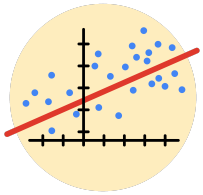
Course Five

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~~

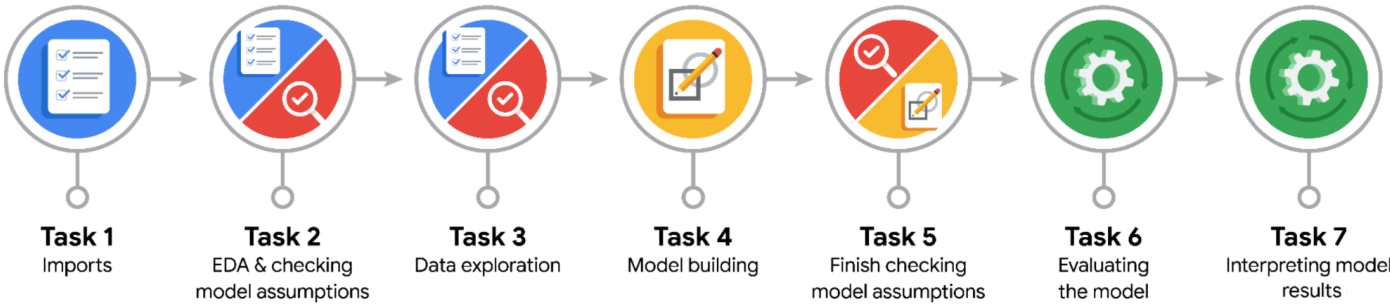
Relevant Interview Questions

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?

New York City Taxi & Limousine Commission (TLC).

- What are you trying to solve or accomplish?

Build a regression model for ride fares based on a variety of variables.

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

Some variables contain outliers.
It has value with 30 mile trip and that's a straight line.

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

Packages for numerics and dataframes.
Packages for visualization.
Packages for date conversions.

PACE: Analyze Stage

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

Understanding data and distribution.
Check outlier, extreme value, and missing value that can impact model.

- Do you have any ethical considerations in this stage?

Check outlier, extreme value, and missing value that can impact model.

PACE: Construct Stage

- Do you notice anything odd?

No. Because using linear regression model.

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

No. The model performance is high on both training and test sets.

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

Packages for OLS, MLR, confusion matrix.

PACE: Execute Stage

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

mean_duration and mean_distance are both highly correlated with the target variable of fare_amount They're also both correlated with each other, with a Pearson correlation of 0.87.

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

The model provides a generally strong and reliable fare prediction that can be used in downstream modeling efforts.

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

Helps identify the values that are currently in use.

- What potential recommendations would you make?

The New York City Taxi and Limousine commission can use these findings to create an app that allows users (TLC riders) to see the estimated fare before their ride begins.

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

No. The model performance is high on both training and test sets.

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

Could improve more model metrics result ?

- Do you have any ethical considerations at this stage?

Recommend based on finding or on model result.