Regression Assignment - PwSkills Java + DSA

What is Simple Linear Regression?

Simple Linear Regression is a method to model the relationship between two continuous variables - one independent variable (X) and one dependent variable (Y). The model is expressed as: Y = mX + c + e.

What are the key assumptions of Simple Linear Regression?

1. Linearity
2. Independence
3. Homoscedasticity
4. Normality
5. No multicollinearity

What does the coefficient m represent in the equation Y = mX + c?

m (the slope) indicates the change in Y for a one-unit change in X.

What does the intercept c represent in the equation Y = mX + c?

c is the predicted value of Y when X is 0.

How do we calculate the slope m in Simple Linear Regression?

m = ((Xi - X)(Yi -)) / ((Xi - X)^2)

What is the purpose of the least squares method in Simple Linear Regression?

It minimizes the sum of squared differences between observed and predicted values.

How is the coefficient of determination (R^2) interpreted in Simple Linear Regression?

R^2 indicates the proportion of variance in Y explained by X. R^2 = 1 means perfect fit, R^2 = 0 means no explanatory power.

Regression Assignment - PwSkills Java + DSA

What is Multiple Linear Regression?

It uses two or more independent variables to predict a dependent variable. Y = b0 + b1X1 + b2X2 + ... + bnXn + e

What is the main difference between Simple and Multiple Linear Regression?

Simple uses 1 independent variable, Multiple uses 2 or more.

What are the key assumptions of Multiple Linear Regression?

1. Linearity
2. Independence
3. Homoscedasticity
4. Normality
5. No multicollinearity

What is heteroscedasticity, and how does it affect the results?

It means non-constant variance of residuals, which can lead to inefficient estimates.

How can you improve a model with high multicollinearity?

1. Remove correlated predictors
2. PCA
3. Ridge/Lasso regression
4. Combine variables

Techniques for transforming categorical variables?

1. One-hot encoding
2. Label encoding
3. Binary encoding
4. Ordinal encoding

Regression Assignment - PwSkills Java + DSA

What is the role of interaction terms?

They capture the combined effect of two variables, not captured individually.

Interpretation of intercept in Simple vs. Multiple Linear Regression?

In Simple LR: Y when X = 0. In Multiple LR: Y when all predictors = 0.

Significance of the slope?

It shows the effect of a 1-unit change in X on Y.

How does the intercept provide context?

It gives the baseline value of Y when predictors are 0.

Limitations of R^2?

1. Doesn't account for overfitting
2. Doesn't imply causation
3. Ignores complexity
4. Misleading in non-linear models

Interpret a large standard error for a coefficient?

Indicates high variability and low confidence in the estimate.

Identifying heteroscedasticity in residual plots?

Look for fanning or narrowing of residuals. It's important to address for valid inferences.

High R^2 but low adjusted R^2?

Indicates overfitting due to non-informative predictors.

Regression Assignment - PwSkills Java + DSA

Why scale variables?

For faster convergence, unbiased coefficient interpretation, and essential for regularization.

What is polynomial regression?

Regression using polynomial terms to fit non-linear data.

How does it differ from linear regression?

Linear fits a straight line, polynomial fits a curve.

When is it used?

When the relationship is non-linear but can be modeled with a polynomial.

General equation?

$Y = b0 + b1X + b2X^2 + ... + bnX^n + e$

Can it be applied to multiple variables?

Yes, called multivariate polynomial regression.

Limitations of polynomial regression?

1. Overfitting
2. Poor extrapolation
3. Computationally expensive
4. Collinearity

Evaluating model fit for polynomial degree?

1. Cross-validation
2. Adjusted R^2

3. AIC/BIC

4. Residual plots

Why is visualization important?

To detect overfitting/underfitting, understand model behavior, and validate curve fit.

How is it implemented in Python?

Use sklearn's PolynomialFeatures with LinearRegression. See code sample in full answer.