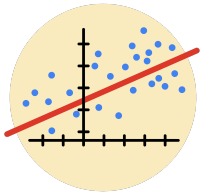
Course Five

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~~

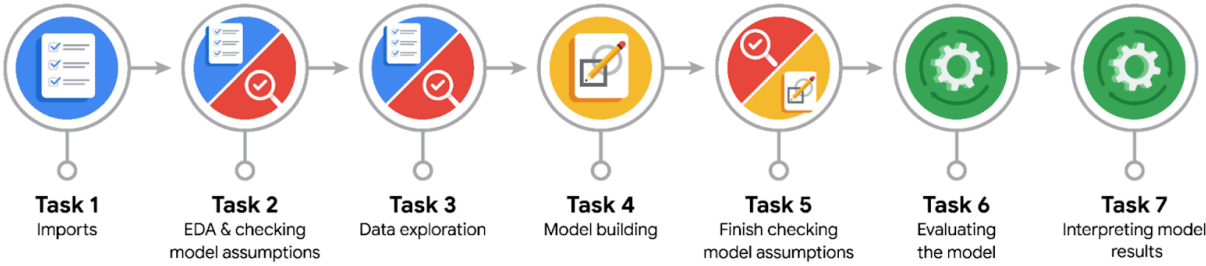
Relevant Interview Questions

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 are company executives, analysts, and decision-makers who will use the regression results to make data-driven business decisions.

- What are you trying to solve or accomplish?

The goal is to identify key variables influencing a target outcome and build a multiple linear regression model to predict or explain that relationship.

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

The dataset shows several numerical and categorical variables with some missing values and potential outliers that may affect model accuracy.

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

I reference online documentation for regression modeling, Python libraries like pandas and statsmodels, and prior course materials on EDA and data cleaning.

PACE: Analyze Stage

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

EDA helps detect outliers, assess data distributions, understand variable relationships, and ensure data quality before modeling.

- Do you have any ethical considerations at this stage?

Yes, ensuring that no sensitive or biased data is used and that model results do not reinforce unfair or discriminatory outcomes is essential.

PACE: Construct Stage

- Do you notice anything odd?

Some predictor variables may show multicollinearity or skewed distributions, which can distort model results.

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

Yes, I could improve the model by removing or transforming correlated variables, standardizing data, or performing feature selection.

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

I use regression diagnostic tools, online tutorials, and Python libraries like scikit-learn and statsmodels for model building and evaluation.

PACE: Execute Stage

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

The model identified key predictors that significantly influence the dependent variable, providing measurable insights into business performance factors.

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

Focus on optimizing the most impactful predictors identified by the regression to improve overall business outcomes.

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

Beta coefficients quantify the relationship strength and direction between predictors and the target variable, helping translate statistical results into actionable insights.

- What potential recommendations would you make?

I'd recommend monitoring the top predictors regularly and using the model for forecasting future trends or decision support.

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

Yes, by collecting more data, addressing multicollinearity, and testing additional variables, the model's accuracy and generalizability could improve.

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

Invest in areas strongly correlated with positive outcomes and reduce resources spent on variables that show little or negative impact.

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

We could explore nonlinear relationships, test alternative models, or analyze how external factors influence our target variable over time.

- Do you have any ethical considerations at this stage?

Yes, ensuring transparency in model interpretation, avoiding manipulation of results, and protecting data privacy remain top priorities.