Pre-Work

Hi everyone! Welcome to Linear Regression, Classification, and Resampling (LCR). To get everyone on the same page, please work through this worksheet.

Predictive vs. Inferential

When we look at our data, we're typically trying to answer one of two main things:

Prediction is all about making educated guesses about the future.

> *Example:* Imagine you're trying to guess what time the bus will arrive tomorrow. You look at past days—on average it came at 8:05 AM—so you predict: "Tomorrow it'll be here around 8:05 AM." You don't necessarily care *why* it's late sometimes, just that your guess is as close as possible.

Inference is about understanding the "why" behind the numbers.

> *Example:* Now imagine you want to know *why* the bus is often late. You notice it tends to run behind schedule more when there's heavy traffic downtown. So you conclude: "Traffic delays cause the bus to arrive later." You're focused on uncovering the *reason* behind the lateness, not just guessing the arrival time.

Test yourself!

1. **"Predictive" model's main goal is to:**
   A. Explain why an effect happens
   B. Forecast or predict new data

2. **"Inferential" model's main goal is to:**
   A. Explain why an effect happens
   B. Forecast or predict new data

Supervised vs. Unsupervised

In data science, we "teach" models in two main ways:

Supervised learning: You train the model on examples where both the inputs (features) and the correct answers (labels) are known.

Example: Imagine you're sorting fruit for a party. You have two baskets already labeled "Apples" and "Oranges," and you show a helper dozens of examples—red round apples in one basket, orange textured oranges in the other. After seeing enough labeled fruit, your helper learns the key differences (color, texture, shape). Now when you hand them a new piece of fruit, they can correctly drop it into the "Apple" or "Orange" basket. **Key point: You teach the model with input (fruit features) and the correct output (Apple vs. Orange), so it can classify new items**.

Unsupervised learning: You train the model on data without any labels and ask it to find patterns or groupings on its own.

Example: Now imagine you've got a huge mixed bag of fruit but **no** labels or baskets. You ask your helper to group the fruit however they see fit. Without any instructions, they might naturally sort by color, putting all red and green items together, then all yellow and orange ones together. Or maybe by size, grouping small berries separately from larger fruits. The helper has found patterns and created clusters entirely on their own. **Key point: The model finds structure or patterns in unlabeled data, letting you discover natural groupings without predefined categories.**

Test yourself!

3. **Which scenario describes unsupervised learning?**

A. You show examples of white shirts in "Lights" and dark jeans in "Darks" so the sorter learns which bin to use.
B. You ask the sorter to group a mixed pile of clothes, however they see fit, finding patterns on their own.
C. You tell them "always put shirts in one bin and pants in another," regardless of color.
D. You have them randomly toss clothes into bins.

4. **Which scenario describes supervised learning?**

A. You show examples of white shirts in "Lights" and dark jeans in "Darks" so the sorter learns which bin to use.
B. You ask the sorter to group a mixed pile of clothes, however they see fit, finding patterns on their own.
C. You tell them "always put shirts in one bin and pants in another," regardless of color.
D. You have them randomly toss clothes into bins.

Variable Type

Variables come in two main flavors:

Continuous variables can take any numeric value within a range (e.g., temperature, test score).

Categorical variables represent discrete groups or categories (e.g., blood type, treatment group).

Test yourself!

5. **Age in years is an example of a:**
   A. Continuous variable
   B. Categorical variable

6. **Diagnosis (i.e., Case, Control) is an example of a:**
   A. Continuous variable
   B. Categorical variable

Choosing a Plot

We will be working with two main plots over the course of this module.

Scatterplots: How two numeric things go together.

Example: Imagine you want to see if kids who study more get better grades. You plot each student as a dot, with "hours studied" on the x-axis and "test score" on the y-axis. If the dots slant upward, it means more study time tends to mean higher scores. **Key takeaways: Look for upward/downward trends, clusters of points, or any outliers that don't follow the pattern.**

Histogram: How a numeric variable is distributed.

*Example***:** Suppose you run a lemonade stand and record how many cups you sell each day for a month. A histogram groups days into "bins" (e.g., 0–10 cups, 11–20 cups, 21–30 cups) and draws a bar for each bin showing how many days fell into that range. You might see most days in the 11–20 range and a few days with very high or very low sales. **Key takeaways: Reveals the shape of your data—whether most values cluster in the middle, lean to one side (skewed), or if there are unusual outliers.**

Test yourself!

7. **To compare the relationship between two numeric variables, you'd use a:**
   A. Scatterplot
   B. Histogram

8. **To compare the distribution of a numeric variable's values, you'd use a:**
   A. Scatterplot
   B. Histogram

Answer key

1. B
2. A
3. B
4. A
5. A
6. B
7. A
8. B