**Statistical Modeling: A 5-Lecture Overview & Decision Framework**

**How to Choose Your Statistical Approach: A Decision Tree**

When faced with a new dataset and a research question, you can follow this conceptual flowchart to determine the best analytical strategy.

**1. Start with Your Response Variable (The Outcome):**

* **Is it Continuous?** (e.g., height, weight, protein expression level)
  + You are in the realm of **Regression**. Start by considering a **Linear Model** (lm).
  + *Proceed to Step 2.*
* **Is it Binary?** (e.g., survived/died, DM/NODM, case/control)
  + You need a **Classification** model. Your default choice is **Logistic Regression** (glm with family="binomial").
  + *Proceed to Step 2.*

**2. Examine Your Data Structure & Predictors:**

* **Are your observations independent?**
  + **Yes:** Standard models (lm, glm, gam) are appropriate.
    - **Is the number of predictors (p) much smaller than the number of samples (n)?**
      * *Go to the* ***Low-Dimensional Workflow*** *(Lecture 1, 2, 4).*
    - **Is the number of predictors large, potentially p > n?** (e.g., gene expression data)
      * *Go to the* ***High-Dimensional Workflow*** *(Lecture 3).*
  + **No, my data is grouped/nested/repeated:**
    - **Are observations nested in clusters?** (e.g., pups within a litter, students in a class)
      * Your observations are not independent. You need a **Multilevel Model for Clustered Data** (lme or lmer). This accounts for the correlation within clusters. *See Lecture 4 Summary.*
    - **Are there repeated measurements on the same subjects over time?** (Longitudinal data)
      * Your observations are not independent. You need a **Multilevel Model for Longitudinal Data** (lme or lmer). This allows you to model individual change trajectories. *See Lecture 3 Summary.*

**3. Check Your Assumptions & Potential Issues:**

* **Do I have missing data?**
  + **Yes:** Do not ignore it. Naive listwise deletion is usually wrong. You must investigate the missingness mechanism. If it's likely Missing At Random (MAR), you need a principled approach like **Multiple Imputation** or **Inverse Probability Weighting**. *See Lecture 5 Summary.*
* **For regression, is the relationship between predictors and the outcome linear?**
  + **Check this by plotting residuals or using a GAM.** If you see clear non-linear patterns, a standard linear model is insufficient. You should use a **Generalized Additive Model (GAM)** or a **Tree-Based Model** to capture these effects. *See Lecture 4 & 5 Summaries.*

**Lecture Summaries & Exam Question Insights**

**Lecture 1 & 2 (A. Abad): Linear & Logistic Regression, Model Selection**

**Core Idea:** These lectures provide the foundation for all statistical modeling. Linear regression models a continuous outcome, while logistic regression models a binary outcome.

**Decision Points & Key Questions:**

* **Linear vs. Logistic:** Your first choice. Is your Y variable continuous (lm) or binary (glm)?
* **Interpreting Coefficients:**
  + **Linear (lm):** A coefficient β₁ for predictor X₁ means: "A one-unit increase in X₁ is associated with a β₁ change in the average value of Y, holding all other predictors constant."
  + **Logistic (glm):** The raw coefficient is on the log-odds scale. You **must exponentiate it** (exp(coef)) to get an **Odds Ratio (OR)**. An OR of 2.5 means: "A one-unit increase in the predictor is associated with the *odds* of the outcome occurring increasing by a factor of 2.5."
* **Overall Model Significance (F-test in lm):**
  + The F-statistic and its p-value at the bottom of the summary(lm\_model) output test the null hypothesis that **all slope coefficients are simultaneously zero** (e.g., H0​:β1​=β2​=⋯=βp​=0). A small p-value (< 0.05) means your model as a whole is significant and explains more variance than a simple intercept-only model.
* **Model Comparison & Selection (anova() vs. AIC):**
  + **anova(model\_1, model\_2):** Use this for comparing two **nested** models (where one is a simplified version of the other). It performs a likelihood ratio test (or F-test for lm) to see if the extra parameters in the more complex model are statistically justified.
  + **AIC (Akaike Information Criterion):** Use this for comparing **any** set of models, even non-nested ones. It balances model fit with complexity. The model with the **lowest AIC** is considered the "best".
  + **Akaike Weights:** These give the probability that each model is the best in the set. If one model has a very high weight (e.g., >0.9), you can be confident in your choice. If several models have comparable weights, it indicates **model uncertainty**, and you should acknowledge that multiple models are plausible.

**Lecture 3 (R. Jelier): High-Dimensionality, Regularization (Lasso/Ridge)**

**Core Idea:** When the number of predictors (p) is large, especially p > n (e.g., genomics), standard regression fails. We need techniques that can handle collinearity and prevent overfitting.

**Decision Points & Key Questions:**

* **Exploring the Data - What are the challenges?**
  + **The key challenge is high-dimensionality (p > n)**. This leads to overfitting (high variance) and makes standard lm unusable.
  + Another challenge is **collinearity**, where predictors are highly correlated. This makes coefficient estimates unstable and difficult to interpret.
* **Lasso vs. Ridge Regression:**
  + **Lasso (alpha=1 in glmnet):** Performs **variable selection** by shrinking some coefficients to exactly zero. Use this when you believe many predictors are truly irrelevant and you want a simpler, more interpretable model.
  + **Ridge (alpha=0 in glmnet):** Shrinks all coefficients towards zero but does not set any to exactly zero. Use this when you believe most predictors contribute to the outcome and you primarily want to stabilize the model in the face of collinearity.
* **Is Over-learning (Overfitting) a Problem?**
  + **Yes, it is the central problem.** A model with too many predictors will perfectly fit the training data but fail on new data.
  + **How do you address it?** By using **cross-validation (cv.glmnet)** to tune the penalty parameter (lambda). CV estimates the test error, allowing you to choose a lambda that balances bias and variance for optimal prediction on unseen data.
* **Interpreting Results:**
  + **How many genes/predictors are used?** In a Lasso model, the number of predictors used is the number of non-zero coefficients at the optimal lambda.
  + **Comparing Lasso Coefficient to Correlation:** A variable might have a strong simple correlation with the response but get a zero coefficient in a Lasso model. This is **not a contradiction**. It happens when that variable's information is redundant (collinear) with other, stronger predictors in the model. Lasso selects one variable from a correlated group and discards the others.

**Lecture 4 (R. Jelier): Beyond Linearity (GAMs)**

**Core Idea:** The assumption of a linear relationship is often too simple. Generalized Additive Models (GAMs) provide a flexible way to capture non-linear effects while maintaining an interpretable, additive structure.

**Decision Points & Key Questions:**

* **When to use a GAM?** When you suspect or find evidence of a non-linear relationship between a predictor and the outcome. You can check this by plotting the residuals of a linear model against the predictor values; any clear pattern suggests non-linearity.
* **How to fit a GAM?** Use the gam() function, and wrap predictors you believe may be non-linear in a smooth function, like s() for a smoothing spline (e.g., gam(Y ~ s(X1, df=4) + X2)). The df (degrees of freedom) controls the "wobbliness" of the curve.
* **Is there evidence for non-linear effects?**
  1. **Visual:** plot(gam\_model) will show the fitted function for each term. If a curve is not a straight line, the effect is non-linear.
  2. **Formal Test:** summary(gam\_model) provides an "Anova for Non-parametric Effects". A significant p-value for a smooth term (e.g., s(X1)) confirms that the non-linear fit is statistically better than a linear one.
* **Comparing GAM to a Linear Model:** Use anova(lm\_model, gam\_model, test="F"). A significant p-value indicates the GAM provides a superior fit and that the non-linearity is real. You can also compare test set performance (e.g., MSE or AUC); a better performance for the GAM is the ultimate justification.

**Lecture 5 (R. Jelier): Trees, Bagging, Boosting, Random Forests**

**Core Idea:** Decision trees are simple, interpretable models that work by splitting the data into rectangular regions. While individual trees have high variance, their performance can be dramatically improved by using **ensemble methods** (Bagging, Random Forests, Boosting).

**Decision Points & Key Questions:**

* **Why use ensemble methods?** A single decision tree is often not a very accurate predictor. Ensemble methods reduce the variance of a single tree by combining the predictions of many trees, leading to much better performance.
* **Bagging vs. Random Forest:**
  + **Bagging** builds many trees on bootstrapped samples of the data. It's effective but the trees can be highly correlated.
  + **Random Forest** improves on bagging by allowing only a random subset of predictors to be considered at each split. This **de-correlates** the trees, which further reduces variance and often leads to better performance.
* **Boosting (GBM):**
  + This is a different ensemble approach that builds trees **sequentially**. Each new tree is fitted to the *residuals* (the errors) of the previous trees. It's a "slow learner" that gradually improves the model where it performs poorly.
  + It is very powerful but can overfit if too many trees are used. Use **cross-validation (gbm.perf)** to find the optimal number of trees.
* **Interpreting Tree Ensembles:**
  + While individual trees are interpretable, ensembles are more like "black boxes". We interpret them using:
    - **Variable Importance Plots (varImpPlot, summary.gbm):** These rank predictors by their total contribution to the model's performance (e.g., by measuring the total reduction in node impurity).
    - **Partial Dependence Plots (plot.gbm):** These show the marginal effect of one or two predictors on the outcome, which can reveal complex non-linear relationships and interactions captured by the model.