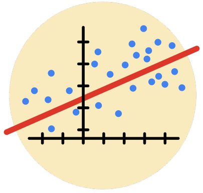
Course Five

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

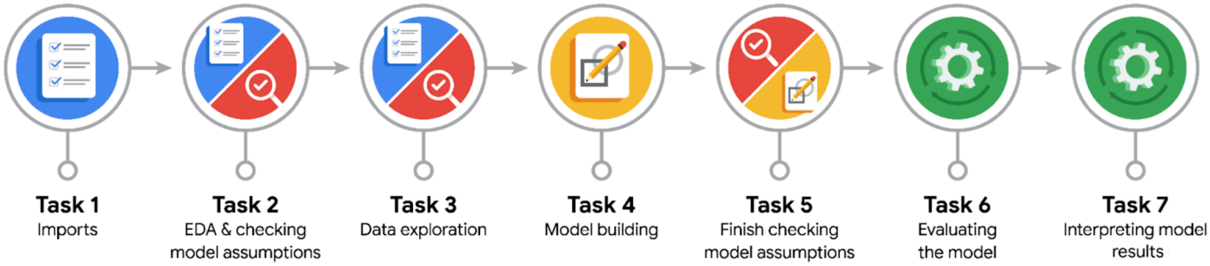
Relevant Interview Questions

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?

For this project, the external stakeholders are the New York City Taxi and Limousine Commission (TLC). They are the agency responsible for licensing and regulating taxi cabs and for-hire vehicles in New York City. Our findings and the regression model will be used by them to potentially estimate fares, understand fare drivers, and inform policy decisions.

- What are you trying to solve or accomplish?

We are trying to build a multiple linear regression model that can accurately predict taxi fares before a ride begins. This involves identifying the key factors (independent variables) in the provided TLC data that significantly influence the total fare amount (dependent variable) and quantifying those relationships. A reliable fare prediction model can enhance transparency for riders and potentially inform strategies for drivers and the TLC.

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

During initial data exploration, I observed several potential relationships. Trip distance appears to have a strong positive correlation with the total fare amount. The pickup and dropoff locations (PULocationID and DOLocationID) likely influence the fare due to varying distances and zone-based pricing. Factors like passenger count, payment type, and the presence of tolls or tips also seem to contribute to the total fare. The datetime features (pickup and dropoff) might reveal patterns related to time of day or day of the week affecting fares. There might also be outliers or unusual values that need investigation.

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

During the Plan stage, I am primarily using the project instructions, the data dictionary to understand the variables, and potentially reviewing materials from previous courses on data exploration and understanding business problems. I might also refer to online resources or documentation about the TLC and their fare structure to gain domain knowledge.

PACE: Analyze Stage

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

1. Outliers and extreme data values can significantly impact linear regression equations. After visualizing data, make a plan for addressing outliers by droping rows, substituting extreme data with average data, and/or removing data values greater than 3 standard deviations.
2. EDA activities also include identifying missing data to help the analyst make decisions on their exclusion or inclusion by substituting values with data set means, medians, and other similar methods.
3. It is important to check for things like multicolinearity between predictor variables, as well to understand their distributions, as this will help me decide what statistical inferences can be made from the model and which ones cannot.
4. Additionally, it can be useful to engineer new features by multiplying variables together or taking the difference from one variable to another. For example, in this dataset you can create a duration variable by subtracting tpep_dropoff from tpep_pickup time.

- Do you have any ethical considerations at this stage?

Avoiding bias in variable selection: Ensuring that the chosen predictor variables are relevant and do not inadvertently discriminate against certain groups (although the provided data might not directly contain sensitive demographic information beyond passenger count).

PACE: Construct Stage

- Do you notice anything odd?

No, the model performance is high on both training and test sets, suggesting that there is little bias in and model and that the model is not overfit. In fact, the test scores were even better than the training scores.

For the test data, an R2 of 0.868 means that 86.8% of the variance in the `fare_amount` variable is described by the model.

The mean absolute error is informative here because, for the purposes of the model, an error of two is not more than twice as bad as an error of one.

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

The model results are great, and it can give great predictions for new data (predict the fair amount before a given taxi ride). However, there is a place for improvement because some independent variables are highly correlated with each other, and this can be bad for drawing statistical inferences about the data from the model. As a result, the model is good for predicting, but it would be wrong to make assumptions for the independent impact of each X feature on the Fare amount.

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

During the Construct stage, I am heavily relying on statistical libraries in Python (like scikit-learn and statsmodels) to build and evaluate the regression model. I would also refer to documentation and tutorials for these libraries, as well as statistical textbooks or online resources explaining regression modeling techniques, assumption checking, and model evaluation metrics.

PACE: Execute Stage

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

1. The model demonstrates strong predictive power for NYC taxi fares, achieving a high R-squared (0.868 on the test set), indicating it explains a large portion of the variance in fares.

2. Key drivers identified during EDA, such as trip_distance, duration (engineered feature), passenger_count, PULocationID, DOLocationID, and potentially total_amount components like tolls, are confirmed as significant predictors by the model.
3. The Mean Absolute Error (MAE) provides a practical measure of the average prediction error in dollars, indicating the typical accuracy users can expect.
4. While highly predictive, the potential presence of multicollinearity (as noted in the Construct stage) means that interpreting the precise independent impact of each individual variable (beta coefficient) on the fare amount requires caution. The model is better suited for prediction than for drawing definitive causal inferences about specific fare components in isolation.

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

1. Implement a Fare Estimation Tool: Deploy the model as a feature in a passenger-facing app or website to provide pre-ride fare estimates, enhancing transparency and customer experience.
2. Benchmarking and Analysis: The TLC can use the model internally to analyze fare patterns, identify potential anomalies or inconsistencies, and establish benchmarks for typical fares on different routes or times.
3. Inform Policy Discussions (with Caution): While precise coefficient interpretation is limited, the model can directionally inform discussions about the potential impact of changes to fare structures (e.g., distance rates, time rates) by simulating outcomes. However, any policy decisions should be backed by further analysis specifically designed to isolate causal effects if multicollinearity is significant.
4. Driver Resource: Provide access to the estimation tool for drivers to help them understand expected fares for different trips.

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

1. Beta coefficients (or regression coefficients) quantify the estimated relationship between each independent variable (predictor) and the dependent variable (fare amount), holding all other variables in the model constant.
2. They indicate the expected increase or decrease in the fare amount for a one-unit increase in the respective predictor. For example, the coefficient for trip_distance tells us the estimated change in fare for each additional mile travelled, assuming factors like duration, passenger count, etc., remain unchanged.
3. Interpreting coefficients helps identify which factors have the most substantial impact on the fare and the direction of that impact (positive or negative).

4. Crucially, this interpretation relies on the model assumptions being met, particularly low multicollinearity. If predictors are highly correlated, the coefficient estimates can be unstable and less reliable for isolating the independent effect of a single variable.

- What potential recommendations would you make?

1. Address Multicollinearity: If precise interpretation of individual predictors' effects is crucial for the TLC, further steps should be taken. This could involve:
2. Removing one or more variables from highly correlated pairs/groups.
3. Combining correlated variables into a single index or feature (like the duration feature already created).
4. Using techniques like Ridge or Lasso regression which can mitigate multicollinearity's impact on coefficients.
5. Incorporate Additional Data: Enhance model accuracy by incorporating external data sources like real-time traffic conditions, weather data, or specific event information (e.g., major concerts, street closures) that might influence trip times and fares.
6. Explore Non-Linearities: Investigate if relationships between predictors and the fare amount are non-linear (e.g., the impact of distance might lessen slightly for very long trips) and consider using transformations or non-linear models if appropriate.
7. Refine Feature Engineering: Experiment with more sophisticated time-based features (e.g., specific rush hour flags, weekend vs. weekday interactions) or location-based features (e.g., airport surcharges explicitly flagged, borough-to-borough interactions).

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

Yes, the model could likely be improved.

Why? While the R^2 is high, it stil has some potential, meaning some variance remains unexplained. More importantly, the potential multicollinearity issue limits the model's inferential capabilities, even if its predictive accuracy is good. Capturing more real-world complexities could further enhance accuracy and utility.

How?

1. Addressing Multicollinearity: As mentioned above (variable selection, feature engineering, regularization). This would improve the reliability of coefficient interpretations.
2. Adding Features: Incorporating external data (traffic, weather, events) is a primary avenue for improvement.

3. Model Complexity: Exploring non-linear models (e.g., Polynomial Regression, Gradient Boosting Machines) might capture complex interactions better than a purely linear model, potentially increasing predictive accuracy.
4. Outlier Treatment: Re-evaluating the strategy for handling outliers might yield improvements.
5. Location Granularity: Using more granular location data or specific zone interactions might capture fare nuances better.

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

1. Operationalize Prediction: Integrate the model into the TLC's operational workflow for fare estimation (e.g., via APIs for third-party apps, internal tools, public website).
2. Establish Monitoring & Retraining: Implement a system to continuously monitor the model's performance on new data. Set triggers for retraining the model periodically or when performance degrades significantly (due to changing traffic patterns, fare structures, etc.).
3. Data Governance: Use the insights from model building to inform data collection strategies. For example, prioritize acquiring reliable traffic or weather data feeds if they prove valuable.
4. Further Investigative Analysis: Use the model as a starting point for deeper dives. If the model flags certain routes or times as having unusually high prediction errors, commission specific analyses to understand the underlying reasons. Use the model to identify areas needing more targeted causal inference studies.

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

- Can we predict trip duration accurately using similar variables?
- What factors are most strongly correlated with tip amount (if available and reliable)? Could we build a separate model for tip prediction?
- How do fare predictions differ significantly between different payment types?
- Can we identify specific pickup/dropoff zone pairs that consistently deviate from predicted fares? Why might that be?
- How does time of day and day of week interact with location to influence fares beyond simple duration/distance?
- Can we model taxi demand (e.g., number of pickups) in different areas based on time, weather, and location features?
- Are there specific driver behaviors (if identifiable in data, though unlikely here) that correlate with fares deviating from the norm?

- Do you have any ethical considerations at this stage?

- Transparency in Deployment: When providing fare estimates to the public, clearly communicate that it is an estimate and provide the MAE or a likely range (e.g., predicted fare +/- $X) to manage expectations about accuracy.
- Fairness: Ensure the model doesn't systematically under/over-predict fares for specific locations or times that might correlate with demographic groups, potentially creating inequities if used for pricing or policy without checks. Monitor performance across different segments.
- Use for Benchmarking: If used to benchmark driver performance or identify anomalous fares, ensure the model is fair and accounts for factors outside a driver's control (e.g., unexpected traffic, passenger route requests). Avoid using it punitively without careful human review.
- Model Interpretability vs. Accuracy Trade-off: Be transparent with stakeholders (TLC) about the limitations in interpreting coefficients if multicollinearity wasn't fully resolved. Don't present coefficients as precise causal effects if the underlying assumptions are weak.
- Data Privacy: Continue to ensure that any deployment or further analysis respects data privacy regulations and anonymization standards, especially concerning location data.