**Project Final – PSYC 8803**

**Kamala Natarajan**

**Objective:** This project final is aimed to explore advanced methods for handling missing data in a dataset focused on forecasts and associated variables. Missing data can bias results and reduce statistical power if not properly addressed. This analysis investigates methods to handle missing data under different mechanisms, such as Completely Missing at Random (CMAR), and Missing Not at Random (MNAR). The methods employed include Maximum Likelihood Estimation (MLE), Bayesian Inference, and Selection Models. Special emphasis is placed on the role of auxiliary variables in improving bias and precision.

Question 1: Table 1: Missing rates for each variable: PHEV has no missing values while the other variables have the indicated missing rates as shown.

A screenshot of a computer

Description automatically generated

Question2: Table 2: Regression Coefficients for the Linear Regression Model without auxiliary variables and omitting NA values and model metrics.A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

**Relationship Between Predictors and Vehicle**

* **Approval**: There is a significant positive relationship between Approval and Vehicle (p < 0.001). This means that higher Approval scores are associated with higher Vehicle scores. Hence, Approval is a meaningful predictor.
* **PHEV**: There is a positive relationship between PHEV and Vehicle, but it is not statistically significant (p > 0.05). This suggests that PHEV may not be a meaningful predictor of Vehicle in this model.
* The model explains 8.63% of the variance in Vehicle (Multiple R-squared: 0.0863). This is relatively low, indicating that other factors not included in the model may have a significant impact on Vehicle.

Question 3: Table 2 and 3: Data showing results for determining Cohens’ D coefficients.

A screenshot of a computer

Description automatically generated

For Vehicle: BoxOffice, dn, crt, cfs, vs all has a cohens'D > 0.2 based on the est.std column. However, age does not have a cohen's D > 0.2.

A group of numbers on a white background

Description automatically generated

For Approval: BoxOffice , dn, crt, cfs, vs all have a cohens'D > 0.2 based on the est.std column. However, age does not have a cohen's D > 0.2.

Table 4: Residual correlation coefficients in the Std.all Column at the very end of the table.

A table of numbers with numbers

Description automatically generated

BoxOffice, dn, crt, cfs, vs all have a cohens'D > 0.2 based on the est.std column from the cohen’s D analysis for both Vehicle and Approval. Age does not have a cohen's D > 0.2 or residual correlation > 0.3 for Vehicle or Approval. For Vehicle, BoxOffice , dn, crt, cfs, vc are catergory B variables since none of the variables had a residual correlation > 0.3 based on the Std. all column results of the residual correlation analysis. For Approval, crt, dn, and cfs have Std.all > 0.3 and cohen's D greater than 0.2 and are hence category A variables. Age is category C variable that is not correlated with missingness or the residuals of Vehicle and Approval.

Question 4: For Approval, dn, crt, cfs are category A variables and hence need to be included in the model to mitigate bias and improve statistical power. For Vehicle and Approval, BoxOffice and vs are catergory B variables. They need to be included for improving precision of estimates. Age is a category C variable for both Approval and Vehicle so it does not need to be included as it won’t help mitigate bias or improve statistical power.

Question 5: I have used brms, rblimp and maximum likelihood: All three methods—Bayesian (brms), Maximum Likelihood (semTools), and rblimp—show consistent significant positive effects of Approval on Vehicle, with effect sizes around 0.30. The effects of PHEV on Vehicle are smaller and consistently positive across all methods, though slightly varying in significance.

Question 6: The chains exhibit good mixing, moving freely across the parameter space without being stuck. The chains overlap and appear as though they are sampling from the same distribution. The trace lines are stable and fluctuate around a consistent mean value without noticeable trends or drifts.

A screenshot of a graph

Description automatically generated A screenshot of a graph

Description automatically generated

Question 7: Table 5: Condensed Results of all CMAR Models shown below.

A table of numbers and symbols

Description automatically generated with medium confidence

**Approval** consistently shows a strong positive effect on Vehicle across all models, with slightly higher estimates in the Bayesian and maximum likelihood CMAR models compared to the simple linear regression. **PHEV**'s effect on Vehicle is significant in CMAR models, but not in the simple linear regression. The effect size is higher in CMAR models.

Question 8: Table 6: Condensed Results table for all 3 models (Screens Shots shown in the end for all three from blimp)A screenshot of a graph

Description automatically generated**CMAR** shows a significantly better fit with much lower DIC2 and WAIC values compared to the Focused and Diffuse Selection models. It also has a lower residual variance. The CMAR model explains a smaller proportion of the variance in Vehicle through the coefficients (13.7%), but the overall model fit (as indicated by DIC2 and WAIC) is substantially better.Although CMAR has slightly lower coefficients for Approval and PHEV, its much lower information criteria values suggest a superior overall performance in modeling the data under the assumption that missing data is completely at random. ​

Screenshots from Blimp: (MNAR and CMAR Models)

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated