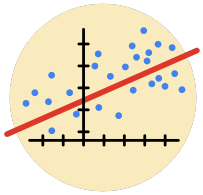
Course Five

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

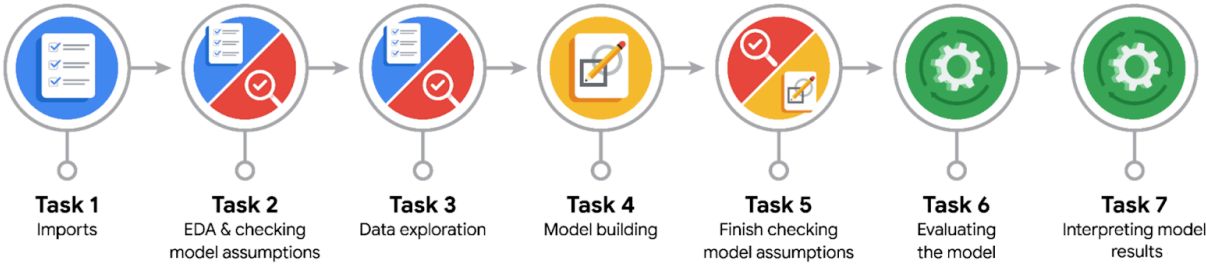
Relevant Interview Questions

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 external stakeholders would be the New York City TLC.

- What are you trying to solve or accomplish?

We are trying to determine which variables (duration, distance, etc.) would affect the fare prices.

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

When I explored the data I see some outliers and some values that are nonsensical, and I would need to impute them all.

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

Mainly the Python libraries for exploratory data analysis and the data dictionary.

PACE: Analyze Stage

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

Firstly, to check for outliers, because they would skew the data a lot. Second, to check values that are nonsensical such as negative distances and negative prices. Third, to check for multicollinearity as it's a no-no in multiple linear regression.

- Do you have any ethical considerations at this stage?

Ethical considerations would include impute the data instead of simply deleting them if they have outliers.

PACE: Construct Stage

- Do you notice anything odd?

Yes I do. Negative prices and durations are very odd.

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

Imputation would be the technique to get rid of the odd aspects.

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

Exploratory data analysis tools and correlation matrices.

PACE: Execute Stage

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

Key insights emerged that the longer the duration and the longer the distance, the higher the fares.

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

Business recommendations would include advertise the riders to use the cab on for longer durations and/or longer distances, that way their fare prices would be higher and it will be beneficial for the business.

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

Because they are the coefficients that a 1-point change in the independent variable or in this case the scaled independent variables affect the dependent variable by how many.

- What potential recommendations would you make?

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

I mean, 87% accuracy (R^2) of around 0.87 is not bad, but not perfect either. However in statistics there is nothing called “perfect” so the model is pretty successful overall.

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

Again, the business recommendations would be put up advertisements for longer trips.

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

Other questions could include, which other models are better? And which other questions do we need to answer?

- Do you have any ethical considerations at this stage?

Ethical consideration is not to overdo the recommendation, as people have the right to take how long they want.