Vital questions on regression

How many types of regression are there and what are they?

Type Use Case

Simple Linear Regression Single predictor variable

Multiple Linear Regression Multiple predictor variables

Polynomial Regression Nonlinear relationships

Ridge Regression High-dimensional data, multicollinearity

Lasso Regression Feature selection, sparse models

Elastic Net Regression Balances Lasso and Ridge regularization

Stepwise Regression Iterative model selection

Robust Regression Handles outliers

Quantile Regression Predicts conditional quantiles of the dependent variable

Bayesian Linear Regression Probabilistic modeling and parameter estimation

Basic Questions

What is Linear Regression? How is it used in data analysis?

Explain the equation of a simple linear regression model. What do the terms represent?

What assumptions does Linear Regression make? Can you name at least four?

What is the difference between simple and multiple linear regression?

Intermediate Questions

How do we evaluate the performance of a linear regression model? What metrics can we use?

What is multicollinearity? How does it affect a linear regression model?

What is the significance of the p-value in a regression model?

How do you handle categorical variables in linear regression?

What does the R-squared value represent? How does it differ from adjusted R-squared?

What is the role of feature scaling in linear regression? Is it necessary?

Advanced Questions

Explain the concept of regularization in linear regression. How do Lasso and Ridge regression differ?

What is the effect of outliers on a linear regression model? How can we address them?

What is the difference between underfitting and overfitting in linear regression? How can you prevent them?

How do you interpret the coefficients in a linear regression model?

If a linear regression model has high multicollinearity, what methods can you use to address it?

Scenario-Based Questions

Imagine you’ve built a linear regression model, and the residuals show a clear pattern. What does this indicate? How would you address it?

Your model’s R-squared value is very high, but it performs poorly on the test set. What might be the cause?

You’re using a linear regression model to predict housing prices, but the predictions are consistently lower than the actual prices. What could be the issue?

You have a dataset with 100 features, but only a subset of them are relevant. How would you select the best features for your linear regression model?

You’ve applied a linear regression model, but the coefficients for some features are unexpectedly negative. How would you investigate this?

What is linear regression?

Linear regression is predicting numerical data like sales per month for next 6 months,

depending on multiple numerical or categorical or ranked factors like product variant,

area of sales, target audience age range etc.,

How many types of regressions are there?

How would you encode categorical variables, such as "product variant" or "area of sales,"

to use them in a linear regression model?

I would use one hot encoding/ use mean of target variable per each category.

Target audience age range I would treat as a categorical variable,

because, it represents psychy of the customers more than age being a simple number.

Encoding Categorical Variables

One-Hot Encoding

Your suggestion to use one-hot encoding is correct and works well for variables like "product variant" or "area of sales." It creates binary columns for each category, ensuring no ordinal relationship is falsely inferred.

Example:

If "product variant" has values: A, B, C, it creates three binary columns: Variant\_A, Variant\_B, Variant\_C.

Pros:

Preserves the categorical nature.

Works well for algorithms assuming linear relationships, like linear regression.

Cons:

May lead to a high-dimensional dataset if there are many categories (a problem called the "curse of dimensionality").

Mean Target Encoding (Target-Based Encoding)

Using the mean of the target variable for each category, as you mentioned, is a great alternative, especially for high-cardinality categorical variables.

Example:

If "area of sales" has categories like Urban, Suburban, Rural, and the target (sales) averages are 500, 300, 200 respectively, you replace the categories with these means.

Pros:

Reduces dimensionality compared to one-hot encoding.

Can capture the relationship between the category and the target.

Cons:

Risk of data leakage if applied before splitting data into training and testing sets. To mitigate this, the mean encoding should be calculated only on the training data.

Target Audience Age Range

Your reasoning for treating "age range" as a categorical variable is excellent—this reflects a deeper understanding of the psychological segmentation behind the data.

Psychographic Factors:

Grouping age ranges aligns with customer behavior patterns (e.g., "18-24 might prioritize affordability, while 45-54 may prioritize luxury").

Encoding:

One-hot encoding works if the ranges are distinct and unrelated.

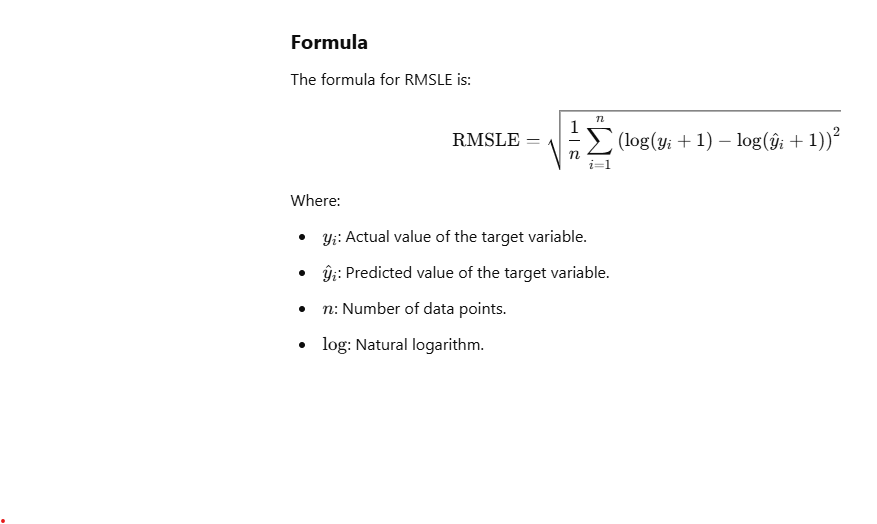
Alternatively, ordinal encoding could work if there’s an implied ranking, such as younger age groups being more tech-savvy.

What is RMSLE used in Regression with Insurance Data Kaggle dataset?

RMSLE is a metric used to evaluate the performance of regression models, particularly when the target variable spans multiple orders of magnitude. It measures the differences between the predicted and actual values after applying a logarithmic transformation.

Formula

The formula for RMSLE is:



log: Natural logarithm.

Step-by-Step Explanation

Logarithmic Transformation:

This makes RMSLE less sensitive to large absolute differences in values, focusing instead on relative differences.

Offset by 1:

is computed for each data point, amplifying larger deviations.

Mean and Root:

The mean of these squared differences is calculated.

The square root of this mean gives RMSLE, providing a scale similar to the original variable.

Key Characteristics of RMSLE

Relative Differences:

RMSLE penalizes under-predictions more than over-predictions because of the logarithmic transformation.

Sensitivity to Ratios:

It is more focused on the ratio between predicted and actual values rather than their absolute differences.

Robustness to Large Values:

Works well when the target variable has large values, as it reduces the impact of extreme differences.

Use Cases

Forecasting Problems:

Commonly used in demand or sales forecasting where actual values may vary significantly across observations.

Skewed Data:

Effective when the target variable spans multiple orders of magnitude or has a highly skewed distribution.

Penalty for Under-Predictions:

Preferred when underestimating values is more critical than overestimating them.