

Machine learning

Regression

Instructor: Hector G. Ceballos

E-mail: ceballos@tec.mx

FH·W-S



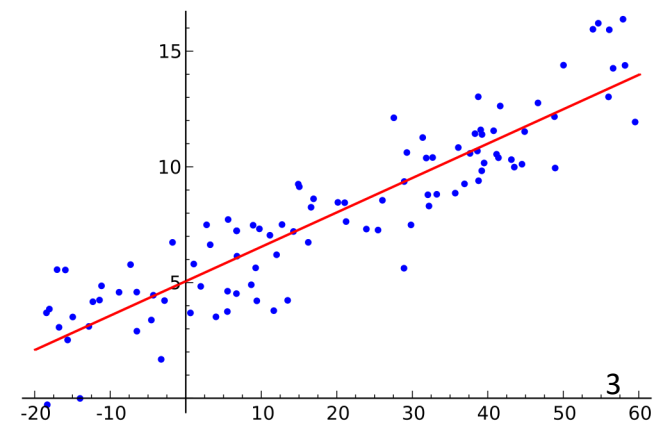
Agenda

- Introduction to Regression Analysis
- Practice 1: Data Exploration [30m]
- Simple Linear Regression
- Practice 2: Learn Simple Linear Regression Models [30m]
- Multiple Linear Regression
- Practice 3: Learn a Multiple Linear Regression Model [45m]

Regression Analysis

- Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features').
- The most common form of regression analysis is linear regression, in which a researcher finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion.

https://en.wikipedia.org/wiki/Regression_analysis



- The model is represented by a (linear) equation, which coefficients are used to explain the influence of the predictors on the outcome variable.
- Regression analysis is widely used for prediction and forecasting.
- In social sciences, regression analysis can be used to infer causal relationships between the independent and dependent variables.

Correlation

- Pearson product-moment correlation coefficient (PPMCC), or "Pearson's correlation coefficient": the quality of least squares fitting to the original data

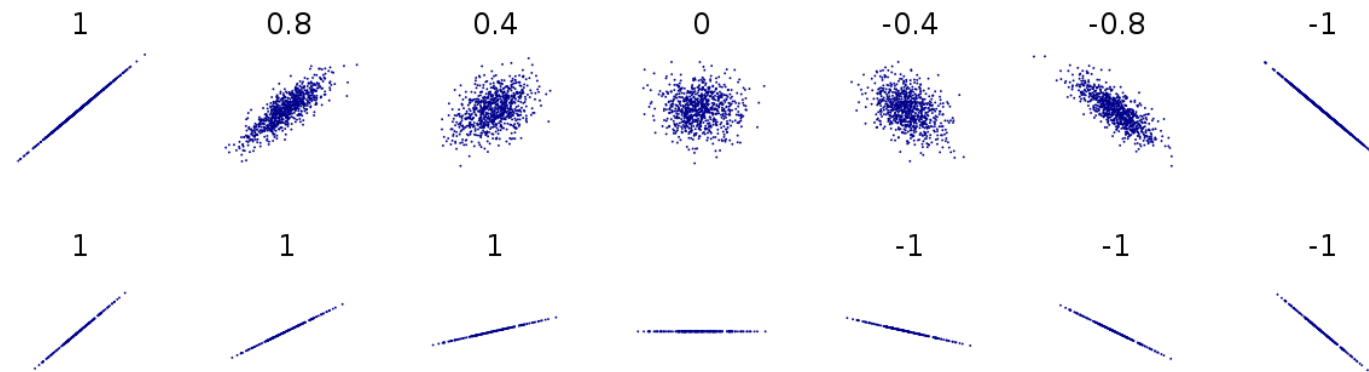
$$r_{xy} \stackrel{\text{def}}{=} \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

where \bar{x} and \bar{y} are the sample means of X and Y , and s_x and s_y are the corrected sample standard deviations of X and Y .

- Non-parametric methods: Spearman's rank correlation coefficient and Kendall's rank correlation coefficient.

Correlation

- Visual representation: Dots represent data points distributed along two dimensions or axes (X and Y).
 - Positive: data points grow in both dimensions.
 - Negative: data points grow in one dimension while they decrease in the other.



- Note that: Correlation is not causation!!

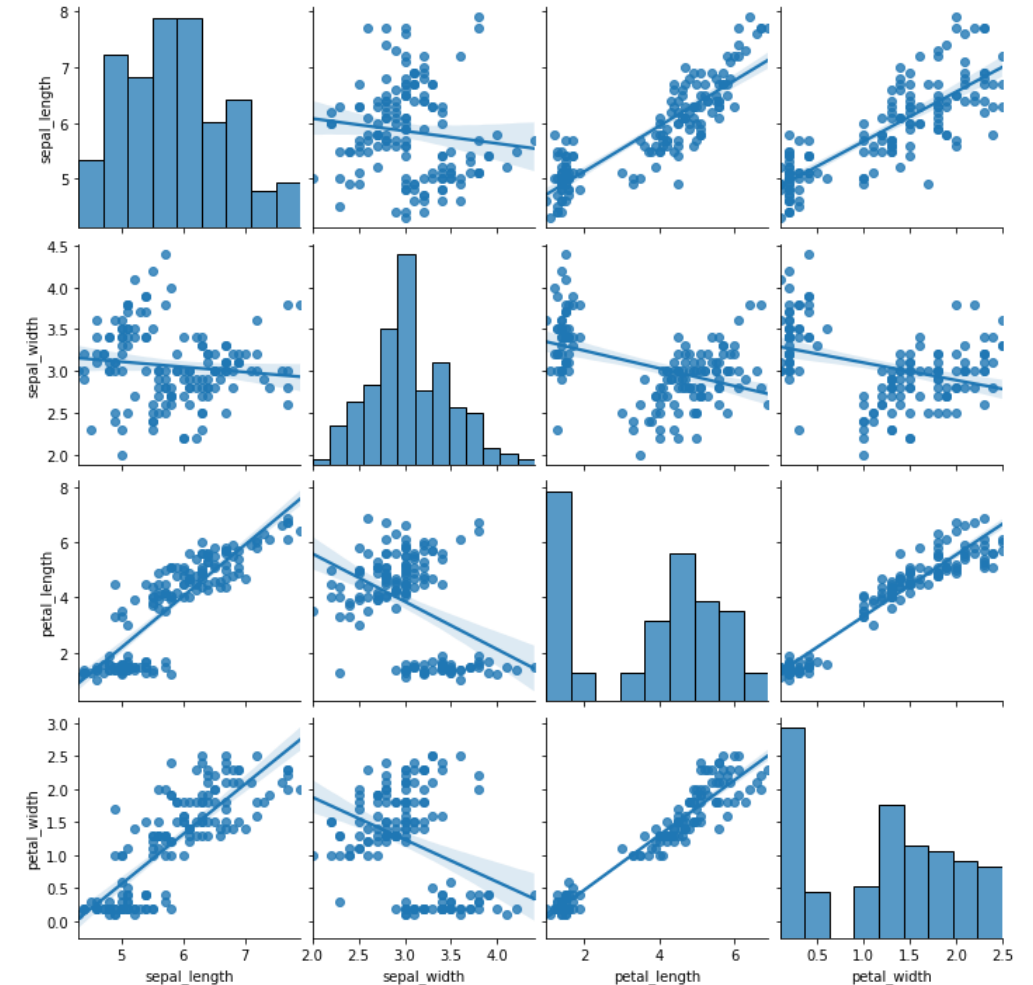
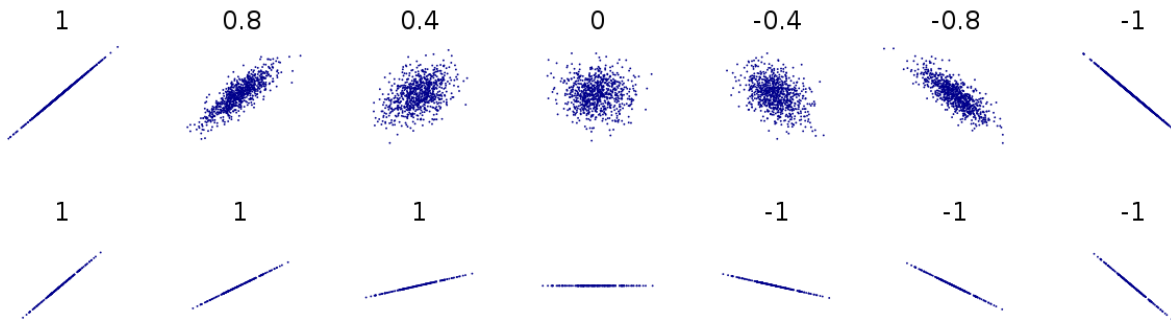
Correlation matrix

- It is used to observe correlation between variables.
- It highlights positive (direct) and negative (inverse) correlation with different colors.

	0	1	2	3	4	5	6	7	8	9
0	1	0.35	0.4	0.46	0.073	-0.23	-0.73	0.48	-0.44	0.015
1	0.35	1	-0.28	0.57	-0.29	0.38	-0.36	0.64	0.25	0.19
2	0.4	-0.28	1	-0.52	0.15	-0.14	-0.093	0.016	-0.43	-0.38
3	0.46	0.57	-0.52	1	-0.23	-0.23	-0.48	0.47	0.28	0.45
4	0.073	-0.29	0.15	-0.23	1	-0.1	-0.15	-0.52	-0.61	-0.19
5	-0.23	0.38	-0.14	-0.23	-0.1	1	-0.03	0.42	0.21	0.095
6	-0.73	-0.36	-0.093	-0.48	-0.15	-0.03	1	-0.49	0.38	-0.35
7	0.48	0.64	0.016	0.47	-0.52	0.42	-0.49	1	0.38	0.42
8	-0.44	0.25	-0.43	0.28	-0.61	0.21	0.38	0.38	1	0.15
9	0.015	0.19	-0.38	0.45	-0.19	0.095	-0.35	0.42	0.15	1


Correlogram


- It combines distribution charts (histogram) and correlation charts.
- It can be used to visually inspect linear correlation between variables.



<https://www.python-graph-gallery.com/111-custom-correlogram>

Hands on: Inspecting the PIMA database


 Dataset




2176


Pima Indians Diabetes Database


Predict the onset of diabetes based on diagnostic measures


 UCI Machine Learning • updated 4 years ago (Version 1)

[Data](#) [Tasks \(3\)](#) [Code \(1,483\)](#) [Discussion \(31\)](#) [Activity](#) [Metadata](#)

[Download \(23 KB\)](#) [New Notebook](#) 

 Usability 8.8

 License CC0: Public Domain

 Tags earth and nature, health, diabetes, healthcare, india

Description

Context

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

<https://www.kaggle.com/uciml/pima-indians-diabetes-database>

Exercise 1. Data exploration on the PIMA database

- Make a new dataframe having only patients with diabetes and remove the Outcome variable.
- Plot the correlation matrix and correlogram of the new dataset.
- Identify the three variables more correlated to Insulin.

Regression model

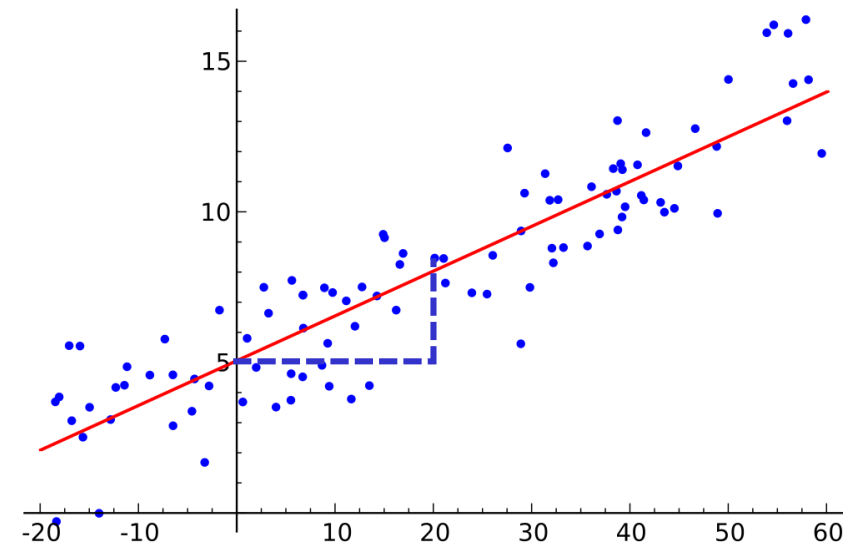
- Regression models involve the following components:
 - The unknown parameters, often denoted as a scalar or vector β .
 - The independent variables, which are observed in data and are often denoted as a vector X_i (where i denotes a row of data).
 - The dependent variable, which are observed in data and often denoted using the scalar Y_i .
 - The error terms, which are not directly observed in data and are often denoted using the scalar e_i .

$$Y_i = f(X_i, \beta) + e_i$$

Linear regression

- In Linear Regression, the model specification is that the dependent variable, y_i is a linear combination of the parameters (but need not be linear in the independent variables).

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n.$$



- The researchers' goal is to estimate the function that most closely fits the data.
 - The form of the function f must be specified based on domain's assumptions.
- Learning methods try to fit the model by minimizing the **residuals** e_i :
 - $e_i = y_i - \hat{y}_i$
 - y_i is the **true** value of the dependent variable.
 - \hat{y}_i is the value of the dependent variable **estimated** by the model.
- One method of estimation is **ordinary least squares**.

Ordinary Least Squares (OLS)

- OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares: minimizing the sum of the squares of the residuals.
- It is common to assess the goodness-of-fit of the OLS regression by comparing how much the initial variation in the sample can be reduced by regressing onto X .

Model Accuracy

- The accuracy of a simple regression model is given by the R^2 and the p-value of the coefficients.
- The coefficient of determination R^2 is defined as a ratio of "explained" variance to the "total" variance of the dependent variable y .
- An R^2 of 1 indicates that the independent variable (regressor) entirely explains the dependent variable (100%).

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

RSS: Residual Sum of Squares
TSS: Total Sum of Squares

$$TSS = \sum (y_i - \bar{y}_i)^2 \quad \text{mean}$$

$$RSS = \sum (y_i - \hat{y}_i)^2 \quad \text{expected}$$

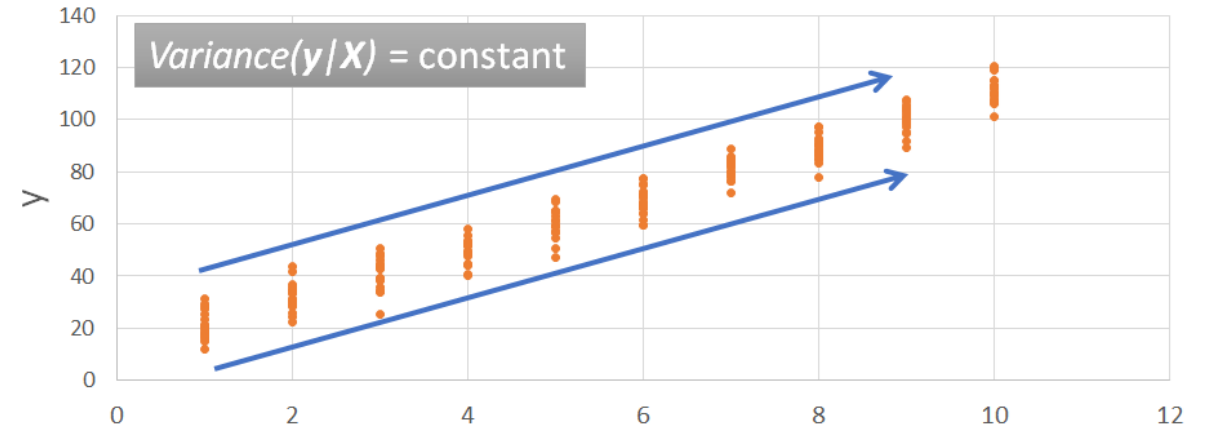
Classic Assumptions

- The sample is representative of the population at large ($n > 30$).
- The independent variables are measured with no error.
- The residuals e_i are uncorrelated with one another (multicollinearity).
- The variance of the residuals e_i is constant across observations (homoscedasticity).

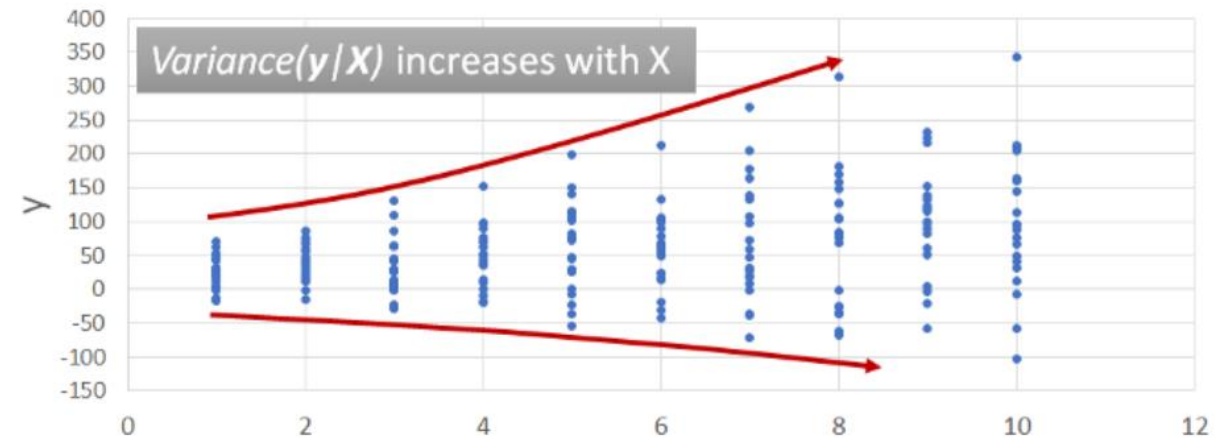
Homoscedasticity

- A sequence (or a vector) of random variables is homoscedastic if all its random variables have the same finite variance (homogeneity of variance).
- Assuming a variable is homoscedastic when in reality it is heteroscedastic results in unbiased but inefficient point estimates and in biased estimates of standard errors, and may result in overestimating the goodness of fit as measured by the Pearson coefficient.

A homoscedastic data set

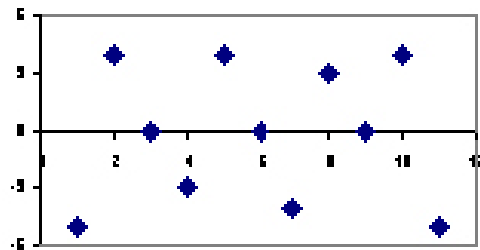


A heteroscedastic data set

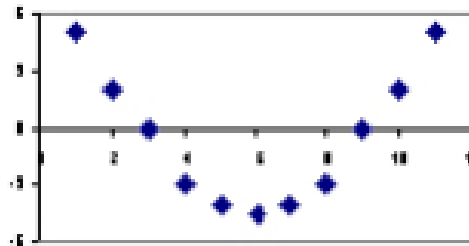


Residual Plots

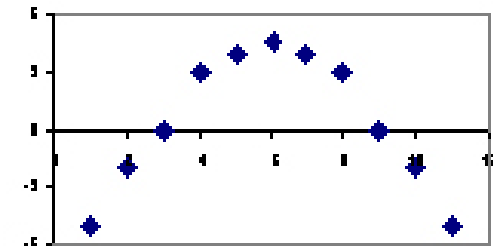
- A residual plot is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis.
- If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a nonlinear model is more appropriate.



Random pattern



Non-random: U-shaped



Non-random: Inverted U

Hands on: Learning a Simple Linear Regression model

- Learn a simple regression model where the Glucose level is used to estimate the insulin level of a diabetic patient.
- Evaluate the accuracy of the model
- Plot the residuals

Exercise 2: Simple Linear Regression

- Train other four single regression models (BMI, Age, SkinThickness, DiabetesPedigreeFunction)
- Plot residuals for each model
- Choose the best model

Multiple Linear Regression

- Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon$$

where, for $i=n$ observations:

- y_i = dependent variable
- x_i = explanatory variables
- β_0 = y-intercept (constant term)
- β_{1-p} = slope coefficients for each explanatory variable
- ϵ = the model's error term (also known as the residuals)

Multicollinearity

- One predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy.
- In this situation, the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data.
- A multivariate regression model with collinear (correlated) predictors can indicate how well the entire bundle of predictors predicts the outcome variable, but it may not give valid results about any individual predictor (making predictions for unknown data points)..

Model accuracy in Multiple Regression models (F-statistic)

- In the case of multiple linear regression, we use another metric: the F-statistic.

$$F = \frac{\frac{TSS - RSS}{p}}{\frac{RSS}{(n - p - 1)}}$$

n is the number of data points
and p is the number of predictors.

- If there is a strong relationship, then F will be much larger than 1. Otherwise, it will be approximately equal to 1.
- Usually, if there is a large number of data points, F could be slightly larger than 1 and suggest a strong relationship. For small data sets, then the F value must be way larger than 1 to suggest a strong relationship.

Variance inflation factor (VIF)

- It quantifies the severity of multicollinearity in an ordinary least squares regression analysis.
- It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity.
- *Interpretation:* If the VIF of a predictor were 5.27 ($\sqrt{5.27} = 2.3$), this means that the standard error for the coefficient of that predictor variable is 2.3 times larger than if that predictor variable had 0 correlation with the other predictor variables.
- $VIF > 10$ indicates multicollinearity is high.

Feature Selection

- Variable selection has two conflicting goals:
 - (a) on the one hand, we try to include as many regressors as possible so that we can maximize the explanatory power of our model,
 - (b) on the other hand, we want as few predictors as possible because more regressors could lead to an increased variance in the prediction.
- When evaluating which variable to keep or discard, we need some evaluation criteria such as: Adjusted R squared or F-statistic.

Adjusted R-Squared

- The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model.
 - The adjusted R-squared increases only if the new term improves the model more than would be expected by chance.
 - It decreases when a predictor improves the model by less than expected by chance.
 - It is always lower than R-squared.

Vars	R-Sq	R-Sq(adj)
1	72.1	71.0
2	85.9	84.8
3	87.4	85.9
4	89.1	82.3
5	89.9	80.7

Feature Selection Methods

- **Forward** elimination starts with no features, and the insertion of features into the regression model one-by-one.
 - First, the regressor with the highest correlation is selected for inclusion, which coincidentally the regressor that produces the largest F-statistic value when testing the significance of the model.
 - The procedure continues until the F statistic exceeds a pre-selected F-value (called F-to-enter) and terminates otherwise.
- **Backward** elimination starts with all regressors in the model.
 - The feature with the smallest F statistic is removed from the model and the procedure continues until the smallest partial F statistic is greater than the pre-selected cutoff value of F, and terminates otherwise.

Feature Selection Methods

- **Stepwise** elimination is a hybrid of forward and backward elimination and starts similarly to the forward elimination method, e.g. with no regressors.
 - Features are then selected as described in forward feature selection, but after each step, regressors are checked for elimination as per backward elimination.
 - The hope is that as we enter new variables that are better at explaining the dependent variable, variables already included may become redundant.

Recursive Feature Elimination (RFE)

- RFE selects features by considering a smaller and smaller set of regressors.
- The starting point is the original set of regressors. Less important regressors are recursively pruned from the initial set. The procedure is repeated until a desired set of features remain.
- That number can either be a priori specified (RFE), or can be found using cross validation (RFECV).
- Regressors are ranked and the top N regressors are selected.

https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

Hands on: Learn Multiple Linear Regression Models

- Learn a model with all variables.
- Detect and eliminate multicollinearity using VIF
- Eliminate non-significant coefficients
- Use a feature selection algorithm

Exercise 3

- Make a new feature selection using the VIF method.
- Build and validate a model with the 4 most correlated variables.
- Apply feature selection starting from all variables.
- Check if you get to the same model with the two methods.
- Describe the best model through a linear equation

Closure

- Show results
- Minute paper