DS 853 Case Study 3

Prepared by:

Marcus Nogueira and Chin Ting Wong

Professor:

Leyla Ozsen

Date:

November 2024

Discussion 1

1. What is the dependent variable?

   The dependent variable is "Amount of Prepaid Card ($)".

2. What are the independent variables?

   The independent variables include "Age", "Days per Month at Starbucks", "Cups of Coffee per Day", and "Income ($1,000s)".

3. Create a correlation matrix for all the variables using the Correlation Tool of Excel's Data Analysis Tool Pack.

	Dollar Amt.	*Age*	*Days*	*Cups*	*Income*
Dollar Amt.	*1*				
Age	*0.21514123*	*1*			
Days	*0.40686371*	*0.03745681*	*1*		
Cups	*0.28622698*	*0.26828876*	*0.58760095*	*1*	
Income	*0.85003234*	*0.17784941*	*0.30543755*	*0.15945113*	*1*

4. Interpret the correlation matrix. The following two bullet points should help organize your answer this question:

   a. Comment on those independent variables that have a high correlation with the dependent variables, if there are any.

      "Income" shows a high positive correlation with "Dollar Amt." (correlation coefficient = 0.85), suggesting that as income increases, the dollar amount of prepaid card also tends to increase. This strong relationship implies that income could be a significant predictor of prepaid card dollar amount.

   b. Comment on those independent variables that are highly correlated with each other, if there are any.

      "Cups" and "Days" have a moderate positive correlation (correlation coefficient = 0.59), which suggests that these variables may have a relationship. For example, customers who visit Starbucks more often may also consume more cups of coffee per day.

5. Conduct a multiple regression analysis using Excel. Include the regression analysis summary output generated by Excel in your report. You can copy/paste it or include a screenshot of the Excel output.

SUMMARY OUTPUT

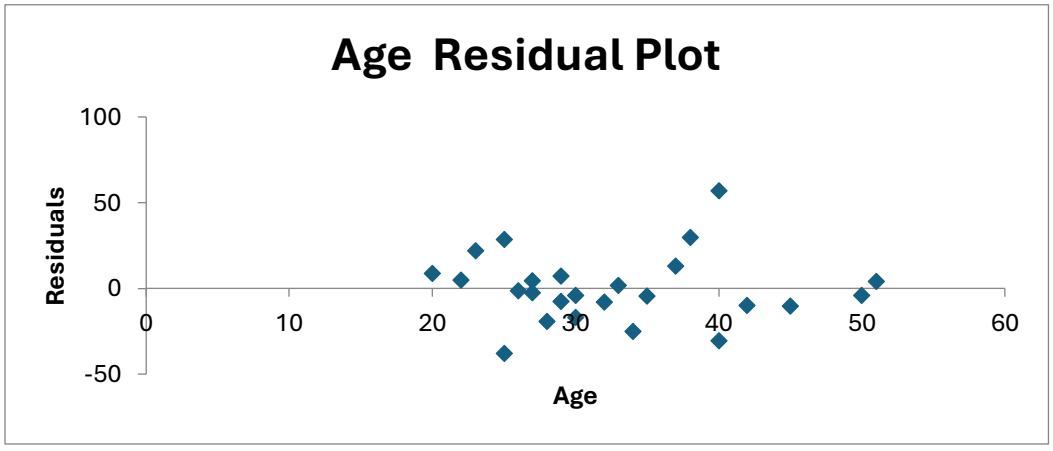
Regression Statistics	
Multiple R	0.868700283
R Square	0.754640181
Adjusted R Square	0.705568218
Standard Error	22.14831563
Observations	25

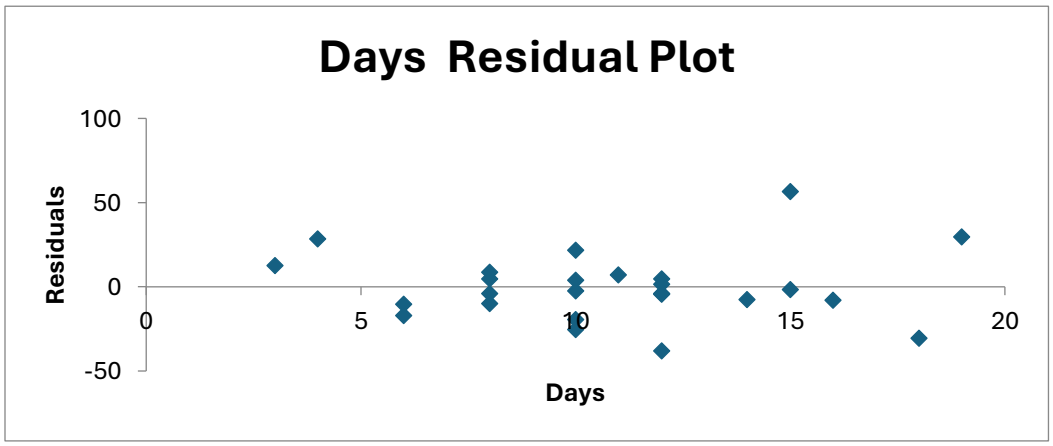
ANOVA

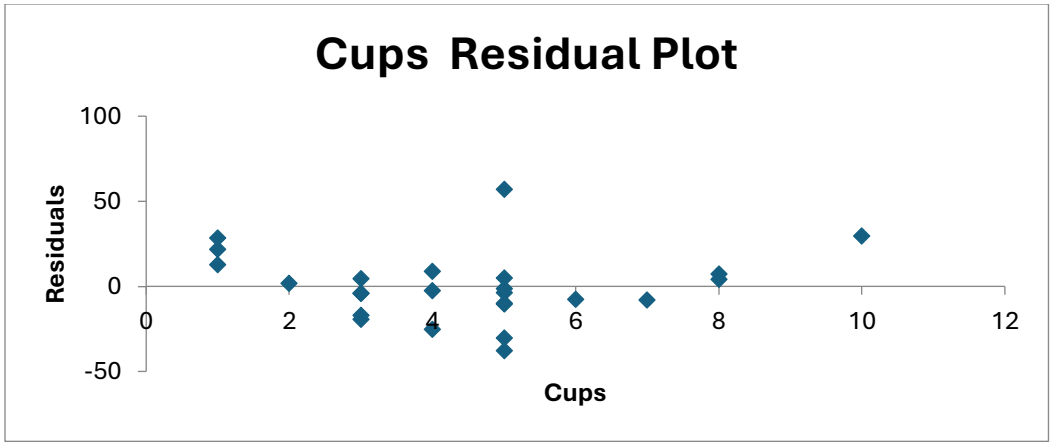
	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	30175.0423	7543.760574	15.37823483	6.75799E-06
Residual	20	9810.957705	490.5478852		
Total	24	39986			

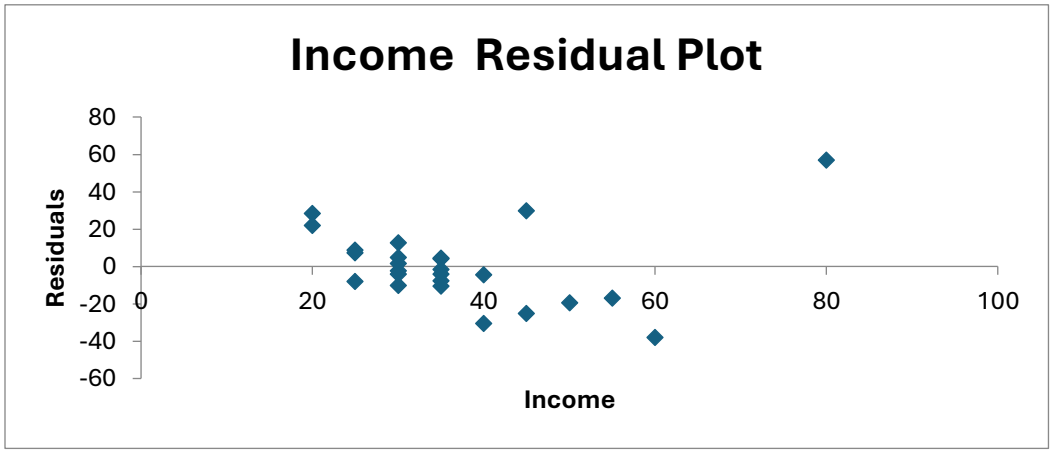
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-83.82574324	22.49444927	-3.726507914	0.001332354	-130.7483422	-36.90314429	-130.7483422	-36.90314429
Age	0.236928518	0.575905816	0.411401502	0.68515331	-0.964389964	1.438247	-0.964389964	1.438247
Days	1.189657003	1.473934142	0.807130366	0.429086253	-1.884915741	4.264229747	-1.884915741	4.264229747
Cups	1.42161108	2.631048846	0.540321052	0.594942115	-4.06666064	6.909882801	-4.06666064	6.909882801
Income	2.406542923	0.359719072	6.69006208	1.64135E-06	1.656182088	3.156903759	1.656182088	3.156903759

6. Include and interpret the residual plot in your report. Are the assumptions for regression met?



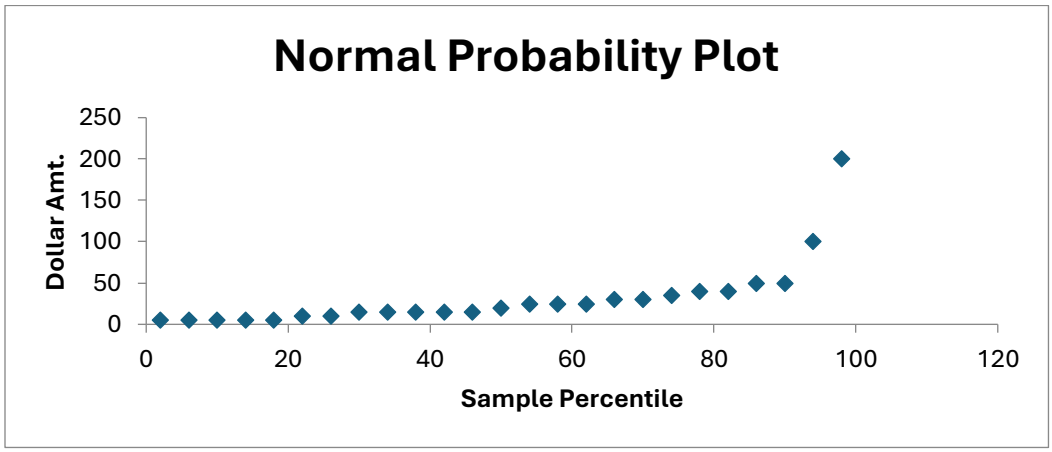






The residual plots suggest that most of the key regression assumptions are met, with some minor concerns. The assumption of linearity is supported by the absence of clear curved patterns in all plots, indicating linear relationships between the predictors and the outcome. Independence also appears to be satisfied, as there are no discernible patterns or trends in the residuals. The assumption of equal variance (homoscedasticity) is mostly met, although there are slight signs of heteroscedasticity in the "Income" and "Cups" residual plots, where the spread of residuals appears somewhat uneven. These issues, however, are not severe enough to invalidate the model. Overall, the assumptions are reasonably satisfied, though the minor violations should be noted as limitations.

7. Include and interpret the normality plot in your report. Are the assumptions for regression met?



The normal probability plot reveals a curved pattern rather than the expected straight-line alignment, indicating a violation of the normality assumption for regression.

8. What is the standard error of the estimate, S_e?

From the regression output above, we know that the standard error of the estimate is 22.1483.

9. Does the size of the standard error of the estimate, S_e, pose practical concerns?

The size of the standard error of the estimate poses practical concerns in the context of predicting Starbucks prepaid card dollar amounts. This value represents the typical deviation between predicted and actual amounts, meaning predictions could be off by approximately $22 on average. Given that many customers load amounts like $25 or $50, this error margin could represent a discrepancy for typical card values.

10. Are there outliers in the data? Examine the S_e and check to see if there are residuals outside (-2 S_e, 2 S_e).

    The residual ranges between -37.8740 and 56.8723, with 24 out of 25, or 96% falling within 2 standard error, which ranges between -44.2966 and 44.2966. With that, we have one outlier in the data.

11. Comment on the statistical significance of each partial regression coefficient.

    Based on the p-values from the regression output, "Income" is statistically significant at the 0.05 significance level with a p-value of 0.000002, indicating that it contributes meaningfully to predicting the "Dollar Amt." In contrast, "Age", "Days", and "Cups" all have p-values greater than 0.05 (0.685153, 0.429086, and 0.594942, respectively), suggesting that these variables are not statistically significant at the 0.05 level and do not have a strong impact on explaining the variation in "Dollar Amt." Therefore, only "Income" has a meaningful relationship with the dependent variable in this model.

12. Comment on the statistical significance of the overall model.

    The statistical significance of the overall regression model can be assessed using the F-statistic and its associated p-value from the ANOVA table. In this case, the F-statistic is 15.38, and the p-value is 6.758E-06. Since the p-value is much smaller than the commonly used significance level of 0.05, we can conclude that the overall model is statistically significant. This indicates that at least one of the predictors ("Age", "Days", "Cups", or "Income") contributes meaningfully to explaining the variation in the dependent variable, "Dollar Amt." Therefore, the regression model as a whole is a good fit for the data.

13. What is the R^2 adjusted of the model?

    The adjusted R^2 of the model is 0.7056.

14. Comment on the R^2 adjusted value.

    The adjusted R^2 of 0.7056 indicates that approximately 70.56% of the variance in the dependent variable ("Dollar Amt.") is explained by the independent variables ("Age", "Days", "Cups", and "Income")

in the model, after adjusting for the number of predictors used. The adjusted R^2 is generally considered a more reliable measure than R^2 when comparing models with different numbers of predictors, as it accounts for the potential overfitting that can occur with additional predictors.

15. Comment on the multiple regression model. More specifically, address the following question: Would you recommend the use of this multiple regression model in predicting the amount of money people spend on their debit card? Why? / Why not?

We would not recommend using this multiple regression model to predict the dollar amount of prepaid card due to several key issues related to both model performance and assumption violations.

While "Income" is statistically significant and strongly correlated with the amount spent, other predictors like "Age", "Days", and "Cups" are not statistically significant, which suggests they do not contribute meaningful predictive value. The inclusion of these non-significant variables introduces unnecessary complexity and noise into the model. Additionally, the standard error of estimate is relatively large, meaning predictions could be highly inaccurate, especially for typical spending amounts. The model's ability to predict accurately is further compromised by the presence of outliers, which may lead to instability in the predictions.

Moreover, the model violates some important regression assumptions. The "normality" assumption is not fully met, as indicated by the curved pattern in the normal probability plot, and there are signs of "heteroscedasticity" in the residuals for "Income" and "Cups", which can undermine the reliability of hypothesis tests and confidence intervals. Given these issues, the model is not suitable for reliable predictions. A simplified approach, focusing solely on "Income" as a predictor, would likely offer more accurate and stable predictions, particularly since it shows strong correlation and statistical significance. Adjusting for assumption violations or exploring alternative modeling techniques could also improve the predictive power of the model.

16. What are your overall conclusions and recommendations as a business consultant? What sales implications might be from this analysis?

Overall, it is found that "Income" is the primary driver of prepaid card spending, and the current model indicates that focusing on higher-income segments could yield the most significant impact. Given that "Income" is the strongest and only statistically significant predictor, it is recommended that Starbucks refines their marketing strategies to target higher-income customers, potentially through tiered loyalty programs and premium prepaid card options. Developing products such as premium cards with additional benefits and auto-reload features for higher-income customers would align well with spending patterns.

Sales implications from this analysis suggest an opportunity to increase revenue by focusing efforts on the high-income demographic, offering premium products and services tailored to their needs. However, the model's high standard error means there is a limited ability to predict exact load amounts, and additional factors that influence spending may not be captured. Therefore, it's crucial to develop income-based marketing campaigns, create premium products, and implement targeted loyalty rewards. To improve future predictions, gathering additional customer data and continuously monitoring performance will help refine these strategies and reduce risks of overlooking potential customer segments.

Discussion 2

1. What is the dependent variable?

   The dependent variable is "Days per months at Starbucks".

2. What are the independent variables?

   The independent variables include "Age", "Income ($1,000's)", and "Cups of Coffee per day".

3. Create a correlation matrix for all the variables using the Correlation Tool of Excel's Data Analysis Tool Pack.

	Days	Age	Cups	Income
Days	1			
Age	0.03745681	1		
Cups	0.58760095	0.26828876	1	
Income	0.30543755	0.17784941	0.15945113	1

4. Interpret the correlation matrix. The following two bullet points should help organize your answer this question:

   a. Comment on those independent variables that have a high correlation with the dependent variables, if there are any.

   There is a moderate positive correlation between "Cups" and "Days" (correlation coefficient = 0.59), suggesting that as the number of cups of coffee increases, the number of days per month a customer visits Starbucks also tends to increase. This relationship indicates that "Cups" could be a relevant predictor for "Days".

   b. Comment on those independent variables that are highly correlated with each other, if there are any.

   Looking at the correlation matrix, it seems like the independent variables have low correlation with each other, since they are all below 0.5.

5. Conduct a multiple regression analysis using Excel. Include the regression analysis summary output generated by Excel in your report. You can copy/paste it or include a screenshot of the Excel output.

SUMMARY OUTPUT

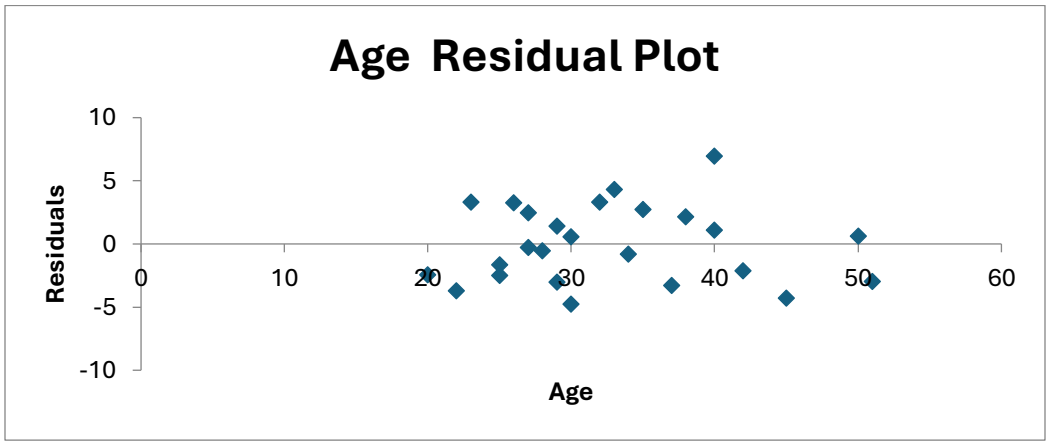
Regression Statistics	
Multiple R	0.644881058
R Square	0.415871579
Adjusted R Square	0.332424662
Standard Error	3.279087289
Observations	25

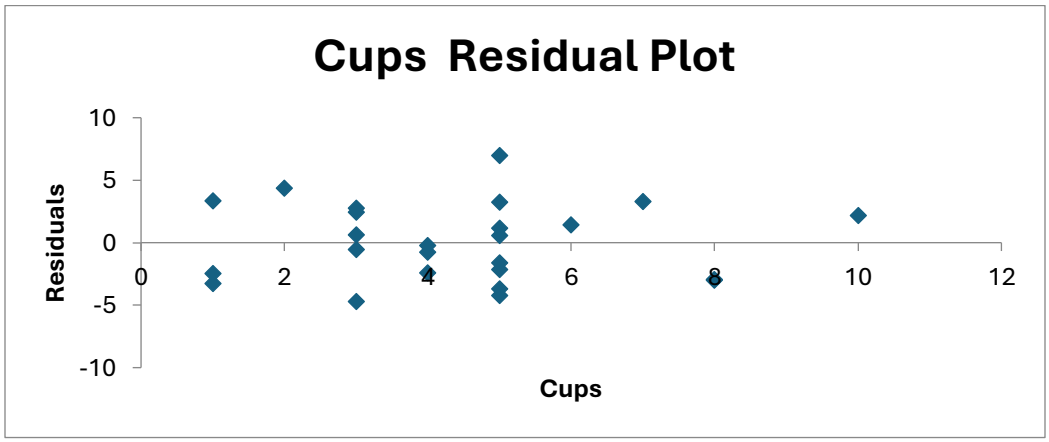
ANOVA

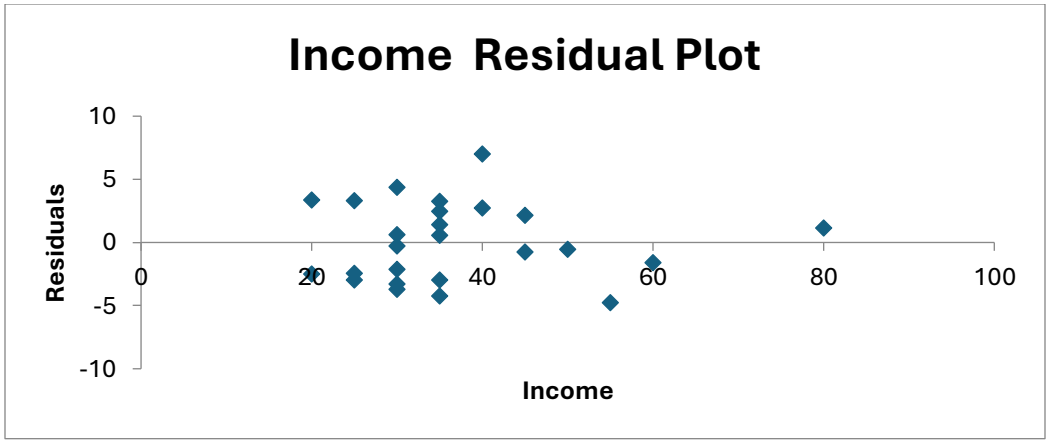
	df	SS	MS	F	Significance F
Regression	3	160.7593176	53.58643921	4.983666177	0.009128891
Residual	21	225.8006824	10.75241345		
Total	24	386.56			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	5.968355082	3.065104593	1.94719459	0.065008201	-0.405878865	12.34258903	-0.405878865	12.34258903
Age	-0.078533985	0.083523585	-0.940261183	0.357774296	-0.25223079	0.095162819	-0.25223079	0.095162819
Cups	1.064407793	0.312702971	3.403894085	0.002673504	0.414106365	1.71470922	0.414106365	1.71470922
Income	0.071611585	0.050912617	1.406558697	0.17418491	-0.034266999	0.177490168	-0.034266999	0.177490168

6. Include and interpret the residual plot in your report. Are the assumptions for regression met?

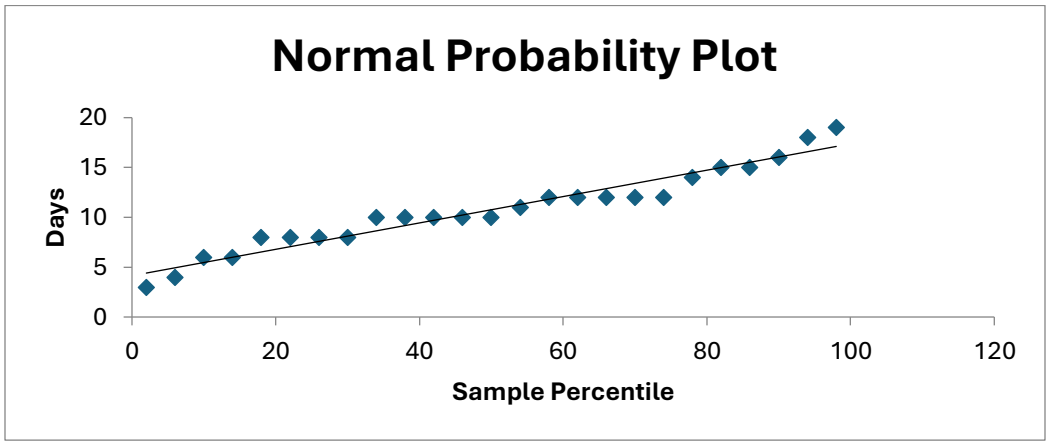






The residual plots for "Age", "Cups", and "Income" suggest that the key regression assumptions are reasonably well met. Linearity is supported by the random scatter of points around zero in all three plots, with no clear patterns indicating non-linearity. Equal variance (homoscedasticity) is satisfied, as the spread of residuals remains consistent across the x-axes, with no funnel or megaphone shape. Independence is also upheld, as the plots show random scatter with no systematic patterns or clustering.

7. Include and interpret the normality plot in your report. Are the assumptions for regression met?



The normal probability plot indicates that the residuals from the linear regression model are approximately normally distributed, as evidenced by the points lying close to an upward slopping straight line.

8. What is the standard error of the estimate, S_e?

   From the regression output above, we know that the standard error of the estimate is 3.2791.

9. Does the size of the standard error of the estimate, S_e, pose practical concerns?

   The size of the standard error of the estimate (S_e) of 3.2791 may pose practical concerns. This value indicates that predictions could be off by approximately 3.28 days on average. Given the context of business decisions, such as revenue projections, customer behavior analysis, staffing, and inventory management—a variation of 3-4 days per month may be substantial. It could lead to inaccurate sales forecasts, suboptimal staffing schedules, imprecise inventory ordering, and unreliable customer behavior modeling.

10. Are there outliers in the data? Examine the S_e and check to see if there are residuals outside (-2 S_e, 2 S_e).

    The residual ranges between -4.7442 and 6.9865, with 24 out of 25, or 96% falling within 2 standard error, which ranges between -6.5582 and 6.5582. With that, we have one outlier in the data.

11. Comment on the statistical significance of each partial regression coefficient.

The statistical significance of the partial regression coefficients reveals that "Cups" is the only independent variable with a significant relationship to the dependent variable, "Days", with a p-value of 0.0027, which is less than the 0.05 significance level. This suggests that the number of cups has a meaningful effect on predicting how many days per month a customer visits Starbucks, with each additional cup increasing the number of days by approximately 1.06. On the other hand, "Age" and "Income" are not statistically significant, with p-values of 0.358 and 0.174, respectively, indicating that they do not have a meaningful effect on "Days" when controlling for other variables in the model.

12. Comment on the statistical significance of the overall model.

The overall model is statistically significant, as indicated by the Significance F value of 0.0091, which is less than the 0.05 significance level. This suggests that at least one of the independent variables ("Age", "Cups", or "Income") significantly contributes to explaining the variation in the dependent variable ("Days"). The F-statistic of 4.98 further supports this conclusion, indicating that the model as a whole is a good fit for the data. Therefore, we can reject the null hypothesis that all regression coefficients are equal to zero, implying that the independent variables collectively have a statistically significant impact on "Days".

13. What is the R^2 adjusted of the model?

The adjusted R^2 value of the model is 0.3324.

14. Comment on the R^2 adjusted value.

The adjusted R^2 indicates that approximately 33.24% of the variability in the dependent variable ("Days") can be explained by the independent variables ("Age", "Cups", and "Income") after adjusting for the number of predictors in the model. While this indicates that the model explains a moderate portion of the variance, it also suggests that there is still a significant amount of unexplained variability. The relatively low adjusted R^2 value indicates that while the model provides some insight, additional variables or improvements to the model may be necessary to increase its explanatory power.

15. Comment on the multiple regression model. More specifically, address the following question: Would you recommend the use of this multiple regression model in predicting the amount of money people spend on their debit card? Why/Why not?

The multiple regression model has several limitations that reduce its effectiveness in predicting the amount of money people spend on their debit cards. The adjusted R^2 of 0.3324 indicates that only about one-third of the variation in spending is explained by the current variables, suggesting that a significant portion of the spending behavior remains unexplained. Moreover, only "Cups of Coffee" is statistically significant (p=0.0027), while the variables "Age" and "Income" lack statistical significance, making them unreliable predictors. Additionally, the standard error of estimate of 3.28 days means that predictions could be off by a substantial margin, which could lead to forecasting errors in the business context. Residual plots show generally acceptable patterns, but some outliers are present, which can further undermine the model's reliability. Given these issues, we would not recommend using this multiple regression model for predicting the amount of money people spend on their debit cards. The model's weak predictive power, combined with high error margins, makes it unsuitable for accurate and dependable forecasting.

16. What are your overall conclusions and recommendations as a business consultant? What marketing implications might be evident from this analysis?

Based on the analysis, it is evident that "Cups of Coffee per day" is the most significant predictor of "Days per month at Starbucks," with a moderate positive correlation (correlation coefficient = 0.59) suggesting that customers who drink more cups of coffee tend to visit Starbucks more frequently. "Income" and "Age" were not statistically significant predictors in this model, meaning they do not have a meaningful impact on visit frequency when considering the number of cups consumed. Given this, Starbucks should consider focusing marketing efforts on promoting higher coffee consumption per visit, rather than targeting income or age demographics.

From a business perspective, it would be beneficial for Starbucks to enhance its loyalty programs and product offerings around increasing cups consumed per visit. Marketing campaigns could emphasize rewards for volume (e.g., buy 3 cups, get 1 free) to encourage customers to increase their daily coffee intake. Additionally, creating bundle offers or offering promotions during specific times of day could encourage customers to buy more drinks, especially targeting customers who typically purchase only one drink per visit.