PROJECT 2: LOGISTIC REGRESSION
MASM22: LINEAR AND LOGISTIC REGRESSION, 2025

Sek Huen Leung, Nancy Truong, Junjie Gu

April 2025

Introduction — Determinants of plasma β-carotene levels

Nierenberg DW, Stukel TA, Baron JA, Dain BJ, Greenberg ER. Determinants of plasma levels of beta-carotene and retinol. American Journal of Epidemiology 1989;130:511-521.

> Observational studies have shown that insufficient food consumption or low plasma concentrations of beta-carotene or other carotenoids may be correlated to an increased risk of getting certain cancers. Few research have examined the variables influencing plasma concentrations of these micronutrients. We performed a cross-sectional study to examine the relationship between individual characteristics, dietary variables, and plasma levels of beta-carotene and other carotenoids. Study subjects (N = 315) were patients who had an elective surgical procedure over three years to biopsy or remove a lesion of the lung, colon, breast, skin, ovary, or uterus that was found to be non-cancerous. We display the data for only one of the analytes. Patients who went through an elective procedure over a three-year period to biopsy or remove a non-cancerous lesion of the lung, colon, breast, skin, ovary, or uterus were the study subjects ($N = 315$). Only one of the analytes' data is shown.
>
> We conclude that human plasma concentrations of these micronutrients vary widely and that a large portion of this variability is related to eating habits and personal characteristics. More research will be required to better understand the physiological relationship between certain individual characteristics and these micronutrients' plasma concentrations.

The datafile carotene.xlsx contains 315 observations on the following 12 variables.

Variable name	Description
age	Age (years)
sex	Sex (1 = Male, 2 = Female)
smokstat	Smoking status (1 = Never, 2 = Former, 3 = Current Smoker)
bmi	Body mass index, BMI = weight/height^2 (kg/m^2)
vituse	Vitamin use (1 = Yes, fairly often, 2 = Yes, not often, 3 = No)
calories	Calories consumed per day (MJ, 2500 kcal $\approx$ 10 MJ)
fat	Fat consumed per day (g)
fiber	Fiber consumed per day (g)
alcohol	Number of alcoholic drinks consumed per week
cholesterol	Cholesterol consumed per day (mg)
betadiet	Dietary β-carotene consumed per day (mg)
betaplasma	Plasma β-carotene (mg/ml)

Our objective is to model the variability of the plasma β-carotene level, betaplasma, using $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$, as a function of one or more of the other variables. We will utilise a linear regression model $Y_i = x_i\beta + \epsilon_i$, where the random errors ϵ_i are assumed to be pairwise independent and $N(0, \sigma^2)$. We will need to use appropriate transformations in order fulfil these model assumptions.

1 Plasma β-carotene and body mass index

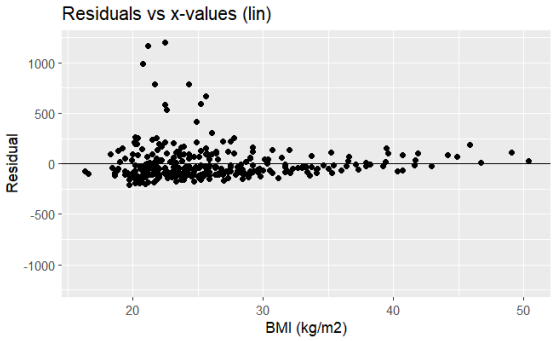
We begin by modelling how plasma β-carotene varies with body mass index (bmi).

1(a)

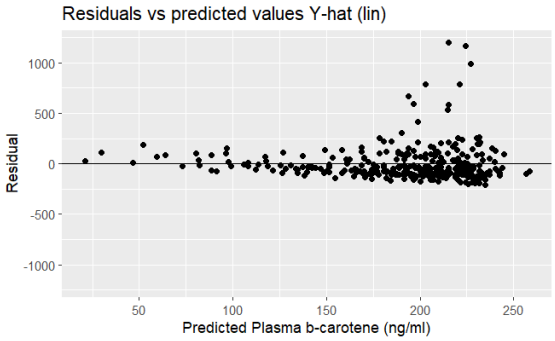
Firstly, we examine the two models, *Lin* and *log*, to determine which one better fits our data.

$$\text{Lin:} \quad \text{betaplasma}_i = \beta_0 + \beta_1 \cdot \text{bmi}_i + \varepsilon_i, \text{where } \varepsilon_i \sim N(0, \sigma^2), i = 1, \ldots, n$$
$$\text{Log:} \quad \ln(\text{betaplasma}_i) = \beta_0 + \beta_1 \cdot \text{bmi}_i + \varepsilon_i, \text{where } \varepsilon_i \sim N(0, \sigma^2), i = 1, \ldots, n$$

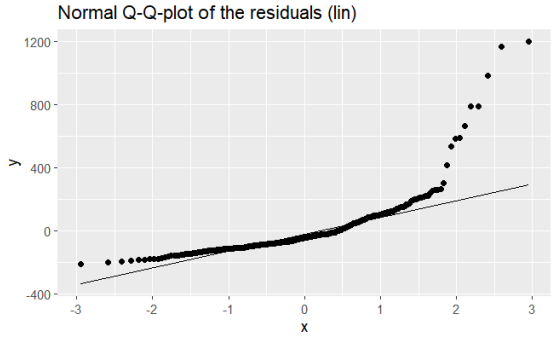
We perform a residual analysis for both models by plotting the residuals against the predicted values, as well as a QQ plot and a histogram for the residuals, to visually assess the model fit and compare their suitability (see Figure 1).



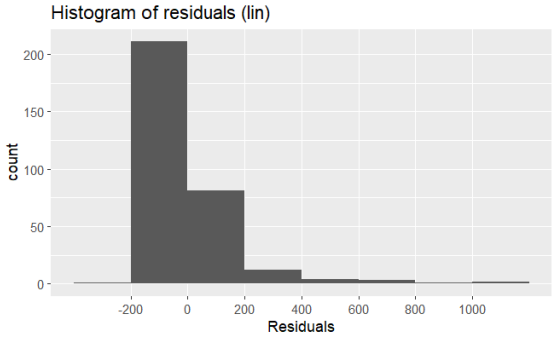
(a) Residuals vs x-values (lin)



(b) Residuals vs predicted values Y-hat (lin)



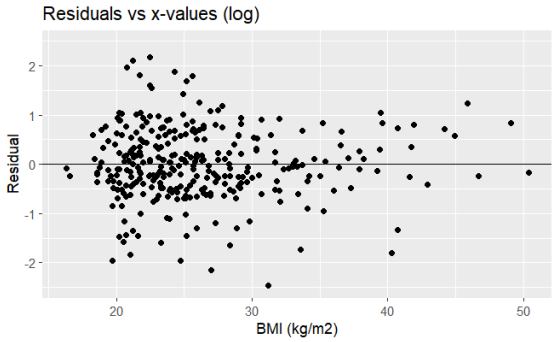
(c) Normal Q-Q plot of the residuals (lin)



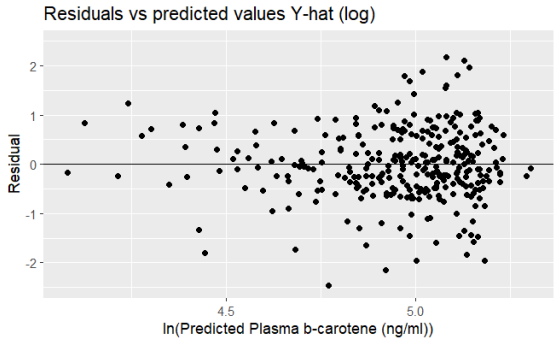
(d) Histogram of residuals (lin)

Figure 1: Residual analysis for the linear model

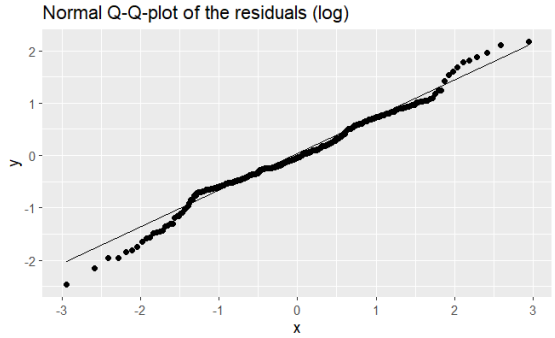
In Figure 1b, we observe a systematic pattern in the residuals, indicating that the residual variance σ^2 is not constant but increases with the predicted values $\hat{Y}_i$. Figure 1c further shows that the residuals are not randomly scattered around zero, suggesting deviations from the assumptions of homoscedasticity and independence. Additionally, the histogram in Figure 1d reveals a skewed residual distribution, indicating a deviation from Gaussian distribution assumption.



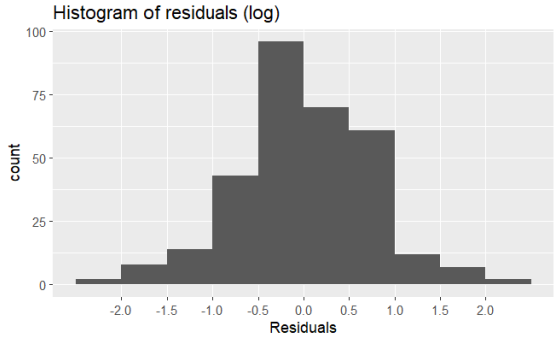
(a) Residuals vs x-values (log)



(b) Residuals vs predicted values Y-hat (log)



(c) Normal Q-Q plot of the residuals (log)



(d) Histogram of residuals (log)

Figure 2: Residual analysis for the logarithm model

Problems mentioned above seem not to be present for *log* model (see Figure 2).

1(b)

We use the Log model and present a table (Table 1) with the estimated β-values and their corresponding 95% confidence intervals.

	Estimate	2.5%	97.5%
(Intercept)	5.89	5.53	6.25
bmi	-0.04	-0.05	-0.02

Table 1: β-estimates and their 95% confidence intervals

We plot the log of plasma β-carotene against BMI (Figure 3), along with the estimated linear regression line, its 95% confidence interval, and a 95% prediction interval for future observations.

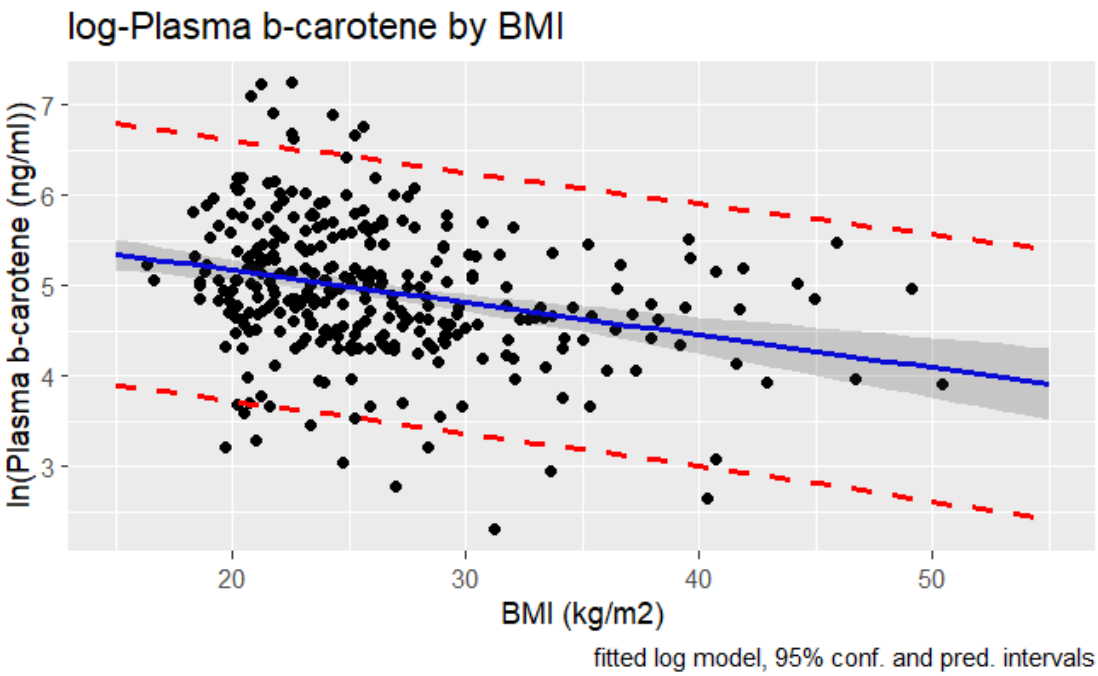


Figure 3: log-Plasma b-carotene by BMI

We transform the relationship back to

$$\text{betaplasma}_i = e^{\beta_0} \cdot e^{\beta_1 \cdot \text{bmi}_i} \cdot e^{\varepsilon_i}$$

and plot plasma β-carotene (ng/ml) against BMI, along with the estimated relationship, its 95% confidence interval, and a 95% prediction interval (see Figure 4).

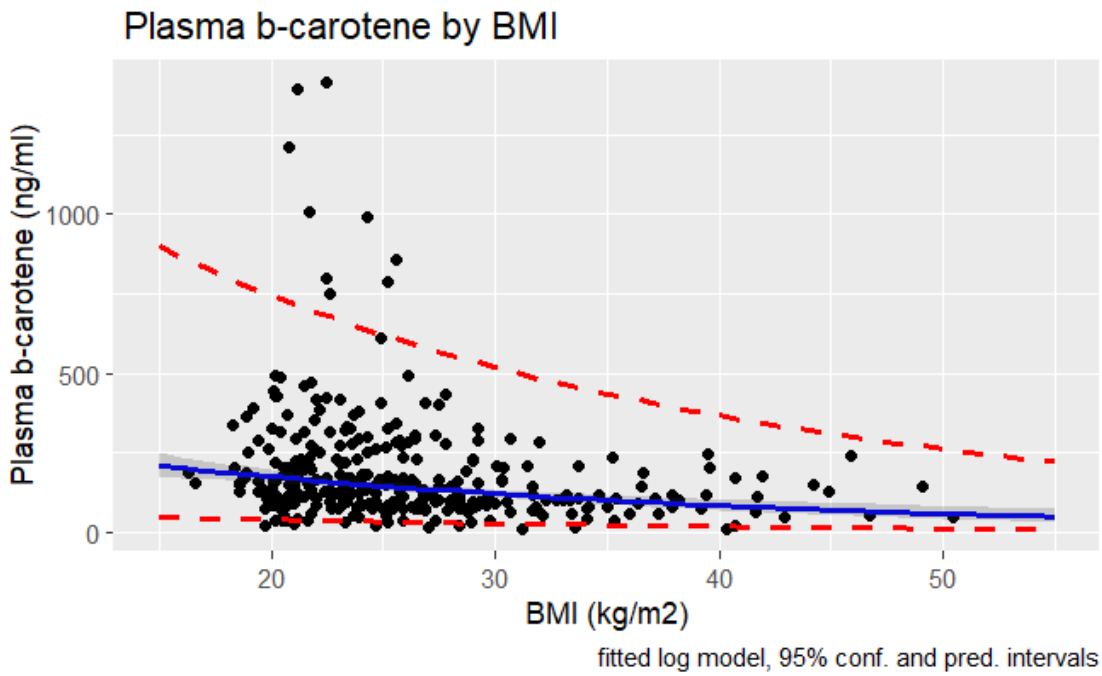


Figure 4: Plasma b-carotene by BMI

1(c)

Let $\mu =$ betaplasma and $t =$ bmi , then our model

$$\ln \mu_0 = \beta_0 + \beta_1 \cdot t_0 \iff \mu_0 = e^{\beta_0} \cdot \left(e^{\beta_1}\right)^{t_0} = a \cdot b^{t_0}$$

Let new bmi $= x_i + \Delta t$, then we have odd ratio:

$$\begin{aligned} \text{odd}_i &= a \cdot b^{t_0 + \Delta t} \\ &= a \cdot b^{t_0} \cdot b^{\Delta t} \\ &= \mu_0 \cdot b^{\Delta t} \end{aligned}$$

which means an additive change (Δt) in t gives relative change ($b^{\Delta t}$) in μ.

	Δt	$b^{\Delta t}$ Estimate	2.5%	97.5%
(i)	1	0.9648	0.9518	0.9779
(ii)	-1	1.0365	1.0226	1.0506
(iii)	-10	1.4317	1.2509	1.6386

Table 2: Relative change-estimates and their 95% confidence intervals

1(d)

To check whether BMI has a significant linear relationship with log plasma β-carotene, we do a two sided t-test on β_1 with the following hypothesis.

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

The test statistic is

$$t = \frac{\hat{\beta}_1}{s\sqrt{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{1,1}}}$$

Which is t-distributed with degree of freedom $n - (p + 1) = 313$. The observed value of the test statistic is -5.23 and the p-value is 3.11×10^{-7}. Since the p-value is less than the significant level 0.05, we can reject the null hypothesis and conclude that BMI has a significant linear relationship with log plasma β-carotene.

2 Plasma β-carotene and smoking habits

2(a)

We turn the categorical variable `smokstat` into a factor variable.

	frequency	β-carotene mean	s.d.	log β-carotene mean	s.d.
Never	157	206.1146	193.14184	5.050849	0.7453628
Former	115	193.4696	191.63952	4.941126	0.7975007
Current Smoker	43	121.3256	78.81163	4.613638	0.6243772

Table 3: Frequency, mean and s.d. corresponding to smoking habits

Never smoke is most suited to use as reference category, since it has the highest frequency (see Table 3) and makes β-estimate more certain.

We show boxplots and violin plots of both plasma β-carotene (Figure 5) and log plasma β-carotene (Figure 6) against `smokstat`.

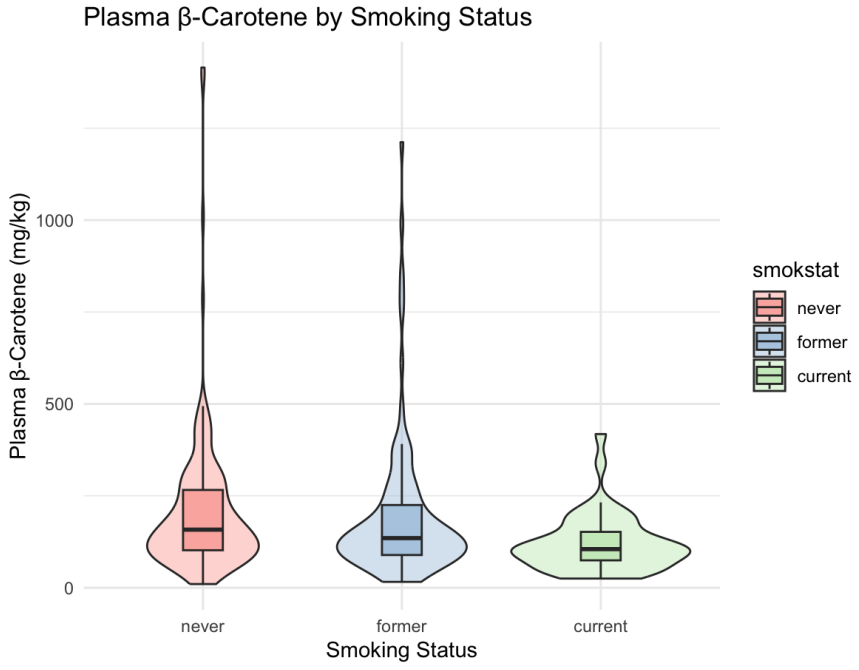


Figure 5: Boxplots of plasma b-carotene against smokstat

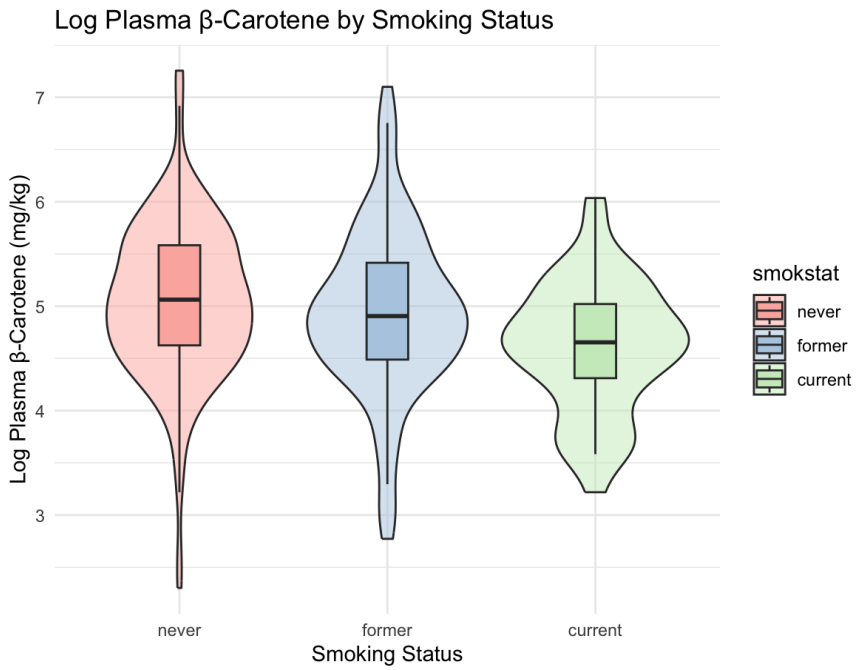


Figure 6: Boxplots of log plasma b-carotene against smokstat

From the plots (Figure 5 and 6), we can see that there are less outliers for log plasma β-carotene. Therefore, we still use log plasma β-carotene as dependent variable.

2(b)

Case 1:

We use "Never" as reference category, then the dummy variables are defined as follows.

$$x_1 = \begin{cases} 1 & \text{if } x_{\text{smokstat}} = \text{Former}, \\ 0 & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if } x_{\text{smokstat}} = \text{Current Smoker}, \\ 0 & \text{otherwise} \end{cases}$$

With $Y = \log(\text{plasmabeta})$, the model would then be expressed as:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \epsilon_i & \text{Never} \\ \beta_0 + \beta_1 + \epsilon_i & \text{Former} \\ \beta_0 + \beta_2 + \epsilon_i & \text{Current Smoker} \end{cases}$$

Variable	parameter	estimate	s.e.
intercept (Never)	β_0	5.05085	0.05986
Former (vs Never)	β_1	-0.10972	0.09207
Current Smoker (vs Never)	β_2	-0.43721	0.12911

Table 4: Parameter estimate and standard error with Never as reference category

Case 2:

We use "Current Smoker" as reference category, then the dummy variables are defined as follows.

$$x_1 = \begin{cases} 1 & \text{if } x_{\text{smokstat}} = \text{Never}, \\ 0 & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if } x_{\text{smokstat}} = \text{Former}, \\ 0 & \text{otherwise} \end{cases}$$

With $Y = \log(\text{plasmabeta})$, the model would then be expressed as:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \epsilon_i & \text{Current Smoker} \\ \beta_0 + \beta_1 + \epsilon_i & \text{Never} \\ \beta_0 + \beta_2 + \epsilon_i & \text{Former} \end{cases}$$

Variable	parameter	estimate	s.e.
intercept (Current Smoker)	β_0	4.6136	0.1144
Never (vs Current Smoker)	β_1	0.4372	0.1291
Former (vs Current Smoker)	β_2	0.3275	0.1341

Table 5: Parameter estimate and standard error with Current Smoker as reference category

From the tables 4 and 5, we can see that the standard errors are larger when we use Current Smoker as reference category. It is because Current Smoker is a small category. Therefore, it is more reasonable to chose Never as reference category.

2(c)

Case 1: We use the model with "Never" as reference category.

smokstat	log(β-carotene) expected	2.5%	97.5%	β-carotene expected	2.5%	97.5%
Never	5.05	4.93	5.17	156.15	138.80	175.68
Former	4.94	4.80	5.08	139.93	121.94	160.57
Current Smoker	4.61	4.39	4.84	100.85	80.52	126.31

Table 6: Perdition and confidence interval in different smokstat with Never as reference category

Case 2: We use the model with "Current Smoker" as reference category. We can see that the expected values

smokstat	log(β-carotene) expected	2.5%	97.5%	β-carotene expected	2.5%	97.5%
Never	5.05	4.93	5.17	156.15	138.80	175.68
Former	4.94	4.80	5.08	139.93	121.94	160.57
Current Smoker	4.61	4.39	4.84	100.85	80.52	126.31

Table 7: Perdition and confidence interval in different smokstat with Current Smoker as reference category

of $\log(\beta - \text{carotene})$ is close the mean in 2(a), which means it is a good fit. However, there is a significant difference between the expected $\beta-$carotene and its mean in 2(a). Since Y is log-normal distributed,

$$\mathrm{E}(\mathbf{Y}) = \exp\big(\mu + \sigma^2/2\big)$$

. We calculate $\exp(\mu)$ which is actually the median. Both model has the same expected value and interval because they have the same fitted line. For example:

$$\hat{y}_{\text{Former}} = \hat{\beta}_0^{(\text{Never})} + \hat{\beta}_{\text{Former}}^{(\text{Never})} = \hat{\beta}_0^{(\text{Current})} + \hat{\beta}_{\text{Former}}^{(\text{Current})}.$$

2(d)

We do a Global F test for whether there are significant differences in log plasma β-carotene between any of the smokstat categories with the following hypothesis.

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_1 : \exists i = 1, 2 : \beta_i \neq 0$$

The test statistic

$$\frac{\text{MS(Regr)}}{\text{MS(Error)}}$$

has F(2,312) distribution when the null hypothesis is true. The observed test statistic is 5.75 and the p-value is 0.00353. Since the p-value is less than 0.05, we reject the null hypothesis and conclude that there are significant differences in log plasma β-carotene between any of the smokstat categories.

3 Multiple linear regression

3(a)

Sex	Frequency
Male	42
Female	273

Table 8: Frequency table of sex

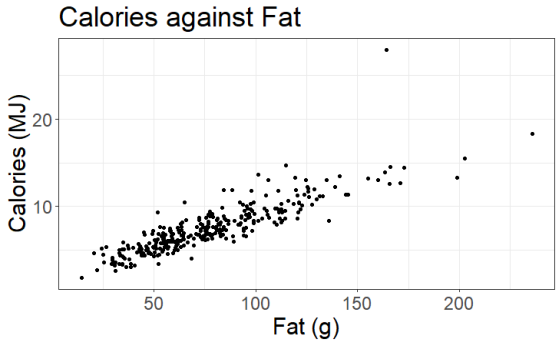
Vituse	Frequency
Often	122
Not Often	82
No	111

Table 9: Frequency table of vituse

We use Female and Often as reference categories since they have the highest frequency.

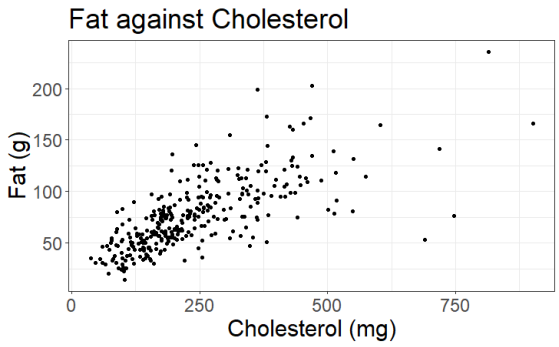
3(b)

The pairs with correlations stronger than ± 0.6 are (calories,fat), (calories,cholesterol) and (fat,cholesterol). In Figure 7, we can see there are some linear relations between these three variables. Some variables may

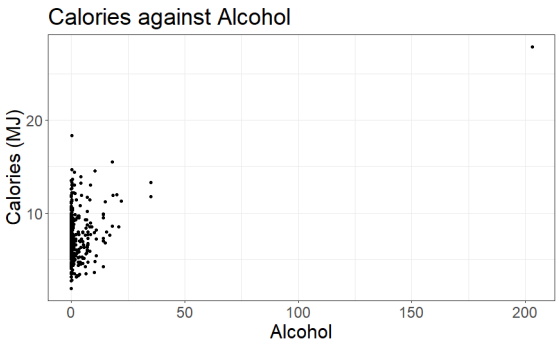


(a) calories against fat

(b) cholesterol against calories



(c) fat against cholesterol



(d) Calories against Alcohol

Figure 7: Pairwise plots of x-variables

not have strong correlations, but have correlations on the extreme values due to eating and drinking habits. We can see that the extreme value of alcohol corresponds to the extreme value of calories.

3(c)

Firstly, we ignore any potential problems and fit a (full) model log β-carotene depends on all the other variables.

We also calculated the VIF (Variance Inflation Factor) or GIVF (Generalised Variance3 Inflation Factor) values for each variable (Table 10).

variable	GVIF	DF	$\text{GVIF}^{(1/(2*DF))}$
bmi	1.069660	1	1.034244
age	1.307586	1	1.143497
calories	13.210244	1	3.634590
fat	8.175794	1	2.859334
cholesterol	2.195956	1	1.481876
fiber	2.504249	1	1.582482
alcohol	2.564752	1	1.601484
betadiet	1.338719	1	1.157030
smokstat	1.178201	2	1.041849
sex	1.287887	1	1.134851
vituse	1.149879	2	1.035531

Table 10: VIF/GVIF of full model

Fat and calories have more than 80% of the variability can be explained using the other x-variables since their VIF are greater than 5. The most problematic x-variable is calories. Calorie is removed and we refit the model without it (Model 3.c).

variable	GVIF	DF	$\text{GVIF}^{(1/(2*DF))}$
bmi	1.067334	1	1.033119
age	1.219329	1	1.104232
fat	2.244437	1	1.498144
cholesterol	2.129296	1	1.459211
fiber	1.465013	1	1.210377
alcohol	1.124159	1	1.060264
betadiet	1.338004	1	1.156721
smokstat	1.177041	2	1.041593
sex	1.287797	1	1.134811
vituse	1.115890	2	1.027792

Table 11: VIF/GVIF of reduced model

In Table 11, we can se all VIF/GVIF is decreased and the VIF of fat is reduced to a more acceptable value. It is expected since fat and calories have strong correlations.

3(d)

Let Y = log(betaplasma), x_1 = age, x_2 = bmi, x_3 = fat, x_4 = fiber, x_5 = alcohol, x_6 = cholesterol, x_7 = betadiet,

$$x_8 = \begin{cases} 1 & \text{if } x_{\text{smokstat}} = \text{Former}, \\ 0 & \text{otherwise} \end{cases}$$

$$x_9 = \begin{cases} 1 & \text{if } x_{\text{smokstat}} = \text{Current Smoker}, \\ 0 & \text{otherwise} \end{cases}$$

$$x_{10} = \begin{cases} 1 & \text{if } x_{\text{sex}} = \text{Male}, \\ 0 & \text{otherwise} \end{cases}$$

$$x_{11} = \begin{cases} 1 & \text{if } x_{\text{vituse}} = \text{Not Often}, \\ 0 & \text{otherwise} \end{cases}$$

$$x_{12} = \begin{cases} 1 & \text{if } x_{\text{vituse}} = \text{Never}, \\ 0 & \text{otherwise} \end{cases}$$

Let $\beta = (\beta_0, \beta_1, \ldots, \beta_{12})^T$, $\mathbf{x} = (1, x_1, \ldots, x_{12})^T$ then our model is

$$Y_i = \mathbf{x}_i^T \beta + \epsilon_i$$

.

Variable	param	β estimate	2.5%	97.5%	e^β estimate	2.5%	97.5%
intercept	β_0	5.5250	4.9917	6.0583	250.8870	147.1883	427.6447
age	β_1	0.0060	0.0003	0.0117	1.0060	1.0003	1.0117
bmi	β_2	-0.0320	-0.0448	-0.0192	0.9685	0.9561	0.9810
fat	β_3	-0.0012	-0.0045	0.0021	0.9988	0.9955	1.0021
fiber	β_4	0.0227	0.0058	0.0397	1.0230	1.0058	1.0405
alcohol	β_5	0.0018	-0.0046	0.0082	1.0018	0.9954	1.0083
cholesterol	β_6	-0.0007	-0.0016	0.0001	0.9993	0.9984	1.0001
betadiet	β_7	0.0559	-0.0028	0.1145	1.0574	0.9972	1.1213
smokFormer(vs Never)	β_8	-0.0718	-0.2385	0.0948	0.9307	0.7878	1.0995
smokCurrent(vs Never)	β_9	-0.2729	-0.5157	-0.0300	0.7612	0.5971	0.9704
sexMale (vs Female)	β_{10}	-0.2011	-0.4497	0.0476	0.8179	0.6378	1.0488
vitNotOften (vs Often)	β_{11}	0.0013	-0.1914	0.1939	1.0013	0.8258	1.2140
vitNo (vs Often)	β_{12}	-0.2658	-0.4454	-0.0862	0.7666	0.6405	0.9174

Table 12: Estimate and confidence interval of β and e^β

	test	H_0	test statistic	dist.	obs test stat	p-value
(i)	t-test	$\beta_2 = 0$	$\frac{\hat{\beta_2}}{s\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{2,2}}}$	t(302)	-4.918	1.44e-06
(ii)	partial F-test	$\beta_i = 0, \forall i \in \{1, \ldots, 12\} \backslash \{2\}$	$\frac{Q/11}{s^2_{\text{full}}}$	F(11,302)	6.2597	2.633e-09
(iii)	partial F-test	$\beta_i = 0, \forall i \in \{1, \ldots, 12\} \backslash \{8, 9\}$	$\frac{Q/10}{s^2_{\text{full}}}$	F(10,302)	8.6928	1.632e-12

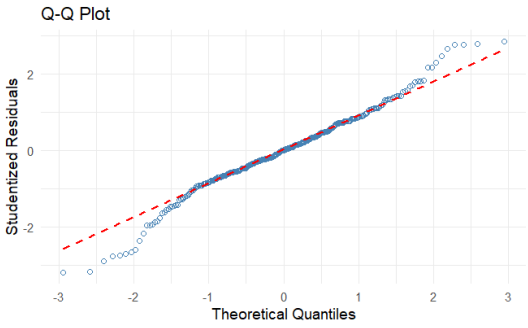
Table 13: test results of different models

For (i), we use t-test because we are trying to test against one parameter. Since the p-vaule is small, we can reject the null hypothesis and conclude that there is a significant relationship between log plasma b-carotene and BMI, given the other variables in the model.

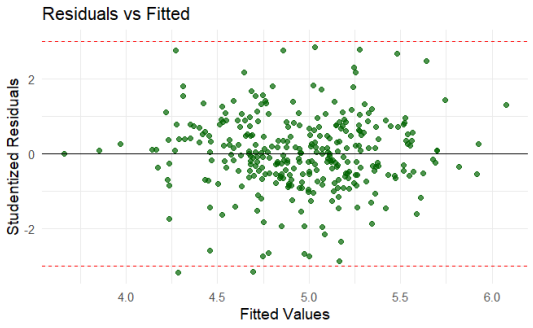
For (ii), we use partial F-test because we are trying to compare two models. Since the p-vaule is small, we can reject the null hypothesis and conclude that Model 3(c) is significantly better than Model 1(b).

For (iii), we use partial F-test because we are trying to compare two models. Since the p-vaule is small, we can reject the null hypothesis and conclude that Model 3(c) is significantly better than Model 2(b).

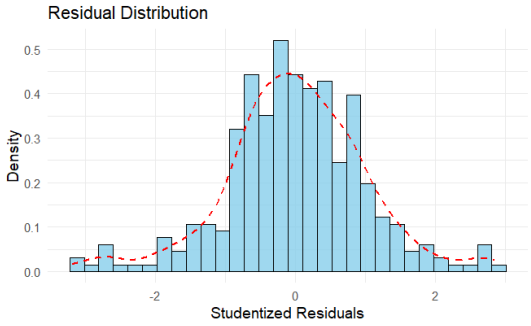
3(e)



(a) Q-Q



(b) Residuals vs Fitted



(c) Residual Distribution

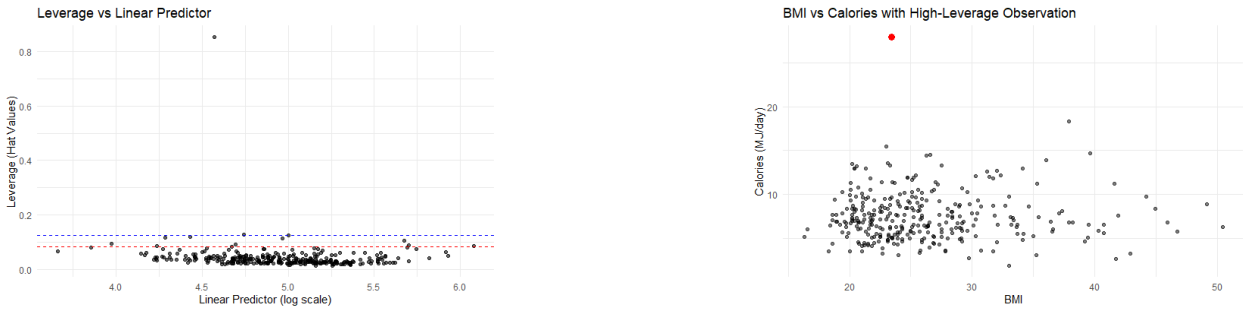
Figure 8: Studentized Residuals

- For the Q-Q Plot (Figure 8a): Most of the points fall near the diagonal, but there are some off in the upper right and lower left corners. Good alignment in the center means that most of the residuals are close to normal, and slight deviation at the two ends means that there may be a few extreme values but with little effect.
- For the Residuals vs. Fitted plot (Figure 8b): Residuals are evenly distributed above and below 0, suggesting that the residuals is almost white noise and there is no significant heteroskedasticity.
- For the Residual Histogram (Figure 8c): The plot is symmetric, single-peaked, this supports the conclusion of the Q-Q plot that the residuals are close to normal.

Conclusion: Overall, the residuals appear to be approximately normally distributed, with only very few extreme values. Minor deviations in the tails do not affect the appropriateness of the model.

3(f)

The horizontal reference line (often set at $\frac{2p}{n}$, where p is the number of parameters) indicates a threshold above which observations are typically considered to have high leverage.



(a) (b)

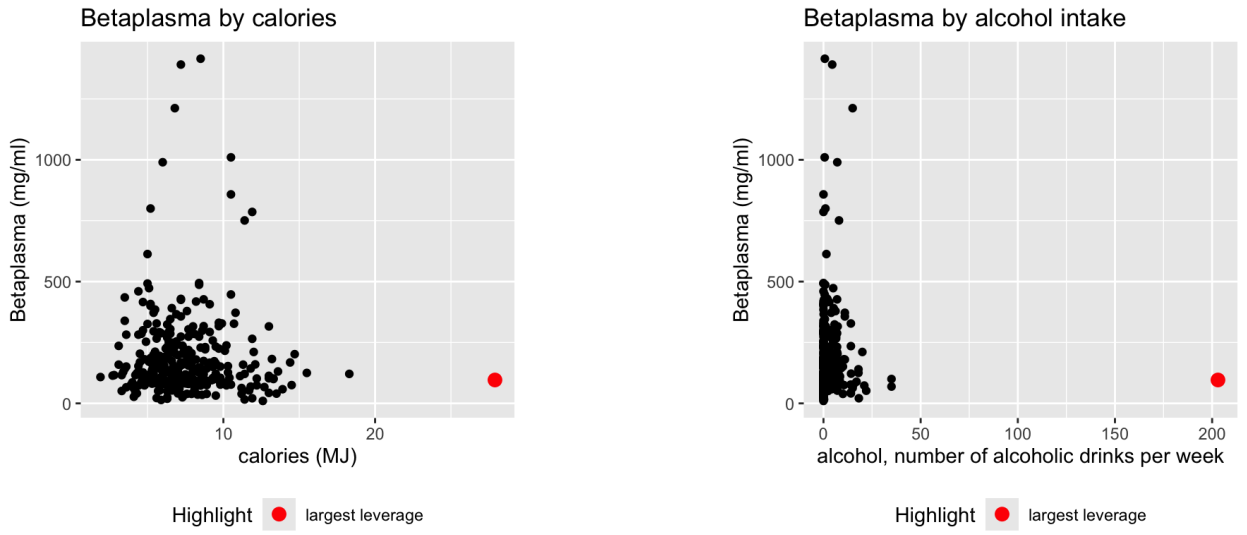
Figure 9: Leverage

Figure(a) has a point significantly beyond the horizontal reference line, suggesting that it has a significantly higher leverage compared to the majority of observations. And figure(b) highlights this highest leverage point in red. This point is identified as one with:

$$\text{Age} = 65, \quad \text{Sex} = \text{male}, \quad \text{BMI} = 23.4, \quad \text{Calories} = 27.9, \quad \text{Fat} = 164.3, \quad \text{Cholesterol} = 603.$$

This person has highest calorie intake (27.9MJ), extremely highest cholesterol and relatively high alcohol consumption, which could destruct model estimates. This could either be a data entry error (for example, wrong units or decimal misplacement) or an actual extreme diet.

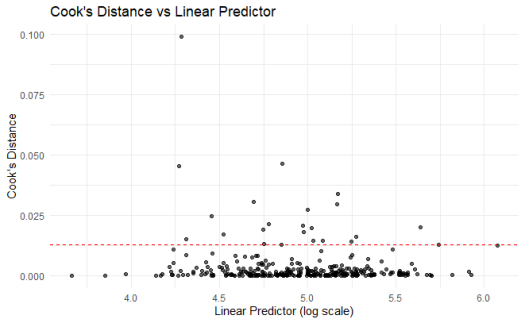
Conclusion: The observation has such a large leverage because its predictor values are atypical relative to the rest of the data, which makes it far away from the center of the predictor space (X- space).



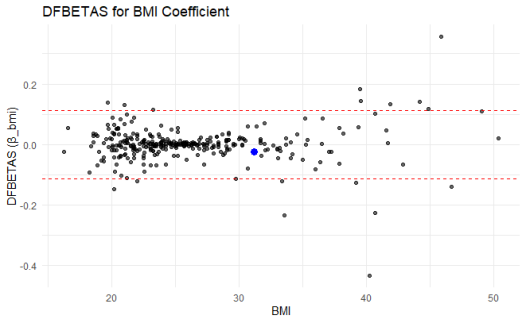
(a) Betaplasma concentration by alcohol intake (b) Betaplasma concentration by calorie intake

3(g)

Cook's Distance for Model 3.c.



(a) Cook's Distance vs Linear Predictor



(b) DFBETAS for BMI Coefficient

Figure 11: Cook's Distance & DFBETAS

The observation that has the highest Cook's Distance is

$$\text{Age} = 40, \quad \text{Sex} = \text{female}, \quad \text{BMI} = 31.2, \quad \text{Calories} = 12.6, \quad \text{Fat} = 165.7, \quad \text{Cholesterol} = 900.7,$$

$$\text{yhat} = 4.285305 \quad \text{r} = \text{-3.184998}, \quad \text{D} = 0.09882828, \quad \text{df6} = \text{-0.89848}.$$

This observation has a large residual, indicating a poor fit relative to the model's prediction. The most affected parameter is β_6, which corresponds to the variable cholesterol. This influence is visually evident in the DFBETAS graph for β_6 against the cholesterol variable (see Figure 12).

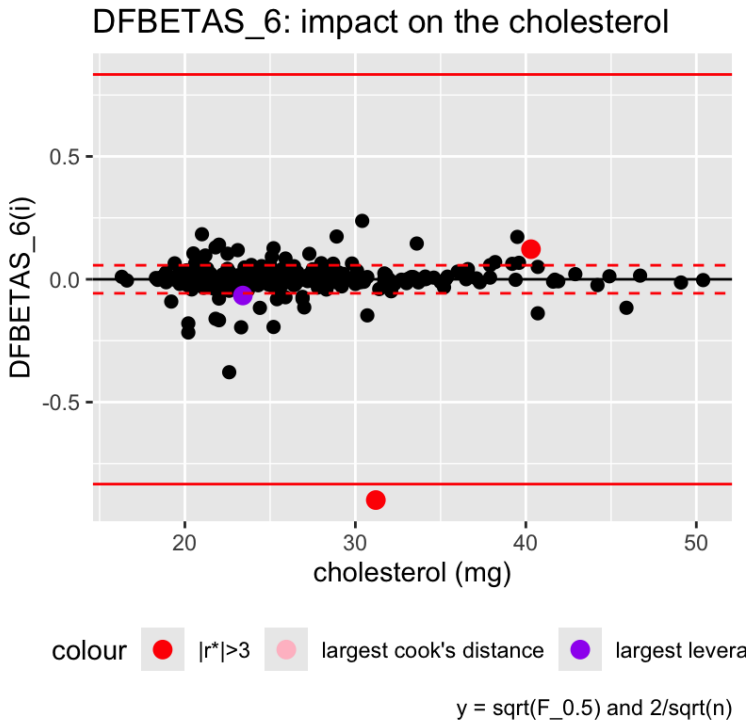


Figure 12: DFBETA$_6$ vs. cholesterol

The observation with the largest Cook's distance is highlighted in pink in the plot (see Figure 12). Note that the same observation also has a residual with an absolute value greater than 3, and such extreme residual are highlighted in red. As a result of this overlap, the highlighting appears red in the figure, indicating that the point is both highly influential and poorly fitted by the model.

Some more details for the observation with the largest leverage,

Statistic	**Value**	**Statistic**	**Value**
Residual ($r_{\log}$)	-0.0286	DFBETA_5	-0.0013
Leverage ($v_{\log}$)	0.8531	DFBETA_6	-0.0660
Cook's Distance (D)	0.0004	DFBETA_7	0.0011
DFBETA_0	0.0040	DFBETA_8	0.0041
DFBETA_1	-0.0034	DFBETA_9	-0.0006
DFBETA_2	-0.0019	DFBETA_{10}	0.0045
DFBETA_3	0.0011	DFBETA_{11}	0.0019
DFBETA_4	-0.0008	DFBETA_{12}	0.0026

Its residual is small, meaning the model fits this point well. Its Cook's Distance is small, indicating it has little to no influence on the overall parameter estimates. Additionally, all DFBETAS are small, meaning that no regression coefficients are meaningfully affected by the inclusion or exclusion of this observation. Therefore, it does not have any alarming influence on the model estimates.

4 Removing the influential observation

4(a)

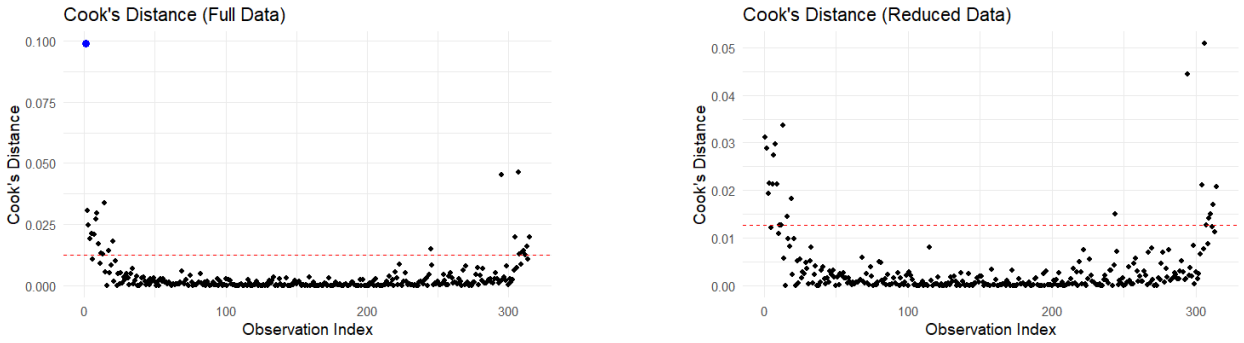
In the full dataset, a reference line at

$$\text{Threshold} = \frac{4}{n} \quad \text{(where } n \text{ is the number of observations)}$$

was added to aid in identifying influential points. One observation exhibited a Cook's distance of approximately 0.10, clearly above the reference line, thereby signaling a high level of influence.

After removing this influential observation, a new model was estimated on the reduced dataset. The corresponding Cook distance plot for the reduced data is shown in the next figure. The reference line in this plot is set at $\frac{4}{n_{\text{reduced}}}$.

The maximum Cook's distance in the reduced data decreased considerably because of the removal of the extreme values.



(a) (b)

Figure 13: Cook's Distance

In Figure 13, the maximum Cook's distance was significantly reduced in the reduced dataset, indicating that the excluded observations had some effects on the fitted model. This supports the decision to exclude them to obtain more stable estimates of the model parameters.

Variable(Full Data)	GVIF	Df	$\text{GVIF}^{1/(2\cdot\text{Df})}$	Variable(Reduced Data)	GVIF	Df	$\text{GVIF}^{1/(2\cdot\text{Df})}$
bmi	1.0673	1	1.0331	bmi	1.0650	1	1.0320
age	1.2193	1	1.1042	age	1.2175	1	1.1034
fat	2.2444	1	1.4981	fat	2.2027	1	1.4842
cholesterol	2.1293	1	1.4592	cholesterol	2.1161	1	1.4547
fiber	1.4650	1	1.2104	fiber	1.4649	1	1.2103
alcohol	1.1242	1	1.0603	alcohol	1.1250	1	1.0606
betadiet	1.3380	1	1.1567	betadiet	1.3455	1	1.1600
smokstat	1.1770	2	1.0416	smokstat	1.1777	2	1.0417
sex	1.2878	1	1.1348	sex	1.2965	1	1.1387
vituse	1.1159	2	1.0278	vituse	1.1169	2	1.0280

Table 14: VIF/GVIF for Full and Reduced Data

From the Table 14, GVIF values did not differ significantly between the full and reduced data sets, indicating that the correlations between the variables remained largely unchanged despite the removal of influential observations.

Conclusion: This suggests that although the influential observations had a significant impact on the Cook distance measure, it had little effect on the interrelationships between the predictor variables.

4(b)

Case 1. Full Dataset:
The original model (null model) has BIC = 732.67. For the reduced data set, the stepwise selection of BIC proceeded as follows:

- **Step 1**: BMI entered first (ΔBIC = -20.64)
- **Step 2**: Fiber added (ΔBIC = -9.44)
- **Step 3**: Cholesterol introduced (ΔBIC = -10.76)
- **Step 4**: Vitamin use (vituse) finalized the model (ΔBIC = -2.90)

The Final Model (Model 4(b)) with BIC = 688.93 is as follows:

$$\begin{aligned}\log(\text{betaplasma}) = \beta_0 + \beta_1 \cdot \text{bmi} + \beta_2 \cdot \text{fiber} + \beta_3 \cdot \text{cholesterol} \\ + \beta_4 \cdot \mathbb{I}(\text{vituse} = \text{Rarely}) + \beta_5 \cdot \mathbb{I}(\text{vituse} = \text{Never}) + \epsilon\end{aligned} \tag{1}$$

where $\mathbb{I}(\text{vituse} = \text{X}) = \begin{cases} 1 & \text{vituse} = \text{X} \\ 0 & \text{otherwise} \end{cases}$, $\epsilon \sim \mathcal{N}(0, 0.4680^2)$

The Estimates of the Parameters are as follows (Figure 15):

Predictor	Estimate	95% CI	p-value
(Intercept)	5.721	(5.308, 6.134)	<0.001
BMI	-0.029	(-0.042, -0.016)	<0.001
Fiber	0.032	(0.018, 0.047)	<0.001
Cholesterol	-0.0012	(-0.0018, -0.0006)	<0.001
Vituse (Rarely)	-0.049	(-0.242, 0.144)	0.614
Vituse (Never)	-0.326	(-0.504, -0.148)	0.0004

Table 15: Final Model(Full Dataset) Estimates with 95% Confidence Intervals

Case 2. Reduced Dataset:
The original model (null model) has BIC = 718.99. For the reduced data set, the stepwise selection of BIC proceeded as follows:

- **Step 1**: BMI entered first (ΔBIC = -19.87)
- **Step 2**: Fiber added (ΔBIC = -10.60)
- **Step 3**: Fat introduced (ΔBIC = -6.11)
- **Step 4**: Vitamin use (vituse) finalized the model (ΔBIC = -4.27)

The Final Model (Model 4(b)) with BIC = 678.13 is as follows:

$$\begin{aligned}\log(\text{betaplasma}) = \beta_0 + \beta_1 \cdot \text{bmi} + \beta_2 \cdot \text{fiber} + \beta_3 \cdot \text{fat} \\ + \beta_4 \cdot \mathbb{I}(\text{vituse} = \text{Rarely}) + \beta_5 \cdot \mathbb{I}(\text{vituse} = \text{Never}) + \epsilon\end{aligned} \tag{2}$$

where $\epsilon \sim \mathcal{N}(0, 0.4552^2)$

The Estimates of the Parameters are as follows (Figure 16):

Predictor	Estimate	95% CI	p-value
(Intercept)	5.721	(5.308, 6.134)	<0.001
BMI	-0.029	(-0.042, -0.016)	<0.001
Fiber	0.032	(0.018, 0.047)	<0.001
Cholesterol	-0.0012	(-0.0018, -0.0006)	<0.001
Vituse (Rarely)	-0.049	(-0.242, 0.144)	0.614
Vituse (Never)	-0.326	(-0.504, -0.148)	0.0004

Table 16: Final Model(Reduced Dataset) Estimates with 95% Confidence Intervals

- The reduced dataset model has a lower final BIC (678.13) compared to the full dataset model (688.93). A lower BIC indicates a better balance between model fit and complexity, suggesting that the reduced dataset that does not include the data that has the largest cook's distance yields a better model.
- In Case 1 (Full Data), the stepwise procedure selected *Cholesterol* as the third predictor, whereas in Case 2 (Reduced Dataset) the predictor chosen at the corresponding step was *Fat*. This difference implies that the variable selection process is sensitive to excluding influential observations.
- The estimated residual standard error is slightly lower in the reduced model (0.4552) than in the full data model (0.4680), indicating that the reduced dataset may have less unexplained variability.
- Model.4(b) maintains greater simplicity (R^2_{Full}=0.1927 and $R^2_{Reduced}$=0.1858) than Model.3(c) (R^2=0.2214) while sacrificing little explanatory power.This model selection process successfully balanced simplicity and interpretability.

Remark:
It is interesting to note that although we obtained model 4(b) in the reduced dataset by using the "step" function. There is still a model with smaller BIC(677.05), as follows:

$$\begin{aligned}\log(\text{betaplasma}) = \beta_0 + \beta_1 \cdot \text{bmi} + \beta_2 \cdot \text{fiber} + \beta_3 \cdot \text{calories} \\ + \beta_4 \cdot \mathbb{I}(\text{vituse} = \text{Rarely}) + \beta_5 \cdot \mathbb{I}(\text{vituse} = \text{Never}) + \epsilon\end{aligned} \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, 4536^2)$

Predictor	Estimate	95% CI	p-value
(Intercept)	5.757	(5.340, 6.174)	<0.001
BMI	-0.030	(-0.042, -0.018)	<0.001
Fiber	0.041	(0.025, 0.057)	<0.001
Calories	-0.052	(-0.081, -0.023)	0.001
Vituse (Rarely)	-0.069	(-0.259, 0.121)	0.474
Vituse (Never)	-0.349	(-0.524, -0.174)	<0.001

Table 17: Final Model 2 (Reduced Dataset) Estimates with 95% Confidence Intervals

5 Fine-tuning the model

5(a)

The final model (Model 5(a)) for the reduced dataset is given by

$$\log(\text{betaplasma}) = \beta_0 + \beta_1 \cdot \text{bmi} + \beta_2 \cdot \text{fiber} + \beta_3 \cdot \text{fat} + \beta_4 \cdot \mathbb{I}(\text{vituse} = \text{Often/Rarely}) + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, 0.4545^2)$.

Predictor	Estimate	95% CI	p-value
(Intercept)	5.392	(4.972, 5.812)	<0.001
BMI	-0.030	(-0.042, -0.017)	<0.001
Fiber	0.035	(0.020, 0.050)	<0.001
Fat	-0.004	(-0.006, -0.002)	0.001
Vitamin Use (Often/Rarely)	0.314	(0.156, 0.472)	<0.001

Table 18: Regression Results with 95% Confidence Intervals

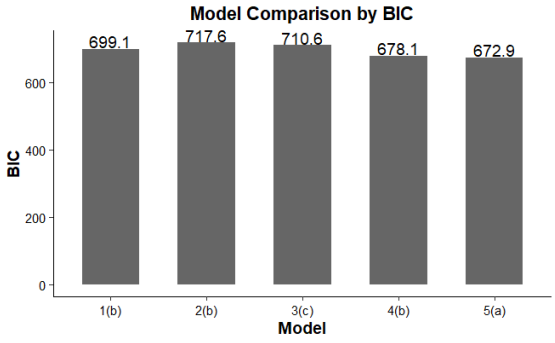
The variable *vituse* was originally represented with three levels. However, the estimate for the 'Rarely' level is not statistically significant (p = 0.458), while the contrast for the 'Never' level is highly significant. So we can reduce the number of parameters that need to be estimated by combining 'Rarely' with 'Often'.

5(b)

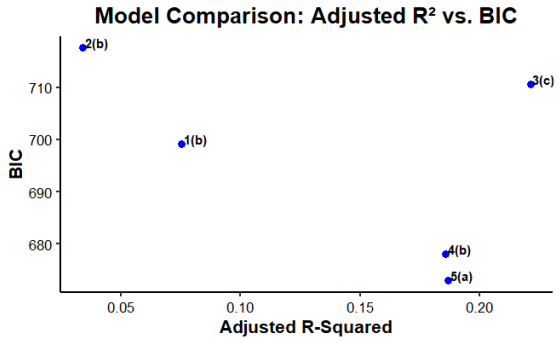
The following Table 19 presents the number of β-parameters, the R^2, adjusted R^2, AIC, and BIC for the five models:

Model	**Num. Params**	**R^2**	**Adj. R^2**	**AIC**	**BIC**
1(b)	2	0.0784	0.0754	687.87	699.12
2(b)	3	0.0402	0.0340	702.61	717.60
3(c)	13	0.2511	0.2214	658.08	710.61
4(b)	6	0.1988	0.1858	651.88	678.13
5(a)	5	0.1974	0.1870	650.44	672.94

Table 19: Model Performance Metrics



(a) Model BIC



(b) Adjusted R^2 vs BIC

Figure 14: Model Comparison

From the Figure 14, we can tell:

- **Models 1(b) and 2(b):** These simpler models have relatively low R^2 values and higher AIC and BIC compared to the others.
- **Model 3(c):** Although it has the highest R^2 (0.2511), it is overly complex with 13 parameters, resulting in a relatively high BIC.
- **Models 4(b) and 5(a):** Both models provide similar explanatory power (with R^2 around 0.198 and adjusted R^2 around 0.186) but differ in complexity. Model 5(a), which merges vitamin use variable into two levels, uses only 5 parameters and has the lowest AIC (650.44) and BIC (672.94).

Conclusion: Based on the lower AIC and BIC values and the simplicity through using fewer parameters without sacrificing significant explanatory power, **Model 5(a)** is the best model.

6 AI Statement

- **Grammarly**: For spell-checking and grammar refinement throughout the document
- **ChatGPT & DeepSeek**: Modify code and debug.

7 Author Contributions

- **Sek Huen Leung**: Derivations, analysis, discussions, programming, visualization, writing report for Part 1 - 3.
- **Nancy Truong**: Derivations, analysis, discussions, programming, visualization, writing report for Part 1 - 3.
- **Junjie Gu**: Derivations, analysis, discussions, programming, visualization, writing report for Part4 & Part5.