# Exercise 1: Linear Regression

## 1. Simple Linear Regression: Academic Example

In this exercise, you will create some simulated data and fit simple linear regression models to it. Make sure to use `rng(1000)` prior to starting part (a) to ensure consistent results.

(a) Using `linspace()` function, create a vector, $X$, containing $N = 50$ equidistant observations between 1 and 100. By default, `linspace()` generates a row vector. Use the transpose to obtain a column vector.

(b) Using the `randn()` function, create a vector, $\epsilon$, containing 100 observations drawn from a $N(0, 30^2)$ distribution, i.e., a normal distribution with zero mean and standard deviation of 30.

(c) Using $X$ and $\epsilon$, generate a vector $Y_0$ according to the function

$$Y_0 = 50 + 7 \cdot X \tag{1}$$

and add the noise to obtain the disturbed output $Y = Y_0 + \epsilon$.

(d) Using `scatter`, create a scatterplot displaying data $X$ versus $Y$. Discuss the relationship you observe.

(e) Using the backslash operator, fit a least squares linear model to predict $Y$ using $X$. First, create the regression matrix using `Phi=[ones(size(X)) X]`. Discuss the model obtained and how $\hat{\beta}_0$ and $\hat{\beta}_1$ compare with $\beta_0$ and $\beta_1$.

(f) Calculate the predictions $\hat{Y}$ on the training data set. Display the least squares line on the scatterplot obtained in (d) using `hold all`. Additionally, draw the population regression line in the plot.

(g) Alternatively, use the Matlab function `fitlm()` to generate a linear regression model object. Plot the results by the use of `plot(myModel)` in a new figure.

(h) Now fit a polynomial regression model that predicts $Y$ using $X$ and $X^2$. Is there any evidence that the quadratic term improves the model fit? Explain your answer.

(i) Repeat (a) to (g) with different noise levels by sampling from $N(0, \sigma^2)$ with $\sigma \in \{40, 60, 100\}$ to generate different $\epsilon$. Describe your results.

(j) Repeat (a) to (g) with different amounts of samples: $N \in \{10, 100, 1000\}$. Describe your results.

## 2. Simple Linear Regression: Real World Problem

This exercise involves the use of simple linear regression on the `carsmall` data set.

(a) Load the `carsmall` data set as a `table` object. Use the `fitlm()` function to perform a simple linear regression with `mpg` as the response and `horsepower` as the predictor. Use the `disp()` function to print the results. Comment on the output. For example:

  (i) Is there a relationship between the predictor and the response?

  (ii) How strong is the relationship between the predictor and the response?

  (iii) Is the relationship between the predictor and the response positive or negative?

  (iv) What is the predicted `mpg` associated with a `horsepower` of 98? What are the associated 95 % confidence and prediction intervals?

(b) Plot the response and the predictor. Use the `plot()` function to display the least squares regression line and the confidence bounds.

(c) Use the `plotResiduals()` and `plotDiagnostics()` functions to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

## 3. Multiple Linear Regression: Real World Problem

This exercise involves the use of multiple linear regression on the `carsmall` data set.

(a) Load the `carsmall` data set as a `table` object and remove the non numeric variables. Use the `corrplot()` function to produce a correlation matrix plot with all remaining variables. What information does the figure provide?

(b) Use the `fitlm()` function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the `disp()` function to print the results. Comment on the output. For instance:

  (i) Is there a relationship between the predictors and the response?

  (ii) Which predictors appear to have a statistically significant relationship to the response? Estimate a model with only `Model_Year` and `Weight` as predictors and compare the results.

  (iii) What does the coefficient for the year variable suggest?

(c) Use the `plotResiduals()` and `plotDiagnostics()` function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

(d) Fit a regression model with interaction effects using the `interaction` argument. Does any interaction appear to be statistically significant?

(e) Try non-linear transformations of the variables, such as $X^2$. Comment on your findings.