoretical null distribution is available.
- Assumptions required for its validity do not hold.

18 Survival Analysis

Survival analysis and cnesored data $\rightarrow$ **time until an event occurs**.

19 Kaplan Meier Survival Curve

- **survival function** defined as $S(t) = P(T > t)$. The Kaplan-Meier estimator of the survival curve becomes

$$S(d_k) = P(T > d_k | T > d_{k-1}) \times \cdots \times P(T > d_2 | T > d_1) P(T > d_1)$$