**IN402 Unit 3 Seminar**

Laurence Burden

* Realpython.com overview
  + What is Regression
    - Searches for relationships between variables
    - Can find how employee salaries depend on experience, role, cities, and more
  + Python packages for regression
    - NumPy
    - Pandas
    - SciKit-Learn
* Multiple regression
  + More practical in real-world
  + Move from linear regression to logistic regression
* Regression is among the most popular data analytics methods
* Supervised learning
  + Use data to compare prediction results
  + Build a regression model using this data
* Unsupervised learning
  + No standard solution to what the answer should be
  + Criteria still exist to evaluate the solution
* When do you need regression?
  + To describe data
  + To predict/forecast data
    - i.e. The stock market
* Linear regression
  + The straight line that best describes the relationship between the independent variable(s) and the dependent variable
  + There are many ways to draw the line, but we are looking for the best fit
    - Use the minimum sum of square errors
      * Uses the sum of all errors squared from predicted vs actual data
    - Looking for the optimal parameters in the line’s slope equation
* Multiple Linear Regression
  + Uses multiple independent variables to find one dependent variable
  + Seek to minimize SSR
* Polynomial Regression
  + Uses a curve instead of a straight line
  + Often seen in growth curves
    - Bacteria, viruses, etc.
* Degrees of freedom = number of parameters – one
* R2 = Ratio of Sum of Squares of Regression and Sum of Squares of Errors
  + Value is between 0 and 1
  + 1 means the line fits the data perfectly
* Underfitting
  + The model doesn’t accurately capture the dependencies among data
  + Outputs a low R2 with known data
* Overfitting
  + This means that the model has learned both the dependencies and random fluctuations within the provided data
  + Complex models are more prone to overfitting
  + Outputs a high R2 with known data
* Adjusted R2 will penalize excess parameters
  + Normal R2 will always be better with more parameters
* Excel package
  + Data analysis tool pack
    - The regression section can give an ANOVA table and adjusted R2
* Using SciKit-Learn
  + First import the LinearRegression() method
  + Assign X and y values
  + Fit the model with the fit() method on the X and y values
  + Use the predict() method to predict based on a new value
  + R2 is found with the score() method
  + Coefficient and intercept can be viewed by calling certain properties of the model
  + The summary() method can output many variables of interest
* P-value
  + Determine the significance of the independent variable within the model
* Many types of regression analysis
  + Linear Regression
  + Polynomial Regression
  + Ridge Regression
    - Regularization technique that addresses multicollinearity and adds a penalty term to the least squares objective function
    - Takes the number of parameters into account from the beginning
  + Lasso Regression
    - Regularization method that addresses multicollinearity by penalizing least square objective function and shrinks some coefficients to zero through variable selection
  + Elastic Net Regression
    - Combines ridge and lasso regression
    - Has an L1 and L2 penalties within the objective function
  + Decision Trees
    - Splits the data based on feature values to make decisions
    - Can handle continuous and categorical data
  + Support Vector Regression
    - Based on support vector machine (SVM) algorithm
    - Finds a hyperplane that maximizes the margin around the predicted values
    - Can handle linear and non-linear relationships
  + Bayesian Regression
    - Uses prior knowledge about the parameters of the regression model
    - Provides a probabilistic framework for estimating model parameters and making predictions
    - Useful for smaller datasets
  + Generalized Linear Models
    - Extends linear regression
    - Uses a distribution other than the normal distribution
    - Can handle binary, count, and categorical outcomes
  + Neural Networks
    - Becoming increasingly popular
    - Requires large amounts of data