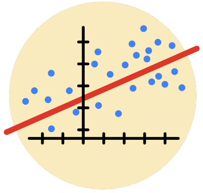
Course Five

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

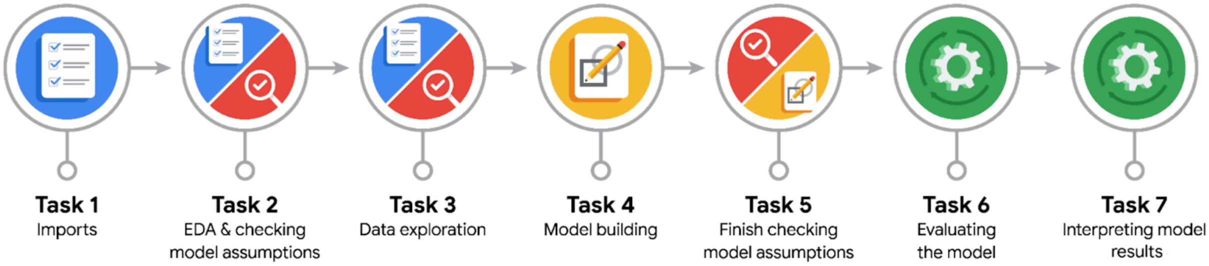
Relevant Interview Questions

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?

Juliana Soto, Finance and Administration Department Head

Titus Nelson, Operations Manager

- What are you trying to solve or accomplish?

We want to estimate the fare before the customer requests a taxi or limousine.

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

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

Python and its function to create linear regression

PACE: Analyze Stage

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

It's important to perform Exploratory Data Analysis to know the variable we-re working and categorize it in continuous or categorical. Also, it's important to determine any missing values that can affect the regression model

- Do you have any ethical considerations at this stage?

Yes, you have to consider if the missing values or any values are relevant to your work or not. Thus, it's important to not work over the original file, instead, you have to create a copy if you decide to remove some arrows.

PACE: Construct Stage

- Do you notice anything odd?

When analyzing our data, we found that there were odds values, such as negatives values in the fare amount, fare amount of almost $1000, and trips with duration of 0 min.

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

We could improved the model by creating three columns to see the relation with fare amount, such as mean duration, mean distance and rush hour. Since it's considered that fare amount is the one that correlates better with our data.

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

PACE: Execute Stage

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

There is R^2 of 83.98%, which is a good standar when creating a regression model. Also, the histograms of the residuals resemble the bell shape, which is a good indicator from our model. Finally, the train data and the test data showed a good relation, only variations from fare amount above 50

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

The TLC can use this model to predict or calculate the fare amount for the cab. Since the expected and actual values fit into the model, Thus, Automatidata can create an application based on this model.

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

The beta coefficients are important because it shows the variables that were used in the model but it's important to determine which variable impacts more on our model.

- What potential recommendations would you make?

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

Yes, there were three variables which had to be removed before the model because they were inconsistent. So, the model could be improved if we ask or collect more data with consistent results.

- 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?

- Do you have any ethical considerations at this stage?