Final Exam STAT 324

library(dplyr) ## Data Transformation --- T

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

library(ggplot2) ## Data Visualization --- V

In the dataset provided, the primary variable of interest is income, which represents the family income in British pounds per week. 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 sfood (share of food expenditures), 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. 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.

## Analysis 1: Exploration Models

| Model # | Formula (new data) | What it captures |
| --- | --- | --- |
| **1** | sfood ~ lincome + kids | Two quantitative predictors. |
| **2** | sfood ~ lincome + age\_group | Quantitative + categorical. |
| **3** | sfood ~ lincome + kids + lincome:kids | Interaction between two quantitative vars. |
| **4** | sfood ~ lincome + age\_group + lincome:age\_group | Quant × categorical interaction. |
| **5** | sfood ~ lincome\_c + kids\_c + lincome\_c:kids\_c | Mean-centered vars + interaction (easier interpretation). |
| **6** | sfood ~ lincome + I(lincome^2) | Polynomial in lincome to allow curvature. |
| **7** | sfood ~ lincome\_gc + age\_group + lincome\_gc:age\_group | **Group-mean-centered** lincome within each age group + interaction. |
| **8** | sfood ~ lincome + kids\_group + lincome:kids\_group | Bin kids into “none / one / two +” and interact with lincome (analogous to the old *air\_time\_group* model). |