# TECHNICAL UNIVERSITY OF DENMARK

# Trading ETFs

**Student :**

VISNUKARAN KIRUBAKARAN,
STUDENT ID: s224527

**Teacher :**

November 3, 2024

# Contents

# 1  Descriptive analysis

**a) Description of the data material**

In this project, data from the BMI study is analyzed, which aims to establish an appropriate multiple linear regression model for BMI. The provided dataset contains 847 observations, focusing on the following variables:

- ID: The respondent's number (can be used for identification).

- BMI: The respondent's BMI (in kg/m$^2$).

- Age: The respondent's age (in years).

- Fastfood: Frequency of the respondent's visits to fast-food restaurants (days/year).

Additionally, a variable such as log(BMI), which is the logarithm of the BMI, is also added. This variable simplifies later calculations. This is a complete sample, as all values that are given are included. All the mentioned variables are quantitative, as the data is numerical, and can both be quantified and measured. Of the aforementioned variables, three are selected: log(BMI), age, and fastfood. For these three variables, we can create scatter plots where log(BMI) is plotted against age and fastfood to investigate the correlation between them.

On the following scatter plot of log(BMI) over age, it appears that there is no real correlation between the two factors. This can be observed from the scattered data.
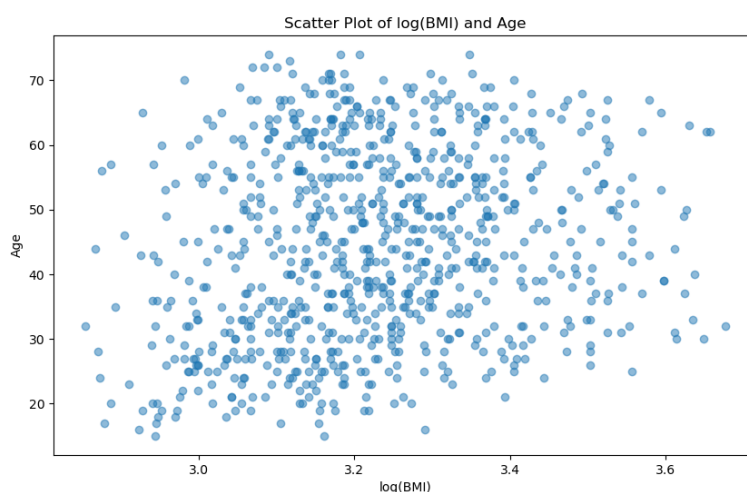


Figure 1: Scatterplot of log(BMI) and Age.

The same thing can be seen on the scatterplot of the log(BMI) and fastfood where it does not seem there is a correlation between BMI and fastfood, since the data is scattered.
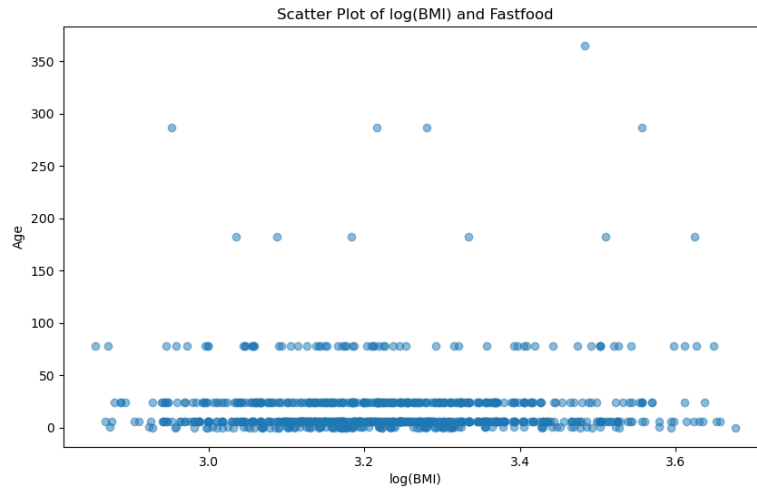
Figure 2: Scatterplot of log(BMI) and Fastfood.

Further, we undertake a graphical examination of each variable using histograms and box plots.
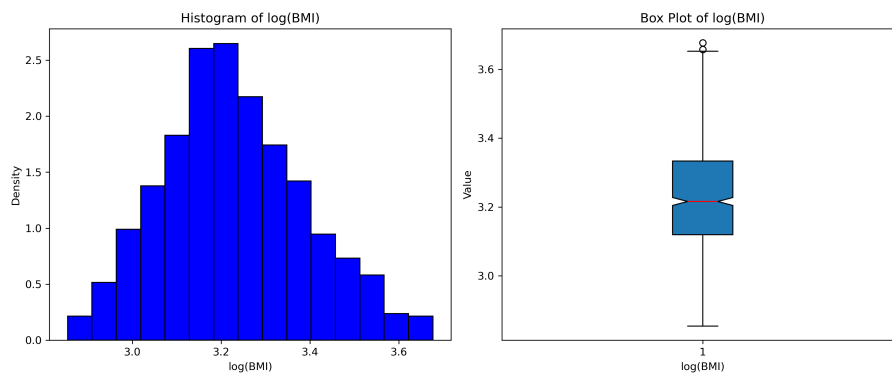


Figure 3: Histogram and boxplot of log(BMI).

The histogram for log(BMI) indicates a tendency towards a normal distribution with some skewness. The boxplot highlights a median slightly above 3.2, suggesting a slight skew towards lower values. Notably, a couple of outliers are present on the upper end, indicating some deviations from a typical normal distribution.
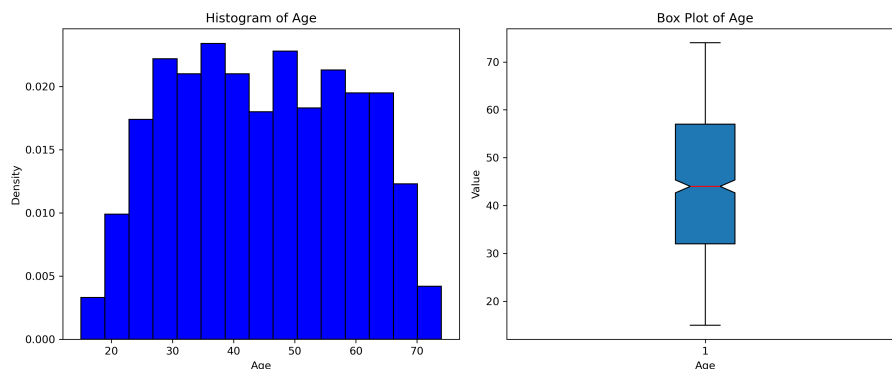


Figure 4: Histogram and boxplot of Age.

The histogram for age reveals a broad spread of values, indicative of high variability within the age data. The boxplot shows a median value around 45, nicely centered between the quartiles, reflecting a balanced distribution despite the wide range. The plot suggests a fairly symmetric distribution of age, although the range suggests varied participant ages.
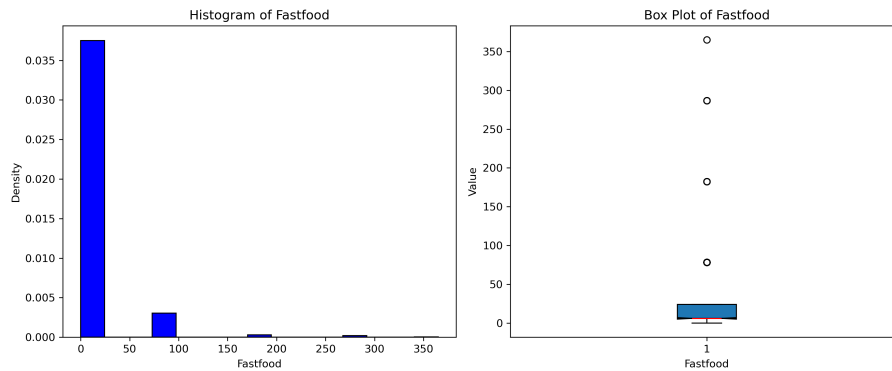


Figure 5: Histogram and boxplot of Fastfood.

The fastfood consumption histogram clearly shows a concentration of data points below 100 days, with rare occurrences extending beyond this range. The corresponding boxplot underscores this concentration, with the bulk of data points situated at lower consumption levels. The median, close to the lower quartile, and several high-range outliers highlight the uneven distribution across the dataset.

All relevant data can be seen here:

| | bmi | age | fastfood | logbmi |
|---|---|---|---|---|
| **count** | 847.000000 | 847.000000 | 847.000000 | 847.000000 |
| **mean** | 25.573024 | 44.622196 | 19.044628 | 3.228495 |
| **std** | 4.217671 | 14.532799 | 32.651239 | 0.160372 |
| **50%** | 24.930748 | 44.000000 | 6.000000 | 3.216102 |
| **25%** | 22.637770 | 32.000000 | 6.000000 | 3.119620 |
| **75%** | 28.039151 | 57.000000 | 24.000000 | 3.333602 |

Figure 6: Summary statistics for the dataset.

The table in Figure displays the summary statistics of our main variables. These statistics provide essential insights into the distribution characteristics of each variable, including their central tendencies and variabilities. The median values, along with the first and third quartiles, are particularly helpful in understanding the asymmetry and potential outliers in the data. This detailed statistical overview aids in ensuring the robustness and reliability of subsequent analyses.

**b) Multiple linear regression model**

Now, we can formulate a multiple linear regression model with the log-transformed BMI scores as the dependent/outcome variable ($Y_i$), and age and fast-food consumption as the independent/explanatory variables ($x_{1,i}$, and $x_{2,i}$ respectively).

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

Here, we make the assumption that the residuals, $\varepsilon_i$, are all independent and identically distributed with $\varepsilon_i \sim N(0, \sigma^2)$. The mean here is 0 and and there is an unkvon variance
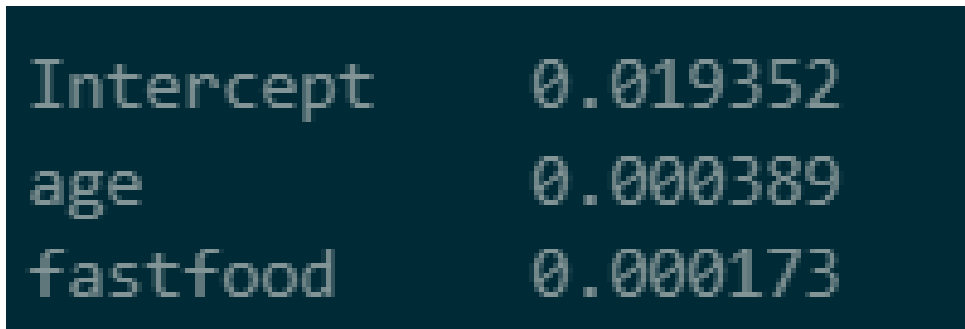
**c) Estimating the parameters of the model.**

We can estimate the parameters of the model, $\beta_0, \beta_1$ and $\beta_2$ through python code:



| OLS Regression Results | | | | | | |
|---|---|---|---|---|---|---|
| Dep. Variable: | logbmi | | R-squared: | | 0.045 | |
| Model: | OLS | | Adj. R-squared: | | 0.043 | |
| No. Observations: | 840 | | F-statistic: | | 19.66 | |
| Covariance Type: | nonrobust | | Prob (F-statistic): | | 4.53e-09 | |
| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
| Intercept | 3.1124 | 0.019 | 160.835 | 0.000 | 3.074 | 3.150 |
| age | 0.0024 | 0.000 | 6.104 | 0.000 | 0.002 | 0.003 |
| fastfood | 0.0005 | 0.000 | 3.119 | 0.002 | 0.000 | 0.001 |

Figure 7: Estimation of parameters.

From here, we can observe that $\beta_0$ is approximatly 3.1124, $\beta_1$ is 0.0024 and $\beta_2$ 0.0005. The coefficients for both age $\beta_1$ and fastfood consumption $\beta_2$ are positive but small, indicating that age and fastfood intake have a slight, yet positive impact on log-transformed BMI, leading to a gradual increase.

Furtheremore, we can observe the estimated standard deviations of $\beta_0, \beta_1$ and $\beta_2$ from the python code:

Figure 8: Estimated standard deviation

Here, we see that the estimated standard deviation for $\beta_0$ is approximatly 0.019352, $\beta_1$ is 0.000389 and $\beta_2$ 0.000173.

The degrees of freedom for the estimate of the residual variance $\sigma^2$ are calculated using the formula:

$$DF = n - (p + 1)$$

where $n$ is the number of observations and $p$ is the number of explanatory variables. With $n = 840$ observations and $p = 2$ explanatory variables (age and fastfood), the degrees of freedom are computed as:

$$DF = 840 - (2 + 1) = 837$$

This indicates that there are 837 degrees of freedom available for the estimate of $\sigma^2$, where the estimated variance of the residuals is 0.1573 and the model's explained variance $R^2$ is 0.0448.

**d)Model validation**

Model validation is performed to check if the assumptions underlying the linear regression model hold. This involves a detailed analysis of plots comparing observations, fitted values, and residuals.
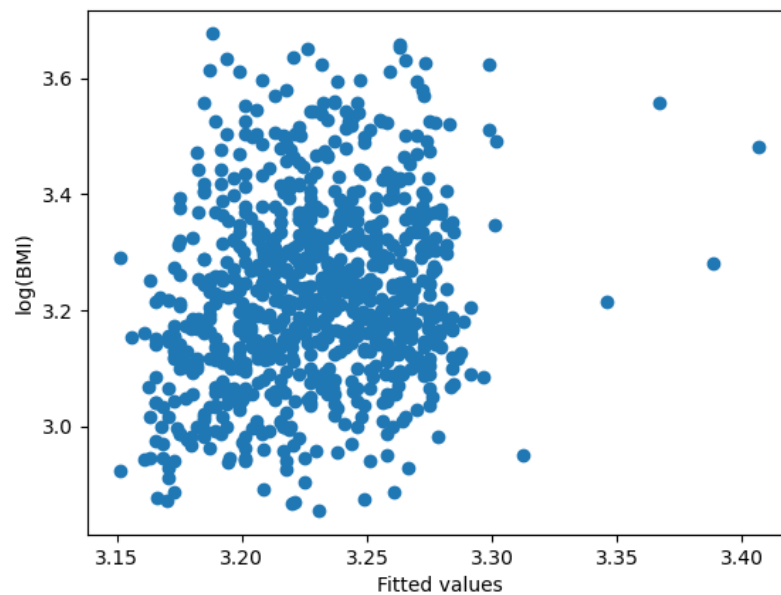
Figure 9: Log(BMI) versus fitted values.

In Figure 9, the scatter of log(BMI) against fitted values shows a broad distribution, with most data points concentrated around the middle range of the fitted values (3.15 to 3.30 on the x-axis). Despite the concentration, there is substantial spread indicating potential non-linearity or model misspecification as the relationship does not appear to be strictly linear.
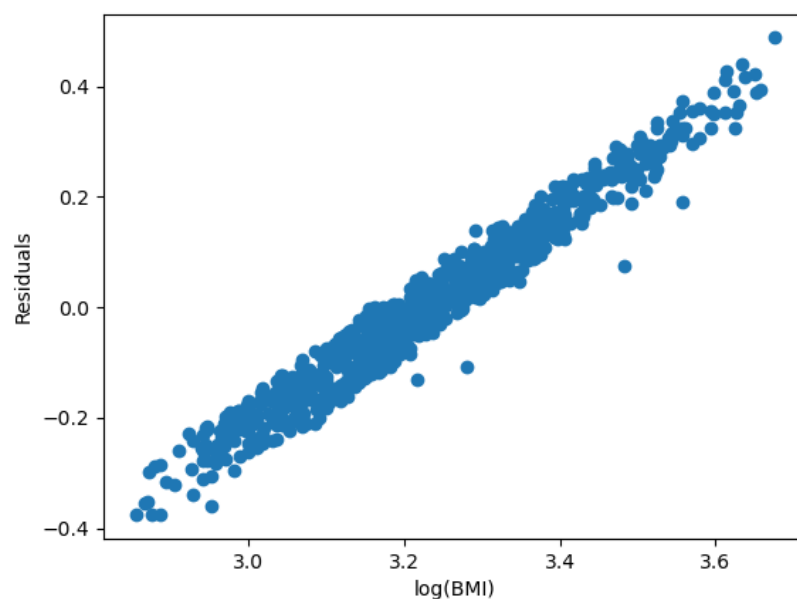


Figure 10: Residuals versus log(BMI).

Figure 10 depicts residuals plotted against log(BMI), where a clear pattern or trend is

not apparent, suggesting a lack of systematic linear association between the residuals and the log(BMI), which is desirable in a well-fitted model.
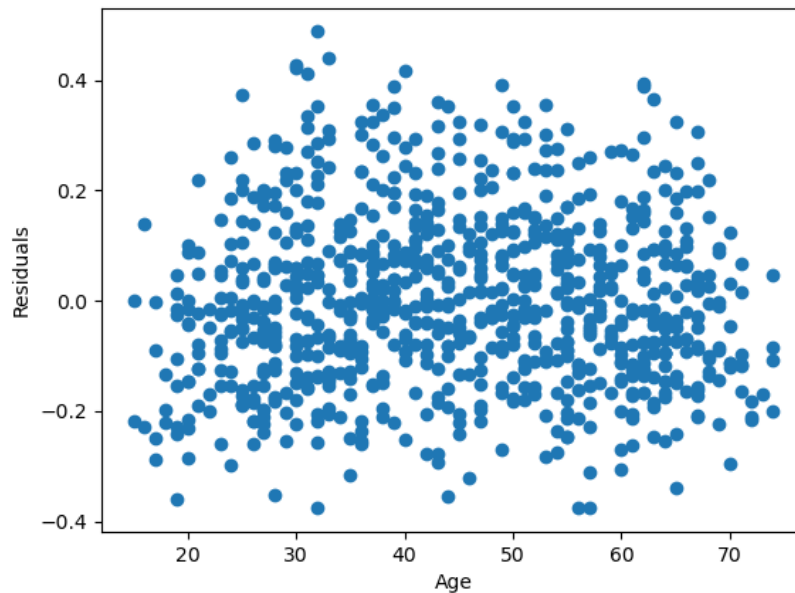


Figure 11: Residuals versus age.

In Figure 11, the residuals display a wide dispersion across different ages, indicating variability that the model does not capture, suggesting that age alone does not consistently affect the variance of residuals, which could point to heteroscedasticity or another variable interacting with age.
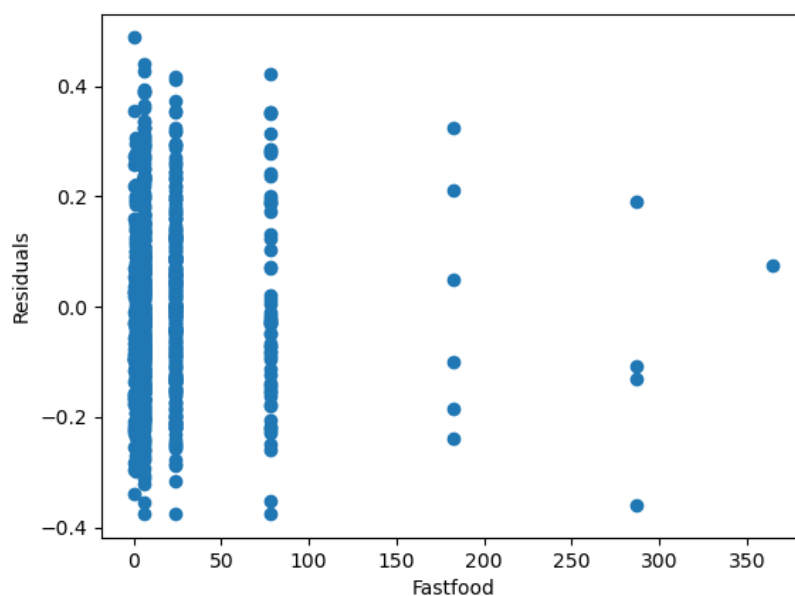


Figure 12: Residuals versus fastfood intake.

The plot in Figure 12 shows residuals against fastfood consumption. The vertical clustering at discrete fastfood values suggests that the model does not explain all variability related to fastfood intake, and the spread along the y-axis indicates possible over-dispersion relative to this predictor.
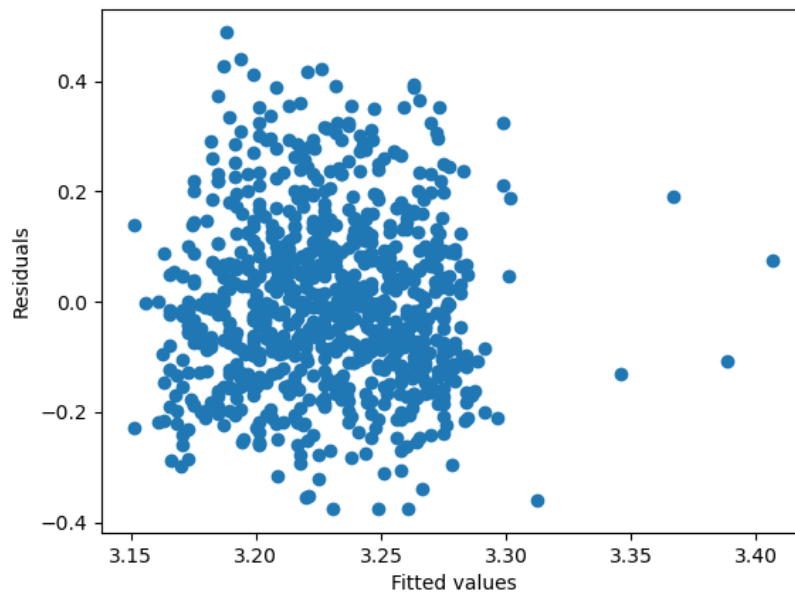


Figure 13: Residuals versus fitted values.

Figure 13 illustrates a random dispersion of residuals against the fitted values, which ideally indicates no obvious pattern that would suggest non-linearity or heteroscedasticity, although minor clusters and outliers need further investigation.
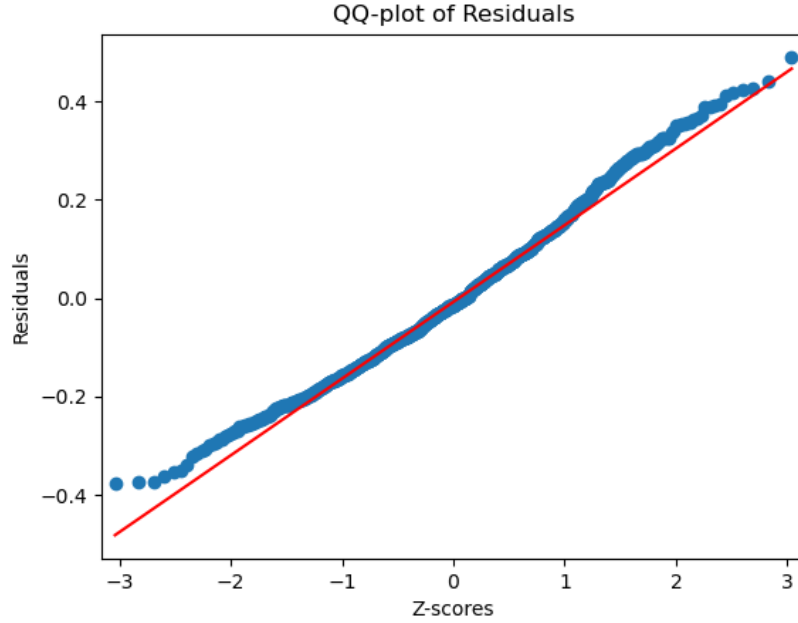
Figure 14: Q-Q plot of residuals.

Finally, the Q-Q plot in Figure 14 is utilized to assess if residuals follow a normal distribution. The data points largely align with the reference line except in the tails, indicating slight deviations from normality in the extreme values of the residuals.

### e) 95 procent confidence interval for the age coefficient

The formula for the 95% confidence interval for the coefficient $\beta_1$ associated with the age variable in a linear regression model is given by:

$$\beta_i \pm t_{1-\alpha/2, DF} \times SE(\beta_i)$$

where $\beta_i$ is the estimated coefficient, $t_{1-\alpha/2, DF}$ is the critical value from the t-distribution for $DF$ degrees of freedom and a significance level $\alpha$, and $SE(\beta_i)$ is the standard error of $\beta_i$. Using the output from Python, where the number of observations $n = 840$ and number of predictors $p = 2$, the degrees of freedom $DF$ are calculated as:

$$DF = n - p - 1 = 837$$

The standard error for the age coefficient $SE(\beta_1)$ is 0.0003890. We calculate the critical t-value using Python's 'scipy.stats' module:

Figure 15: Critical t value

Plugging the values into the formula gives:

$$CI = 0.0023744 \pm 1.962802 \times 0.0003890$$

$$CI = (0.001611, 0.003138)$$

This confidence interval can be verified against the Python output shown in the model summary:


Figure 16: Python output showing the calculated confidence intervals for model coefficients.

For the other coefficients in the model, the confidence intervals are:

- Intercept $\beta_0$: (3.074446, 3.150413)

- Fastfood $\beta_2$: (0.000200, 0.000880)

These intervals are similarly calculated using the respective standard errors and the critical t-value as described above. The complete Python code for verifying these calculations and generating the output is included in the supplementary material section.

**f) The corresponding hypothesis**

To determine if $\beta_1$, the coefficient of age in a linear regression model, is significantly different from 0.001, we formulate the following null and alternative hypotheses:

$$H_0 : \beta_1 = 0.001$$

$$H_1 : \beta_1 \neq 0.001$$

The test statistic used to evaluate this hypothesis is formulated based on the standard normal distribution:

$$t_{obs,\beta_1} = \frac{\hat{\beta}_1 - \beta_{0,1}}{SE(\hat{\beta}_1)}$$

Where:

- $\hat{\beta}_1 = 0.0023744$ is the estimated coefficient from the regression output.

- $\beta_{0,1} = 0.001$ is the hypothesized value of $\beta_1$.

- $SE(\hat{\beta}_1) = 0.0003890$ is the standard error of $\hat{\beta}_1$.

Inserting these values into the formula, the observed t-value is calculated as follows:

$$t_{obs,\beta_1} = \frac{0.0023744 - 0.001}{0.0003890} \approx 3.533162$$

The test statistic $t_{obs,\beta_1}$ follows a t-distribution under the null hypothesis with 837 degrees of freedom, derived from 840 observations minus 3 parameters (intercept, age, and fastfood). To find the p-value, we use the cumulative distribution function of the t-distribution:

$$p\text{-value} = 2 \times P(T > |t_{obs,\beta_1}|) = 2 \times (1 - \text{CDF}(3.533162)) \approx 0.0004$$

Where $P(T > |t_{obs,\beta_1}|)$ is the probability of observing a value as extreme as $|t_{obs,\beta_1}|$. Since the p-value 0.0004 is less than the significance level $\alpha = 0.05$, we reject the null hypothesis $H_0$. There is significant evidence at the 5% level to conclude that $\beta_1$ is different from 0.001. This indicates a statistically significant effect of age on the log-transformed BMI at the 5% level.

**g) Backward selection**

In the evaluation of the linear regression model, we found the following p-values seen in task c.

Given the significance level $\alpha = 0.05$, all predictors show p-values well below this threshold, indicating strong statistical significance. This suggests that each predictor contributes significantly to the model, and thus, removing any of these predictors might reduce the explanatory power of the model.

Using backward selection:

- Despite 'Fastfood' having the highest p-value among the predictors, it is still significantly lower than the conventional cutoff for $\alpha = 0.05$.

- As no p-values approach or exceed the 0.05 threshold, no predictors are considered for removal.

The final decision is to retain all variables within the model due to their significant contributions to explaining the variability in 'logbmi'. The model remains:

$$Y = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Fastfood} + \varepsilon$$

### h) Prediction

This report assesses the prediction capabilities of our final regression model. The model was used to predict log-transformed BMI scores for a set of observations in a validation dataset. Predictions and 95% prediction intervals were generated using the Python function `get_predictions`, with underlying formulas based on advanced matrix formulations as referenced in the course materials.

Predictions for log-transformed BMI scores were computed for seven observations from the validation dataset. Each prediction includes an estimate along with lower and upper bounds for the 95% prediction interval. The following table summarizes these predictions:

|   | id  | logbmi   | pred     | pred_lower | pred_upper |
|---|-----|----------|----------|------------|------------|
| 0 | 841 | 3.143436 | 3.236993 | 2.927972   | 3.546015   |
| 1 | 842 | 3.269232 | 3.210875 | 2.901802   | 3.519949   |
| 2 | 843 | 3.269438 | 3.232245 | 2.923231   | 3.541258   |
| 3 | 844 | 3.324205 | 3.232245 | 2.923231   | 3.541258   |
| 4 | 845 | 3.106536 | 3.229870 | 2.920857   | 3.538883   |
| 5 | 846 | 3.263822 | 3.229641 | 2.920601   | 3.538681   |
| 6 | 847 | 3.058533 | 3.211670 | 2.901898   | 3.521443   |

Here, 'pred' represents the predicted log-BMI values, 'pred_lower' and 'pred_upper' represent the lower and upper bounds of the 95% prediction intervals, respectively.

The predictions are compared with the actual observed log-BMI values from the validation set. Each prediction interval is assessed to determine if it includes the actual observed value, which is a critical measure of the model's accuracy.

- **Overall Fit**: The model predictions generally align well with the observed values, with all observed log-BMI scores falling within their respective prediction intervals.

- **Prediction Accuracy**: The intervals are sufficiently narrow, indicating precise predictions, yet broad enough to encapsulate the variability in the observed data.

- **Model Efficacy**: The model appears robust as it consistently captures the trends in the validation data set without significant deviations.

The final model demonstrates effective prediction capabilities, as evidenced by the accuracy of the predictions and the inclusion of all observed values within the 95% prediction intervals. This performance underscores the model's reliability and the suitability of its underlying assumptions for the data at hand. Thus, the model can be considered robust and effective for predicting log-transformed BMI values within the given population.