# FDA - Assignment 01

# Yashal Jamil – 26591

# Linear Regression Salary

1. ***How can we interpret the coefficient of "years of experience"? What does a positive coefficient indicate?***

The coefficient of "years of experience" represents the expected change in salary for each additional year of experience. A positive coefficient means that as experience increases, salary also increases.

1. **Why is it important to check the distribution of salary and years of experience before applying regression?**

Before applying regression, it is essential to examine the distribution of salary and years of experience to ensure the model assumptions hold. If the data is skewed, contains extreme values, or does not follow a linear pattern, it can distort the results and affect the accuracy of predictions.

1. **What happens if we use all the data for training and none for testing? How does it impact model evaluation?**

Using all data for training without reserving any for testing prevents evaluating the model’s ability to generalize. This can cause overfitting, where the model excels on training data but fails on new observations. Without a test set, it may learn patterns specific to the training data rather than general relationships, making future predictions unreliable.

1. **What does it mean if the model's predicted salaries are consistently lower or higher than actual salaries?**

It indicates systematic bias. This could mean that an important factor influencing salary is missing from the model, or that the relationship between experience and salary is not purely linear.

1. **How does increasing or decreasing the training dataset size affect model performance?**

Increasing the training dataset size generally leads to a more robust and accurate model, improves model performance by making predictions more accurate and reducing overfitting. Decreasing the dataset leads to high variance and poor predictions. However accurate data is always better than more data.

1. **If another variable, such as "certifications," is added to the dataset, how might it impact the regression results?**

Including certifications can enhance the regression model if it significantly impacts salary. However, if certifications strongly correlate with years of experience, it may lead to multicollinearity.

1. **If the dataset contained outliers (e.g., a person with 50 years of experience but a very low salary), how would that affect the model?**

Outliers can significantly impact the regression model by skewing the line of best fit, leading to misleading coefficients and inaccurate predictions. They increase errors, reduce overall model accuracy, and violate key regression assumptions such as normality and homoscedasticity.

1. **If the dataset only contains a small number of observations, what problems might arise when training the model?**

A small dataset increases variance and the risk of overfitting, making predictions unreliable. It also leads to biased estimates and weak statistical conclusions, limiting the model's generalizability for real-world applications.

1. **How does the LinearRegression().fit() function work in training the model?**

This function works by finding the best-fit line that minimizes the difference between actual and predicted salaries. It does this using the Ordinary Least Squares (OLS) method, which calculates regression coefficients in a way that minimizes the sum of squared residuals. values

1. **If the dataset had outliers (e.g., a CEO with 40 years of experience earning $1M), how would that affect the model?**

If the dataset includes a CEO with 40 years of experience earning $1 million, the regression model may be heavily influenced by this extreme value. Since linear regression minimizes squared errors, high salaries can pull the regression line upward, potentially leading to an overestimation of salaries for other individuals with high experience.

1. **If the company wanted to predict salary for interns (0 years experience), would this model still be reliable? Why or why not?**

The model may not reliably predict salaries for interns (0 years of experience) as it is based on the experience-salary relationship. Since interns are not included in the training data, predicting their salaries would require extrapolation, which can lead to inaccuracies.

# Multi Linear Regression

1. **How can this model help the company decide how much to invest in RCD, Administration, and Marketing?**

The model helps the company decide investment levels in R&D, Administration, and Marketing by identifying which category has the strongest correlation with profit, helping the company allocate its budget to maximize returns.

1. **What does the mean and standard deviation of RCD Spend suggest about the dataset?**

The mean R&D Spend is $70,543.76, and the standard deviation is $43,042.35, indicating high variability in R&D investments among startups, showing a high spread around the mean, implying inconsistent spending patterns.

1. **The Marketing Spend has a large range (difference between min and max values). How might this impact our model?**

A large range in Marketing Spend (from -185,350 to 682,176) suggests the presence of extreme values or outliers. This can skew the regression model, making predictions less reliable.

1. **If the dataset had categorical features like Startup Industry, how would we handle them in this model?**

Categorical features like Startup Industry can be handled by encoding them into numerical values like 1& 0.

1. **What does the correlation matrix tell us about which spending category influences profit the most?**

R&D.

1. **Why do we use train\_test\_split() before fitting the model? What is the default split ratio?**

We use train\_test\_split() to divide the data into training and testing sets, ensuring the model learns patterns without overfitting. The default split ratio is 75% training and 25% testing.

1. **What happens if we remove Administration spending from the model? How would it affect predictions?**

It would have minimal impact on predictions, as its low correlation coefficient (0.0056) suggests almost no relationship with profit.

1. **If the test set accuracy is much lower than the training set accuracy, what might be the reason?**

If test accuracy is much lower than training accuracy, the model is likely overfitting, meaning it memorized training data instead of learning general trends.

1. **If a new startup has zero spending in Marketing but high RCD and Administration spending, would this model still make accurate predictions?**

Yes, the model can still make accurate predictions since R&D Spend has a strong correlation with Profit (0.87) and a high coefficient (0.8451), making it the key driver. Marketing Spend, though moderately correlated (0.37), has a low coefficient (0.0159), so its absence may slightly reduce accuracy but not drastically affect predictions.

1. **The company wants to maximize profit. Should they focus more on increasing RCD spending or Marketing spending based on the model’s results?**

R&D spending.

1. **Could this model be used to predict the profit of a startup in a different country? Why or why not?**

This model may not reliably predict profit for a startup in a different country because the impact of R&D Spend, Administration, and Marketing Spend on profit can vary significantly across regions. Differences in market dynamics, cost structures, consumer preferences, and economic conditions could alter these relationships. Since the model is trained on data from one specific country, applying it elsewhere without recalibration could lead to misleading predictions.

1. **If you were advising this startup, what additional variables would you suggest adding to improve predictions?**

Monthly Sales Volume, Economic conditions, Customer Retention Rate, Number of Employees.

# Advanced Linear Regression

1. **What role do the coefficients (slopes) play in Linear Regression? How would you interpret them in the context of our dataset?**

The coefficients (slopes) in Linear Regression indicate how much the Performance Index changes when an independent variable increases by one unit while holding others constant. For example, a coefficient of 3 for Hours Studied means each additional hour increases performance by 3 points.

1. **What are the independent (predictor) and dependent (target) variables in this dataset?**

Independent variables are Hours Studied, Previous Scores, Extracurricular Activities, Sleep Hours, and Sample Question Papers Practiced. The dependent variable is

Performance Index.

1. **What happens when we increase the number of features (independent variables) in a Linear Regression model? How does it affect accuracy?**

Increasing the number of features in a Linear Regression model can improve accuracy if they are relevant. However, adding unnecessary or highly correlated variables can lead to overfitting, reducing performance on new data.

1. **What does the output summary of the model tell us about the statistical significance of each feature?**

The model summary provides statistical significance through p-values—a low p-value (typically <0.05) suggests that the variable significantly impacts student performance, while a high p-value suggests little to no effect.

1. **If we observe that the model has high training accuracy but low testing accuracy, what does that indicate?**

It indicates overfitting —the model memorizes training data but fails to generalize.