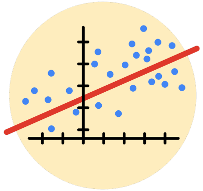
# Course Five

## Regression Analysis: Simplifying Complex Data Relationships



### Instructions

Use this PACE strategy document to record decisions and reflections as you work through this end-of-course project. As a reminder, this document is a resource that you can reference in the future, and a guide to help you consider responses and reflections posed at various points throughout projects.

### Course Project Recap

Regardless of which track you have chosen to complete, your goals for this project are:

- □ Complete the questions in the Course 5 PACE strategy document
- □ Answer the questions in the Jupyter notebook project file
- □ Build a multiple linear regression model
- □ Evaluate the model
- □ Create an executive summary for team members

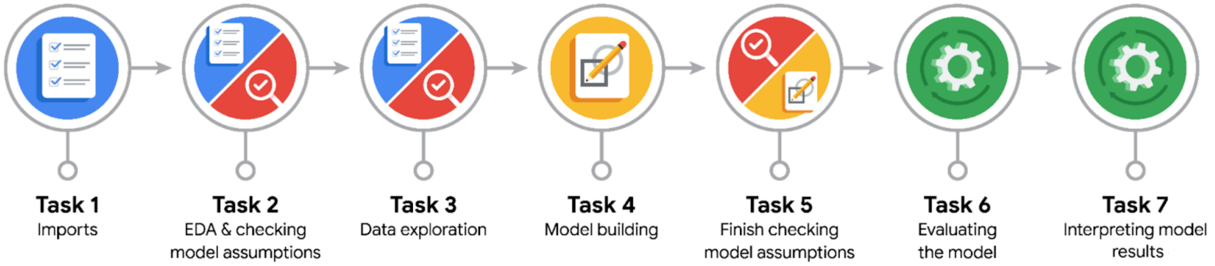
### Relevant Interview Questions

Completing the end-of-course project will empower you to respond to the following interview topics:

- Describe the steps you would take to run a regression-based analysis
- List and describe the critical assumptions of linear regression
- What is the primary difference between $R^2$ and adjusted $R^2$?
- How do you interpret a Q-Q plot in a linear regression model?
- What is the bias-variance tradeoff? How does it relate to building a multiple linear regression model? Consider variable selection and adjusted $R^2$.

## Reference Guide

This project has seven tasks; the visual below identifies how the stages of PACE are incorporated across those tasks.



## Data Project Questions & Considerations

### PACE: Plan Stage

- Who are your external stakeholders for this project?

The New York City TLC team members.

- What are you trying to solve or accomplish?

This project aims to develop a multiple linear regression to predict taxi ride fares through variables.

- What are your initial observations when you explore the data?

Looking through the whole dataset, the total columns,  and understanding every single variable. Checking if there are missing values in any variable, the correct format, and data types.

- What resources do you find yourself using as you complete this stage?

The New York City TLC data and Jupyter Notebook.

## PACE: Analyze Stage

- What are some purposes of EDA before constructing a multiple linear regression model?

Purpose of conducting EDA:

- Getting a general and in-depth understanding of the dataset and variables.
- Descriptive stats are useful such as mean, mode, standard deviation, quartile, etc.
- Other tasks such as handling missing values, outliers, formatting, etc. on variables.
- Checking model assumptions is important to determine if the chosen model will be a good candidate.

- Do you have any ethical considerations at this stage?

There might be concern about having negative values, and outliers in some variables that need to be immediately addressed.

## PACE: Construct Stage

- Do you notice anything odd?

One thing noticeable is the r-squared value between training & test sets. While the training set has an r-squared of 84%, the test set has a value of 87%. This might account for data leakage problem.

- Can you improve it? Is there anything you would change about the model?

We could account for some variables to be more influential on the target variable. Additionally, we might address the problem stated above, which leads to the data leakage problem while training the model, as well as address for model's assumptions.

- What resources do you find yourself using as you complete this stage?

New York TLC dataset and Jupyter Notebook for Python codes.

## PACE: Execute Stage

- What key insights emerged from your model(s)?

- Comparing metrics (R-squared, MAE, RMSE) after developing the model on both training and test sets shows a bit of difference. Metrics like MAE & RMSE on the test set are better than the training set since we account for the smallest values (2.13 vs. 2.18, and 3.78 & 4.23 respectively).
- However, the r-squared is higher on the test set (86.8%) vs. training (84%). This number shows the proportion of variance in `fare_amount` can be explained by the predictor variables. Since this value is pretty high on the test set, it might indicate a data leakage problem.
- Most of the assumptions met the model's condition, except we accounted for variables (`mean_distance` & `mean_duration`). These are highly correlated to the `fare_amount`, but they are also correlated to each other (multicollinearity).
- In this model, `mean_distance` is the most influential variable to the `fare_amount`. Specifically, if we increase every 3.57 miles traveled, we expect the fare amount will increase by $7.13 on average. On reduced, if we increase every one mile traveled, the fare amount will increase by $2.00 on average.

- What business recommendations do you propose based on the models built?

Since the model archives pretty well in results, we might recommend the New York TLC deploy and develop an app to predict the fare of taxi rides based on these results and findings.

However, since this just was the first version of the model, and also accounted for some problems. Thus, our Automatidata team might choose to account for several model options to determine if they can achieve and address higher results and performance than the multiple regression one.

- To interpret model results, why is it important to interpret the beta coefficients?

It's important to interpret the beta coefficients to understand the effect of predictor variables on the target variable. Specifically, the increase/decrease of the target variable will be based on the predictor ones by one unit of increasing, and how much influence there is.

- What potential recommendations would you make?

- Do you think your model could be improved? Why or why not? How?

- What business/organizational recommendations would you propose based on the models built?

- Given what you know about the data and the models you were using, what other questions could you address for the team?

We might want to request the New York TLC to collect more taxi trip data since it might account for boosting the model's performance. Besides, the multiple regression addresses some issues (data leakage, multicollinearity, etc) that the team needs to carefully consider. Since this model won't be the last, the team can decide to develop some models that might perform and yield higher achievement results.

- Do you have any ethical considerations at this stage?