Regression

Assignment Questions

1. What is Simple Linear Regression?

Ans: Simple Linear Regression is a foundational analytics mechanism that quantifies the relationship between one independent variable and one dependent variable to forecast outcomes with an optimized straight-line fit.

2. What are the key assumptions of Simple Linear Regression?

Ans:

- **Linear linkage:** The dependent metric moves in a straight-line relationship with the independent variable.
- **Error neutrality:** The residuals are centered around zero without directional bias.
- **Constant variance:** The spread of errors remains consistently distributed across the prediction range (homoscedastic).
- **Error independence:** Each observation's error operates independently, with no sequential influence.
- **Normal error profile:** Residuals follow an approximately normal distribution, enabling reliable inference.

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

Ans: In plain terms, the m parameter operationalizes the rate of change(slope of line) in the output variable.

More simply: m tells you how much Y grows (or drops) whenever X increases by one unit.

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

Ans: In business-friendly terms, the intercept "c" signifies the baseline output value when the input variable X is positioned at zero.

In short: it's the starting value of Y before X has any impact.

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

Ans: The slope **m** is computed by operationalizing the change in **Y** for every unit change in **X**.

Short formula perspective

$$m = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$$

Plain meaning
We look at how much X moves away from its average and how much Y moves away from its average, aggregate those co-movements, and benchmark them against the total variation in X.

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

Ans: The least-squares method is leveraged to **minimize the overall error** between the predicted line and the actual data points. In operational terms, it ensures the regression line is **the best-fit line** by reducing the squared distance between observations and predictions, thereby optimizing model accuracy and analytical reliability.

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

Ans: The coefficient of determination (R^2) quantifies how effectively our regression line explains the variance in the outcome variable.

In simple terms: it's the percentage of the target's behavior that is successfully accounted for by the predictor.

Example:

$R^2 = 0.80 \rightarrow$ roughly 80% of the output variation is clarified by the model, indicating strong explanatory alignment.

8. What is Multiple Linear Regression?

Ans: Multiple Linear Regression is essentially a data-driven forecasting framework that models how one outcome is influenced by several independent variables simultaneously.

In simpler terms:

It's a method used to predict a value (like price, marks, sales, etc.) using more than one input factor at the same time, enabling more holistic and accurate decision insights.

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

Ans: From a high-level analytics perspective, the distinction is fairly straightforward:

Simple Linear Regression operationalizes a predictive relationship using **one** independent variable to forecast the target metric.

Multiple Linear Regression extends the decision framework by leveraging **several** independent drivers simultaneously to deliver more comprehensive predictive insights.

10. What are the key assumptions of Multiple Linear Regression?

Ans:

• **Linearity** – the output variable is expected to maintain a straight-line relationship with each predictor, enabling predictable trend modeling.

• **Independence** – the error terms should not influence each other, ensuring clean, de-correlated insights across observations.

• **Constant variance (Homoscedasticity)** – the spread of prediction errors remains stable across all ranges of the predictors, ensuring consistent performance quality.

• **Normality of errors** – the residuals are assumed to follow a normal distribution, supporting statistically robust conclusions.

• **No strong multicollinearity** – predictors shouldn't excessively overlap in meaning, so each variable delivers unique business value rather than redundant signals.

11. What is heteroscedasticity, and how does it affect the results of a Multiple Linear Regression model ?

Ans: Heteroscedasticity basically means the spread of errors isn't constant across all values of the predictors. In other words, the prediction errors get bigger or smaller in different parts of the data, instead of staying steady.

Impact on Multiple Linear Regression:

When this imbalance occurs, it distorts the reliability of coefficient estimates and inflates or deflates standard errors, which ultimately undermines the trustworthiness of p-values and can lead to sub-optimal decision-making based on the model's outputs.

12. How can you improve a Multiple Linear Regression model with high multicollinearity?

Ans:

- **Remove highly-correlated inputs** to de-duplicate information and stabilize coefficients

- **Use regularization frameworks (Ridge/Lasso)** to strategically penalize redundant variables

- **Apply dimensionality-reduction (like PCA)** to compress overlapping predictors into cleaner components

13. What are some common techniques for transforming categorical variables for use in regression models?

Ans:

- **Label encoding** – assign numeric IDs to each category to streamline model consumption.
- **One-hot encoding** – spin up binary flags for every category, enabling the algorithm to capture category presence without implying order.
- **Ordinal encoding** – map categories to a ranked scale when there's a natural priority sequence.
- **Target (mean) encoding** – replace categories with their average target value to extract predictive signal while maintaining model efficiency.

14. What is the role of interaction terms in Multiple Linear Regression?

Ans: In a data-driven context, interaction terms essentially capture the combined impact of two predictors when their joint influence on the target variable isn't purely additive.

In plain terms: sometimes two factors together change the outcome differently than each alone. Interaction terms help the model account for that partnership effect, enabling more granular and strategically aligned predictions.

15. How can the interpretation of intercept differ between Simple and Multiple Linear Regression?

Ans: In **Simple Linear Regression**, the **intercept** is the predicted value of y when the single x is 0.

In **Multiple Linear Regression**, the **intercept** is the predicted value of y when **all independent variables** are 0. It's less intuitive because having all x equal to 0 might not make real-world sense.

16. What is the significance of the slope in regression analysis, and how does it affect predictions?

Ans: In regression analysis, the **slope** shows how much the **dependent variable** (what you want to predict) changes for each 1-unit change in the **independent variable** (the input).

- **Positive slope:** As the input increases, the output increases.
- **Negative slope:** As the input increases, the output decreases.

It directly affects predictions because it determines the **direction and rate of change** in your forecast.

17. How does the intercept in a regression model provide context for the relationship between variables ?

Ans: The intercept in a regression model is the starting point: it shows the expected value of the outcome (dependent variable) when all input factors (independent variables) are zero. It gives context by anchoring the relationship.

18. What are the limitations of using R^2 as a sole measure of model performance?

Ans: R^2 alone has limitations:

1. **Doesn't show errors** – it won't tell you if predictions are far off.
2. **Can be misleading** – adding more variables can increase R^2 even if they don't help.
3. **Not for non-linear models** – may give wrong impression if the relationship isn't linear.
4. **Ignores overfitting** – a high R^2 doesn't mean the model will work on new data.

19. How would you interpret a large standard error for a regression coefficient ?

Ans: A large standard error for a regression coefficient means there's a lot of uncertainty about the estimated effect of that variable. In simple terms, the coefficient might not be reliable, and small changes in the data could lead to big changes in the estimate.

20. How can heteroscedasticity be identified in residual plots, and why is it important to address it ?

Ans: Heteroscedasticity shows up in residual plots when the spread of residuals (errors) gets wider or narrower as the predicted values change—like a funnel shape.

It's important to fix because it violates regression assumptions, making predictions and confidence intervals less reliable.

21. What does it mean if a Multiple Linear Regression model has a high R^2 but low adjusted R^2?

Ans: It means your model **fits the training data well** (high R^2) but **has too many unnecessary variables** (low adjusted R^2), so it may **not generalize well** to new data.

22. Why is it important to scale variables in Multiple Linear Regression ?

Ans: Scaling variables in Multiple Linear Regression is important because it ensures all features are on a similar scale, which:

1. **Prevents one feature from dominating the model just because it has larger values.**
2. **Improves convergence when using optimization algorithms like gradient descent.**
3. **Makes interpretation easier if you compare coefficients.**

23. What is polynomial regression ?

Ans: Polynomial regression is a type of regression analysis where the relationship between the input (independent variable) and output (dependent variable) is modeled as a polynomial instead of a straight line.

In simple words: it helps fit curvy lines to data instead of just straight lines, so it can capture more complex patterns.

For example:

- **Linear: $y = 2x + 3$(straight line)**
- **Polynomial: $y = 2x^2 + 3x + 1$(curve)**

24. How does polynomial regression differ from linear regression ?

Ans: Polynomial regression is like linear regression but with a twist: instead of fitting a straight line, it fits a curve to the data. Linear regression = straight line, polynomial regression = curved line to capture more complex patterns.

25. When is polynomial regression used ?

Ans: Polynomial regression is used when the relationship between the input (independent variable) and output (dependent variable) is curved, not straight. It helps model trends that a simple straight-line (linear) regression cannot capture.

26. What is the general equation for polynomial regression ?

Ans: The general equation for polynomial regression is:

$$y = b_0 + b_1x + b_2x^2 + b_3x^3 + \cdots + b_nx^n$$

In simple words:

- **y= predicted value**
- **x= input variable**
- **$b_0, b_1, \ldots, b_n$= coefficients**
- **n= degree of the polynomial**

It's basically like linear regression but includes powers of $x$to capture curves in the data.

27. Can polynomial regression be applied to multiple variables ?

Ans: Yes. Polynomial regression can be applied to multiple variables by creating polynomial terms for each variable and their combinations. This is called multivariate polynomial regression. It models non-linear relationships between several inputs and the output.

28. What are the limitations of polynomial regression ?

Ans: Polynomial regression has a few key limitations:

1. **Overfitting risk – High-degree polynomials can fit the training data too closely, capturing noise instead of true patterns.**
2. **Poor extrapolation – Predictions outside the data range can be wildly inaccurate.**
3. **Complexity grows fast – Higher degrees make the model harder to interpret.**
4. **Sensitive to outliers – Extreme values can distort the curve significantly.**

29. What methods can be used to evaluate model fit when selecting the degree of a polynomial ?

Ans: You can evaluate the fit of a polynomial model using these methods:

1. **Visual check – Plot the curve and see if it follows the data without overfitting.**
2. **Train/Test split – Check how well the model predicts new data.**
3. **Cross-validation – Split data multiple times to see how consistent the model performs.**
4. **Error metrics – Use RMSE, MAE, or R^2 to measure prediction accuracy.**
5. **Adjusted R^2 – Accounts for extra terms, helps avoid overfitting.**

Keep it balanced: enough degree to fit, but not too high to overfit.

30. Why is visualization important in polynomial regression ?

Ans: Visualization in polynomial regression is important because it helps you see the relationship between variables, check if the model fits the data well, and detect patterns, trends, or overfitting easily. It makes complex data intuitive.

31. How is polynomial regression implemented in Python?

Ans: from sklearn.linear_model import LinearRegression

from sklearn.preprocessing import PolynomialFeatures

from sklearn.pipeline import make_pipeline

```
# Example: degree 2 polynomial

model = make_pipeline(PolynomialFeatures(degree=2), LinearRegression())



# Fit the model

model.fit(X_train, y_train)



# Predict

y_pred = model.predict(X_test)
```