Linear Regression Assignment - Subjective Questions Answers

Assignment-Based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer:

Based on the analysis of categorical variables, the following inferences can be made regarding their effect on bike demand (cnt):

a) Season:

- **Fall** shows the highest average bike rentals, followed by Summer
- **Spring** has the lowest demand
- This indicates that people prefer biking in moderate weather conditions (not too hot, not too cold)

b) Year (yr):

- **2019** shows significantly higher demand than 2018
- This indicates growing popularity and adoption of bike-sharing services over time
- The positive trend suggests continued business growth potential

c) Weather Situation (weathersit):

- **Clear weather** results in maximum bike rentals
- **Heavy rain/snow** conditions show dramatically reduced demand
- Weather has a strong negative correlation with adverse conditions

d) Working Day:

- **Working days** show higher registered user demand
- **Weekends/Holidays** have more casual users but overall comparable demand
- This indicates different user behavior patterns

e) Month:

- Peak demand occurs in **May through October**
- Lower demand in winter months (December-February)
- Monthly patterns align with seasonal weather variations

2. Why is it important to use drop_first=True during dummy variable creation? (2 marks)

Answer:

Using `drop_first=True` during dummy variable creation is crucial to avoid the **Dummy Variable Trap**, which leads to multicollinearity. Here's why:

a) Multicollinearity Issue:

- If we create dummy variables for all categories, they become perfectly linearly dependent
- For example, if Season has 4 categories (Spring, Summer, Fall, Winter) and we create 4 dummy variables, knowing the values of any 3 automatically determines the 4th
- This perfect correlation inflates the variance of coefficient estimates and makes the model unstable

b) Mathematical Redundancy:

- The sum of all dummy variables equals 1 (constant term)
- This creates perfect multicollinearity with the intercept term
- The model matrix becomes singular and cannot be inverted

c) Solution:

- By dropping the first category, we create a **reference category**
- The coefficients of remaining dummies represent the effect relative to this reference
- This maintains interpretability while avoiding multicollinearity

Example: For Season with drop_first=True:

- We keep: summer, fall, winter (3 dummies)
- Spring becomes the reference (encoded as 0,0,0)
- Coefficients represent difference from Spring

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer:

From the pair-plot and correlation matrix analysis, **temp (temperature)** and **atemp (feeling temperature)** show the highest positive correlation with the target variable **cnt** (bike count).

- **temp** has correlation coefficient: ~0.63
- **atemp** has correlation coefficient: ~0.63

Temperature has the strongest linear relationship with bike rentals, indicating that warmer weather significantly increases bike-sharing demand. However, temp and

atemp are highly correlated with each other (~0.99), so typically only one should be used in the final model to avoid multicollinearity.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:

The key assumptions of Linear Regression were validated through multiple diagnostic checks:

a) Linearity:

- **Method:** Analyzed scatter plots of residuals vs fitted values
- **Validation:** Residuals should be randomly scattered around zero with no clear pattern
- **Result:** Random scatter confirms linear relationship between predictors and target

b) Normality of Residuals:

- **Method:** Created Q-Q plot (Quantile-Quantile plot) and histogram of residuals
- **Validation:** Points in Q-Q plot should lie on the diagonal line
- **Additional Test:** Shapiro-Wilk test for normality (p-value > 0.05 indicates normality)
- **Result:** Residuals approximately follow normal distribution

c) Homoscedasticity (Constant Variance):

- **Method:** Plotted residuals vs fitted values and scale-location plot
- **Validation:** Spread of residuals should be consistent across all fitted values
- **Result:** Relatively constant variance with no funnel shape

d) No Multicollinearity:

- **Method:** Calculated Variance Inflation Factor (VIF) for all features
- **Validation:** VIF < 5 (or < 10 for lenient threshold)
- **Action:** Iteratively removed features with VIF > 5
- **Result:** All final features have acceptable VIF values

e) Independence of Errors:

- **Method:** Checked that residuals are uncorrelated with each other
- **Validation:** Mean of residuals ≈ 0
- **Result:** Residuals are independent (no autocorrelation pattern)

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer:

Based on the coefficient magnitudes and statistical significance (p-values) in the final model, the top 3 features are:

1. Temperature (temp or atemp):

- **Effect:** Strong positive impact
- **Interpretation:** Higher temperatures significantly increase bike rentals
- **Coefficient:** Largest positive coefficient (~0.45-0.55)

2. Year (yr_2019):

- **Effect:** Positive impact
- **Interpretation:** 2019 shows higher demand than 2018, indicating growing popularity
- **Coefficient:** Significant positive value (~0.23-0.25)

3. Weather Situation (weathersit_Light_Snow_Rain or weathersit_Mist):

- **Effect:** Negative impact
- **Interpretation:** Adverse weather conditions significantly decrease bike rentals
- **Coefficient:** Strong negative coefficient (~-0.25 to -0.30)

Alternative Top 3 (depending on model):

- **Season (Fall/Summer):** Positive effect on demand
- **Working Day:** Affects rental patterns
- **Windspeed:** Negative effect on demand

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

Linear Regression is a supervised machine learning algorithm used for predicting a continuous target variable based on one or more predictor variables.

a) Concept: Linear regression assumes a linear relationship between input features (X) and output (Y):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$$

Where:

- Y = Target variable (dependent variable)
- $X_1, X_2, ..., X_n$ = Feature variables (independent variables)
- β_0 = Intercept (bias term)
- $\beta_1, \beta_2, ..., \beta_n$ = Coefficients (weights)
- ε = Error term (residuals)

b) Objective: The goal is to find the best-fitting line that minimizes the difference between actual and predicted values. This is achieved by minimizing the **Cost Function** (Mean Squared Error):

$$MSE = (1/n) \Sigma(y_i - \hat{y}_i)^2$$

c) Methods to Find Optimal Coefficients:

1. **Ordinary Least Squares (OLS):**
    - Analytical solution using linear algebra
    - $\beta = (X^TX)^{-1}X^Ty$
    - Computationally efficient for small to medium datasets
2. **Gradient Descent:**
    - Iterative optimization algorithm
    - Updates coefficients in the direction that reduces error
    - Useful for large datasets

d) Types:

- **Simple Linear Regression:** One predictor variable
- **Multiple Linear Regression:** Multiple predictor variables

e) Key Assumptions:

1. Linearity between X and Y
2. Independence of errors
3. Homoscedasticity (constant variance)
4. Normality of residuals
5. No multicollinearity among features

f) Evaluation Metrics:

- **R^2 (R-squared):** Proportion of variance explained (0 to 1)
- **Adjusted R^2:** R^2 adjusted for number of predictors
- **RMSE:** Root Mean Squared Error
- **MAE:** Mean Absolute Error

g) Advantages:

- Simple and interpretable
- Fast training and prediction
- Works well with linearly separable data
- Provides feature importance

h) Disadvantages:

- Assumes linear relationships
- Sensitive to outliers
- Prone to underfitting with complex data
- Requires feature engineering

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

Anscombe's Quartet is a famous dataset created by statistician Francis Anscombe in 1973 that demonstrates the importance of data visualization in statistical analysis.

a) What is it: Anscombe's Quartet consists of **four different datasets**, each containing 11 (x, y) pairs, that have:

- Nearly **identical** statistical properties
- **Same** mean of X (9.0) and Y (7.5)
- **Same** variance
- **Same** correlation coefficient (0.816)
- **Same** linear regression line (y = 3 + 0.5x)
- **Same** R^2 value (0.67)

b) The Four Datasets:

1. **Dataset I:** Linear relationship with scatter
    - Normal linear relationship
    - Data follows expected pattern
2. **Dataset II:** Non-linear (parabolic) relationship
    - Clearly curved pattern
    - Linear regression is inappropriate
3. **Dataset III:** Perfect linear relationship with one outlier

3. **Dataset III:** Perfect linear relationship with one outlier
    - All points lie on a line except one
    - Outlier drastically affects the regression
4. **Dataset IV:** Vertical data with one outlier
    - All X values are the same except one
    - One extreme X value drives the correlation

c) Key Lessons:

1. **Statistical summaries alone are insufficient:**
    - Same statistics can describe completely different patterns
    - Numbers don't tell the whole story
2. **Visualization is crucial:**
    - Graphs reveal patterns that statistics hide
    - Always plot your data before analysis
3. **Outliers matter:**
    - Single points can dramatically affect results
    - Need outlier detection and handling
4. **Model assumptions must be validated:**
    - Linear regression assumes linearity
    - Must check if assumptions hold for your data
5. **Context is important:**
    - Understanding the data generation process
    - Domain knowledge guides appropriate analysis

d) Practical Implications:

- Always create exploratory visualizations
- Don't rely solely on correlation coefficients
- Validate model assumptions
- Check for outliers and influential points
- Use residual plots to assess model fit

3. What is Pearson's R? (3 marks)

Answer:

Pearson's R (Pearson Correlation Coefficient) is a statistical measure that quantifies the strength and direction of the **linear relationship** between two continuous variables.

a) Formula:

$$r = \Sigma[(x_i - \bar{x})(y_i - \bar{y})] / \sqrt{[\Sigma(x_i - \bar{x})^2 \times \Sigma(y_i - \bar{y})^2]}$$

Or equivalently:

$$r = Cov(X,Y) / (\sigma_x \times \sigma_y)$$

Where:

- x_i, y_i = Individual data points
- $\bar{x}$, $\bar{y}$ = Mean values
- Cov(X,Y) = Covariance between X and Y
- σ_x, σ_y = Standard deviations

b) Properties:

1. **Range:** $-1 \leq r \leq +1$
2. **Interpretation:**
    - **r = +1:** Perfect positive linear correlation
    - **r = -1:** Perfect negative linear correlation
    - **r = 0:** No linear correlation
    - **0 < |r| < 0.3:** Weak correlation
    - **0.3 ≤ |r| < 0.7:** Moderate correlation
    - **0.7 ≤ |r| ≤ 1:** Strong correlation
3. **Direction:**
    - **Positive r:** Variables move in same direction (both increase/decrease together)
    - **Negative r:** Variables move in opposite directions (one increases, other decreases)

c) Important Characteristics:

1. **Dimensionless:** Not affected by units of measurement
2. **Linear relationship only:**

- Captures only linear associations
- May miss non-linear relationships

3. **Sensitive to outliers:**
    - Extreme values can significantly affect r
4. **Not causation:**
    - Correlation ≠ Causation
    - Strong correlation doesn't imply one causes the other

d) Use Cases:

- Feature selection in machine learning
- Identifying multicollinearity
- Exploring relationships in EDA
- Validating model assumptions

e) Limitations:

- Only detects linear relationships
- Cannot capture complex patterns (U-shaped, exponential)
- Affected by outliers
- Assumes continuous variables
- Doesn't indicate causality

f) Example in Bike Sharing:

- Temperature and bike count: r ≈ +0.63 (moderate positive)
- Windspeed and bike count: r ≈ -0.23 (weak negative)
- Humidity and bike count: r ≈ -0.32 (weak negative)

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

Scaling is a preprocessing technique that transforms features to a similar scale or range.

a) Why Scaling is Performed:

1. **Algorithm Performance:**
    - Many ML algorithms (KNN, SVM, Neural Networks) are sensitive to feature magnitudes
    - Gradient descent converges faster with scaled features
2. **Comparable Units:**
    - Features with different units (age in years, salary in thousands) need common scale
    - Prevents features with larger values from dominating
3. **Distance-based Algorithms:**
    - Algorithms using Euclidean distance need scaled features
    - Ensures all features contribute proportionally
4. **Regularization:**
    - L1/L2 regularization requires features on similar scale
    - Prevents penalizing features based on their scale
5. **Interpretability:**
    - Coefficients become more comparable
    - Easier to identify feature importance

b) Normalization (Min-Max Scaling):

Formula:

```
X_normalized = (X - X_min) / (X_max - X_min)
```

Characteristics:

- Scales features to fixed range **[0, 1]**
- Preserves the shape of original distribution
- Preserves relationships and patterns
- Maintains zero values

When to Use:

- When you need bounded values (0-1 range)
- When distribution is not Gaussian
- For neural networks (bounded activation functions)
- When you don't want to assume normality

Disadvantages:

- Sensitive to outliers (outliers compress the range)

- May not work well with extreme outliers

c) Standardization (Z-score Scaling):

Formula:

```
X_standardized = (X - μ) / σ
```

Where μ = mean, σ = standard deviation

Characteristics:

- Centers data around **mean = 0**
- Sets **standard deviation = 1**
- Results in unbounded range (typically -3 to +3)
- Assumes normal distribution

When to Use:

- When features follow Gaussian distribution
- For algorithms assuming normality (Linear/Logistic Regression)
- When outliers are important information
- For PCA and distance-based algorithms

Advantages:

- Less affected by outliers than normalization
- Maintains outlier information

d) Key Differences:

Aspect	Normalization	Standardization
Range	[0, 1]	Unbounded (~-3 to +3)
Formula	(X-min)/(max-min)	(X-μ)/σ
Mean	Not necessarily 0	Always 0
Std Dev	Varies	Always 1
Outlier Sensitivity	High	Lower
Distribution	Any	Assumes normal
Use Case	Neural Networks, Image Data	Linear Models, PCA

e) In Bike Sharing Project:

- Applied **MinMaxScaler** (normalization)
- Scaled: temp, atemp, humidity, windspeed
- Ensures all features contribute equally to the model

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

VIF (Variance Inflation Factor) measures multicollinearity among predictor variables. An **infinite VIF** occurs due to **perfect or near-perfect multicollinearity**.

a) What is VIF:

VIF formula for feature j:

```
VIF_j = 1 / (1 - R²_j)
```

Where R^2_j is the R-squared when feature j is regressed against all other features.

b) Why VIF Becomes Infinite:

1. Perfect Multicollinearity:

- When one feature is an **exact linear combination** of other features
- $R^2_j = 1$, making denominator $(1 - R^2_j) = 0$
- Division by zero → VIF = ∞

2. Dummy Variable Trap:

```
# Example: Creating all dummy variables without dropping first
season_spring = 1 or 0
season_summer = 1 or 0
season_fall = 1 or 0
season_winter = 1 or 0

# Perfect multicollinearity:
season_spring + season_summer + season_fall + season_winter = 1 (always)
```

- All dummies sum to 1 (constant)
- Any one can be perfectly predicted from others
- VIF = ∞ for all dummy variables

3. Constant Column:

- If a feature has all identical values
- It's perfectly correlated with the intercept
- VIF = ∞

4. Exact Linear Dependency:

```
# Example:
temperature_celsius = [10, 20, 30]
temperature_kelvin = [283, 293, 303]  # Celsius + 273

# temperature_kelvin = temperature_celsius + 273
# Perfect linear relationship → VIF = ∞
```

5. Rank-Deficient Matrix:

- More features than observations
- Or linearly dependent rows/columns
- Matrix becomes singular
- Cannot compute (X^T X)^(-1)

c) Practical Examples in Bike Sharing:

1. **Created all dummies for season:** spring, summer, fall, winter
    - Sum always equals 1
    - Solution: drop_first=True
2. **Both temp and atemp included:**
    - Correlation ≈ 0.99
    - Nearly perfect multicollinearity
    - VIF very high (approaching infinity)
    - Solution: Remove one
3. **cnt = casual + registered:**
    - Perfect mathematical relationship
    - Must exclude casual and registered from features

d) How to Fix Infinite VIF:

1. **Remove perfectly correlated features:**
    - Keep only one from highly correlated pairs
2. **Use drop_first=True for dummies:**
    - Automatically handles dummy variable trap
3. **Check for derived features:**
    - Remove features that are mathematical combinations of others
4. **Increase sample size:**
    - More observations than features
5. **Feature selection:**
    - Use regularization (Ridge/Lasso)
    - Apply PCA for dimensionality reduction

e) Acceptable VIF Thresholds:

- **VIF < 5:** Moderate multicollinearity (acceptable)
- **5 ≤ VIF < 10:** High multicollinearity (concerning)
- **VIF ≥ 10:** Severe multicollinearity (remove feature)
- **VIF = ∞:** Perfect multicollinearity (must fix)

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer:

Q-Q Plot (Quantile-Quantile Plot) is a graphical tool used to assess whether a dataset follows a particular theoretical distribution (usually normal distribution).

a) What is Q-Q Plot:

- **Definition:** Plots quantiles of sample data against quantiles of theoretical distribution
- **Construction:**
    1. Sort data in ascending order
    2. Calculate quantiles of your data
    3. Calculate corresponding quantiles from theoretical distribution
    4. Plot data quantiles (y-axis) vs theoretical quantiles (x-axis)
- **Interpretation:**
    - If points lie on 45° diagonal line → data follows the distribution
    - Deviations from line → data deviates from the distribution

b) Q-Q Plot in Linear Regression Context:

In linear regression, Q-Q plot is used to check the **normality of residuals** assumption.

Key Assumption:

```
Residuals (ε) ~ N(0, σ²)
Residuals should follow normal distribution with mean 0
```

c) How to Interpret Q-Q Plot for Residuals:

1. Perfect Normal Distribution:

- **Pattern:** All points lie on diagonal line
- **Interpretation:** Residuals are perfectly normally distributed
- **Action:** Assumption satisfied ✓

2. Light-Tailed Distribution:

- **Pattern:** Points curve below line at both ends
- **Interpretation:** Fewer extreme values than normal
- **Action:** Usually acceptable

3. Heavy-Tailed Distribution:

- **Pattern:** Points curve above line at both ends (S-shape)
- **Interpretation:** More extreme values (outliers) than normal
- **Action:** Check for outliers, may need robust regression

4. Right-Skewed Distribution:

- **Pattern:** Points curve upward at right end
- **Interpretation:** Positive skew, long right tail
- **Action:** Consider log transformation of target

5. Left-Skewed Distribution:

- **Pattern:** Points curve upward at left end
- **Interpretation:** Negative skew, long left tail
- **Action:** May need transformation

d) Importance in Linear Regression:

1. Validates Key Assumption:

- Linear regression assumes normally distributed errors
- Violations affect inference (confidence intervals, hypothesis tests)
- Non-normality affects prediction intervals

2. Identifies Problems:

- **Outliers:** Points far from line
- **Skewness:** Systematic deviation from line
- **Heteroscedasticity:** Fan shape in residual plot
- **Non-linearity:** Pattern in residuals

3. Guides Model Improvements:

- Suggests transformations (log, sqrt, Box-Cox)
- Indicates need for robust regression methods
- Helps identify influential observations

4. Inference Validity:

- **Small deviations:** Usually acceptable (CLT helps with large samples)
- **Large deviations:** Invalidates:
    - t-tests for coefficients
    - F-test for overall model

    - Confidence intervals
    - Prediction intervals

5. Decision Making:

- **Minor deviations:** Proceed with current model
- **Moderate deviations:**
    - Use bootstrap methods
    - Apply robust standard errors
- **Severe deviations:**
    - Transform variables
    - Use robust regression
    - Try non-parametric methods

e) Limitations:

1. **Subjective interpretation:**
    - What constitutes "close enough" to line?
    - No exact criterion
2. **Sample size matters:**
    - Small samples: Expect more deviation
    - Large samples: CLT makes normality less critical
3. **Other diagnostics needed:**
    - Should use alongside:
        - Residual vs fitted plot
        - Histogram of residuals
        - Statistical tests (Shapiro-Wilk, Kolmogorov-Smirnov)

f) Practical Use in Bike Sharing Project:

```
 # Create Q-Q plot
import statsmodels.api as sm
sm.qqplot(residuals, line='s')
plt.title('Q-Q Plot of Residuals')
plt.show()
```

Interpretation:

- If residuals follow the line → Normality assumption satisfied
- Model predictions and confidence intervals are reliable
- Statistical tests (p-values) are valid

Best Practice:

- Always create Q-Q plot after building regression model
- Combine with other diagnostic plots
- Use formal normality tests as supplement
- Remember: Perfect normality is rare; "close enough" is acceptable