**\*\* CODE IN THE SAS FILE \*\***

**2) (i) Fully investigate and discuss any issues of multicollinearity between the ten potentials**

**explanatory variables listed above. (15 marks)**

🡪 **(a) Checking Correlation Among Variables:**

/\*Interpretation of Multicollinearity from Correlation Matrix:

The Pearson correlation matrix identifies potential multicollinearity issues.

The correlation coefficient (r) with a value greater than 0.7 or less than -0.7 shows a strong relationship between two independent variables,

which can result in multicollinearity in regression models.

Key Observations from the Correlation Matrix:

High Correlation Pairs (r > 0.7) – Strong Multicollinearity Risk:

a) Footprint & Otherarea (r = 0.845) - Strong correlation, redundancy.

b) Otherarea & House (r = 0.864) - Highly correlated; can cause instability if both are present.

c) House & Ratevalue (r = 0.795) - Indicates collinearity between property-related variables.

d) House & Footprint (r = 0.824) - These variables are measuring the same factors.

e) Car & Garden (r = 0.878) - Indicates redundancy; can remove one of them.

f) Otherarea & Garden (r = 0.728) - Strong collinearity between property-related variables.

g) Car & Otherarea (r = 0.722) - Affects model independence.

Implications of Multicollinearity:

a) Redundant Information:

- Variables like Footprint, Otherarea, and House are highly correlated, i.e., they hold the same information.

Including all of them in the model can create instability.

- Car and Garden are highly correlated, i.e., having both of them in the model might be redundant.

b) Impact on Regression Models:

- Multicollinearity increases standard errors, thus making coefficient estimates unreliable.

- High values of VIF will determine the level of multicollinearity.

- Interpretation of the model becomes cumbersome if predictors are highly correlated.

c) Possible Solutions:

- Drop one of the highly multicollinear variables (e.g., Otherarea or House, Car or Garden).

Next Step

- Perform multicollinearity check using VIF (Variance Inflation Factor).

**(b)** **Compute Variance Inflation Factor (VIF):**

Meaning of Variance Inflation Factor (VIF) Values:

VIF approximates how much the multicollinearity increases the variance of a regression coefficient.

a) VIF < 5 - Low multicollinearity (fine)

b) VIF 5 - 10 - Moderate multicollinearity (problem)

c) VIF > 10 - High multicollinearity (trouble)

Key VIF Output Observations

a) High Multicollinearity (VIF > 7)

i) House (7.6645), ii) Otherarea (7.5716), iii) Garden (7.5386)

These are high multicollinearity values; thus, these variables carry similar information.

b) Moderate Multicollinearity (5-7 VIF)

i) Car (6.4629)

ii) Footprint (5.4001)

These variables also possess some collinearity problems, particularly with House and Otherarea.

Conclusion:

- House, Otherarea, and Garden show maximum multicollinearity.

- Car and Footprint also create problems of collinearity.

- Reduction of multicollinearity will improve model stability, interpretability, and predictive validity.

**(c) Compute Tolerance:**

Tolerance measures how much a predictor variable is independent of other variables in a regression model.

A low tolerance (<0.1) indicates severe multicollinearity, meaning the variable is highly correlated with others and should be removed or adjusted.

Computing tolerance helps ensure accurate coefficient estimation and improves model reliability.

Conclusion:

a) The low tolerance values (< 0.2) for House, Garden, Car, and Otherarea suggest that multicollinearity is an issue.

b) To address this, attempt to drop or group collinear variables (e.g., using PCA or stepwise regression) to improve model stability.

c) The R² = 0.7154 for the model shows high explanatory power, but multicollinearity could distort coefficient estimates.

**(d) Compute Condition Index (CI):**

Condition Index (CI) detects multicollinearity by identifying linear dependencies among the variables.

CI = n (a certain number) indicates that there is severe multicollinearity, making the model unstable.

CI calculation helps determine if transformation or removal of variables is needed to make the regression estimates correct.

Conclusion:

Multicollinearity exists in the dataset, particularly involving Garden, Otherarea, House, and Footprint.

**(e) Remedy Multicollinearity:**

Will remove Garden, Otherarea, House, and Footprint attributes as by computing the correlation, VIF, Tolerance & CI,this 2 columns are largely affecting our data,

which we need to improve to get the needed output.

And perform stepwise regression on the selected variables to determine the best predictors while mitigating multicollinearity.

**Conclusion:**

Step 1: Pubs

R-Square: 0.3142 (31.42% variance explained).

C(p): 4.5539 (good fit).

Significance: F-value = 17.41, p-value < 0.0002 (Pubs is significant).

Step 2: Employ

R-Square: 0.3959 (39.59% variance explained).

C(p): 1.7210 (better fit).

Significance: F-value = 5.01, p-value = 0.0314 (Employ is significant).

Insights:

- Pubs is the most influential predictor.

- Model explains 39.59% of variation in Ratevalue.

- All variables are significant at the 0.1500 level.

**2) (ii) Fit appropriate multiple regression models to these data using both forward selection and**

**backward elimination procedures, in each case using an associated significance level of 5%. For**

**each procedure, state and briefly justify the decision taken at each step, and identify the**

**explanatory variables present in the final reduced model.**

**Briefly discuss the composition of the final models obtained with these two different**

**approaches in the light of the collinearity analysis conducted in part (i). (15 marks)**

🡪 **(a) Forward Selection:**

Forward Selection starts with an empty model and adds the most significant predictor (p-value < 0.05) at each step.

This continues until no new variable significantly improves the model.

Stepwise Process:

Step 1: Start with no predictors.

Step 2: Add the variable with the lowest p-value (< 0.05) first.

Step 3: Continue adding variables one at a time until no more variables meet the threshold (SLENTRY = 0.05).

Interpretation:

a) Important Predictors: Pubs and Employ were selected as the most important predictors of Ratevalue.

b) Model Fit: The model that eventually fitted explains 39.59% of the variance (R² = 0.3959), which is a moderate fit.

c) Negative Influence:

- More Pubs reduces Ratevalue (-2.67 effect).

- Employ increased also reduces Ratevalue (-0.93 effect).

d)Statistical Significance:

- Pubs (p = 0.0002) and Employ (p = 0.0314) are both 5% significant.

- No other variable met the selection criteria.

e) Model Stability: The condition number (5.78) reflects no severe multicollinearity, and therefore the model is stable.

**(b) Backward Elimination:**

Backward Elimination starts with all explanatory variables and removes the least significant variable (p-value > 0.05) in each step.

The process continues until only significant variables remain.

Stepwise Process:

Step 1: Start with all predictors.

Step 2: Remove the variable with the highest p-value (> 0.05).

Step 3: Repeat until all remaining variables are statistically significant.

Interpretation:

a) Significant Predictors: `Pubs` and `Employ` alone were retained as the sole predictors having a significant effect on `Ratevalue`.

b) Fit of the Model: The resulting model explains 39.59% of the variance (R² = 0.3959), which is indicative of a moderate fit.

c) Negative Effect:

- Having greater `Pubs` decreases `Ratevalue` (-2.67 effect).

- More `Employ` decreases `Ratevalue` (-0.93 effect).

d) Statistical Significance:

- `Pubs` (p = 0.0212) and `Employ` (p = 0.0314) are 5% level significant.

- `Social`, `Food`, and `Car` were dropped since their p-values were extremely high (>0.05).

e) Model Stability: The condition number (5.78) reflects no severe multicollinearity, and the model is stable.

**(c) Checking Multicollinearity using VIF:**

Final Interpretation:

(a) Model Summary:

- Number of Observations Used: 40

- R² = 0.3959 (Moderate fit)

- Adjusted R² = 0.3632

- F-Statistic = 12.12, p < 0.0001 (Overall model is statistically significant)

(b) Final Model from Forward and Backward Selection:

- Predictors: Pubs, Employ

- Coefficients:

(i) Intercept: 38.70 (p < 0.0001)

(ii) Pubs: -2.67 (p = 0.0212)

(iii) Employ: -0.93 (p = 0.0314)

(c) Interpretation:

- Pubs: A rise in the number of pubs is associated with a decrease in Ratevalue (ß = -2.67). This suggests that more pubs mean less rating.

- Employ: Higher employment is also associated with a decrease in Ratevalue (ß = -0.93).

- Statistical Significance: Both predictors are significant at the 5% level.

- Model Stability:

- VIF for Pubs and Employ = 1.45 (No serious collinearity issues).

- Condition Number = 5.78 (Stability indicated).

(d) Comparison of Methods:

- Both Forward Selection and Backward Elimination resulted in the same final model.

- Food, Social, and Car were removed due to high p-values (> 0.05), i.e., they don't play an important role in predicting Ratevalue.

**(e) Conclusion:**

- The final model explains 39.59% of the variance in Ratevalue.

- Pubs and Employ are the most important predictors.

- No severe multicollinearity was found, which ensures a stable model.

- The model provides useful insight into the predictors of Ratevalue.

**2) (iii) Further investigate model selection using the Cp method for comparing all possible models.**

**Comment specifically on the two reduced models identified in (ii) above.**

**Hence, using the data alone, choose a final, reduced model, justifying your choice. Are any other**

**models of the selected size also reasonable candidates for selection? Explain. (10 marks)**

🡪 **(b) Analysis of Mallows’ Cp Results:**

Understanding Mallows’ Cp:

Mallows’ Cp balances model fit and complexity, where Cp ˜ p (parameters).

- Cp < p: Strong fit with minimal bias.

- Cp ˜ p: Adequate fit.

- Cp > p: Possible underfitting.

Model Analysis:

1. Pubs Employ (p = 3, Cp = 1.7210, R² = 39.59%)

- Best 2-variable model with the lowest Cp and highest R².

- Strong fit with minimal bias and good explanatory power.

2. Food Employ (p = 3, Cp = 2.0691, R² = 39.00%)

- Similar to Pubs Employ but slightly less efficient.

3. Pubs Employ Car (p = 4, Cp = 2.2170, R² = 42.13%)

- Best 3-variable model with the highest R².

- Adding Car improves variance explained but increases complexity.

4. Food Employ Car (p = 4, Cp = 2.7910, R² = 41.16%)

- Good fit but slightly weaker than Pubs Employ Car.

5. Pubs Car (p = 3, Cp = 3.0281, R² = 37.38%)

- Least competitive model with the lowest R².

Key Takeaways:

- Best 2-variable model: Pubs Employ (Cp = 1.7210, R² = 39.59%).

- Best 3-variable model: Pubs Employ Car (Cp = 2.2170, R² = 42.13%).

- Trade-off: Adding Car to Pubs Employ improves R² by 2.54% but increases complexity.

- Consistency: Pubs Employ aligns with selection methods, confirming its robustness.

**(c) Comment on Reduced Models:**

The forward and backward selection methods identified Pubs and Employ as the final predictors.

Final Model:

Ratevalue = 38.70 - 2.67 Pubs - 0.93 Employ

R² = 0.3959 (Explains 39.59% variance)

Cp = 1.7210 < p = 3 (Indicates an excellent fit with minimal bias)

Significance: Pubs (p = 0.0212), Employ (p = 0.0314)

Comparison:

Best among 2-variable models (Beats Food Employ and Pubs Car in both Cp and R²).

Slightly less explanatory than Pubs Employ Car (R² = 0.4213, Cp = 2.2170) but simpler and more interpretable.

**(d) Choosing the Final Model:**

Final Choice: Pubs Employ

Justification:

Lowest Cp (1.7210 < 3) ? Unbiased and well-fitted.

Parsimonious (Only 2 predictors) ? Simpler and interpretable.

Sufficient Fit (R² = 0.3959) ? Adequate explanatory power.

No Multicollinearity (VIF = 1.45) ? Stable coefficients.

Why Not Pubs Employ Car?

Higher R² (0.4213) but minimal improvement (+2.54%).

Adds complexity without strong justification (Car was not significant in part ii).

**(e) Are Other 2-Predictor Models Viable?**

Food Employ (Cp = 2.0691, R² = 0.3900) - Close contender but slightly weaker.

Pubs Car (Cp = 3.0281, R² = 0.3738) - Not competitive due to higher Cp and lower R².

**Conclusion:**

Best Model: Pubs Employ (Lowest Cp, highest R², strong statistical support).

Alternative: Food Employ (Slightly weaker but reasonable).

Pubs Car: Not a strong candidate.

**2) (iv) Consider again the reduced model obtained by backward elimination. State the fitted regression equation for this model. Interpret and briefly discuss each of the partial regression coefficients and the relevant related information in the output by SAS. (13 marks)**

🡪 **(a) Reduced model obtained from Backward Elimination:**

**(b) Fitted Regression Equation:**

Ratevalue = 38.70 - 2.67\*Pubs - 0.93\*Employ

Intercept: 38.70

Coefficient for Pubs: -2.67

Coefficient for Employ: -0.93

**(c) Interpretation of Partial Regression Coefficients:**

- Intercept (38.70463):

Baseline Ratevalue when Pubs and Employ are zero.

Highly significant (p < 0.0001).

- Pubs (-2.67492):

Each additional pub decreases Ratevalue by 2.67492 (p = 0.0212).

Significant negative effect.

- Employ (-0.93016):

Each unit increase in employment decreases Ratevalue by 0.93016 (p = 0.0314).

Moderate negative effect.

**(d) Model Fit and Diagnostics:**

- R² = 0.3959, Adj R² = 0.3632: Model explains 39.59% of the variance.

- F = 12.12 (p < 0.0001): Model is highly significant.

- Condition Number = 5.7802, VIF = 1.45: No multicollinearity issues.

**(e) Backward Elimination Process:**

Step 0: All variables (Pubs, Food, Social, Employ, Car) included. R² = 0.4250, Cp = 6.0000.

Step 1: Social removed (p = 0.8928 > 0.05). R² = 0.4247, Cp = 4.0185.

Step 2: Food removed (p = 0.6541 > 0.05). R² = 0.4213, Cp = 2.2170.

Step 3: Car removed (p = 0.2165 > 0.05). R² = 0.3959, Cp = 1.7210.

Summary: Variables were systematically removed based on p-values exceeding 0.05, leaving only significant predictors (Pubs, Employ).

**(f) Type II SS and F Values:**

- Intercept: Type II SS = 7237.40672, F = 114.83, p < 0.0001—dominant contribution to the model.

- Pubs: Type II SS = 365.29477, F = 5.80, p = 0.0212—significant unique contribution.

- Employ: Type II SS = 315.51567, F = 5.01, p = 0.0314—significant but slightly less impactful than Pubs.

**(g) Model Validation:**

**2) (v) Investigate the overall validity of the reduced model obtained by backward elimination by**

**undertaking suitable diagnostic analyses involving the studentised and/or deleted residuals**

**and/or the fitted values. Note that you are not required to produce plots involving residuals and**

**the surviving explanatory variables. (10 marks)**

🡪 **Interpretation:**

**(a) Model Overview:**

- Fit: R² = 0.3959 (39.59% variance explained), F = 12.12 (p < 0.0001), model is significant.

- Error: Root MSE = 7.93902, Dependent Mean = 22.12500.

- Coefficients: Intercept = 38.70463 (p < 0.0001), Pubs = -2.67492 (p = 0.0212), Employ = -0.93016 (p = 0.0314).

**(b) Diagnostic Analysis:**

Linearity:

- Sum of residuals = 0, Rstudent spread (e.g., -1.5410 to 6.5277) across fitted values (10.6786–35.0996) shows no clear trend.

- Exception: Obs 1 (Rstudent = 6.5277, Pred = 31.3789) is extreme but isolated.

- Result: Assumption holds, minor concern from Obs 1.

Normality:

- Mean = 0.0607 (p = 0.7645), Skewness = 3.4055, Kurtosis = 16.9642.

- Tests: KS (p < 0.010), CvM (p < 0.005), AD (p < 0.005) reject normality.

- Result: Violated, driven by Obs 1 (6.5277).

Homoscedasticity:

- Rstudent variance higher at Pred > 25 (SD ˜ 2.0) vs. <25 (SD ˜ 1.0).

- PRESS = 2828.35774 > SSE = 2332.03974 indicates instability.

- Result: Violated, heteroscedasticity present.

Outliers/Influence:

- Outliers: Obs 1 (Rstudent = 6.5277 > 3), 5/40 > |1| (within normal range).

- Influence: Obs 1 (Cook’s D = 0.791 > 0.1), Obs 18 (0.108 > 0.1), DFFITS Obs 1 = 2.2450 (>0.548).

- Result: Obs 1 is a major outlier and influential.

**Conclusion:**

- Validity: Robust fit (R² = 0.3959), but normality and homoscedasticity violations, due to Obs 1, affect inference reliability.

- Recommendation: Exclude Obs 1 and re-run (code above) to improve model assumptions.

**2) (vi) Considering again the reduced model obtained by backward elimination, identify any potentialinfluential observations (use the value of the variable ID to specify each such observation).**

**Further investigate the two most extreme of these potential influential observations with respect**

**to their effect on the model by considering the corresponding values of the leverage H, the**

**deleted residual, the covariance ratio C and their DFBETAS values. Note that you are not required**

**to construct a plot of observed versus fitted values for these observations. (14 marks)**

**🡪 (b) Identify Potential Influential Observations**

Criteria:

- Cook’s D: > 4/n = 4/40 = 0.1 (significant influence).

- Hat Diag H: > 2p/n = 23/40 = 0.15 (high leverage).

- Rstudent: > |2| (outlier), > |3| (extreme).

- DFFITS: > 2\*sqrt(p/n) ˜ 0.548 (influence on fit).

From Output Statistics (ID = Obs):

- Cook’s D > 0.1:

- ID 1: 0.791

- ID 18: 0.108

- Hat Diag H > 0.15:

- ID 18: 0.2277

- ID 3: 0.2126

- |Rstudent| > 2:

- ID 1: 6.5277

- DFFITS > 0.548:

- ID 1: 2.2450

- ID 18: 0.5706

Potential Influential Observations:

- ID 1: Cook’s D = 0.791, Rstudent = 6.5277, DFFITS = 2.2450 (extreme across metrics).

- ID 18: Cook’s D = 0.108, H = 0.2277, DFFITS = 0.5706 (influential due to leverage and fit).

- ID 3: H = 0.2126 (high leverage), but Cook’s D = 0.061, Rstudent = 0.8208 (less extreme).

Conclusion: ID 1 and ID 18 are the most potentially influential based on multiple criteria.

ID 3 is noted for leverage but less impactful.

**(c) Investigate the Two Most Extreme Influential Observations:**

We’ll analyze ID 1 and ID 18 (most extreme by Cook’s D and combined influence) using:

- Leverage (Hat Diag H)

- Deleted Residual (Rstudent)

- Covariance Ratio (Cov Ratio)

- DFBETAS (Intercept, Pubs, Employ)

(a) ID 1 (Obs 1):

- Data: Ratevalue = 65.0, Pred = 31.3789, Residual = 33.6211

- Metrics:

- Leverage (H): 0.1058 (< 0.15, not high leverage).

- Deleted Residual (Rstudent): 6.5277 (> 3, extreme).

- Cov Ratio: 0.1166 (< 1, variance increases).

- DFBETAS:

- Intercept: 2.0126 (> 2/sqrt(40) ˜ 0.316, large).

- Pubs: -1.7460 (large negative).

- Employ: 0.2248 (moderate).

- Interpretation:

- H: Low leverage; influence isn’t from extreme predictor values.

- Rstudent: Extreme outlier; prediction underestimates by 33.6211 units.

- Cov Ratio: Very low (< 0.9), indicating model precision worsens significantly without ID 1.

- DFBETAS: Large shifts in Intercept (up) and Pubs (down) if removed; Employ less affected.

- Effect: ID 1 pulls the Intercept down and Pubs coefficient up, heavily skewing the fit.

(b) ID 18 (Obs 18):

- Data: Ratevalue = 18.0, Pred = 10.6786, Residual = 7.3214

- Metrics:

- Leverage (H): 0.2277 (> 0.15, high leverage).

- Deleted Residual (Rstudent): 1.0509 (< 2, not an outlier).

- Cov Ratio: 1.2839 (> 1, variance decreases).

- DFBETAS:

- Intercept: -0.3021 (moderate).

- Pubs: 0.5279 (moderate).

- Employ: -0.2052 (small).

**- Interpretation:**

- H: High leverage; extreme Pubs and/or Employ values influence fit.

- Rstudent: Moderate residual; not an outlier.

- Cov Ratio: > 1 suggests slight precision gain without ID 18 (not extreme, < 1.3).

- DFBETAS: Moderate shifts; Intercept increases, Pubs decreases if removed, Employ minimally affected.

- Effect: ID 18 adjusts coefficients subtly, driven by leverage rather than residual size.

**2) (vii) Based on a comprehensive business analysis, the company will fully investigate a property as apotential purchase if it can reasonably be anticipated that the total income for one year (52**

**weeks) would be greater than or equal to the property's (freehold) price.**

**"The Good King James" has just come onto the market with an advertised freehold price of**

**£515,000. Its rateable value is £13,000, local unemployment is running at 10% and the property**

**has an accessible outside space of 100 m2.**

**Given that the company has satisfactorily investigated the various influential points identified in**

**(vi) and found all the data to be completely valid, use the reduced model obtained by backward**

**elimination to advise the company as to whether it should further investigate "The Good King**

**James", commenting on the reliability of your analysis.**

**What is the maximum price that the company should consider offering (assuming the necessary**

**further investigations are satisfactory) if it wishes to have 95% assurance that its potential**

**investment in the property is well founded? (8 marks)**

🡪 **Advice:**

- Prediction: Ratevalue = -47.68469 (£ thousands), Annual Income = -£2,479,603.88.

- Decision: Do not investigate "The Good King James" (income < £515,000).

- Reliability: Model (R² = 0.3959) is moderately reliable, but negative prediction is unrealistic, suggesting extrapolation beyond data range (Ratevalue 12–65).

Influential points validated, yet model fails to predict positive income.

**Maximum Price:**

- LCL95 = -52.74969, Income = -£2,742,983.88.

- Max Price = £0 (95% assurance not met).

- Note: Model fails to reflect actual Ratevalue = 13 (Income = £676,000).