I. Explanations and Evidence
   a. The whole point of statistic modeling is to ask the question: what can I say about the world in light of some evidence?
   b. Selection bias – possible way in which confounding can be present in an experiment
      i. Hospital example: those with hospital visits have worse health
         1. Possible interpretation is going to the hospital makes you sick
         2. More realistic interpretation is biased sample of people in hospital who were sick to begin with
      ii. Marijuana example: marijuana smokers perform worse on cognitive tests
         1. Marijuana makes you stupid
         2. Dumber people smoke marijuana to begin with
   c. Randomize and intervene
      i. Abstract ideal
      ii. Gold standard of placebo controlled, double blind trial
      iii. Good evidence through a perfected lens – does not occur regularly
   d. Endogeneity and confounding
      i. Confounding: some third variable can affect both the predictor and response variables
      ii. Exogenous: "outside the system" variable that is not caused by something in the system (in diagram, no arrow leads to an exogenous variable)
         1. Flip of coin → treatment vs. control → clinical outcome
         2. Flip of coin is exogenous, completely random
      iii. Endogenous: causal relationship where third variable can cause both predictor and response
         1. x→y z goes to both x and y
         2. x is no longer exogenous, z is exogenous
         3. Endogenous variables are caused by something outside the system
   e. Natural experiments vs. real experiments
      i. Natural experiments - Clean experiments that can protect from selection bias, costs, and ethical dilemmas intervening in the system
      ii. Israeli school system: no classroom can have more than 20 kids, so 41 vs. 40 kids in an age group determines class size
         1. Number of kids in age group is exogenous
         2. Avoids selection bias of small class sizes given to smarter/richer kids
      iii. Real experiments can be costly and difficult to administer through intervention
   f. Regression Models
      i. Statistical Models → general strategy to handling issues of selection bias, confounding, and inability to effectively create experiments or observe natural experiments
II. Exploring multivariate data
   a. Plots and summaries
      i. Histograms/standard deviations (one variable)
         1. Standard deviation is the average error
         2. Histogram allows visual estimation of the standard deviation
      ii. Box plots/dot plots (two variables)
         1. One grouping and one quantitative variable
      iii. Scatterplots (two variables)

1. Two quantitative variables

iv. Lattice plots (three variables)
1. One grouping and two quantitative variables
2. Ex. SAT Score vs. GPA vs. College

b. Simple grouping models
i. Essentially calculating the group means

c. Ordinary least squares regression models
i. Models decompose observed value and split into fitted and residual value
ii. Observed value = fitted value + residual

d. Fitted values/residuals
i. Takes observed value and splits into fitted and residual value
ii. Fitted value – part of observed value that CAN be predicted by the x variable
iii. Residual – part of observed value NOT predicted by the x variable

e. Plug in prediction (bronze level prediction)
i. Have past data and regression model $\hat{y}_i = \hat{B}_0 + \hat{B}_1 * x_i$
ii. Have new data x*
iii. Plug in prediction is $y^* = \hat{B}_0 + \hat{B}_1 * x^*$

f. Interpreting coefficients of regression model ($\hat{B}_0$ and $\hat{B}_1$), summarizing the trend
i. $\hat{y}_i = \hat{B}_0 + \hat{B}_1 * x_i$
ii. $\hat{B}_0$ is predicted value of response when predictor is 0
iii. $\hat{B}_1$ is predicted rate of change of response as predictor changes
1. Small is a shallow line, slowly changing y with x
2. Large is a steep line, rapidly changing y with x

g. Taking the x-ness out of y
i. First introduction to statistical adjustment (silver level adjustment)
ii. Food example
1. cheapest meal in Austin is simply the restaurant with the lowest price
2. cheapest meal in Austin accounting for food score is the restaurant with the lowest residual from the model plotting price vs. food rating (lowest price compared to what is expected)

iii. Strips the observed value of the part that is predicted by x, leaving the residual
1. Observed – fitted = residual
2. $y_i - \left[\hat{B}_0 + \hat{B}_1 * x_i\right] = e_i$

iv. Knowing information always reduces uncertainty
1. Residuals are more closely grouped together than y data
2. Leads to r^2 later on
3. Can quantify how much that data is worth by looking at the variance in the original values and the variance in the residual values

h. Transformations for Nonlinear models
i. Logs and power laws
1. Brain weight and body weight for mammals
a. X and y data and squished in bottom left corner
b. Taking logs of both x and y is an "un-squishing" operation
2. Fits a power law relationship between x and y
3. Creates a geometric, not linear, relationship between predictor and response

4. $\log(y_i) = \hat{B}_0 + \hat{B}_1 * \log(x_i) + e_i \rightarrow y_i = K x_i^{\hat{B}_1} * e^{e_i}$

ii. Polynomials and adding x terms of higher order to fit a more linear representation
1. Cars sold and months on the job example: fitting an 8th order polynomial overfits data and results in a complex model that is not useful for extrapolation
2. Simple models that explain the data may not fit the data as well, but simplicity and ability to generalize to bigger data ranges outweigh the cons
3. Models should be no more complex than they need to be
4. Highly predictive models - models that generalize well to future cases and tend to be the models that are no more complex than they need to be to explain the system

III. Predictable and unpredictable variation
a. Coverage intervals
i. Interval that contains x% of data is an x% coverage interval
ii. Defined generically for any set of numbers
iii. Important to understand confidence interval, which comes later
b. Naïve prediction intervals (silver level prediction)
i. Takes average error associated with past data into account in prediction
ii. Does not take into account the uncertainty associated with the predicted portion (b0 and b1)
iii. Quantify past error with standard deviation
1. $\hat{y}_i = \hat{B}_0 + \hat{B}_1 * x_i + e_i$
2. Have new x*
3. $y^* = \hat{B}_0 + \hat{B}_1 * x^{*} \; ^{+}/_{-} \; t * \sigma$
iv. Ways to quantify confidence of naïve prediction:
1. Empirical coverage: go and count past data that lie inside vs. outside the interval
2. Simple rule of thumb: t=1 → 65-70%, t=2 → 95%
c. R^2 and the decomposition of variance
i. TV = PV + UV
ii. Percentage of variation in y predicted by x
iii. Generalization of correlation concept that can be applied to multiple regression (r, the correlation, is not useful for multivariable prediction)

IV. Quantifying uncertainty (part 1: focus on parameter and prediction uncertainty)
a. Sampling distribution
i. Core concept of all statistical inference
ii. Impossible to understand anything at more than a superficial level without understanding the sampling distribution
b. Standard errors
i. Standard deviation of the sampling distribution
ii. Describes spread or average error of sampling distribution
iii. Useful for quantifying error in estimates and to construct confidence intervals
c. Confidence intervals – Questions like: How sure am I that my estimate of slope is getting at the true slope?
i. Simple definition: range of plausible values in light of the data
ii. Coverage interval of sampling distribution
iii. Mathematical definition: frequentist coverage property: “truth in advertising” or “assembly line” property

1. If you quote the 95% confidence interval every time, you will get it right 95% of the time

d. Theoretical ideals that take repeated samples
e. Bootstrapping
   i. Cannot take the repeated samples from the population that are the ideal
   ii. Fake multiple samples by taking repeated samples from particular data set (not population) with resampling, meaning that we get ties and omissions to replicate and simulate the variability we would get with resampling from the entire population
   iii. Assumes sample is broadly representative of population
f. Normal Linear Regression Model
   i. Linear assumption of $\hat{y}_i = \hat{B}_0 + \hat{B}_1 * x_i + e_i$
   ii. Added assumption of e_i~N(0,σ) about the data generation process
      1. Normality of residuals
      2. Constant variance of residuals across all values of predictor
      3. Independence of residuals
   iii. Data is different because of the residuals, so make assumptions about the way residuals are generated to estimate the error in the data
   iv. Normal Linear Regression Model assumptions lead to formulas for standard errors and confidence intervals
g. Cross-validation
   i. Purpose: estimate the prediction or generalization error of a statistical model
   ii. Interested in how well the model will generalize to future data sets – what is the future forecasting error likely to be?
   iii. Mechanics
      1. Train/test splits
      2. Fit the model
      3. Predict response values of predictor values in test split with model
      4. Evaluate outcome by comparing predicted responses with actual responses
      5. Loop over many splits and models to avoid sample bias

V. Grouping variables
   a. Aggregation paradox
      i. Come to wrong conclusion when inappropriately aggregate across groups
      ii. Grouping variable can be a confounder between y and x that leads to an aggregation paradox
   b. Dummy variables: expressed in baseline/offset form
      i. Indicator with value 0 or 1 (or another integer if more than 2 options)
   c. Interaction terms
      i. Change the slope of the model based on grouping variable

VI. Multiple regression
   a. Partial slope
      i. $\hat{y}_i = \hat{B}_0 + \hat{B}_1 * x_{1i} + \hat{B}_2 * x_{2i} + e_i$
      ii. $\hat{B}_1$ and $\hat{B}_2$ are partial slopes
   b. Statistical adjustment (gold level statistical adjustment)
      i. $\hat{B}_1$ and $\hat{B}_2$ split effects of x_1 and x_2 on y from other predictors
      ii. General strategy of what we do when we have confounders – put confounders into multiple regression equation

c. Collinearity
    i. Predictors are themselves correlated (they change together)
    ii. The effects of collinear variables make it difficult to isolate the effect of a single predictor

VII. Hypothesis testing
a. Neyman-Pearson testing: 6 steps
b. Permutation test
    i. Shuffling cards