Analysis of Consumer Spending in America

Calvin Kapral

Problem introduction:

Does the spending of money on non-essential items coincide with eating out more, and likewise, does spending more on essential items imply eating out less? In a dynamic economy, spending habits can change monthly, weekly, or even daily. Because of these changes, it is important to note and examine the changes between different categories of goods and understand not only the economic factors behind spending habits, but also the impact of spending habits on one good on spending habits of another good. We are interested in spending in the accommodation and food services industry, characterized by bars, hotels, takeout food, and restaurants.

Dataset:

Our dataset, Percent Change in Consumer Spending, was sourced from Kaggle, the popular data science platform with many datasets available to the public. This dataset contains 4730 observations with 11 features. The first two features are "State FIPS code" and "Date", neither of which are relevant in this analysis, as the first identifies geographical locations and the second is only relevant in a time series. The remaining 9 features are percent changes in consumer spending in different industry sectors.

We are only interested in some of these features, so we take a subset of these 9 features for our analysis. Specifically, we are interested in spending on accommodation and food services, so it is our response variable (denoted ACF_spending). We take another 5 of these features for our predictor variables:

1. All merchant category spending (All_merchant)
2. Arts, entertainment, and recreation spending (AER_spending)
3. General merchandise stores, apparel, and accessories spending (GEN_spending)
4. Grocery and food store spending (GRF_spending)
5. Health care and social assistance spending (HCS_spending)

This selection of features ensures that a variety of different sectors are being covered. For example, health care and social assistance spending is a necessary item to have and one could hypothesize that if you cannot afford health care, you cannot afford to go out to eat at a restaurant. An interesting point to look at is the relationship between grocery and food store spending; is spending money at the grocery store substitutable for getting takeout? That is to say, is there an inverse relationship between the two? There is also the inclusion of less essential goods such as general merchandise and apparel stores or entertainment and recreation. We hypothesize that as spending on the arts, entertainment, recreation, and apparel increase, so will our response variable, as affording to do fun activities or

buying new clothes and accessories would likely mean that one could afford to go out to a restaurant and eat a meal.

Initial Analysis:

For our first steps we utilized R to create a summary of variables of interest which returned the following output.

```
 All_merchant      ACF_Spending      AER_Spending
Min.   :-44.000   Min.   :-75.700   Min.   :-83.60
1st Qu.: -1.270   1st Qu.:-24.900   1st Qu.:-40.50
Median : 11.800   Median : -2.120   Median :-10.50
Mean   :  8.517   Mean   : -7.359   Mean   :-12.81
3rd Qu.: 19.500   3rd Qu.: 12.800   3rd Qu.: 12.70
Max.   : 45.700   Max.   : 47.700   Max.   :107.00
 GRF_Spending      HCS_Spending      GEN_Spending
Min.   :-12.60   Min.   :-86.100   Min.   :-49.100
1st Qu.: 12.32   1st Qu.: -6.938   1st Qu.:  3.248
Median : 17.30   Median :  8.090   Median : 16.800
Mean   : 17.46   Mean   :  4.510   Mean   : 14.082
3rd Qu.: 22.10   3rd Qu.: 20.000   3rd Qu.: 27.400
Max.   : 81.00   Max.   :152.000   Max.   : 67.300
```

With first glance we can see that all variables have both negative minimum values (potential refunds or adjustments) and positive maximum values which reflects diverse consumer behavior. On top of this the median values suggest that spending on essential categories like Grocery Spending is relatively consistent, whereas discretionary categories like food service Spending and Art and entertainment Spending show wider variability. Then looking into the distribution of individual variables we observe Food Service Spending has a wide range (-75.7 to 47.7), indicating variability in dining out behavior across consumers. While a variable such as Grocery Spending has a more consistent spread (min: -12.6, max: 81.0) and a higher median, suggesting that essential grocery expenditures are less variable.

After analyzing the overall summary, a correlation matrix was then generated.

	All_merchant	ACF_Spending	AER_Spending	GRF_Spending	HCS_Spending	GEN_Spending
All_merchant	1.0000000	0.8981462	0.79315679	0.19301567	0.79480075	0.9020270
ACF_Spending	0.8981462	1.0000000	0.83653467	0.07199690	0.73271678	0.8285839
AER_Spending	0.7931568	0.8365347	1.00000000	0.02084595	0.61819093	0.7011928
GRF_Spending	0.1930157	0.0719969	0.02084595	1.00000000	0.08935423	0.2233188
HCS_Spending	0.7948007	0.7327168	0.61819093	0.08935423	1.00000000	0.7260542
GEN_Spending	0.9020270	0.8285839	0.70119281	0.22331881	0.72605416	1.0000000

The correlation matrix provides strong evidence supporting the hypothesis that spending on non-essential items is positively associated with dining out. The strong correlations between Food Service Spending and other non-essential categories, such as **Arts and Entertainment Spending and General Spending**, reflect this relationship. On the other hand, the weak correlations between Grocery Spending and Food Service Spending suggest that spending on essential items, such as groceries, does not significantly impact dining out behavior. These findings reinforce the idea that essential and non-essential spending habits are largely independent of each other.

Initial preview of the full regression model:

A multiple linear regression model was created in R and following is its summary.

The regression model demonstrates a strong fit, with an R^2 value of 0.8563 and an adjusted R^2 of 0.8561, indicating that approximately 85.63% of the variability in the response

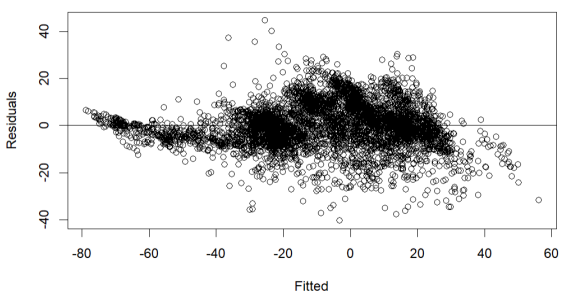
```
Residuals:
    Min      1Q  Median      3Q     Max
-40.322  -5.556  -0.117   5.818  44.799

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -11.313146   0.318083 -35.567  < 2e-16 ***
All_merchant   0.824878   0.027814  29.657  < 2e-16 ***
AER_Spending   0.235420   0.006884  34.196  < 2e-16 ***
GRF_Spending  -0.162403   0.014483 -11.213  < 2e-16 ***
HCS_Spending   0.043635   0.008517   5.123 3.12e-07 ***
GEN_Spending   0.183447   0.017096  10.730  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.415 on 4724 degrees of freedom
Multiple R-squared:  0.8563,    Adjusted R-squared:  0.8561
F-statistic:  5628 on 5 and 4724 DF,  p-value: < 2.2e-16
```

variable is explained by the predictors. The F-statistic (5628, $p < 2.2e-16$) confirms the model's statistical significance.

All predictors are highly significant ($p<0.001$), with logical coefficients. All merchant category codes spending ($\beta=0.8249$) has the largest positive impact, reflecting its strong influence on total spending. Non-essential spending categories, such as Arts and Entertainment **Spending** ($\beta=0.2354$) and General Merchandise Spending ($\beta=0.1834$), also contribute positively, indicating their alignment with increased overall spending. In contrast, Grocery Spending ($\beta=-0.1624$) has a significant negative effect, suggesting a trade-off between grocery spending and discretionary expenses like dining out, which makes sense as a meal bought a grocery store results in less meals needing to be bought from a restaurant or other food service. Health Spending ($\beta=0.0436$) has a small but positive contribution. The residual standard error (9.415) and centered residuals indicate a well-fitting model with no significant skewness. These findings support the hypothesis that non-essential spending coincides with increased dining out, while essential spending may lead to reduced discretionary expenditures.

Model diagnostics, transformations, and correlated errors

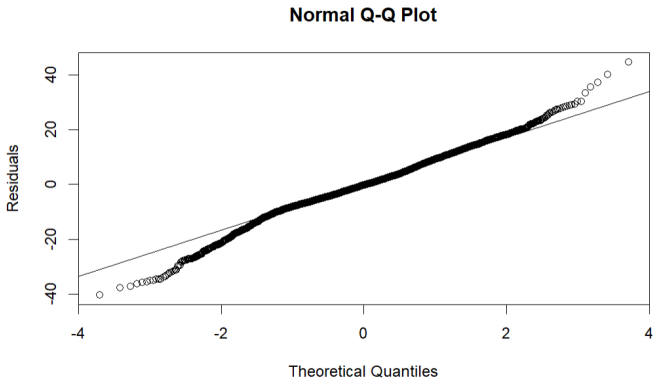
In using this model to predict spending on the accommodation and food services industry, it is important to verify the assumptions we have placed on our model. One of these assumptions is our assumption of constant variance, or homoskedasticity. We verify that this holds by plotting our fitted values against the residuals with the hope that they are randomly and uniformly dispersed about a residual value of 0, without being skewed upward or downward.




```
Call:
lm(formula = abs(residuals(fit)) ~ fitted(fit))

Residuals:
   Min     1Q Median     3Q    Max
-8.649 -4.431 -1.253  2.777 38.433

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.51749    0.09140   82.25   <2e-16 ***
fitted(fit)  0.04515    0.00379   11.91   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.986 on 4728 degrees of freedom
Multiple R-squared:  0.02914,   Adjusted R-squared:  0.02893
F-statistic: 141.9 on 1 and 4728 DF,  p-value: < 2.2e-16
```

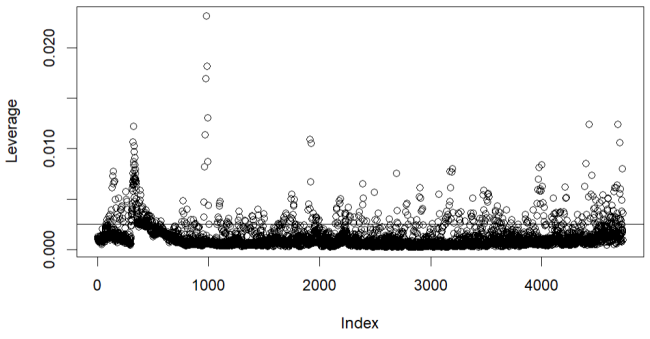
The plot on the left clearly shows a violation of this assumption, as the left side of the plot is much narrower than the middle of the plot. Additionally, the regression on the right corroborates this claim and shows that the fitted values are a statistically significant predictor for the residuals, which is problematic for the model, as coefficient interpretations and hypothesis testing may be unreliable.

Another important assumption is normality. The assumption is that our model's residuals follow a normal distribution. Without this assumption, there are consequences such as biases in the parameters and the model being a poor fit for a different distribution.



The Q-Q Plot is evidence that the residuals do NOT follow a normal distribution. A deviation of the residuals from the line can be seen at the bottom left and top right of the graph, and this deviation is significant enough to indicate that the assumption of normality may be violated. To back this up, the Shapiro-Wilk normality test is performed, and a p-value of 2.2e-16 is achieved, meaning that the null hypothesis of a normal distribution is REJECTED, which checks out with the plot.

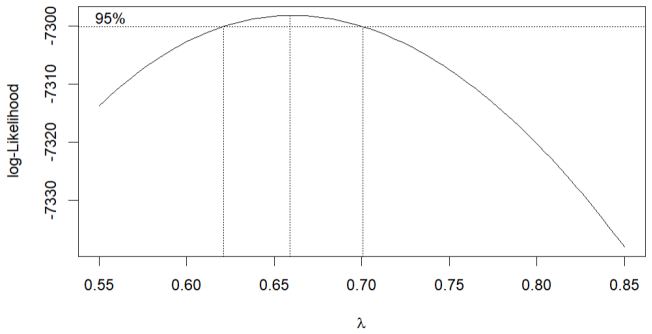
It is also useful to check if these results are caused by any outlying or influential points. We first check to see if there are any points with large leverage that could potentially be influential.

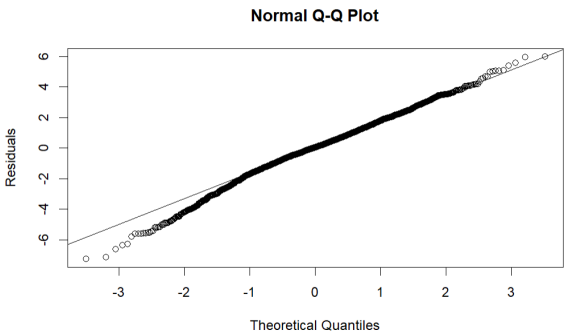


We see that there are many points above the threshold of what we consider to be "high leverage". However, this is extremely misleading: our threshold is a function of our sample size (2*p/n), which is very large, so there are bound to be many points above the threshold (475, to be exact). In fact, the maximum Cook distance of any point is 0.0206925 in

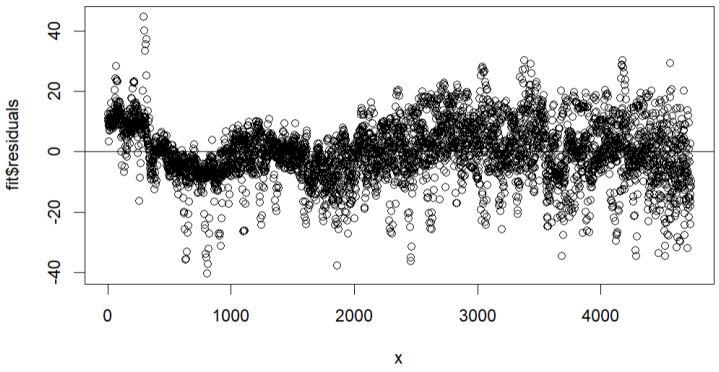
observation 4705, implying that none of these points are considered ”influential points”. Ultimately, we end up with one outlier through analysis of the jackknife residuals with a threshold of qt(0.05/(2*n), df=n-p-1, lower.tail = FALSE) = 4.409918.

Because of the lack of normality in our residuals, we consider the possibility of a Box-Cox transformation for a better fit and a solution to normality.





The graph on the left indicates that our 95% confidence interval for the lambda that maximizes the likelihood function is in the range of approximately 0.625 and 0.70. Thus, a natural conclusion is to perform the Box-Cox transformation where lambda=2/3. However, as we can see on the right, the normality has not changed as we can see a large deviation at the tail ends of the line. This is supported by the fact that the Shapiro-Wilk returns a p-value of 9.617e-10, rejecting the hypothesis that the residuals in this new model are normally distributed. There is some improvement, however, in the assumption of homoskedasticity: at the 0.05 level, the fitted values are no longer a statistically significant predictor for the residuals.



Another important assessment is the relationship between the errors. More specifically, are these errors correlated with one another? This can be analyzed by plotting the residuals as a function of the index. While it is a large sample and it may be harder to interpret a plot such as the one below than if the sample was much smaller, there still seems to be some clear patterns with points moving together. Thus, we suspect autocorrelated errors may be at play. To correct this, we use a generalized least squares model. We specifically hope to return a 95% confidence interval of the correlation value that includes 0 in it. Unfortunately, we return a 95% confidence interval of the correlation (0.288, 0.378) which does not include 0. This means that even with our attempt at corrections through the generalized least squares model, our errors still exhibit some degree of correlation.

Additionally, this GLS model changes the significance of grocery and food spending (denoted grf_spending below), as it now has a p-value of 0.3859, far exceeding our desired significance level of 0.05.

	Value <chr>	Std.Error <chr>	t-value <chr>	p-value <chr>
(Intercept)	-15.007222	0.3987355	-37.63704	0.0000
all_merchant	0.763837	0.0286977	26.61668	0.0000
aer_spending	0.197891	0.0071275	27.76464	0.0000
gen_spending	0.229015	0.0179113	12.78604	0.0000
grf_spending	0.014952	0.0172424	0.86714	0.3859
hcs_spending	0.042498	0.0083102	5.11399	0.0000

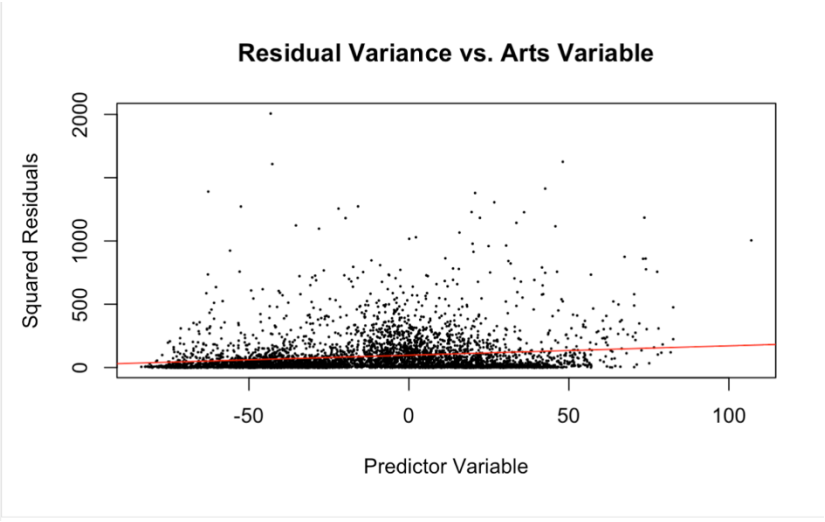
Does the variance of the errors depend on a predictor variable?

The studentized Breusch-Pagan test was conducted to assess whether the variance of the residuals depends on the specific predictor variable. In this test, a higher BP value is evidence of heteroskedasticity, though there is no specific cutoff. A p value smaller than 0.05 is evidence that we can reject the null hypothesis that the variables are homoscedastic.

1. Arts and entertainment saw a BP of 124.44 and a p value of 2.2e^-16
2. All spending saw a BP of 76.416 and a p value of 2.2e^-16
3. Grocery and food store spending saw a BP of 15.521 and a p value of 8.161e^-5
4. Health care spending saw a BP of 47.233 and a p value of 6.303e^-12
5. General merchandise spending saw a BP of 19.62 and a p value of 9.955e^-6

The results indicate that for all tested predictor variables, the null hypothesis of **homoskedasticity** (constant variance of errors) is rejected. This means that the variance of the errors changes systematically with respect to each predictor variable. Among the predictors, arts and entertainment spending as well as all spending exhibit the strongest influence on the residual and variance, as reflected by their high Breusch-Pagan statistics and extremely low p-values. This means that the model's assumptions about constant error variance are violated, and the standard errors of the coefficients might be biased.

We can also take the arts spending variable, and plot it against the residual variance to get an idea of what is happening: The scatterplot shows that the spread of squared residuals increases as the predictor variable aer_spending moves away from the zero (both in positive and negative directions suggesting **heteroskedasticity**, meaning the variance of the residuals is not constant across the range of the predictor.



To solve this, we create a weighted least squares model. The purpose is to assign weights to the observations that are inversely proportional to the variance of the residuals, ensuring that the observations with higher error variance have less influence on the model fit. We mad ethe weights the inverse of the squared residuals from the original model, and then apply the weights to a new model. In the output, we see that the weighted residuals range from -4.29 to 4.76 which indicates the spread of the errors after applying the weights. When testing how the predictor variables contribute to accommodation and food service spending,

1. All spending sees a coefficient of 0.829 with a p value of 2.2e^-16, indicating a strong positive effect
2. Arts and entertainment spending sees a coefficient of 0.2353 with a p value of 2.2r^-16, indicating a smaller but positive influence.
3. General; spending sees a coefficient of 0.1835 with a p value of 2.2e^-16, indicating a small positive effect
4. Grocery and food store spending sees a coefficient of -0.162 with a p value of 2.2e^-16, indicating a small, negative effect.
5. Health care spending sees a coefficient of 0.0436 and a p value of 3.12e^-7, indicating a super small positive effect.

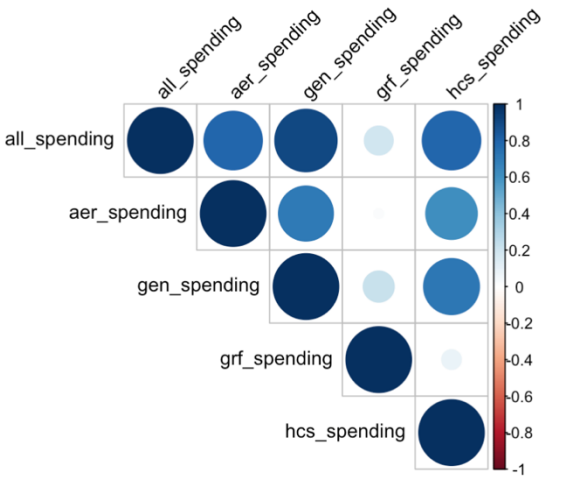
All predictors are statistically significant ($p<0.05$) implying that they all contribute to predicting accommodation and food service spending. The strongest predictors are all spending and entertainment and arts spending with positive correlations. Grocery and food store spending has a negative relationship with the response variable which intuitively makes sense because people spending more on food services like restaurants should spend less on groceries in many cases. Overall, the WLS model adjusts for heteroskedasticity, leady to more reliable coefficient estimates and standard errors.

Assessing potential multicollinearity:

Multicollinearity occurs when predictor variables in a regression model are highly correlated with each other. This can inflate the variance of the coefficient estimates, making it difficult to determine the individual contribution of each predictor to the model. The Variance Inflation Factor quantifies the extent of multicollinearity in a regression model. VIF is calculated for each predictor variable, and the interpretation is as follows. VIF = 1: no multicollinearity, 1 < VIF < 5: mild multicollinearity, VIF > 5: high multicollinearity, and if VIF > 10, there is severe multicollinearity.

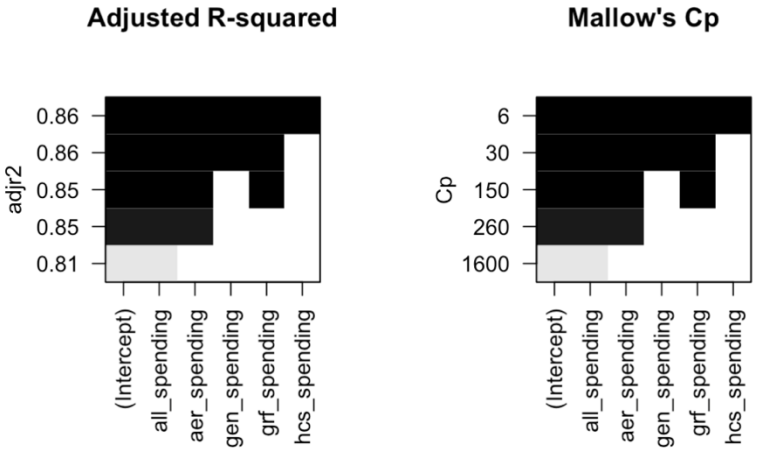
1. All spending saw VIF of 9.2, indicating high multicollinearity
2. Arts and entertainment spending saw a VIF of 2.85, indicating mild multicollinearity
3. General spending saw a VIF of 5.46, indicating high multicollinearity
4. Grocery store spending saw a VIF of 1.12, indicating very low to no multicollinearity
5. Health care spending saw a VIF of 2.76, indicating low multicollinearity.

We can look at a correlation plot to get a visual of this. We can see that there are high correlations between all spending, general spending, and arts and entertainment spending which is represented by the dark blue cells, and grocery spending has very weak relationships with the other variables and seems to be very independent.



Exhaustive Model Selection Search

The goal of exhaustive model selection is to identify the best subset of predictor variables that maximize the model's performance based on criteria such as Adjusted R^2 (measures the proportion of variance explained by the predictors, adjusted for the number of predictors) and Mallow's Cp (assess model fit by penalizing unnecessary predictors). Using the regsubsets() function in R, the best model based on adjusted R^2 was found to include all predictors with a high adjusted R^2 value of 0.86. Exploring mallow's CP shows that the best model if the one including all the predictors as the value was 6, which is close to the number of predictors used, in this case 5. We can visually see this with the plots on the right that when all 5 predictors are included in the model, the adjusted R^2 and Mallow's CP values are the most significant.



Backward and Forward Selection:

To see if our model was able to be improved upon through a different form of selection we applied backwards and forward selection. For backward selection what we did was take a full model and see which predictors should be removed based on their statistical significance. For forward selection we take a null model, otherwise understood as an empty model, and add predictors to it based on its statistical significance.
What we found was that both the backward selection as well as the forward selection produced the same model as what we had created before. As a result all of the model diagnostics were exactly the same as what we found previously. This told us that even through a different form of selection we still come to the same model because all of the features are extremely important to predicting the target feature, in this case that is accommodation and food service spending otherwise known as ACF spending.

```
> summary(forward_model)

Call:
lm(formula = acf_spending ~ all_spending + aer_spending + grf_spending +
    gen_spending + hcs_spending, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-40.322  -5.556  -0.117   5.818  44.799

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -11.313146   0.318083 -35.567  < 2e-16 ***
all_spending   0.824878   0.027814  29.657  < 2e-16 ***
aer_spending   0.235420   0.006884  34.196  < 2e-16 ***
grf_spending  -0.162403   0.014483 -11.213  < 2e-16 ***
gen_spending   0.183447   0.017096  10.730  < 2e-16 ***
hcs_spending   0.043635   0.008517   5.123 3.12e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.415 on 4724 degrees of freedom
Multiple R-squared:  0.8563,    Adjusted R-squared:  0.8561
> summary(backward_model)

Call:
lm(formula = acf_spending ~ all_spending + aer_spending + gen_spending +
    grf_spending + hcs_spending, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-40.322  -5.556  -0.117   5.818  44.799

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -11.313146   0.318083 -35.567  < 2e-16 ***
all_spending   0.824878   0.027814  29.657  < 2e-16 ***
aer_spending   0.235420   0.006884  34.196  < 2e-16 ***
gen_spending   0.183447   0.017096  10.730  < 2e-16 ***
grf_spending  -0.162403   0.014483 -11.213  < 2e-16 ***
hcs_spending   0.043635   0.008517   5.123 3.12e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.415 on 4724 degrees of freedom
Multiple R-squared:  0.8563,    Adjusted R-squared:  0.8561
F-statistic:  5628 on 5 and 4724 DF,  p-value: < 2.2e-16
```

Conclusions:

We found that All merchant category codes spending (all_spending), Arts, entertainment, and recreation (AER) spending, General merchandise stores (GEN) and apparel and accessories (AAP) spending, and Health care and social assistance (HCS) spending all had positive relationships with the target variable ACF spending. Grocery and food store (GRF) spending had a negative relationship with the target variable. Our full model achieved a r squared value of 0.8561. This tells us that 85.61% of the variation within ACF spending can be explained by our model. Every predictor variable was statistically significant at the 0.05 level. But even though the model seems to be extremely strong there are still some drawbacks. We struggled with showing evidence for our assumption of

normality. There was very apparent heteroscedasticity. And due to the fact that every single feature had to do with spending there was multicollinearity in some of the features. However even through these drawbacks our model does seem to answer our problem that we proposed at the beginning. The spending of money on non-essential items does coincide with eating out more. It also answers the second part of the question that spending on essential items implies eating out less.