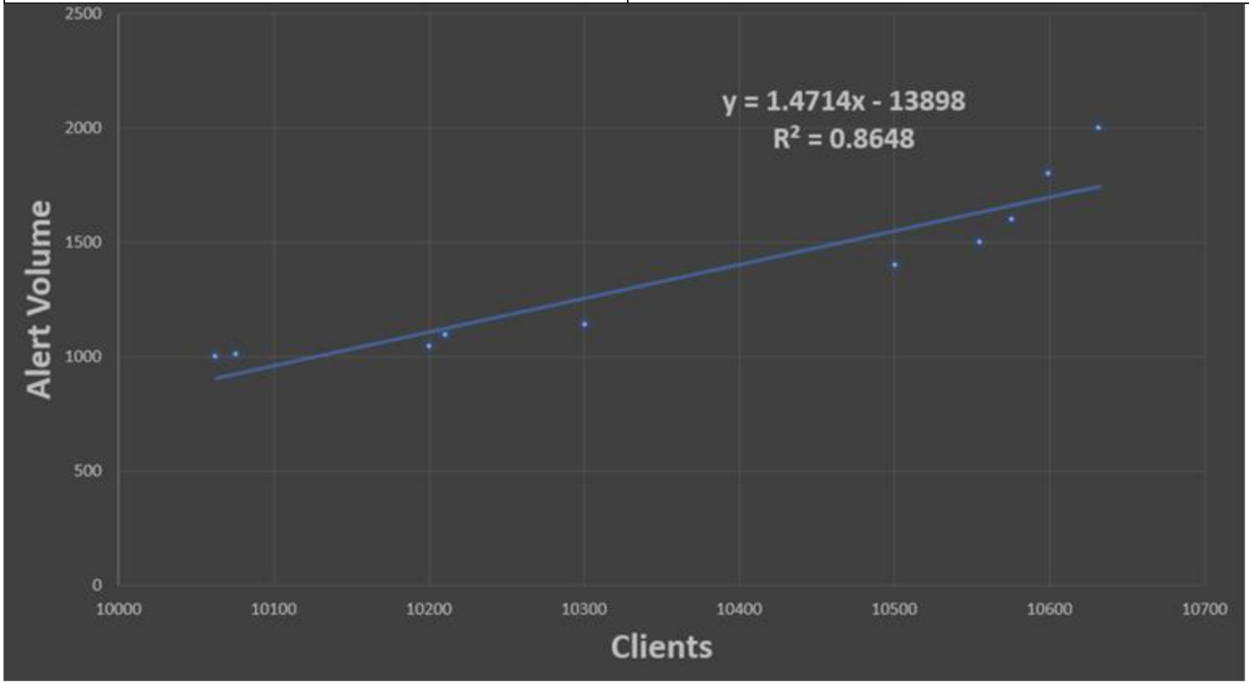
Lisa Coombs
Ex 5.5 Introduction to Predictive Analysis
March 18, 2025

Step 1) Understanding Regression

Characteristics	Linear Regression	Logistic Regression
What they do	Tries to predict a number	Tries to predict outcomes that are yes/no, true/false or something similar
How they work	It draws a straight line through the data points that shows the overall trend	It uses an S-shaped curve to show the probability of an outcome happening
What they interpret	The numbers given tell you how much the result changes when one factor changes (either the independent or dependent variable)	These numbers tell you how much a factor increases or decreases the odds of one of the outcomes happening
When to use	Best to use when you want a predicted number. This method uses a straight line to make its predictions	Best to use when you want a decision. This method uses a curve to turn the prediction into a probability

Step 2) More on Linear Regression
Interpreting the Measurement and Assessing Model Fitness

Variables Defined	Dependent Variable = Alert Volume Independent Variable = Clients
Rise of Regression Line	There is a positive relationship between the variables because the line rises from the left to the right
Steepness of Regression Line (Strength)	The incline of this line is not steep, it gently slopes up from left to right. This indicates a weaker relationship between the variables
Data Points and R-Squared value	The R-Squared value is 0.86 which is really close to 1. The data points fall close to the regression line. R-Squared values that are closer to 1 indicate a better predictive model



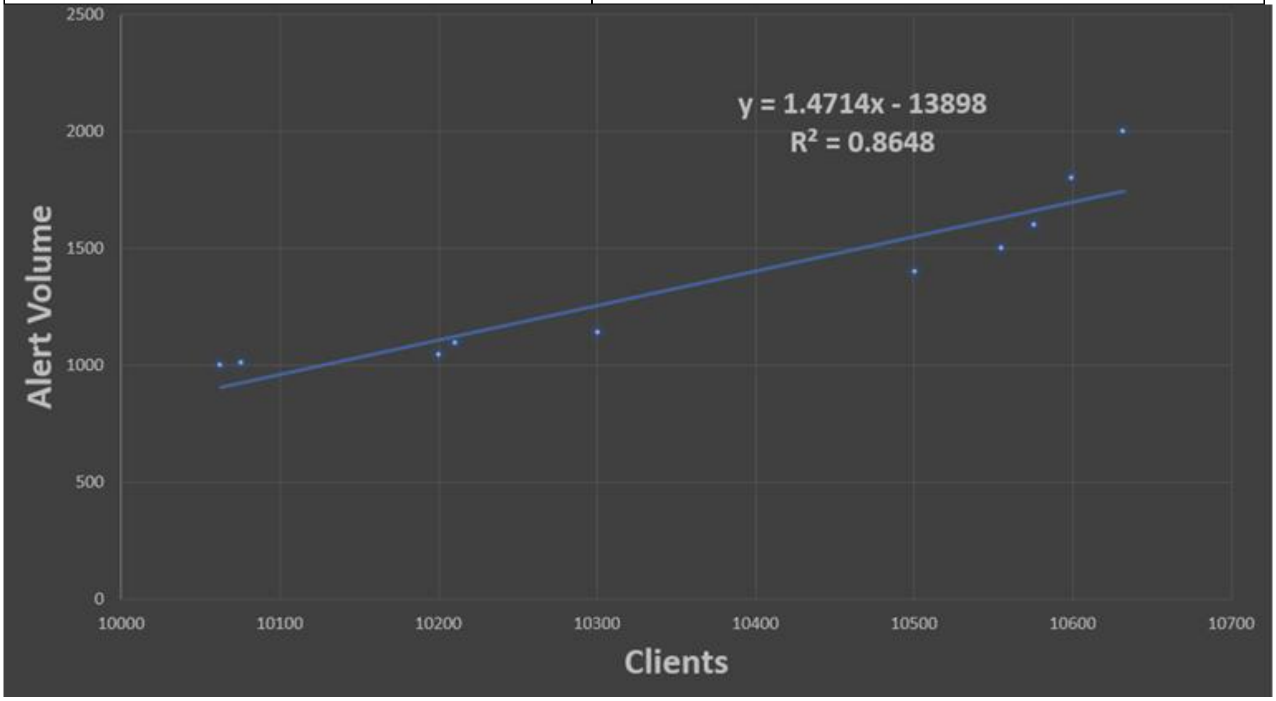
Lisa Coombs
Ex 5.5 Introduction to Predictive Analysis
March 18, 2025

Step 3) Differentiating between models

	Scenario A – Financial Institution	Scenario B – Online Movie Provider
Predictive Model	Regression - Linear	Classification – Random Forest
Variables	Dependent Variable = Global Oil Prices, Independent Variable = Unemployment rates of top 20 GDP countries	Dependent Variable = whether the customer is likely to watch RomCom with Sandler and Barrymore (yes/no) Independent Variable = customers' viewing habits
Explanation	I want to look at how changes in the unemployment rates of the top 20 GDP countries relate to global oil prices. Using a linear regression model will tell me how much the oil price might change with a one-unit change in the unemployment rate.	Random Forests are used when large data sets contain more variables than decision trees. Each decision tree within the forest will predict an outcome to the same question "is this customer likely to watch RomCom with Sandler and Barrymore?"

Step4) Bias in the Data from Step 2 Linear Regression Model

Describe relationship between variables	There is a positive correlation between the number of fraud alerts and the number of clients, the equation y=1.4714x – 13898 means that as the number of clients increase, so does the number of fraud alerts, by 1.47.
Assess the fitness of model in predicting alert volume based on the number of clients	The R-squared value (strength of relationship) indicates a strong positive relationship between clients and fraud alerts, but doesn't consider any other influencing factors like: time frame or geographic scope.
Potential bias in Data	The dataset does not include a wide range of periods with significantly different fraud alert rates. This could mean that if this model covers a period when there was high fraud activity, it could lead to an overestimation of fraud rates for new clients


y = 1.4714x - 13898
R² = 0.8648
Alert Volume
Clients