**Introduction**

A study collected data on plasma beta-carotene (PBC) levels in 196 patients who underwent an elective surgery to biopsy or remove non-cancerous tissue. Using these data, I analyzed the effects of biometric and lifestyle variables on PBC levels, measured in nanograms per milliliter. The predictor variables are as follows:

Age: the age of the patient in years

BMI: the patient’s weight in kilograms divided by the square of their height in meters

Daily Calories: the patient’s daily caloric intake

Fat: The patient’s daily fat intake in grams

Weekly Alcohol: the number of alcoholic drinks consumed by the patient per week

CHL: the patient’s daily cholesterol intake in milligrams

DBC: the patient’s daily dietary beta-carotene intake in micrograms

Gender: The gender of the patient

Vitamin: the frequency with which the patient uses vitamins (never, occasional, regular)

Smoking: the smoking status of the patient (never, former, current)

**Data visualization**

There was one observation where PBC was zero. Given that it is extremely unlikely that a patient had no beta carotene in their plasma, I assumed this was erroneous and removed it. After this, PBC ranged from 14 to 1415 and appeared to have a right-skewed distribution (figure 1A). Thus, I anticipated non-normality of the residuals, making transformations of the response necessary. However, I refrained from performing transformations until inspecting the appropriate diagnostic plots. All continuous predictors except Age possessed outliers (figure 1B-H). Moreover, there appeared to be a patient with a much higher alcohol intake than any other patient (figure 1F). I expected this observation to be highly influential. Finally, there appeared to be multicollinearity between calories and fat (figure 2).

A group of graphs showing different types of data

Description automatically generated with medium confidence

Figure 1: Boxplots for the predictor (A) and all continuous response variables (B-H).

A graph showing the amount of calories in daily

Description automatically generated

Figure 2: Scatterplot of fat vs daily calories.

The categorical predictors were not evenly distributed (tables 1-3). For example, there were far fewer males than females (table 1). This sometimes resulted in low sample sizes for some combinations of values (e.g., males who occasionally use vitamins; table 4). This could potentially cause issues when adding interactions.

Table 1: Frequency table for patient gender.

|  |  |
| --- | --- |
| Male | Female |
| 42 | 272 |

Table 2: Frequency table for patient vitamin use.

|  |  |  |
| --- | --- | --- |
| None | Occasional | Regular |
| 111 | 82 | 121 |

Table 3: Frequency table for patient smoking status.

|  |  |  |
| --- | --- | --- |
| Never | Former | Current |
| 156 | 115 | 43 |

Table 4: Cross-tabulation for patient gender (vertical) and vitamin use (horizontal).

|  |  |  |  |
| --- | --- | --- | --- |
|  | Never | Occasional | Regular |
| Male | 24 | 5 | 13 |
| Female | 87 | 77 | 108 |

Table 5: Cross-tabulation for patient gender (vertical) and smoking status (horizontal).

|  |  |  |  |
| --- | --- | --- | --- |
|  | Never | Former | Current |
| Male | 13 | 22 | 7 |
| Female | 143 | 93 | 36 |
|  |  |  |  |

Table 6: Cross-tabulation for patient vitamin use (vertical) and smoking status (horizontal).

|  |  |  |  |
| --- | --- | --- | --- |
|  | Never | Former | Current |
| Never | 46 | 44 | 21 |
| Occasional | 40 | 30 | 12 |
| Regular | 70 | 41 | 10 |

**Full model**

First, I fit a model containing all predictors. Two issues stood out. First, the qq-plot indicated non-normality of the residuals (figure 3), as anticipated. Second, the gvif indicated that calories had excessive multicollinearity (gvif2/(2\*df) = 13.3). I first addressed the multicollinearity by removing calories, which resolved the issue. I then log-transformed PBC to address non-normality. Though there were still slight departures at the tails on the qq-plot (figure 4), the transformation sufficiently resolved the issue. Additionally, this transformation did not introduce any new issues.

A graph of a function

Description automatically generated with medium confidence

Figure 3: QQ-plot for the initial model’s residuals. The line indicates the expected trend if the residuals are normally distributed.

A graph showing the amount of quantity of the amount of the amount of the amount of the amount of the amount of the amount of the amount of the amount of the amount of the amount of the

Description automatically generated

Figure 4: QQ-plot for the revised model. The line indicates the expected trend if the residuals are normally distributed.

**Model selection**

To build a model, I used a stepwise selection algorithm, evaluating models using AIC. I began with no predictors and considered new variables one at a time, both in the forwards and backwards directions. The best model was as follows:

where . I then considered every two-way interaction using the same algorithm. The best model was as follows:

where .

**Final model diagnostics**

After fitting the final model, I checked to ensure model assumptions were met. The fitted values plotted against the studentized residuals showed no patterns (figure 5). The same was true for the continuous predictors plotted against studentized residuals (figure 6). Additionally, the categorical predictors plotted against studentized residuals showed no trends (figure 7).

A diagram of a graph

Description automatically generated with medium confidence

Figure 5: Fitted values plotted against studentized residuals. The red line indicates a value of zero for the residuals.

A group of black dots

Description automatically generated

Figure 6: Predictors plotted against studentized residuals plots for continuous predictors. The red lines indicate a value of zero for the residuals.

A row of boxes with different sizes and numbers

Description automatically generated with medium confidence

Figure 7: Categorical predictors plotted against studentized residuals.

Surprisingly, there were no influential observations (figure 8). As with the previous check, the qq-plot showed that the residuals were approximately normally distributed (figure 9). Finally, there was no excessive multicollinearity (table 7).

A graph with numbers and lines

Description automatically generated

Figure 8: Cook’s distance for each observation. The top 3 values are labelled with their observation number.

A graph showing the amount of quantity of the amount of the amount of the amount of the amount of the amount of the amount of the amount of the amount of the amount of the amount of the

Description automatically generated

Figure 9: QQ-plot for the final model. The line indicates the expected trend if the residuals are normally distributed.

Table 7: Gvif2/(2\*df) for each predictor and interaction.

|  |  |
| --- | --- |
| Term | Gvif2/(2\*df) |
| BMI | 1.06 |
| Sex | 1.25 |
| Age | 1.26 |
| Fat | 1.28 |
| Fiber | 1.49 |
| Vitamin\*Smoking | 2.11 |
| Smoking\*DBC | 4.00 |
| Vitamin | 4.49 |
| Vitamin\*DBC | 4.76 |
| Smoking | 4.78 |
| DBC | 4.82 |

**Conclusion**

The model explained approximately 30% of the variability. This is fairly low, implying the model may not be sufficient for reliable prediction. Most terms in the final model were significant at the α = 0.05 level (table 8). DBC, Age, and Vitamin\*Smoking were the only exceptions.

Table 8: Type 1 ANOVA table for the final model. Variables with significant

effects are bolded.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Term | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| **BMI** | **1** | **13.46** | **13.46** | **32.87** | **< 0.01** |
| **Vitamin** | **2** | **9.44** | **4.72** | **11.53** | **< 0.01** |
| **Fiber** | **1** | **6.57** | **6.57** | **16.05** | **< 0.01** |
| **Fat** | **1** | **4.73** | **4.73** | **11.57** | **< 0.01** |
| **Smoking** | **2** | **3.49** | **1.75** | **4.27** | **0.01** |
| **Smoking\*DBC** | **2** | **2.91** | **1.45** | **3.55** | **0.03** |
| **Sex** | **1** | **1.76** | **1.76** | **4.31** | **0.04** |
| **Vitamin\*DBC** | **2** | **2.65** | **1.33** | **3.24** | **0.04** |
| DBC | 1 | 1.38 | 1.38 | 3.37 | 0.07 |
| Vitamin\*Smoking | 4 | 3.35 | 0.84 | 2.05 | 0.09 |
| Age | 1 | 1.17 | 1.17 | 2.86 | 0.09 |

Exact changes in average response value based on isolated changes in each predictor are difficult to obtain due to the necessary log-transformation. However, we can make conclusions about trends (table 9):

1. An increase in BMI is associated with a decrease in PBC on average.
2. Occasional and regular vitamin users have lower PBC on average than patients who do not take vitamins, though neither individual effect is significant.
3. A higher fiber intake is associated with higher PBC on average.
4. A higher fat intake is associated with a decrease in PBC on average.
5. Former smokers have lower PBC on average than patients who have never smoked. Interestingly, current smokers have higher PBC on average than patients who have never smoked.
6. Males have lower PBC on average than females.
7. Increases in DBC are associated with larger increases in average PBC when patients regularly take vitamins. Moreover, increases in DBC are associated with smaller increases in average PBC in patients who currently smoke. However, though these effects are technically significant, the effect sizes are so small, relative to other effect sizes, they are unlikely to have any practical significance.

Table 9: Coefficient t-test table for term coefficients with significant overall effects.

Though DBC did not have a significant effect, it is included here since there are

significant interactions including it.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Term | Estimate | Std. Error | t value | Pr(>|t|) |
| DBC | 6.68E-06 | 5.38E-05 | 0.12 | 0.90 |
| BMI | -0.03 | 0.01 | -4.87 | < 0.01 |
| Vitamin (Occasional) | -0.07 | 0.20 | -0.35 | 0.72 |
| Vitamin (Regular) | -0.12 | 0.18 | -0.67 | 0.50 |
| Fiber | 0.02 | 0.01 | 2.48 | 0.01 |
| Fat | -0.0022 | 0.00 | -1.85 | 0.07 |
| Smoking (Former) | -0.23 | 0.18 | -1.32 | 0.19 |
| Smoking (Current) | 0.13 | 0.24 | 0.55 | 0.58 |
| Sex (Male) | -0.19 | 0.12 | -1.64 | 0.10 |
| Vitamin (Occasional)\*DBC | 5.69E-05 | 6.86E-05 | 0.83 | 0.41 |
| Vitamin (Regular)\*DBC | 0.00015 | 6.09E-05 | 2.50 | 0.01 |
| Smoking (Former)\*DBC | -3.72E-05 | 5.41E-05 | -0.69 | 0.49 |
| Smoking (Current)\*DBC | -0.00022 | 9.08E-05 | -2.48 | 0.01 |

This study reveals some interesting trends in PBC based on several biometric and lifestyle factors. Namely, all the included variables are significant except Age and DBC, and alcohol intake, caloric intake, and cholesterol intake are not worth including in the model. Moreover, though the main effect of DBC is not significant, there are significant interactions with vitamin use and smoking, though the effect sizes are unlikely to have any practical significance. Finally, it is unlikely that this model can be reliably used to predict PBC in new patients.