Final Exam STAT 324

## 1. Introduction & Research Goal

In the dataset provided, the response variable is sfood which represents the family income in British pounds per week. sfood—the household share of total expenditure devoted to food—is a classic Engel-curve outcome: it naturally varies with purchasing power (income) and household composition (kids, age group), making it an intuitive, policy-relevant response for modeling how spending priorities shift across socioeconomic strata. This variable serves as the response variable in the analysis, indicating it’s the outcome that the analysis seeks to explain or predict. The dataset includes a range of explanatory variables, each representing a different share of weekly expenditures as a proportion of the total. These include sfuel (share of fuel expenditures), sclothes (share of clothing expenditures), salcohol (share of alcohol expenditures), stransport (share of transportation expenditures), and sother (share of other expenditures). These variables provide insight into how household spending is distributed across various categories.

Additionally, the dataset contains variables that capture the total expenditure (totexpend) of a household in British pounds per week, the age of the household head (age), and the number of children in the household (kids), which is categorized as either 1 or 2. These variables can be used to understand the demographic and economic factors that might influence the family income. The inclusion of ltotexpend and lincome, which are the natural logarithms of total expenditure and income respectively, suggests that the relationship between the logarithm of income and expenditures could be explored. Similarly, agesq, the square of the age of the household head, might be used to investigate non-linear relationships between age and income.

The variable high\_sfuel is a binary variable that categorizes the share of fuel expenditures (sfuel) into “high” if it’s greater than the median, and “low” otherwise. This variable provides a simple way to compare households with high and low fuel expenditures. The variable kids\_binary is another binary variable that categorizes the number of children (kids) into “one” if there is one child, and “two” otherwise. This variable simplifies the analysis by reducing the number of children to two categories.

The variable sclothes\_group categorizes the share of clothing expenditures (sclothes) into three groups: “low”, “medium”, and “high”. This variable allows for an analysis of the impact of different levels of clothing expenditures on the response variable. The variable age\_group categorizes the age of the household head (age) into three groups: “young”, “middle-aged”, and “old”. This variable allows for an analysis of the impact of different age groups on the response variable. stransport measures the fraction of a household’s total outlays devoted to transportation. A larger stransport value means transportation absorbs a bigger slice of the family budget. Finally, the variable salcohol\_group categorizes the share of alcohol expenditures (salcohol) into four groups: “low”, “medium”, “high”, and “very high”. This variable allows for an analysis of the impact of different levels of alcohol expenditures on the response variable. Overall, this dataset provides a comprehensive view of household finances, with a focus on how various factors relate to family income.

library(dplyr) ## Data Transformation --- T  
library(ggplot2) ## Data Visualization --- V  
library(ggfortify) ## Data Visualization --- V  
library(MASS) ## Data Analysis --- A  
library(leaps) ## Data Analysis --- A  
library(lmtest) ## Data Analysis --- A  
library(broom) ## Data Analysis --- A  
library(car) ## Data Analysis --- A

## 2. Data Description

Briefly describe the 1,519-household dataset, noting each key variable (shares, totals, logs, groups) and highlighting categorical groupings created for fuel, clothing, alcohol, kids, and age. Emphasize that these groupings enable flexible modeling of nonlinear or subgroup-specific patterns.

## 3. Data Preparation

Outline the cleaning and transformation steps: creation of binary and multi-level factors, grand-mean and group-mean centering, and polynomial terms. Note that these steps reduce multicollinearity and prepare predictors for interaction analyses.

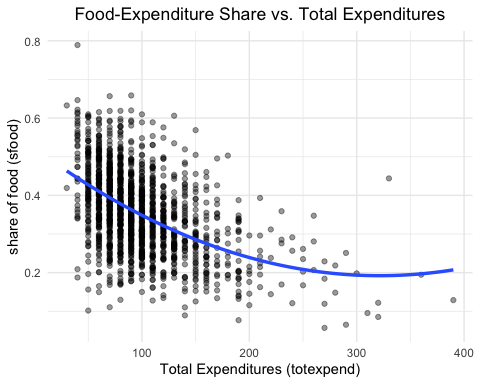
mod\_1\_expend\_df <- expend\_df %>%  
 mutate(high\_sfuel = if\_else(sfuel > median(sfuel), "high", "low")) %>%   
 ## mutate binary variable for kids  
 mutate(kids\_binary = if\_else(kids == 1, "one", "two")) %>%  
 ## mutate sclothes into three groups  
 mutate(sclothes\_group = case\_when(  
 sclothes < 0.08 ~ "low",  
 sclothes >= 0.08 & sclothes < 0.2 ~ "medium",  
 sclothes >= 0.2 ~ "high"  
 )) %>%  
 ## mutate age into three groups  
 mutate(age\_group = case\_when(  
 age < 30 ~ "young",  
 age >= 30 & age < 50 ~ "middle-aged",  
 age >= 50 ~ "old"  
 )) %>%   
 ## mutate share alcoholo intp four groups  
 mutate(salcohol\_group = case\_when(  
 salcohol < 0.1 ~ "low",  
 salcohol >= 0.1 & salcohol < 0.2 ~ "medium",  
 salcohol >= 0.2 & salcohol < 0.3 ~ "high",  
 salcohol >= 0.3 ~ "very high"  
 ))

mod\_2\_expend\_df <- mod\_1\_expend\_df %>%   
 group\_by(sclothes\_group) %>%   
 mutate(totexpend\_gc = totexpend - mean(totexpend, na.rm = TRUE)) %>%   
 ungroup() %>%   
 mutate(stransport\_c = stransport - mean(stransport, na.rm = TRUE)) %>%   
 mutate(totexpend\_c = totexpend - mean(totexpend, na.rm = TRUE))

## 4. Exploratory Visualizations

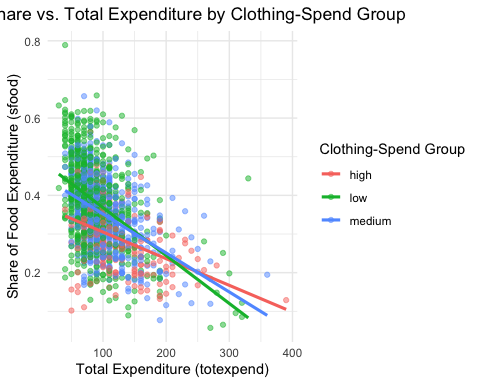
List the three plots: (i) quadratic sfood–totexpend curve, (ii) stratified lines by sclothes\_group, and (iii) group-mean-centered lines. Explain that these visuals reveal curvature, group-specific slopes, and within-group effects that motivate the formal models.

ggplot(mod\_1\_expend\_df, aes(x = totexpend, y = sfood)) +  
 geom\_point(alpha = .4) +  
 geom\_smooth(method = "lm", formula = y ~ poly(x, 2), se = FALSE, linewidth = 1.2) +  
 labs(  
 title = "Food-Expenditure Share vs. Total Expenditures",  
 x = "Total Expenditures (totexpend)", y = "share of food (sfood)"  
 ) +  
 theme\_minimal() +  
 theme(plot.title = element\_text(hjust = 0.5))



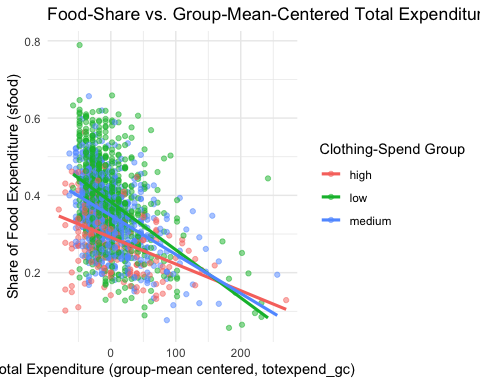
ggplot(mod\_1\_expend\_df, aes(x = totexpend, y = sfood, colour = sclothes\_group)) +  
 geom\_point(alpha = .5) +  
 geom\_smooth(method = "lm", se = FALSE, linewidth = 1.1) +  
 labs(  
 title = "Food-Share vs. Total Expenditure by Clothing-Spend Group",  
 x = "Total Expenditure (totexpend)",  
 y = "Share of Food Expenditure (sfood)",  
 colour = "Clothing-Spend Group"  
 ) +  
 theme\_minimal() +  
 theme(plot.title = element\_text(hjust = 0.5))

`geom\_smooth()` using formula = 'y ~ x'



ggplot(mod\_2\_expend\_df, aes(totexpend\_gc, sfood, colour = sclothes\_group)) +  
 geom\_point(alpha = .5) +  
 geom\_smooth(method = "lm", se = FALSE, linewidth = 1.1) +  
 labs(  
 title = "Food-Share vs. Group-Mean-Centered Total Expenditure",  
 x = "Total Expenditure (group-mean centered, totexpend\_gc)",  
 y = "Share of Food Expenditure (sfood)",  
 colour = "Clothing-Spend Group"  
 )+  
 theme\_minimal()

`geom\_smooth()` using formula = 'y ~ x'



## 5. Investigation into Hypothesized Models

| Model # | Formula (new data) | What it captures |
| --- | --- | --- |
| **1** | sfood ~ totexpend + stransport | Two quantitative predictors (totexpend, stransport). |
| **2** | sfood ~ totexpend + sclothes\_group | Quantitative predictor (totexpend) plus a categorical grouping (sclothes\_group). |
| **3** | sfood ~ totexpend \* stransport | Interaction between two quantitative variables (does the food-share response to total spending vary with transportation share?). |
| **4** | sfood ~ totexpend \* sclothes\_group | Interaction between a quantitative predictor and a categorical group (does the slope for total spending differ across clothing-spend groups?). |
| **5** | sfood ~ totexpend\_c \* stransport\_c | Mean-centered versions of both quantitative predictors plus their interaction (easier interpretation, lower multicollinearity). |
| **6** | sfood ~ totexpend + I(totexpend^2) | Polynomial term in totexpend to allow curvature in the Engel-type relationship. |

options(scipen = 9)  
# Model 1: two quantitative predictors  
model\_1 <- lm(sfood ~ totexpend + stransport, data = mod\_2\_expend\_df)  
summary(model\_1)

Call:  
lm(formula = sfood ~ totexpend + stransport, data = mod\_2\_expend\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.34207 -0.05958 -0.00171 0.05944 0.34724   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.49738518 0.00601237 82.73 <2e-16 \*\*\*  
totexpend -0.00106824 0.00005289 -20.20 <2e-16 \*\*\*  
stransport -0.26818263 0.02169086 -12.36 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.08802 on 1516 degrees of freedom  
Multiple R-squared: 0.2998, Adjusted R-squared: 0.2989   
F-statistic: 324.6 on 2 and 1516 DF, p-value: < 2.2e-16

# Model 2: quantitative + categorical  
model\_2 <- lm(sfood ~ totexpend + sclothes\_group, data = mod\_2\_expend\_df)  
summary(model\_2)

Call:  
lm(formula = sfood ~ totexpend + sclothes\_group, data = mod\_2\_expend\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.26288 -0.05959 -0.00372 0.05966 0.35551   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.41718123 0.00902285 46.236 < 2e-16 \*\*\*  
totexpend -0.00104803 0.00005571 -18.812 < 2e-16 \*\*\*  
sclothes\_grouplow 0.05823221 0.00714871 8.146 7.80e-16 \*\*\*  
sclothes\_groupmedium 0.03969529 0.00719663 5.516 4.08e-08 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.09041 on 1515 degrees of freedom  
Multiple R-squared: 0.2619, Adjusted R-squared: 0.2604   
F-statistic: 179.1 on 3 and 1515 DF, p-value: < 2.2e-16

# Model 3: interaction between two quantitative variables  
model\_3 <- lm(sfood ~ totexpend \* stransport, data = mod\_2\_expend\_df)  
summary(model\_3)

Call:  
lm(formula = sfood ~ totexpend \* stransport, data = mod\_2\_expend\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.35672 -0.05904 -0.00170 0.05858 0.34605   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.52390017 0.00815340 64.255 < 2e-16 \*\*\*  
totexpend -0.00130562 0.00007232 -18.054 < 2e-16 \*\*\*  
stransport -0.44871219 0.04351491 -10.312 < 2e-16 \*\*\*  
totexpend:stransport 0.00151469 0.00031725 4.774 0.00000198 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.0874 on 1515 degrees of freedom  
Multiple R-squared: 0.3102, Adjusted R-squared: 0.3088   
F-statistic: 227.1 on 3 and 1515 DF, p-value: < 2.2e-16

# Model 4: interaction between quantitative and categorical predictor  
model\_4 <- lm(sfood ~ totexpend \* sclothes\_group, data = mod\_2\_expend\_df)  
summary(model\_4)

Call:  
lm(formula = sfood ~ totexpend \* sclothes\_group, data = mod\_2\_expend\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.26343 -0.05967 -0.00414 0.05925 0.36094   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)  
(Intercept) 0.3742137 0.0160458 23.322 < 2e-16  
totexpend -0.0006902 0.0001238 -5.575 0.0000000293390  
sclothes\_grouplow 0.1181707 0.0179590 6.580 0.0000000000646  
sclothes\_groupmedium 0.0786933 0.0191051 4.119 0.0000401149063  
totexpend:sclothes\_grouplow -0.0005504 0.0001492 -3.688 0.000234  
totexpend:sclothes\_groupmedium -0.0003196 0.0001549 -2.063 0.039284  
   
(Intercept) \*\*\*  
totexpend \*\*\*  
sclothes\_grouplow \*\*\*  
sclothes\_groupmedium \*\*\*  
totexpend:sclothes\_grouplow \*\*\*  
totexpend:sclothes\_groupmedium \*   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.09006 on 1513 degrees of freedom  
Multiple R-squared: 0.2686, Adjusted R-squared: 0.2661   
F-statistic: 111.1 on 5 and 1513 DF, p-value: < 2.2e-16

# Model 5: centered variables with interaction  
model\_5 <- lm(sfood ~ totexpend\_c\*stransport\_c, data = mod\_2\_expend\_df)  
summary(model\_5)

Call:  
lm(formula = sfood ~ totexpend\_c \* stransport\_c, data = mod\_2\_expend\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.35672 -0.05904 -0.00170 0.05858 0.34605   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.35543825 0.00225264 157.788 < 2e-16 \*\*\*  
totexpend\_c -0.00110515 0.00005308 -20.820 < 2e-16 \*\*\*  
stransport\_c -0.29921743 0.02249617 -13.301 < 2e-16 \*\*\*  
totexpend\_c:stransport\_c 0.00151469 0.00031725 4.774 0.00000198 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.0874 on 1515 degrees of freedom  
Multiple R-squared: 0.3102, Adjusted R-squared: 0.3088   
F-statistic: 227.1 on 3 and 1515 DF, p-value: < 2.2e-16

# Model 6: polynomial (quadratic) term for lincome  
model\_6 <- lm(sfood ~ totexpend + I(totexpend^2), data = mod\_2\_expend\_df)  
summary(model\_6)

Call:  
lm(formula = sfood ~ totexpend + I(totexpend^2), data = mod\_2\_expend\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.32547 -0.06132 -0.00242 0.05961 0.34396   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.5220832648 0.0113530290 45.986 < 2e-16 \*\*\*  
totexpend -0.0020540326 0.0001791857 -11.463 < 2e-16 \*\*\*  
I(totexpend^2) 0.0000031969 0.0000006141 5.206 0.000000219 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.09154 on 1516 degrees of freedom  
Multiple R-squared: 0.2427, Adjusted R-squared: 0.2417   
F-statistic: 243 on 2 and 1516 DF, p-value: < 2.2e-16

## 6. Model Selection Procedures

### a. Stepwise Regression

Explain the stepwise AIC search and best-subsets (adjusted-R^2) approach run on an expanded predictor set. Note that these automated tools cross-check the hand-built models and suggest a parsimonious final candidate.

# Start with a full model  
full\_model <- lm(sfood ~ totexpend + stransport + sclothes + income + age, data = mod\_2\_expend\_df)  
  
# Stepwise model using both directions (backward & forward)  
step\_model <- stepAIC(full\_model, direction = "both", trace = TRUE)

Start: AIC=-7575.72  
sfood ~ totexpend + stransport + sclothes + income + age  
  
 Df Sum of Sq RSS AIC  
<none> 10.284 -7575.7  
- income 1 0.11202 10.396 -7561.3  
- age 1 0.23523 10.519 -7543.4  
- totexpend 1 1.17247 11.457 -7413.7  
- sclothes 1 1.18893 11.473 -7411.5  
- stransport 1 1.79314 12.077 -7333.6

# Summary of final selected model  
summary(step\_model)

Call:  
lm(formula = sfood ~ totexpend + stransport + sclothes + income +   
 age, data = mod\_2\_expend\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.29929 -0.05481 0.00088 0.05324 0.32287   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.47600995 0.01089572 43.688 < 2e-16 \*\*\*  
totexpend -0.00078142 0.00005950 -13.134 < 2e-16 \*\*\*  
stransport -0.34159139 0.02103120 -16.242 < 2e-16 \*\*\*  
sclothes -0.32106782 0.02427632 -13.226 < 2e-16 \*\*\*  
income -0.00016029 0.00003948 -4.060 0.00005167650 \*\*\*  
age 0.00165043 0.00028055 5.883 0.00000000496 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.08244 on 1513 degrees of freedom  
Multiple R-squared: 0.387, Adjusted R-squared: 0.3849   
F-statistic: 191 on 5 and 1513 DF, p-value: < 2.2e-16

#### i. Multicollinearity Check

car::vif(step\_model)

totexpend stransport sclothes income age   
 1.474901 1.095745 1.183036 1.298180 1.062008

#### ii. Outliers, Leverage, and Influence

Note that standardized residuals, Cook’s distance, and leverage values are extracted and the most influential 15 cases tabulated; any problematic observations are flagged for further scrutiny.

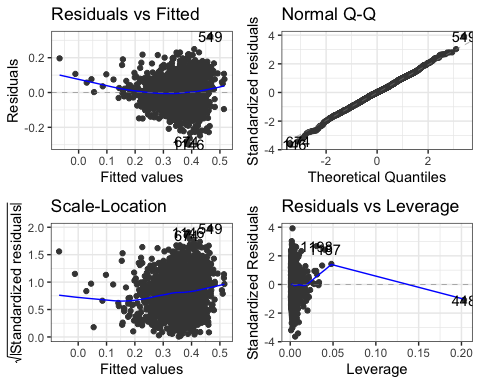
# Add standardized residuals  
augmented\_df <- augment(step\_model) ## from broom package  
  
rownames(augmented\_df) <- NULL  
  
augmented\_df %>%   
 dplyr::select( -totexpend, -stransport, -sclothes, -income, -age,-sfood) %>%  
 arrange(desc(.cooksd)) %>%   
 head(15) %>%   
 round(3)

# A tibble: 15 × 6  
 .fitted .resid .hat .sigma .cooksd .std.resid  
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
 1 0.266 -0.08 0.203 0.082 0.05 -1.08  
 2 -0.067 0.196 0.04 0.082 0.041 2.43  
 3 0.226 0.218 0.031 0.082 0.038 2.68  
 4 0.382 -0.28 0.009 0.082 0.018 -3.42  
 5 0.262 0.115 0.048 0.082 0.017 1.43  
 6 0.389 -0.299 0.006 0.082 0.013 -3.64  
 7 0.157 0.191 0.013 0.082 0.012 2.33  
 8 0.367 -0.216 0.01 0.082 0.012 -2.64  
 9 0.214 -0.148 0.02 0.082 0.011 -1.82  
10 -0.012 0.107 0.037 0.082 0.011 1.32  
11 0.28 0.215 0.007 0.082 0.008 2.62  
12 0.406 0.171 0.01 0.082 0.007 2.08  
13 0.345 -0.192 0.008 0.082 0.007 -2.34  
14 0.412 -0.227 0.005 0.082 0.007 -2.76  
15 0.466 0.323 0.003 0.082 0.007 3.92

#### iii. Diagnostics & Assumption Tests

List the residual tests performed (RESET for functional form, Durbin-Watson for autocorrelation, Shapiro for normality, Breusch-Pagan for heteroskedasticity). State that plots and p-values guide whether assumptions hold or require remediation.

autoplot(step\_model) +   
 theme\_bw()



resettest(step\_model) # From lmtest package

RESET test  
  
data: step\_model  
RESET = 11.923, df1 = 2, df2 = 1511, p-value = 0.000007284

dwtest(step\_model) # From lmtest package

Durbin-Watson test  
  
data: step\_model  
DW = 2.019, p-value = 0.6446  
alternative hypothesis: true autocorrelation is greater than 0

shapiro.test(resid(step\_model)) # Base R

Shapiro-Wilk normality test  
  
data: resid(step\_model)  
W = 0.99867, p-value = 0.302

bptest(step\_model) # From lmtest package

studentized Breusch-Pagan test  
  
data: step\_model  
BP = 57.253, df = 5, p-value = 0.00000000004485

### b. Model Selection Best Subset

# Fit all subsets model  
subset\_model <- regsubsets(  
 sfood ~ totexpend + stransport + sclothes + income + age, data = mod\_2\_expend\_df,  
 nvmax = 5  
)  
  
plot(x = subset\_model, ## Specify Model Fit  
 scale = "adjr2" ## Specify Selection Metric  
)

