A Regression Analysis of Nutrient in Branded Cookies

………..

2025-10-21

## Description

This dataset contains detailed nutritional and product information for branded cookie and biscuit products sold in the United States. Sourced from the USDA’s Branded Foods Database, the data represent real-world commercial products, making it particularly relevant for consumer nutrition analysis, food marketing research, or health-oriented product development studies. All nutrient values are standardized per 100 grams, enabling fair and consistent comparisons between products regardless of packaging or serving size.

## Sources

1. Nutrition Resource. 2025. *USDA Branded Food Products Database*. Accessed October 6, 2025. <http://www.nutritionresource.com/>

## Variables

1. name: categorical -> The commercial name of the cookie or biscuit product.
2. brand\_owner: categorical -> The parent company or manufacturer that owns the brand.
3. brand: categorical -> The specific brand under which the product is sold.
4. calories\_serving: quantitative -> Total calories provided per 100 grams of the product.
5. proteins: quantitative -> Grams of protein per 100g.
6. carbohydrates: quantitative -> Grams of carbohydrates per 100g.
7. fats: quantitative -> Grams of total fat per 100g.
8. fibers: quantitative -> Grams of dietary fiber per 100g.
9. total\_sugars: quantitative -> Grams of sugar per 100g, including both added and naturally occurring sugars.
10. gain\_type: categorical -> Primary grain ingredient classification(e.g., wheat, oats, mixed, other)

* “Wheat” – if the ingredient list includes wheat flour, enriched flour, or whole wheat.
* “Oats” – if the main grain is oats (e.g., whole grain oats, oat flour), and no wheat is present.
* “Mixed” – if both wheat and oats are present.

## Exploratory Data Analysis

Before building statistical models, it’s essential to understand the structure and patterns within your data. **Exploratory Data Analysis (EDA)** is the process of investigating datasets to summarize their main characteristics, identify potential issues, and uncover insights that can guide further analysis.

For the cookies dataset, which includes standardized nutritional and product information from the USDA’s Branded Foods Database, EDA allows us to:

1. Examine the raw data to get an overview of its structure and content
2. Compute summary statistics to understand distributions and relationships between variables
3. Create visualizations to reveal trends, outliers, or patterns not immediately visible from tables alone

Let’s begin by taking a first look at the raw data using glimpse().

[1] "Number of rows: 37"

[1] "Number of columns: 10"

Rows: 37  
Columns: 10  
$ name <chr> "MINI MARIA COOKIES", "ALMOND BUTTER GRANOLA CUPS, AL…  
$ brand\_owner <chr> "Goya Foods, Inc.", "General Mills, Inc.", "Bloom Pac…  
$ brand <chr> "GOYA", "NATURE VALLEY", "BLOOM'S", "BLOOM'S KOSHER P…  
$ calories\_serving <dbl> 452, 486, 467, 467, 447, 467, 417, 515, 481, 480, 387…  
$ proteins <dbl> 6.45, 8.57, 6.67, 6.67, 2.63, 6.67, 5.56, 6.06, 3.70,…  
$ carbohydrates <dbl> 80.6, 60.0, 70.0, 76.7, 68.4, 60.0, 69.4, 63.6, 74.1,…  
$ fats <dbl> 9.68, 25.70, 20.00, 15.00, 18.40, 20.00, 12.50, 27.30…  
$ fibers <dbl> 3.2, 5.7, 3.3, 6.7, 2.6, 6.7, 0.0, 3.0, 3.7, 4.0, 3.5…  
$ total\_sugars <dbl> 22.6, 22.9, 16.7, 23.3, 31.6, 26.7, 36.1, 24.2, 40.7,…  
$ grain\_type <chr> "WHEAT", "OATS", "WHEAT", "WHEAT", "MIXED", "OATS", "…

At this point, we know that the following are numerical variables and can be used for statistical summaries, visualizations, and regression modeling:

* calories\_serving
* proteins
* carbohydrates
* fats
* fibers
* total\_sugars

The categorical variables in the dataset are:

* name
* brand\_owner
* brand
* grain\_type

We also display a random sample of 10 rows of the 37 rows corresponding to different cookies and biscuits. Remember due to the random nature of the sampling, you will likely end up with a different subset of 10 rows.

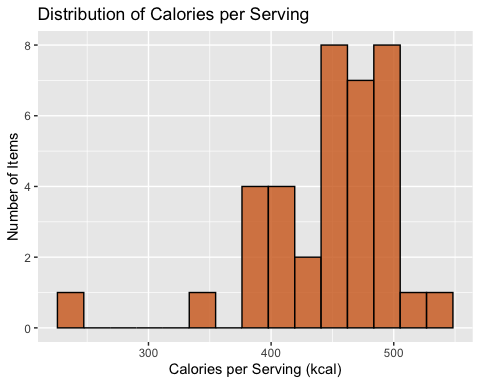
# A tibble: 10 × 10  
 name brand\_owner brand calories\_serving proteins carbohydrates fats fibers  
 <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
 1 CHOCO… H E Butt G… CENT… 481 3.7 74.1 18.5 3.7  
 2 BREAK… Harris-Tee… HARR… 460 8 74 16 6   
 3 SOFT … Partake Fo… PART… 484 3.23 67.7 22.6 3.2  
 4 VANIL… McCain Foo… OREO 433 6.67 56.7 20 0   
 5 12 SA… McKee Food… LITT… 447 2.63 68.4 18.4 2.6  
 6 ANGRY… Shrewd Foo… ANGR… 380 6 62 10 0   
 7 CARAM… Mars Choco… TWIX 498 3.32 63.1 23.3 0   
 8 ALMON… General Mi… NATU… 486 8.57 60 25.7 5.7  
 9 BONNI… Mr. Bevera… BONN… 391 10.9 59.4 17.2 7.8  
10 CHOCO… Partake Fo… PART… 452 3.23 58.1 22.6 6.5  
# ℹ 2 more variables: total\_sugars <dbl>, grain\_type <chr>

Let’s now perform the second step in an EDA: computing summary statistics displaying single numerical values that summarize a large number of values. Examples of summary statistics include the mean, the median, the standard deviation, and various percentiles. Let’s use the convenient tidy\_summary().

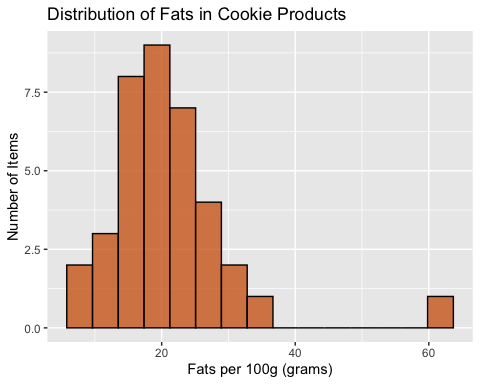
# A tibble: 8 × 11  
 column n group type min Q1 mean median Q3 max sd  
 <chr> <int> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 calories\_ser… 37 <NA> nume… 235 419 448. 460 484 536 55.0   
2 proteins 37 <NA> nume… 2.63 3.57 6.10 6.67 7.14 13.3 2.40  
3 carbohydrates 37 <NA> nume… 21.4 58.1 61.9 63.4 70 80.6 12.5   
4 total\_sugars 37 <NA> nume… 0 22.9 26.8 26 34.3 46.5 9.72  
5 fats 37 <NA> nume… 6.67 16 21.0 20 25 60.7 9.25  
6 grain\_type 11 MIXED char… NA NA NA NA NA NA NA   
7 grain\_type 10 OATS char… NA NA NA NA NA NA NA   
8 grain\_type 16 WHEAT char… NA NA NA NA NA NA NA

The average calories per serving in cookies is around 448 kcal, with most items falling between 419 and 484 kcal. Carbohydrates have a high mean of about 62g and range widely from 21.4g to 80.6g, showing significant variation. Total sugars also vary, with an average of 26.8g and some items containing no sugar at all, while others reach up to 46.5g. Protein content is generally lower, averaging 6.1g, and ranges from 2.6g to 13.3g, indicating that most items are not high in protein. Fats have a moderate average of 21g, but with a maximum of 60.7g, a few items are much higher in fat than the rest. These patterns suggest that carbohydrates, sugars, and fats are likely the strongest contributors to calorie content, and the presence of outliers in fats and sugars may affect regression models. Normalization or transformation of skewed features could improve predictive performance.

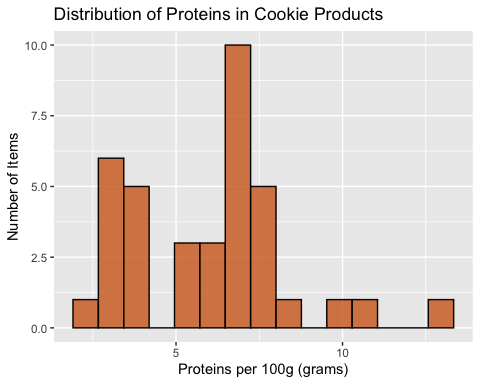
Let’s now perform the last of the three common steps in an exploratory data analysis: creating data visualizations. Let’s first create univariate visualizations. These are plots focusing on a single variable at a time.



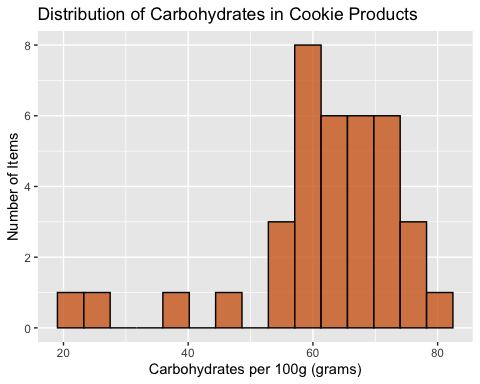
The histogram of calories per serving shows that most items have between 350 and 550 kilocalories. The distribution is right-skewed, as indicated by the long tail on the higher end.



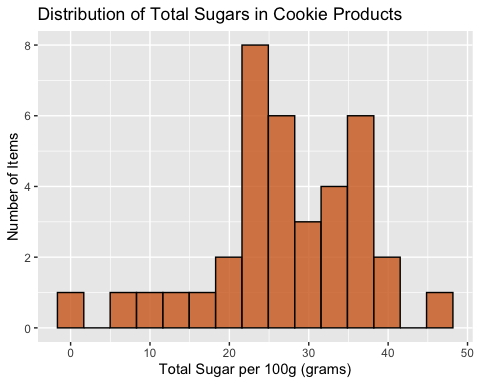
The histogram of fats per 100g shows that most items contain between 0 and around 40 grams of fat. The x-axis extends beyond 60 due to a few items with exceptionally high fat content. The distribution appears symmetric, with peaks around 20 grams.



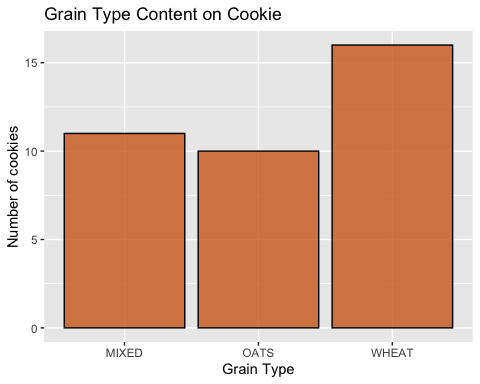
The histogram of proteins per 100g shows values spread across the x-axis, ranging from 0 to over 12 grams. The distribution appears bimodal, with noticeable peaks around 4 and 7 grams.



The histogram of carbohydrates per 100g shows values ranging from around 20 to over 80 grams. The distribution appears roughly symmetric, with a slight left skew and a noticeable peak around 61 grams.

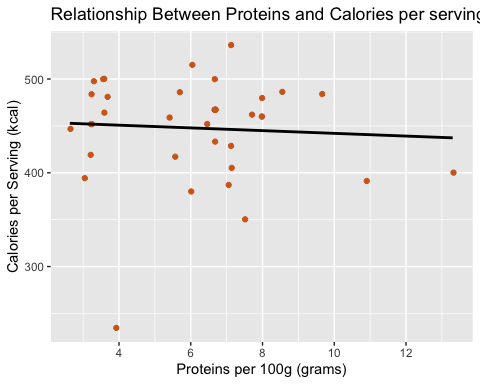


The histogram of total sugars per 100g shows values spread across the x-axis, ranging from 0 to close to 50 grams. The distribution appears bimodal, with noticeable peaks around 24 and 36 grams.

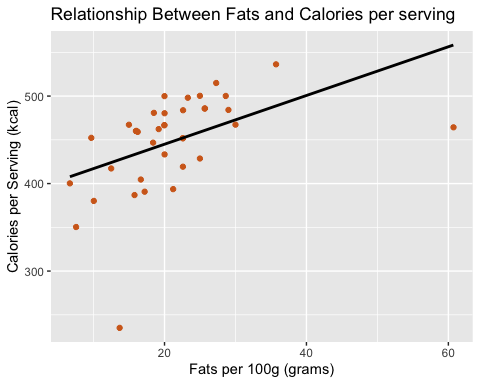


Finally, in the last of the univariate visualizations, the bar plot of Grain Type shows that the majority of cookies are made from WHEAT, with a count of over 15. MIXED and OATS grain types follow closely behind, with similar but slightly lower counts.

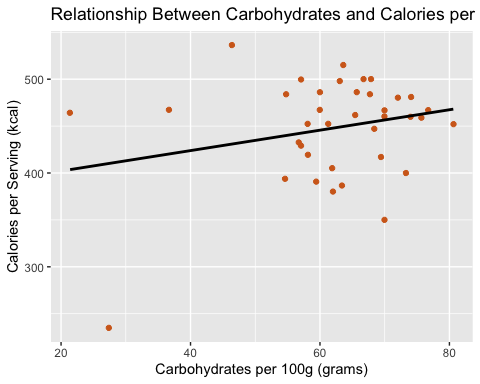
Let’s now continue our EDA by creating multivariate visualizations. Unlike the univariate histograms and barplot in the earlier, multivariate visualizations show relationships between more than one variable. This step is especially useful for identifying trends, correlations, or potential group differences in cookie characteristics.



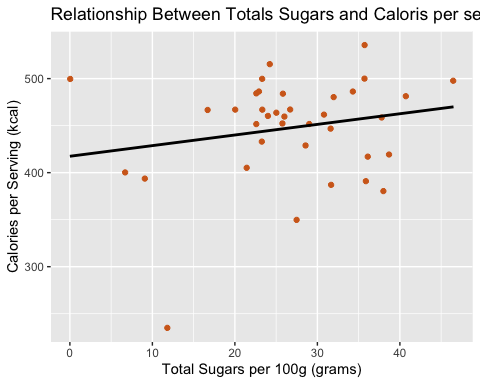
There appears to be a moderately negative linear relationship between protein content (per 100g) and calories per serving in the cookie dataset: as protein increases, calories per serving tend to decrease. Two clusters are noticeable, one around 3 grams of protein and another between 5 to 8 grams both with a wide spread of calories, mostly above 400 kcal. Additionally, there’s a visible outlier around 4 grams of protein and under 300 calories, which may be influencing the overall trend.



Observe there is a strong positive linear relationship between fat content (per 100g) and calories per serving in the cookie dataset: as fat increases, calories per serving also tend to increase. Most data points cluster between 8 and 30 grams of fat, with corresponding calorie values generally between 350 and 500 kcal. There are also noticeable outliers one around 14 grams and another above 60 grams of fat that may be affecting the overall trend.



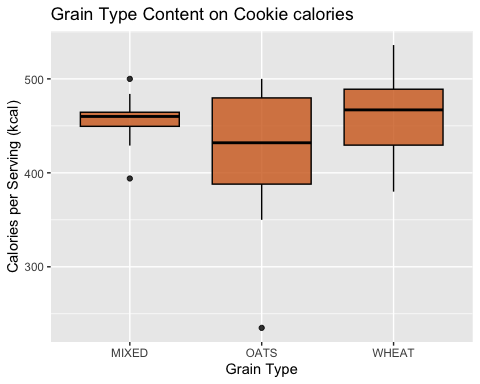
Also, there is a moderately positive linear relationship between carbohydrate content (per 100g) and calories per serving in the cookie dataset: as carbohydrates increase, calories per serving also tend to increase. Most data points are clustered toward the higher end of the x-axis, between 55 and 80 grams of carbohydrates, with calorie values generally ranging from 380 to 500 kcal. A few outliers are present, with the most notable one around 28 grams of carbohydrates and a lower calorie value.



There is a weak positive linear relationship between total sugar content (per 100g) and calories per serving in the cookie dataset: as total sugars increase, calories per serving also tend to increase. Most data points are clustered toward the higher end of the x-axis, between 20 and 40 grams of total sugars, with calorie values generally ranging from 350 to over 500 kcal. A few outliers are present, including a notable one with 0 grams of total sugars and a high calorie value of around 500 kcal.

Overall, the scatterplots reveal that most nutrient components including fats, carbohydrates, and total sugars tend to have a positive relationship with calories per serving. In other words, as the quantity of these ingredients increases, so do the calories. An exception to this pattern is protein, which shows a moderate negative relationship, suggesting that higher protein content does not necessarily result in higher calorie values and may even be linked to slightly lower ones.

Now, shifting our focus from continuous nutrients to a categorical ingredient, we’ll explore the relationship between grain type and calories per serving to understand how the type of grain used in cookies might influence their caloric content.



The plot shows that cookies made with WHEAT grain are more frequent and tend to have a higher calorie content, with a median around 480 kcal, compared to the OATS and MIXED groups, which have lower medians of approximately 430 kcal and 460 kcal, respectively. The interquartile range for OATS is also notable, spanning from about 385 to 480 kcal. Additionally, there are visible outliers in both the MIXED and OATS categories.

## Simple Linear Regression Models Using Calories per Serving as the Response Variable

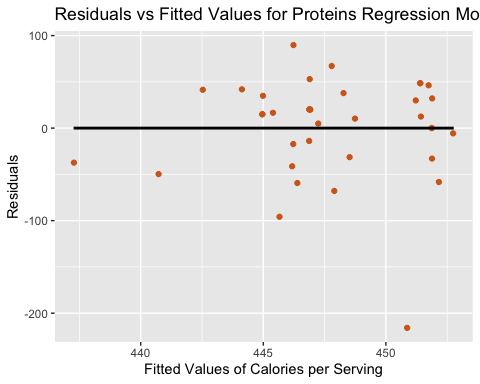
In this section, we explore the relationship between calories per serving and each of the key nutritional variables using simple linear regression models. Specifically, we use proteins, total sugars, fats, and carbohydrates all measured per 100g as explanatory variables, along with the categorical variable grain\_type. For each model, calories per serving is treated as the response variable, and we assess how well each predictor explains variation in calorie content.

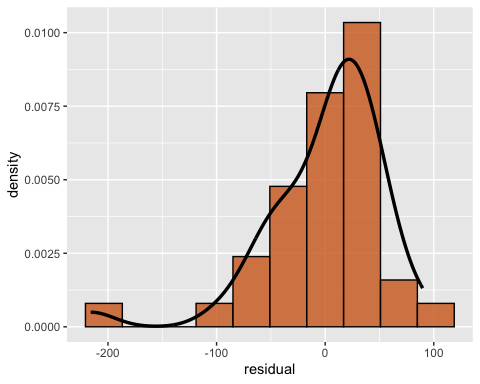
We begin by fitting a model with protein content as the predictor. Then, we separately examine models using total sugars, fats, and carbohydrates. Lastly, each model is followed by an evaluation of the residual plots, and a discussion of how well the assumptions hold.

# A tibble: 2 × 7  
 term estimate std\_error statistic p\_value lower\_ci upper\_ci  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 intercept 457. 25.2 18.1 0 405. 508.   
2 proteins -1.45 3.86 -0.377 0.709 -9.28 6.38

[1] 0.004036822

# A tibble: 1 × 1  
 cor  
 <dbl>  
1 -0.0635





When cookies have more protein, they tend to have slightly fewer calories per serving, but the effect is weak and not statistically significant. The regression model using protein content as the predictor yields an R-squared value of 0.004, meaning that protein explains only 0.4% of the variation in calories.

The slope is -1.453, indicating that for every additional gram of protein, the calories per serving decrease by about 1.45 kcal, on average. However, the p-value for the slope is 0.709, which is not statistically significant at any conventional level of 0.05. This means we do not have strong evidence to suggest a meaningful relationship between protein content and calories per serving.

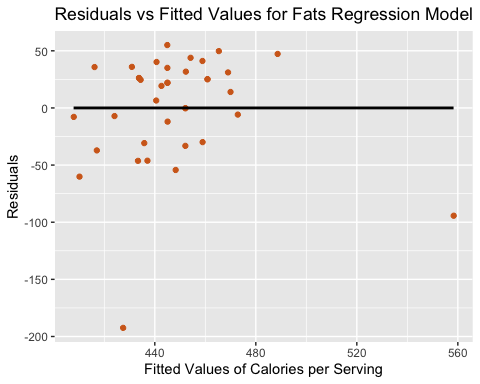
The residuals appear to be mostly almost all falling between -100 and 100, and generally supports all the LINE assumptions.

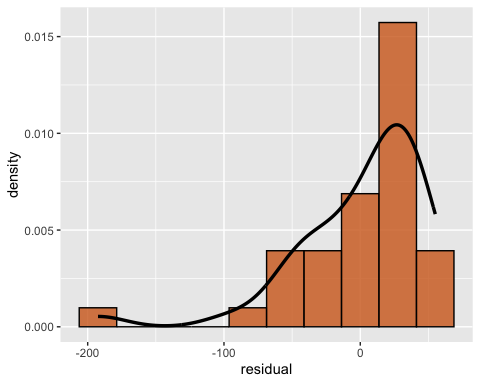
* Linearity: The residual plot shows a random scatter around zero with no clear pattern, suggesting that the linear model is appropriate for this data.
* Independence: While independence cannot be fully assessed from the residuals alone, there are no signs of autocorrelation or dependence. Since the data are not time-series or spatial, independence is reasonably assumed.
* Normality: The histogram of residuals appears approximately symmetric and bell-shaped, supporting the assumption that residuals are normally distributed.
* Equal Variance (Homoscedasticity): The spread of residuals across fitted values of 430 to 460 is not perfectly constant, but the variation is not extreme. This suggests mild heteroscedasticity, but not enough to invalidate the model.

# A tibble: 2 × 7  
 term estimate std\_error statistic p\_value lower\_ci upper\_ci  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 intercept 389. 20.3 19.2 0 348. 431.   
2 fats 2.78 0.888 3.14 0.003 0.982 4.59

[1] 0.2194338

# A tibble: 1 × 1  
 cor  
 <dbl>  
1 0.468





When cookies contain more fat, they tend to have higher calorie content per serving. The simple linear regression model using fat content (per 100g) as the explanatory variable produces an R-squared value of 0.22, meaning that fats explain approximately 22% of the variation in calories per serving.

The slope estimate is 2.785, indicating that for each additional gram of fat, the number of calories per serving increases by an average of 2.79 kcal. The p-value for the slope is 0.003, which is statistically significant at the 0.05 level. This provides strong evidence of a meaningful linear relationship between fat content and calorie count in this dataset.

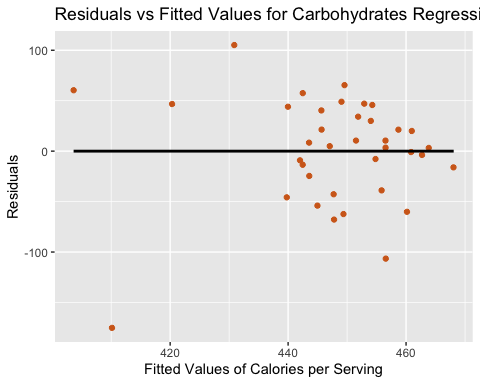
The residuals appear to be mostly almost all falling between -50 and 50, and generally supports all the LINE assumptions.

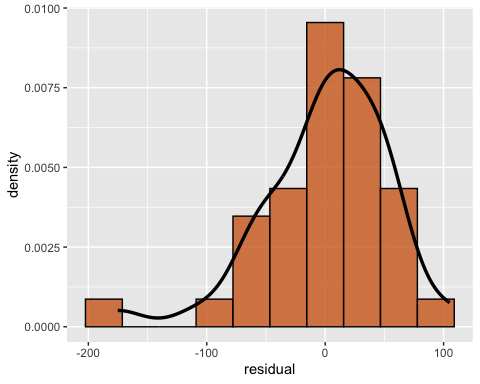
* Linearity: The residuals are randomly scattered around zero in the residual plot, with no obvious curves or systematic patterns. This supports the assumption of a linear relationship between fat and calories.
* Independence: While independence cannot be confirmed solely from visual inspection, there are no signs of autocorrelation or grouping. Given the nature of the dataset (not time-series or spatial), independence is reasonably assumed.
* Normality: The histogram of residuals is approximately bell-shaped around zero, suggesting that the residuals are roughly normally distributed.
* Although the spread of residuals is not perfectly uniform across the range of fitted values (mostly between 430 and 480 kcal), and a few outliers are present, there is no strong evidence of heteroscedasticity.

# A tibble: 2 × 7  
 term estimate std\_error statistic p\_value lower\_ci upper\_ci  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 intercept 380. 45.4 8.37 0 288. 473.   
2 carbohydrates 1.09 0.72 1.51 0.139 -0.373 2.55

[1] 0.06134769

# A tibble: 1 × 1  
 cor  
 <dbl>  
1 0.248





When cookies contain more carbohydrates, they tend to have higher calorie content per serving. The simple linear regression model using carbohydrates content (per 100g) as the explanatory variable produces an R-squared value of 0.061, meaning that fats only explain approximately 6.1% of the variation in calories per serving.

The slope estimate is 1.089, indicating that for each additional gram of carbohydrates, the number of calories per serving increases by an average of 1.089 kcal. The p-value for the slope is 0.139, which is not statistically significant at the 0.05 level. This provides somehow evidence of a linear relationship between carbohydrates content and calorie count in this dataset.

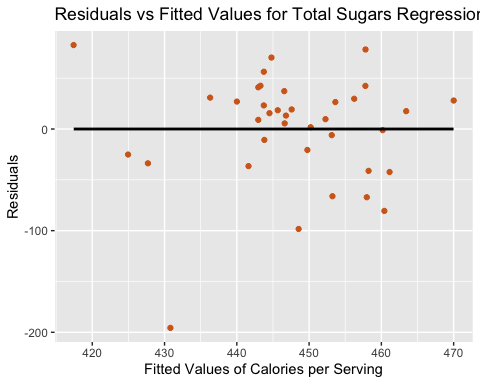
The residuals appear to be mostly almost all falling roughly between -80 and 80, and generally supports all the LINE assumptions.

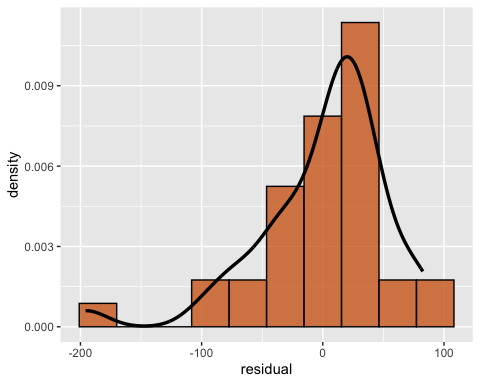
* Linearity: The residuals are randomly scattered around zero in the residual plot, with no obvious curves or systematic patterns. This supports the assumption of a linear relationship between fat and calories.
* Independence: While independence cannot be confirmed solely from visual inspection, there are no signs of autocorrelation or grouping. Given the nature of the dataset (not time-series or spatial), independence is reasonably assumed.
* Normality: The histogram of residuals is approximately bell-shaped and centered around zero, suggesting that the residuals are roughly normally distributed.
* Equal Variance (Homoscedasticity): Although the spread of residuals is not perfectly uniform across the range of fitted values are on the far end mostly between 430 and 470 kcal, and a few outliers are present, there is no strong evidence of heteroscedasticity.

# A tibble: 2 × 7  
 term estimate std\_error statistic p\_value lower\_ci upper\_ci  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 intercept 417. 26.7 15.6 0 363. 472.   
2 total\_sugars 1.13 0.937 1.21 0.236 -0.772 3.03

[1] 0.0398975

# A tibble: 1 × 1  
 cor  
 <dbl>  
1 0.200





When cookies contain more total sugars, they tend to have slightly higher calorie content per serving. The simple linear regression model using total sugars content (per 100g) as the explanatory variable yields an R-squared value of 0.040, indicating that total sugars explain only about 4% of the variation in calories per serving.

The slope estimate is 1.13, meaning that for each additional gram of total sugars, the number of calories per serving increases by an average of 1.13 kcal. However, the p-value is 0.236, which is not statistically significant at the conventional 0.05 level. This means there is insufficient evidence to conclude a strong linear relationship between total sugar content and calorie count in this dataset.

The residuals mostly fall between -100 and 100 kcal, and the diagnostics generally support the LINE assumptions:

* Linearity: The residuals are randomly scattered around zero in the residual plot, showing no clear curvature or pattern. This suggests the linear model is appropriate.
* Independence: While independence cannot be definitively assessed from the plot alone, there are no visible signs of autocorrelation or clustering. Given that the data are not time-series or spatial, independence is reasonably assumed.
* Normality: The histogram of residuals is roughly bell-shaped and centered close to zero, indicating that the residuals are approximately normally distributed.
* Equal Variance (Homoscedasticity): Although the spread of residuals is not perfectly constant across fitted values, the variation appears reasonably between 435 and 465 kcal. A few outliers are present, but there is no strong evidence of heteroscedasticity.

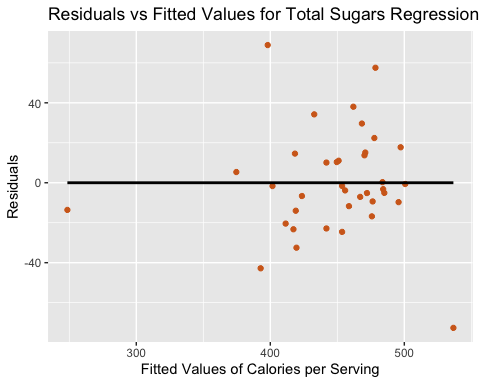
## Multiple Regression: Two numerical explanatory variables

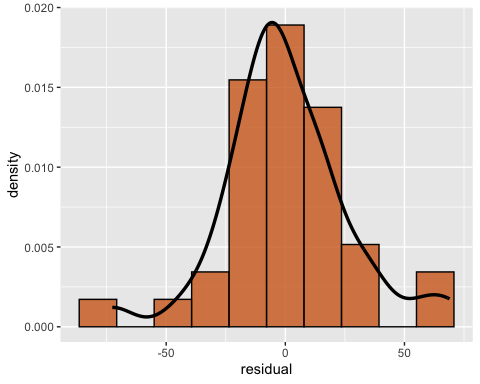
**Model 1**

After examining simple linear regression models using individual predictor, we now extend our analysis to a multiple linear regression models using two quantitative explanatory variables: fats and carbohydrates. This allow us to assess the combined effect of both ingredients on the calories per serving, building a more accurate model that better explains the variation in calorie content across cookies.

# A tibble: 3 × 7  
 term estimate std\_error statistic p\_value lower\_ci upper\_ci  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 intercept 37.6 40.3 0.933 0.357 -44.3 119.   
2 fats 6.68 0.649 10.3 0 5.37 8.00  
3 carbohydrates 4.36 0.48 9.09 0 3.38 5.33

[1] 0.7723622





The regression equation is:

* calories\_serving\_hat = 37.582 + 6.685 \* Fats + 4.358 \* Carbohydrates

In this model, the coefficient for Fats is 6.685, meaning that for every additional gram of fat per 100 grams of cookie, the calories per serving increase by approximately 6.69 kcal, holding carbohydrates constant. The coefficient for Carbohydrates is 4.358, indicating that for every additional gram of Carbohydrates, the calories per serving increase by approximately 4.36 kcal, holding Fats constant. Also, both fats and Carbohydrates have p-values less than 0.001, showing they are statistically significant predictors of calories. The model has an R-squared values of 0.772, indicating that 77.2% of the variability in calories per serving is explained by the combined linear effects of facts and carbohydrates. This suggest a strong overall fit.

The residuals mostly fall between -40 and 40 kcal, and the diagnostics generally support the LINE assumptions:

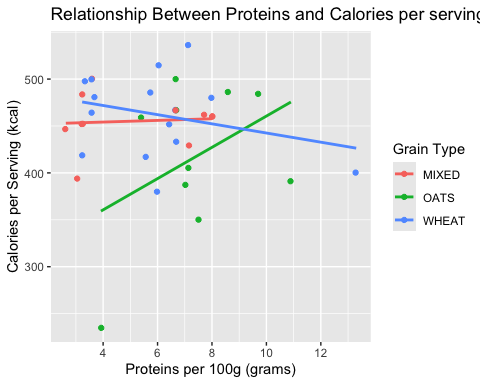
* Linearity: The residuals are randomly scattered around zero in the residual plot, showing no clear curvature or pattern. This suggests the linear model is appropriate.
* Independence: While independence cannot be definitively assessed from the plot alone, there are no visible signs of autocorrelation or clustering. Given that the data are not time-series or spatial, independence is reasonably assumed.
* Normality: The histogram of residuals is bell-shaped and centered at zero, indicating that the residuals are normally distributed.
* Equal Variance (Homoscedasticity): Although the spread of residuals is not perfectly constant across fitted values, the variation appears reasonably between 400 and 500 kcal. A few outliers are present, but there is no strong evidence of heteroscedasticity.

## Multiple Regression: one numerical and one categorical explanatory variables

In the next analysis, we explore how the calories per serving (100 grams) in cookies are influenced by two variables: protein content and grain type. To assess this relationship, we apply two multiple regression models: an interaction model and a parallel slopes model.

## Interaction Model: Model 2

This model allows both the intercepts and the slopes to vary across the levels of grain\_type. The effect of protein on calories per serving is not assumed to be the same for all grain types.



# A tibble: 6 × 7  
 term estimate std\_error statistic p\_value lower\_ci upper\_ci  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 intercept 451. 39.5 11.4 0 370. 531.   
2 proteins 0.882 7.07 0.125 0.902 -13.5 15.3   
3 grain\_type: OATS -157. 76.3 -2.06 0.048 -312. -1.18  
4 grain\_type: WHEAT 40.8 52.5 0.776 0.444 -66.4 148.   
5 proteins:grain\_typeOATS 15.8 11.1 1.42 0.166 -6.92 38.5   
6 proteins:grain\_typeWHE… -5.77 8.88 -0.65 0.521 -23.9 12.3

[1] 0.2394287

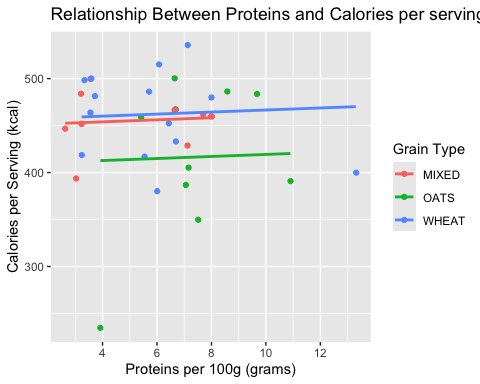
The regression equation is:

* calories\_serving\_hat = 450.660 + (-156.727 \* Doats) + 40.750 \* Dwheat + 0.882 \* X + 15.781 \* X \* Doats + (-5.771 \* X \* Dwheat)

In this model, for OATS grain type, each additional gram of protein is associated with an increase of 16.67 calories per serving (0.882 + 15.781), compared to the reference grain type. Also, for WHEAT grain type, each additional gram of protein is associated with a decrease of 4.89 calories per serving (0.882 − 5.771), compared to the reference grain type. Finally, for MIXED grain type, each additional gram of protein is associated with an increase of 0.882 calories per serving, but this effect is not statistically significant (p = 0.902).

## Parallel Model: Model 3

We can simplify the previous model by removing the interaction effects. The model still represents different grain type groups with different regression lines by allowing different intercepts but all the lines have the same slope: they are parallel as shown.



# A tibble: 4 × 7  
 term estimate std\_error statistic p\_value lower\_ci upper\_ci  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 intercept 450. 26.0 17.3 0 397. 503.   
2 proteins 1.09 3.98 0.273 0.786 -7.00 9.17  
3 grain\_type: OATS -41.2 25.0 -1.65 0.109 -92.1 9.73  
4 grain\_type: WHEAT 6.09 21.3 0.286 0.777 -37.2 49.3

[1] 0.1285255

The regression equation is:

* calories\_serving\_hat = 449.612 + (-41.194 \* Doats) + 6.086 \* Dwheat + 1.086 \* X

In this model, each additional gram of protein is associated with an increase of 1.09 calories per serving, regardless of grain type. Cookies made with OATS have, on average, 41.19 fewer calories per serving than cookies made with the reference grain type like MIXED, when protein content is held constant. Also, cookies made with WHEAT have, on average, 6.09 more calories per serving than the reference grain type.

**Model Evaluation**

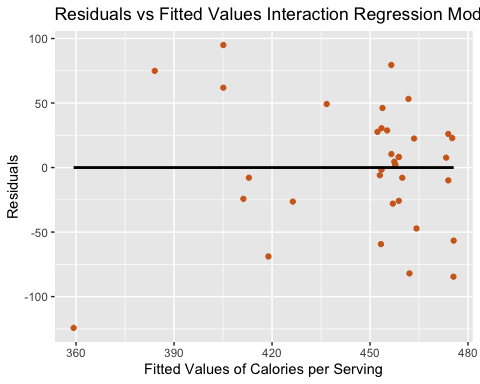
In the interaction model, only one coefficient grain\_type: OATS was statistically significant (p = 0.048), suggesting that cookies made with oats tend to have lower calorie content per 100 grams compared to the reference grain type. However, the interaction terms proteins:grain\_typeOATS (p = 0.166) and proteins:grain\_typeWHEAT (p = 0.521) were not statistically significant, indicating little evidence that the effect of protein on calorie content varies across grain types. The overall model fit was modest, with an R-squared value of 0.239, meaning the model explains about 23.9% of the variability in calories per 100 grams of cookie products.

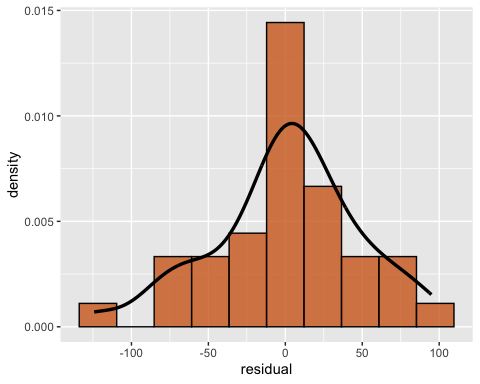
The parallel model, which excludes the interaction terms and assumes a consistent protein effect across all grain types, produced a lower R-squared value of 0.129. This means it explains only 12.9% of the variability in calorie content, suggesting a poorer fit. In this model, the protein variable was not statistically significant (p = 0.786), and none of the grain type categories were significant at the 0.05 level.

Based on these results, the interaction model is preferred. Although its individual interaction terms are not significant, it still provides a better overall fit to the data as reflected by the higher R-squared value (0.239 vs. 0.129) and captures more variation in calorie content among cookie products.

**Interaction Model: Fit Evaluation**

# A tibble: 1 × 1  
 sum\_of\_squared\_residuals  
 <dbl>  
1 82760.





The interaction model, which includes both protein content and grain type along with their interaction, explains approximately 23.9% of the variability in calories per serving (R-squared = 0.239) in cookies. This indicates a moderate but not strong fit. The sum of squared residuals is 82,760.27, representing the total amount of variation in calorie content per 100 grams of cookie that remains unexplained by the model.

Among the predictors, only the grain type “OATS” shows a statistically significant effect (p = 0.048), suggesting that cookies made primarily with oats tend to have lower calorie content compared to those made with the reference grain type. However, the interaction terms and the main effect of protein are not statistically significant, indicating that the relationship between protein and calorie content does not differ meaningfully across grain types. The relatively large residual variation and limited number of significant predictors suggest that additional variables may be necessary to better explain differences in calorie content among cookie products.

## Comparison between Model 1 and Model 2

Model Comparition

| Predictors | Rsquared | KeyFindings | Interpretation |
| --- | --- | --- | --- |
| fats, carbohydrates | 0.772 | Fats: Coefficient = 6.685, p < 0.001 has a significant effect, Carbohydrates: Coefficient = 4.358, p < 0.001 has a significant effect, Intercept: Not significant (but not critical) | This model explains 77.2% of the variance in calories\_serving, and both predictors are statistically significant. This suggests that fats and carbohydrates are strong predictors of cookie calories. |
| proteins, grain\_type, and interaction terms | 0.239 | Most terms are not significant, Only grain\_type:OATS with p = 0.048, Interaction terms (proteins × grain) are not significant | This model explains only 23.9% of the variance in calories\_serving. Most predictors are not statistically significant, suggesting that protein and grain type are much weaker predictors of calories compared to fats and carbohydrates. |