BAX 442 001 002 - Advanced Statistics - Homework 6
Amit Gangane, Varun Rao, Shivani Tayade, Sravya Bhaskara, Karishma Mehta

Clustering and Mix-Regression

This report entails the analysis of smartphone customer segmentation using clustering and mixture regression. The objective was to obtain an optimal number of customer segments using clustering, understanding clustering penalty along with estimating elasticities for each segment to understand price sensitivity.

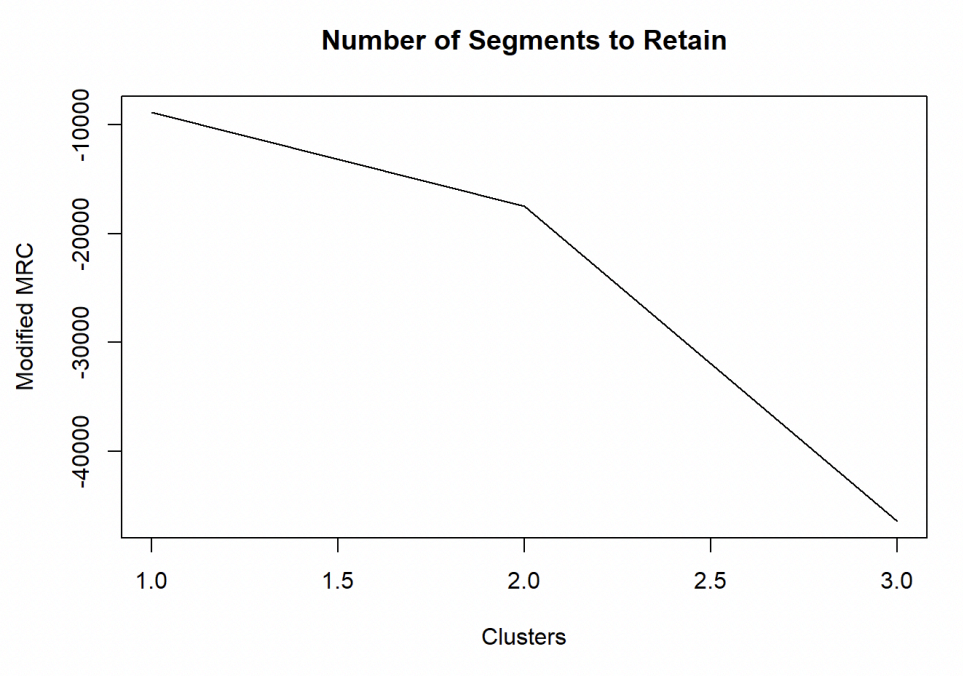
Clustering penalty is an adjustment added to standard goodness‑of‑fit criteria (AIC/BIC) to counter the natural tendency of these metrics to favor models with more clusters. In our context, as the number of clusters increases, the log‑likelihood tends to improve (i.e., become less negative), which may lead to over-segmentation. By subtracting a penalty term based on the relative sizes of clusters (with $\alpha_k = n_k/N$), we obtain a modified metric that more appropriately balances model fit with parsimony.

1. Clustering Approaches in Our Analysis

We experimented with two approaches:

1. **K-means Clustering Followed by Cluster-wise Regression:**
   - *Procedure:* We standardized the predictor variables and applied K-means for k = 1, 2, …, 10. For each candidate k, we ran separate regressions of log(price) on log-transformed predictors (e.g., log(height), log(hand size), etc.) within each cluster.
   - *Evaluation:* We manually computed the overall AIC from the sum of squared residuals and then subtracted the clustering penalty. The resulting plots show both the actual AIC and the modified AIC versus k. In our results, the penalty in this approach appeared rather severe—sometimes favoring a one-cluster model—which indicates that the balance between fit and complexity can be sensitive to the method used.
2. **Mixture Regression via flexmix:**
   - *Procedure:* We used the flexmix package to fit mixture regression models directly on the data, with the number of latent segments (k) varied from 2 to 10. For example, with k = 2, our model converged after 68 iterations with cluster sizes of 1079 and 1921, a log-likelihood of 6535.867, and AIC/BIC of –13005.73 / –12807.52.
   - *Evaluation:* We observed that as k increases beyond 5, the model often fails to converge reliably (with k = 4 also showing instability). When comparing the actual and modified AIC/BIC values from mixture regression, the clustering

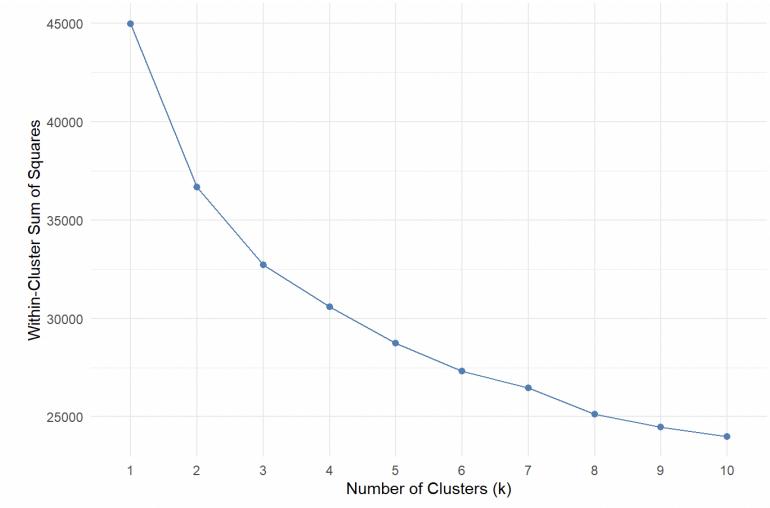
penalty tends to be less extreme than in the K-means approach, although modified BIC still sometimes over-penalizes the number of clusters.

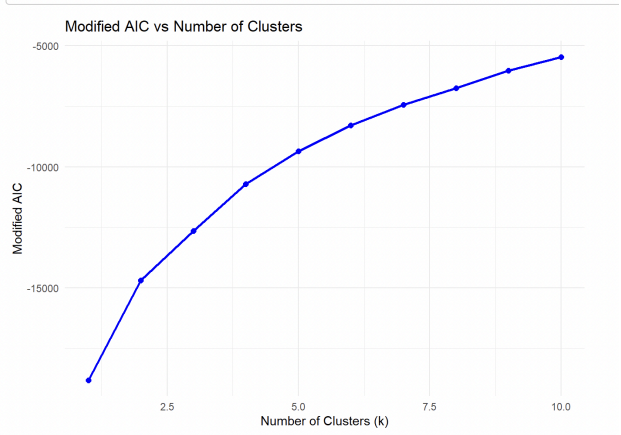


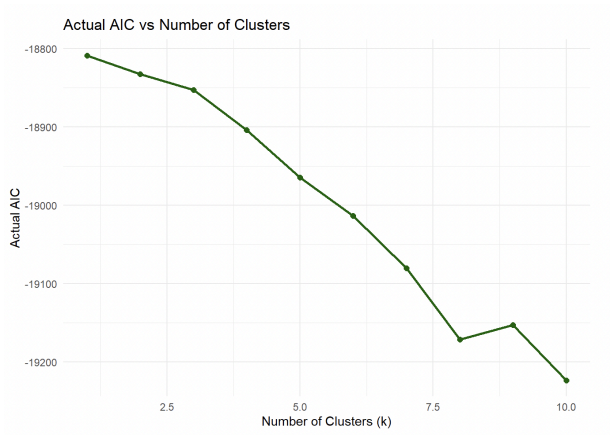
-

2. Selecting the Optimal Number of Clusters

Based on our mixture regression results and supplementary analysis using the Elbow (WSS) scree plot for K-means, we observed that three clusters (k = 3) provide a good balance between model fit and parsimony. Although a model with k = 4 may have a slightly lower modified AIC, its convergence is less stable, and the WSS plot also supports k = 3. For modified BIC, the penalty can be overly harsh—occasionally favoring a single-cluster solution—so our selection of k = 3 is based on both stability and interpretability considerations.







3. Interpreting Estimated Elasticities

Using the retained model (k = 3 via mixture regression), we examined the estimated coefficients (elasticities) for each segment. Because our dependent variable is log(price) and most predictors are log-transformed (or log(1+x) transformed), each coefficient represents the percentage change in price for a 1% change in the predictor. The key findings are:

- **Cluster 1:**
  - *Significant Variables:* Apple, Samsung, Screen Size, and Reading.
  - *Insights:*
    - The positive coefficient on the Apple dummy (≈0.165) indicates that customers in this segment pay a premium for Apple products.
    - Screen Size is critical; a 1% increase in screen size corresponds to roughly a 1.4% increase in price.
    - Reading usage is significant, implying that customers who frequently use their phones for reading are willing to pay more.
- **Cluster 2:**
  - *Significant Variables:* Apple, Samsung, Screen Size, and Days_Ago.
  - *Insights:*
    - Similar brand premium effects are observed with a strong preference for Apple.
    - Here, the elasticity on Screen Size is even larger (about 1.9%), meaning that this group values larger displays highly.
    - The significance of Days_Ago suggests these customers prefer newer smartphones, as they are willing to pay extra for more recent models.

- **Cluster 3:**
  - *Observations:*
    - Although p-values are not available for most estimates in this cluster, the magnitude of the coefficients for Apple, Samsung, and Screen Size is substantial.
    - This cluster also indicates a premium for well-known brands and shows that Screen Size remains the dominant factor (with an elasticity of around 0.9%).

4. Managerial Insights Based on the 4Ps

- **Price:**
  All clusters exhibit a clear willingness to pay a premium for the Apple brand, with Samsung also contributing positively. Notably, Screen Size is a key differentiator in price determination. Cluster 2, in particular, appears more responsive to increases in screen size—suggesting that emphasizing high-quality, larger displays could enable premium pricing.
- **Product (Design) & Promotion:**
  - *Cluster 1:*
    - With reading-related features being significant, product design improvements that enhance readability (e.g., better display resolution and brightness) may capture this market segment. Promotional campaigns highlighting these features could be especially effective.
  - *Cluster 2:*
    - The importance of Days_Ago indicates that customers in this segment are attracted to newer models. Advertising that stresses cutting-edge technology and recent advancements would likely resonate with this group.
  - *Cluster 3:*
    - Despite limited significance testing, the substantial coefficients on brand and screen size suggest similar preferences. Consistent emphasis on premium design and high-performance screens is recommended.
- **Place:**
  Since the dataset does not include detailed information on distribution channels, it is challenging to extract specific insights regarding place. Future analyses incorporating distribution data would be needed to derive actionable strategies in this area.

This report synthesizes the complete analytical process—from applying clustering algorithms and running regression models to modifying the standard information criteria and interpreting elasticities—using the results from our analysis. The findings provide both a statistically sound

segmentation (favoring k = 3) and actionable insights for pricing, product design, and promotion strategies.