1: Introduction to Statistical Learning

```
$ echo "Data Science Institute"
```

Welcome!

- So far, we've focused primarily on coding - now we explore the relationship between coding and statistics as this will allow us to answer questions such as "should we spend more of the advertising budget on TV or the internet"?
- This learning module will include definitions, mathematical concepts and approaches that may be new for most participants
- Live learning sessions will focus on theory, while the homework and assignments will focus on coding and applications
- Work periods will cover the homework from that week - Learning Support will review Homework 1 & 2 this week
- The learning curve will feel steep - this is expected - don't be hard on yourself if it takes time to sink in

Rules of Engagement

- This session is an overview and we will go into more detail in later sessions.
- Questions are encouraged - ask as we go - this is your time to understand these concepts
- Listen to and learn from each other, ask questions on Slack between sessions
- Anything else?

Session 1 - 3 distinct sections

What is Statistical Learning?

Types of Statistical Learning

Applying Statistical Learning

Think and share:

What is Statistical Learning?

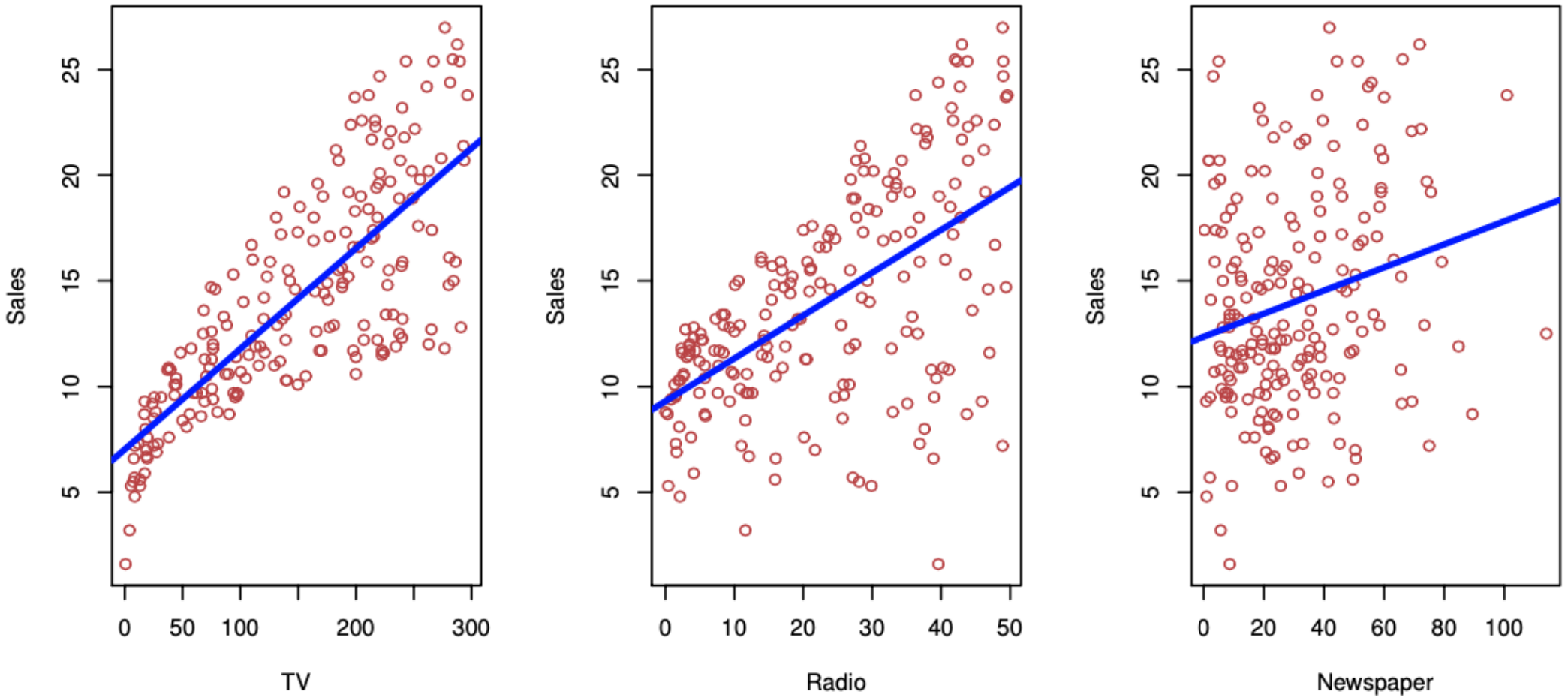
What is Statistical Learning?

Imagine we work in a marketing agency. Our client wants to know whether the money spent on advertising is leading to sales, and which advertising channels results in the most sales. They want to increase their sales and need to determine how to spend the money the right way to drive that increase. Suppose we want to figure out the **relationship between how the advertising budget is spent and sales** in order to increase sales for a client.

- There are three types of advertising: TV, radio, and newspaper which can be labelled $X_1, X_2, and\ X_3$ respectively.
- The advertising budgets (in thousands of dollars) are independent, or ***predictor variables***, which we label X (the horizontal axis).
- The number of sales (in thousands of units) is the dependent, or ***response variable***, which we label Y (the vertical axis).

What is Statistical Learning?

The sales in relation to each of the advertising budgets are shown along with a simple fitted line for the relationships.



What is Statistical Learning?

We want to find the relationship between the predictor variables (budget) and the response variable (sales). This relationship can be described as a function f. In reality the relationship is complicated, and cannot be perfectly described. We are using this function to model the relationship. The difference between the actual value and the estimation of that value can be described as a random error term ϵ.

This relationship between X and Y can be written as:

$$Y = f(X_1, X_2, X_3) + \epsilon$$

Statistical learning is summarized by the set of approaches which are used to estimate f.

Types of Statistical Learning

Prediction vs Inference

There are two main reasons why we want to model to estimate f:

1. If we want to know what sales can be expected for a given advertising budget? ***What response is expected given a set of predictors.*** This is **prediction**.
2. If we want to know to what extent sales volume is related to the advertising budget? ***How the response variable is affected by changes in the predictors.*** This is **inference**.

Types of Statistical Learning

Prediction

Prediction problems focus on the response Y. They can arise when the ***predictor variables X are known but the response Y is not easily obtained***. We use "^" to denote estimates. That is, $\hat{Y}$ is an estimate for Y and $\hat{f}$ is an estimate for f.

The accuracy of our prediction, $\hat{Y}$, depends on two types of errors: 1) those that we can potentially control, influence or **reduce** and 2) those that we cannot control or **reduce**. The **reducible error** is the error that we need to focus on as an analyst. But there is alway some **irreducible error**: the random error associated with the true response $Y = f(x) + \epsilon$. (Even if $\hat{f} = f, \hat{Y}$ will still have error associated with its prediction since ϵ is not a function of X.)

Our focus is on making predictions for Y using $\hat{f}$.

Types of Statistical Learning

Inference

Inference problems focus on predictors X. They can arise when both the ***predictor variables*** X ***and the response*** Y ***are known*** and we want to know how they are related.

The accuracy of our inference depends on how exactly we can estimate $\hat{f}$. It may depend on: 1) understanding which predictors are more important than others, 2) how does the response change (positively or negatively) given changes in the predictors, and 3) does the response change linearly or non-linearly given changes in the predictors.

Our focus is on finding the true form of f.

Applying Statistical Learning

How do we estimate f?

Assume that we have n observations in our data set. The standard approach is to split the data set into training data and testing data.

- **training data** is used to train or teach the model we are using to estimate f.
- **testing data** is used to test the accuracy of the resulting estimate for f on new data.

Applying Statistical Learning

Accuracy-Interpretability Trade-Off

- The models we will introduce have different levels of restrictiveness or flexibility.
- Choosing a model on the basis of flexibility will depend on the problem at hand.
  - ***If we are interested in inference, restrictive models are much more interpretable*** (i.e., the relationship between the predictors and the response is more clear)
  - ***If we are only interested in prediction accuracy, flexible models ♦ might ♦ perform better.***
- Flexible models will not always provide better predictions since they are prone to overfitting!

Applying Statistical Learning

Supervised vs Unsupervised Learning

- **Supervised learning** involves models for predicting a response based on predictor variables.
    - Examples of supervised learning models are linear regression and classification. These models are the primary focus of this learning module.
- **Unsupervised learning** refers to models used to investigate features associated with observations
    - There is no response variable to predict, instead the goal is to understand the relationship between variables or observations.
    - An example of this is clustering.
    - Future modules explore this.

Applying Statistical Learning

Regression vs Classification Problems

- Variables can be either qualitative or quantitative.
  - **Quantitative** variables have numerical values (ex: age, monetary value, etc.)
  - **Qualitative** variables are categorical values (ex: {small, medium, large} or {yes, no})
- ***Problems that involve quantitative response variables are ♦ regression ♦ problems.***
- ***Problems that involve qualitative response variables are ♦ classification ♦ problems.***
- This is a bit of a generalisation (logistic regression is a classification method but its output is numerical so can be thought of as a regression method as well)

Applying Statistical Learning

Assessing Model Accuracy

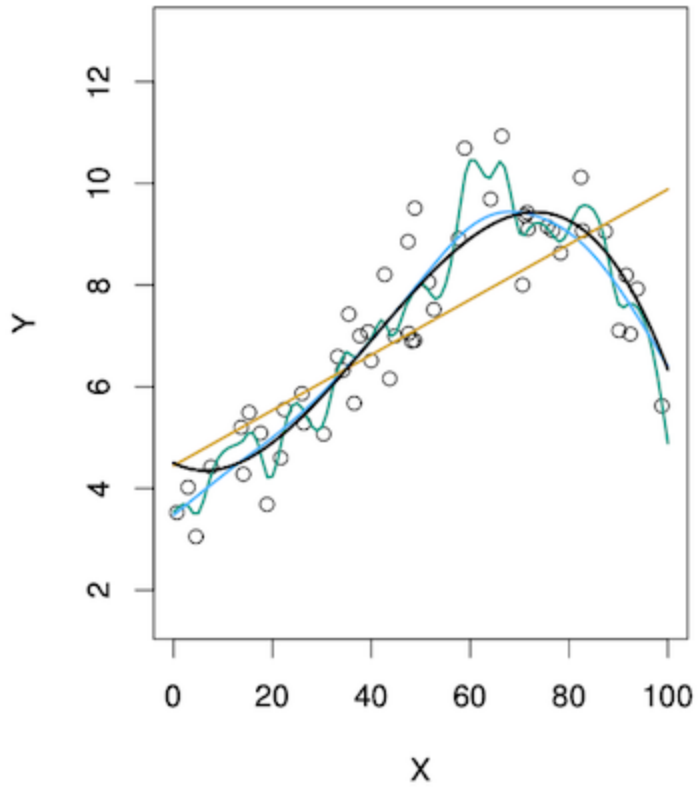
There is no model that works the best on all data sets so we need a way to assess the quality of the model's fit to the data.

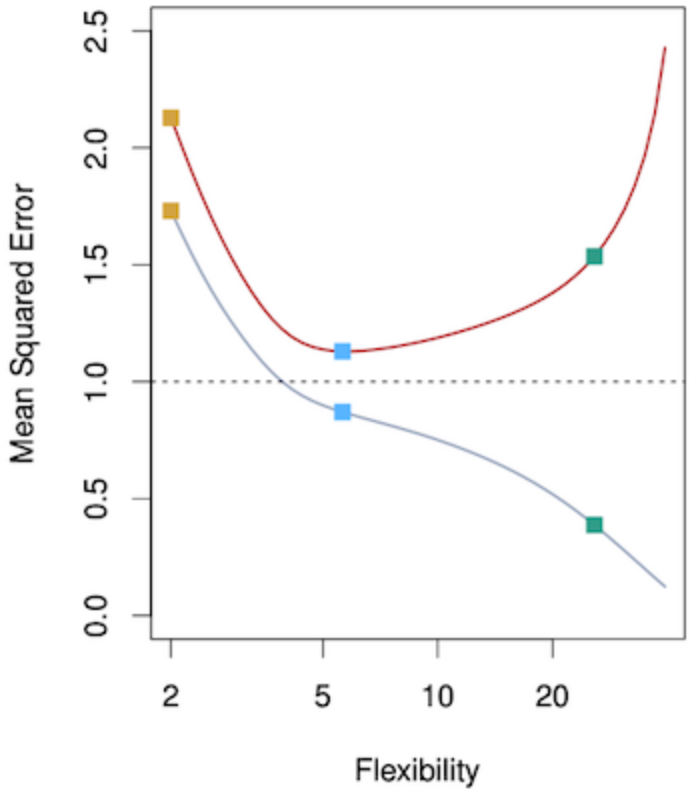
- Recall that we use the training data set to fit the model.
- Then the test data set is used to see how well the model performs on new data by computing some test statistic.
- One way to do this is to look at how different our prediction is from the actual.
- For regression, the **mean squared error** (MSE) is most common which measures how close the predicted responses are to the true responses.
- We can compute the MSE for both the training data and the test data.
    - The training MSE will usually be lower (better) than the test MSE.
    - We want to ***choose the model that minimizes the test MSE*** since we only really

Applying Statistical Learning

Assessing Model Accuracy

The first plot shows three different models fit to data, and the second compares the training Mean Squared Error (MSE) and test MSE against the flexibility of the model.





Applying Statistical Learning

Assessing Model Accuracy

In the described plots:

- The least flexible model (orange) exhibits the worst training and test MSE, indicating a poor fit to the data.
- The most flexible model (green) shows the lowest training MSE but suffers from a much higher test MSE, suggesting that the model is overfit.
- A model of intermediate flexibility (blue) achieves the lowest test MSE, making it the preferred choice among the three.

This textual description aims to convey the essential information without relying on specific layout instructions. For actual visual representations, including the plots directly as images is the best practice, as shown above.

Applying Statistical Learning

Breakout Room: What if no test observations are available?

In your breakout rooms, think, and share what happens if test observations are not available.

Hint: Google and ChatGPT are your friends!

Applying Statistical Learning

Assessing Model Accuracy

What if no test observations are available?

- As we saw, there is no guarantee that the model that minimises the training MSE will minimise the test MSE.
- ***As model flexibility increases, the training MSE decreases*** but the test MSE may not.
- A model is said to be overfit if a less flexible model would results in a lower test MSE.
- An overfit model is picking up on patterns in the training data that are not in the test data.
- We will cover many approaches for estimating the test MSE from training data.

Applying Statistical Learning

The Bias-Variance Trade-Off

In order to minimise the expected test error we need to use statistical learning methods that result in low bias and low variance.

- **Variance** is the amount $\hat{f}$ would change if we fit it using a different training set.
  - More flexible models have higher variance since they fit the training data more closely.
- **Bias** is the error from approximating a complicated relationship with a simpler model.
  - More restrictive models have higher bias since they make more assumptions about the form of f.# Applying Statistical Learning

Applying Statistical Learning

The Bias-Variance Trade-Off

The test MSE will always exhibit a U-shaped curve as a function of model flexibility. This is because the expected test MSE for some observation x_0 is the sum of:

- $\mathrm{Var}(\hat{f}(x_0))$: the variance of $\hat{f}(x_0)$ (the variance of the predicted response for the test observation x_0 given many $\hat{f}$ fit on different training sets).
- $[\mathrm{Bias}(\hat{f}(x_0))]^2$: The squared bias of $\hat{f}(x_0)$.
- $\mathrm{Var}(\epsilon)$: the variance of the error terms ϵ.

Applying Statistical Learning

The Bias-Variance Trade-Off

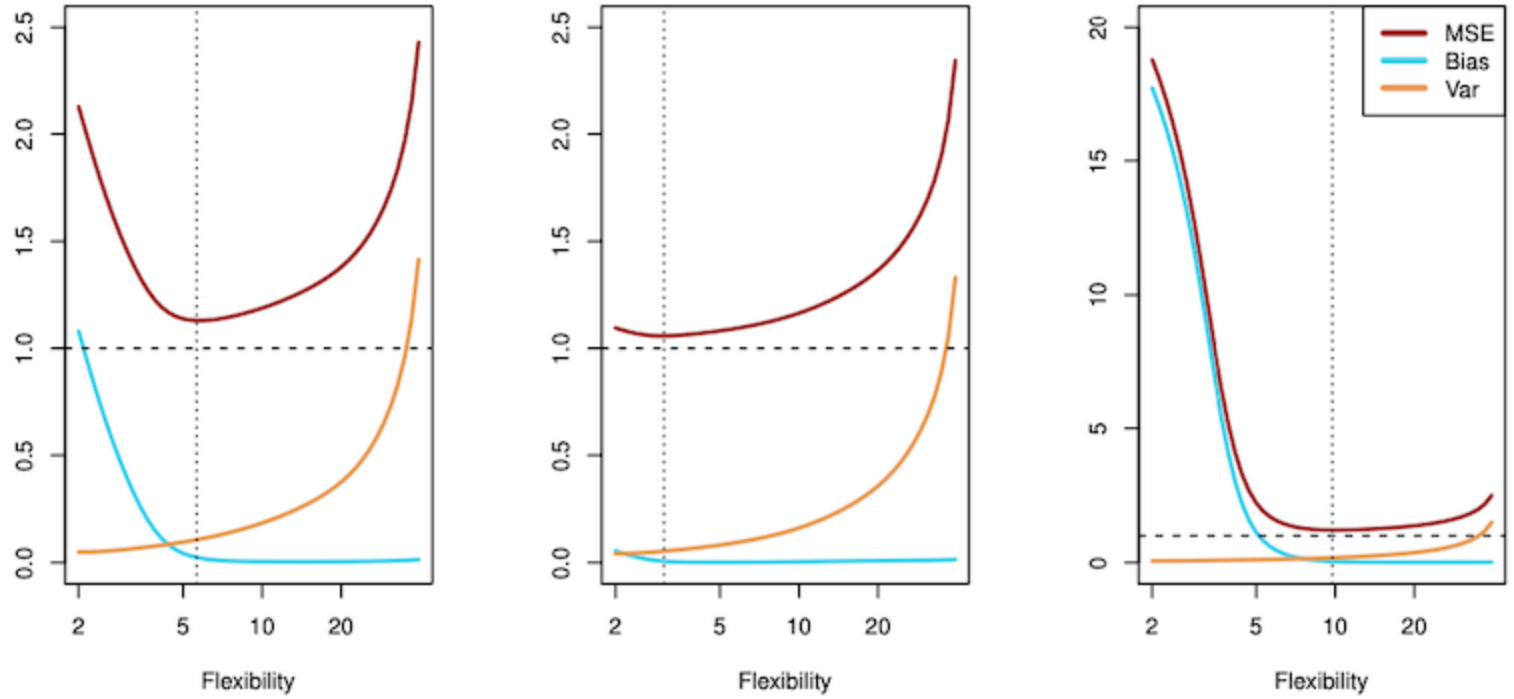
The rate of change of the bias versus variance determines whether the test MSE will decrease or increase with flexibility.

- Initially, as the flexibility of the model increases, the bias decreases faster
  - The test MSE declines
- At some point the increasing flexibility has little impact on the bias but the variance increases significantly.
  - The test MSE increases.
  - This results in a U-shaped curve for test MSE vs model flexibility.

Applying Statistical Learning

The Bias-Variance Trade-Off

The plot illustrates the bias (blue), variance (orange), variance of the error (dashed line), and the test Mean Squared Error (MSE) (red) as functions of model flexibility for three different data sets.



Applying Statistical Learning

The Bias-Variance Trade-Off

Key points from the plot:

- The MSE is the sum of the bias squared, variance, and the variance of the error (irreducible error).
- The MSE cannot be smaller than the variance of the error, which represents the irreducible error.
- The model that is closest to linear shows that the test MSE starts increasing immediately with increased flexibility, indicating a trade-off between bias and variance.

Applying Statistical Learning

Classification Model Accuracy

So far we have discussed model accuracy in the context of regression but the same ideas apply to classification.

- The most common approach for assessing model accuracy is the **training error rate** which is the ***proportion of misclassified training observations***.
- The **test error rate** is what we are actually interested in and hoping to minimise.
- The bias-variance trade-off is what controls the test error rate in the classification context as well.

References

Chapter 2 of the ISLP book:

James, Gareth, et al. "Statistical Learning." An Introduction to Statistical Learning: with Applications in Python, Springer, 2023.