Introduction to ML
Assignment 6 and 7 (3 points)

SLR and MLR on Auto Dataset

This exercise should be solved in a jupyter notebook. You can use `statsmodels` or `scikit-learn` to solve it (However, `scikit-learn` does not provide `summarize` table, you need to build it yourself).

Please, read online documentation to understand how both libraries work.

- Scikit-Learn OLS function: LinearRegression.
- Stats Model Linear OLS function: LinearRegression

Please, be careful with your visualizations and use seaborn to make qualitative plots. Gallery.

Consider the `Auto` dataset and perform the following tasks:

1. 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:
   - Is there a relationship between the predictor and the response?
   - How strong is the relationship between the predictor and the response?
   - Is the relationship between the predictor and the response positive or negative?
   - What is the predicted `mpg` associated with a `horsepower` of 98?
   - What is the value of the coefficient of determination (R^2)? Interpret its meaning in this context.
   - According to the p-value of the predictor, is the `horsepower` statistically significant at the 5% level?
   - Compare the standard error of the estimate to the mean of `mpg`. What does this tell you about the model's accuracy?
2. Plot the response and the predictor in a new set of axes `ax`. Use the `ax.axline()` method or the `abline()` function to display the least squares regression line.

3. Compare the residuals Vs. Fitted values with a scatter plot. Use this plot to checks for linearity and homoscedasticity (constant variance of residuals). Interpretation: Residuals should be scattered randomly around zero. What does a curved pattern suggest?

4. Produce a scatterplot matrix which includes all of the variables in the data set. You can use 'PairGrid' for it: PairGrid

5. Compute the matrix of correlations between the variables using the `DataFrame.corr()` method. Find a fancy way to display this matrix.

6. 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:

   - Is there a relationship between the predictors and the response? Look at the R^2 value and the p-values of the coefficients to answer this question.
   - Which predictors appear to have a statistically significant relationship to the response? Focus on the p-values associated with each predictor.
   - What does the coefficient for the `year` variable suggest? Interpret its sign and magnitude.

7. Produce a residual plot of the linear regression fit. Comment on any problems you see with the fit. Do your plot suggest any unusually large outliers? In your opinion, which of the MSE, RMSE and MAE would be more convenient to use?

8. Fit some models with interactions. Do any interactions appear to be statistically significant?

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

Linear regression assumptions

(a) What is the bias of an estimator ? Max two sentences.

Consider Multi-linear regression model :

$$y = X\beta + \epsilon$$

(b) Show that the OLS estimator is unbiased.

(c) What is the **homo**skedasticity assumption ? Give the maths and intuition.

(d) Derive the variance of $\hat{\beta}$ under **homo**skedasticity assumption.

(e) If the zero conditional mean and **homo**skedasticity hold, how are the residuals distributed ?

Conceptual questions

1. Which metric uses squared errors to more heavily penalize large deviations?

   A. Mean Absolute Error
   B. Median Absolute Error
   C. Adjusted R^2
   D. Mean Squared Error

2. In a residual plot, a systematic increasing or decreasing trend typically indicates:

   A. Heteroscedasticity
   B. An uncaptured non-linear relationship
   C. A well-fitted model
   D. Absence of feature interactions

3. What does MAE measure?

   A. The average of squared errors
   B. The average of absolute errors
   C. The square root of average errors
   D. The sum of squared errors

4. What is the advantage of using RMSE over MSE?

   A. It is less sensitive to outliers
   B. It does not require a square root
   C. It is always lower than MAE
   D. It is in the same units as the target variable calculation

5. A feature interaction means that:

   A. The two features are statistically correlated, indicating multicollinearity
   B. The model's performance improves only if both features are included as a product term
   C. The features must be scaled or standardized before creating the interaction term to avoid numerical instability
   D. The relationship between one feature and the target variable changes across different levels of another feature

6. If the residuals of a model are normally distributed around zero with no apparent pattern, this suggests:

A. The model is biased

B. The model assumptions are not satisfied

C. The model assumptions are generally satisfied

D. The explanatory variables are redundant

7. Which of the following metrics is the most robust to outliers?

   A. MSE

   B. RMSE

   C. R^2

   D. MAE

8. A residual is defined as:

   A. The difference between the predicted and actual value

   B. The square of the prediction error

   C. The difference between the observed value and the mean of the target variable

   D. The squared difference between the observed and predicted values

9. Which metric is expressed in the same units as the target variable?

   A. MSE

   B. RMSE

   C. R^2

   D. Adjusted R^2

10. You evaluate a regression model to predict a price in euro and obtain and obtain: MAE = 3.0. Which interpretation is correct?

    A. On average, the model's predictions deviate by 3 euros (in absolute value) from the actual values.

    B. The model's predictions constantly deviate by 3 euros (in absolute value) from the actual values.

    C. The mean of the squared errors of the model is 3 euros.

    D. The model systematically overestimates the actual values by 3 euros.

11. You evaluate a regression model to predict a price in euro and obtain: MSE = 25.0 Which interpretation is correct?

    A. On average, the model's errors are 25 euros.

    B. The square root of the mean of the squared errors is 25 euros.

    C. The model has a bias of 25 euros in its predictions.

D. The mean of the squared errors of the model is 25 square euros, indicating sensitivity to large errors.

12. You evaluate a regression model to predict a price in euro and obtain and obtain: RMSE = 4.0. Which interpretation is correct?

    A. On average, the absolute errors of the model are 4 euros.

    B. The mean of the squared errors is 4 square euros.

    C. The square root of the mean of the squared errors is 4 euros, giving more weight to large errors than MAE.

    D. The model systematically underestimates the actual values by 4 euros.

13. Which plot is most suitable for visualizing the distribution of residuals?

    A. Box plot

    B. Scatter plot of explanatory variables

    C. Histogram of coefficients

    D. Residuals vs. fitted values plot

14. What does a U-shaped pattern in a residual plot typically indicate?

    A. The model is well-specified

    B. There is a non-linear relationship not captured by the model

    C. The model is overfitted

    D. The features are highly correlated

15. What is the bias of an estimator?

    (a) The difference between the estimator's expected value and the true parameter value

    (b) The variance of the estimator

    (c) The mean squared error of the estimator

    (d) The estimator's sensitivity to outliers

16. Under which condition is the Ordinary Least Squares (OLS) estimator unbiased?

    (a) If the residuals are normally distributed

    (b) If the homoskedasticity assumption holds and the model is correctly specified

    (c) If the X matrix is orthogonal

    (d) If the conditional expectation of errors $E[\epsilon|X] = 0$

17. Which assumption is violated if the variance of the residuals is not constant?

    (a) Normality

    (b) Homoskedasticity

    (c) Linearity

    (d) Exogeneity

18. What does the acronym BLUE stand for in the context of the OLS estimator?

    (a) Biased but Linear Useful Estimator

    (b) Basic Linear Uncorrelated Estimator

    (c) Best Linear Unbiased Estimator

    (d) Best Limited Unbiased Error

19. What is the effect of heteroskedasticity on the OLS estimator?

    (a) It becomes biased

    (b) It remains unbiased but loses its efficiency (no longer has minimum variance)

    (c) It becomes inconsistent

    (d) It no longer follows a normal distribution