First Project

Your first project assignment is from the book. Individually, you should solve the following questions using **Linear Regression**: Exercises **3.7.8** and **3.7.9**. You should submit your solution as **a report together with the Python codes** that you used. (**Deadline: 06.03.2025, Perşembe**)

3.7.8

8. This question involves the use of simple linear regression on the `Auto` data set.

   (a) Use the `sm.OLS()` function to perform a simple linear regression with `mpg` as the response and `horsepower` as the predictor. Use the `summarize()` 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 in a new set of axes `ax`. Use the `ax.axline()` method or the `abline()` function defined in the lab to display the least squares regression line.

   (c) Produce some of diagnostic plots of the least squares regression fit as described in the lab. Comment on any problems you see with the fit.

3.7.9

9. This question involves the use of multiple linear regression on the `Auto` data set.

   (a) Produce a scatterplot matrix which includes all of the variables in the data set.

   (b) Compute the matrix of correlations between the variables using the `DataFrame.corr()` method. `.corr()`

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

      i. Is there a relationship between the predictors and the response? Use the `anova_lm()` function from `statsmodels` to answer this question.

      ii. Which predictors appear to have a statistically significant relationship to the response?

      iii. What does the coefficient for the `year` variable suggest?

   (d) Produce some of diagnostic plots of the linear regression fit as described in the lab. 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?

   (e) Fit some models with interactions as described in the lab. Do any interactions appear to be statistically significant?

   (f) Try a few different transformations of the variables, such as $\log(X)$, $\sqrt{X}$, X^2. Comment on your findings.