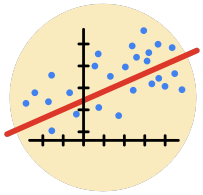
Course Five

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

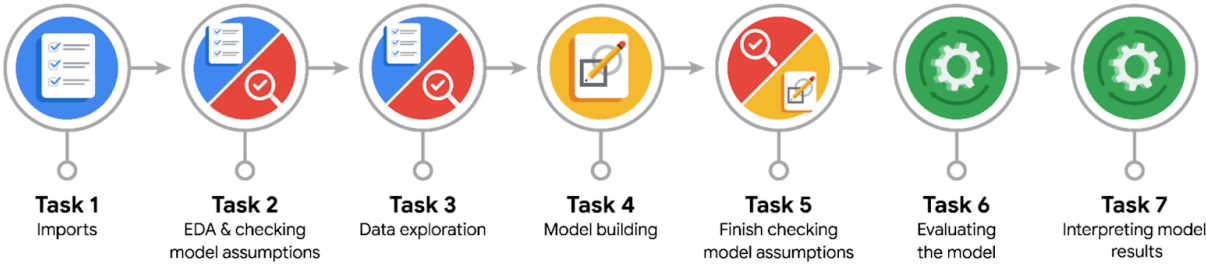
Relevant Interview Questions

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 for this project are the **New York City Taxi and Limousine Commission (NYC TLC)**. They have contracted Automatidata to build a multiple linear regression model that predicts taxi fares.

- What are you trying to solve or accomplish?

The goal is to develop a **multiple linear regression model** to predict taxi fares based on various features in the dataset, such as trip distance, time of day, and payment method. This will help NYC TLC provide accurate fare estimates for taxi rides.

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

Initial exploration of the data reveals the presence of **outliers**, **missing values**, and **unusual values** (e.g., trips with zero distance). The dataset also includes a mix of **categorical and numerical variables** that need to be processed properly before modeling.

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

- **Pandas** for data manipulation and exploration.
- **Matplotlib/Seaborn** for initial visualization and identifying patterns in the data.
- **Sklearn's StandardScaler and OneHotEncoder** for preprocessing.

PACE: Analyze Stage

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

Identifying patterns, correlations, and potential outliers.

Understanding the distribution of data and checking assumptions (e.g., linearity, normality, homoscedasticity).

Determining if any variables need transformation (e.g., log transformation) or encoding (e.g., categorical variables via one-hot encoding).

- Do you have any ethical considerations at this stage?

Ensuring that the data does not reinforce any **biases** (e.g., if certain neighborhoods have higher fares, ensure the model doesn't reinforce inequality).

Handling **personal data responsibly** (if included), ensuring compliance with data privacy laws.

Transparency in feature selection and model interpretation to avoid decisions that disproportionately impact any group.

PACE: Construct Stage

- Do you notice anything odd?

Outliers in trip distance and fare amount that could skew model performance.

Some features, like `duration`, are highly correlated with the target variable, which is expected.

The presence of **categorical features** like day of the week, which need to be properly encoded before constructing the model.

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

Handling outliers through appropriate imputation or removal.

Trying different **feature engineering** approaches, such as creating interaction terms between variables.

Testing with different models (e.g., Ridge or Lasso regression) to compare performance and check for overfitting.

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

- **Sklearn's LinearRegression, OneHotEncoder, StandardScaler**, and the pipeline for combining preprocessing steps with model training.
- **Cross-validation** for evaluating model performance.

PACE: Execute Stage

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

The features with the largest impact on the predicted taxi fare are **trip distance**, **duration**, **mean distance**, and **tip amount**.

The model provides reasonable predictions with minimal error, and most residuals are distributed around zero.

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

NYC TLC can use the model to improve the accuracy of their **fare estimation system**, providing better price transparency for riders.

They can also explore **incentives for longer rides** or rides during specific times (e.g., during rush hours) based on fare estimation patterns.

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

The beta coefficients tell us the **magnitude and direction** of the impact each feature has on the target variable. Understanding this helps us interpret which factors drive higher or lower fares and how sensitive the fare is to changes in each feature.

- What potential recommendations would you make?

Implement the fare prediction model into TLC's systems, providing fare estimates to riders before rides are taken. Further optimize the model by periodically **retraining** it with more recent data, ensuring it adapts to any changes in riding patterns or fare structures.

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

Handling multicollinearity between variables. Incorporating more **complex features** like weather conditions or traffic levels, which could affect ride duration and fares. Testing other models like **random forests** or **gradient boosting** for better accuracy.

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

Use the model for better **resource allocation** (e.g., optimizing the distribution of taxis during peak hours). Analyze fare patterns to develop more **targeted promotions** for customers based on high-demand areas or times.

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

How would external factors like **traffic congestion** or **public transportation availability** affect fare predictions? Can we use the data to predict **ride demand** at specific times or locations?

- Do you have any ethical considerations at this stage?

Yes, ensuring that the fare model doesn't disproportionately increase fares for **disadvantaged communities** or reinforce any biases.

Transparency in how the fare estimates are communicated to customers is essential to maintain trust.