LUND UNIVERSITY

PROJECT 2: LOGISTIC REGRESSION

MASM22: LINEAR AND LOGISTIC REGRESSION, 2025

Nancy Truong, Xinxu Yao

May 2025

Contents

Introduction

This project analyzes the relationship between low Plasma β-carotene which is categorical variable and other 11 variables using logistic regression models. We first build one logistic regression model between low Plasma β-carotene and vitamin use and one between Plasma β-carotene and bmi. Then we build a multiple logistic regression model and make a variable selection. After that we compare the several multiple models we get and find out two best model with criterion function AIC and BIC, then we make residual analysis and perform a goodness-of-fit test. Finally we find the "best" model and make a conclusion with it.

1 Low plasma β-carotene and vitamin use

1.a Cut-off value

We will investigate whether the plasma β-carotene concentration is low or not and model the probability of having a low concentration dependent on dietary and/or background factors. To do this, we require an appropriate cut-off value. One that effectively differentiates between those at high and low risk for developing specific medical conditions will be used.

The cut-off value is 0.42μmol/l (micromoles per liter). Unfortunately, our data is from 1989, when the concentrations were measured in nanograms per milliliter (ng/ml).

We convert 0.42μmol/l to ng/ml using the molar mass of β-carotene ($C_{40}H_{56}$), which is approximately 536.9 g/mol [1].

The conversion formula is,

$$\begin{aligned} \text{ng/ml} &= 10^{-9} \times 10^{3}\text{g/l} = \mu\text{g/l} \\ &= \mu\text{mol/l} \times \text{Molar mass (g/mol)}. \end{aligned}$$

Substituing the values,

$$0.42\mu mol/l \approx 0.42 \times 536.9\text{ng/l} \approx 225ng/l$$

And we calculated the quartile for the dataset, the result is around 230, which is close to 225, this implies that we get a correct result.

1.b Frequencies and proportions of "low" and "high" plasma β-carotene

We create a new variable, lowplasma_01, which is 1 if the plasma β-carotene concentration (betaplasma) is below the cut-off value, and 0 otherwise. See the frequency table (Table 1), which shows the number and proportion (in percentage) of observations with high and low plasma β-carotene concentrations.

Table 1: Number and percentage of observations with high and low plasma β-carotene concentrations

	Frequency	Percentage (%)
High	80	25.4
Low	235	74.6

1.c The relationship between low plasma β-carotene concentration and vitamin use

We examine the relationship between low plasma β-carotene concentration and vitamin use.

We calculate the probabilities and the corresponding odds of having low plasma β-carotene concentration (% Low) for the three vitamin use categories, shown in the Table 2. In addition, the table also includes the ratio odds for each with "Often" as the reference category because it is the largest category.

Table 2: Plasma β-carotene concentration: frequency, probability of low levels, odds, and odds ratios by vitamin use category

Vitamin use	High	Low	% Low	Odds	OR
Often	46	76	0.623	1.65	1.00
Notoften	21	61	0.744	2.90	1.76
No	13	98	0.883	7.54	4.56

1.d Logistic model for low plasma β-carotene concentration and vitamin use

We fit a logistic model, (Model 1.d), for lowplasma_01 with vituse as the explanatory variable.

$$\log \text{odds}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} = \begin{cases} \beta_0 & \text{Often} \\ \beta_0 + \beta_1 & \text{Not often} \\ \beta_0 + \beta_2 & \text{No} \end{cases}$$

Table 3: Regression coefficients and odds ratios with 95% confidence intervals

Variable	Parameter	Estimate (β_i)	95% CI for β_i	Estimate ($\exp(\beta_i)$)	95% CI for $\exp(\beta)$
(Intercept)	β_0	0.50	(0.14, 0.87)	1.65	(1.15, 2.40)
Not often	β_1	0.56	(-0.04, 1.19)	1.76	(0.96, 3.30)
No	β_2	1.52	(0.86, 2.24)	4.56	(2.36, 9.36)

The odds, odds ratios (OR), and the probabilities of having low plasma β-carotene concentration of each vitamin use category in Table 2 are derived by the exponential of parameter estimates e^{β_i} in Table 3 as follows,

$$\begin{aligned} \text{odds}_i &= e^{\beta_0} \cdot e^{\beta_i}, \\ \text{OR} &= e^{\beta_i}, \\ p_i &= \frac{e^{\beta_0+\beta_i}}{1+e^{\beta_0+\beta_i}} \end{aligned}$$

, for $i = 1, 2$.

Using the regression model, we calculate the linear predictor (Logit), the odds, and the probability of having a low plasma β-carotene concentration, along with their respective 95% confidence intervals, for each of the three vitamin use categories.

Table 4: Predicted logit values, odds, and probabilities (with 95% confidence intervals) for low plasma β-carotene for vitamin use categories

Vitamin Use	Logit	95% CI (Logit)	Odds	95% CI (Odds)	% Low	95% CI (% Low)
Often	0.50	(0.14, 0.87)	1.65	(1.15, 2.38)	0.623	(0.534, 0.704)
Not Often	1.07	(0.57, 1.56)	2.90	(1.77, 4.77)	0.744	(0.639, 0.827)
No	2.02	(1.44, 2.60)	7.54	(4.23, 13.44)	0.883	(0.809, 0.931)

We get similar results of probabilities in Table 4 as compared to those in Table 2.

1.e Parameters, odds, probabilities and CI

We now want to test if there are any significant differences between different categories of vitamin use; in other words, we need to do a global test. So we do a LR-test in Model 1.d against the null model with the following hypotheses,

$$\begin{aligned} &H_0 : \beta_1 = \beta_2 = 0 \\ &H_1 : \exists i = 1, 2 \ \ \textit{such that} \ \ \beta_i \neq 0 \end{aligned}$$

The test statistic, and it's asymptotic distribution when H_0 is true,

$$D_0 - D \sim \chi^2_\alpha(2).$$

The observed test statistic is 21.83785 and the p-value is $1.811218 \cdot 10^{-5}$. Since the p-value is substantially less than $\alpha = 0.05$, we reject the null hypothesis and conclude that there are significant differences in the odds (or the probability of having a low plasma β-carotene concentration) between any of the vitamin use categories.

2 Low plasma β-carotene and BMI

We will now examine the relationship between low plasma β-carotene and BMI using a simple logistic regression.

2.a Low plasma β-carotene vs BMI and odd ratio

We plot lowplasma_01 against bmi and overlay it with a moving average line.

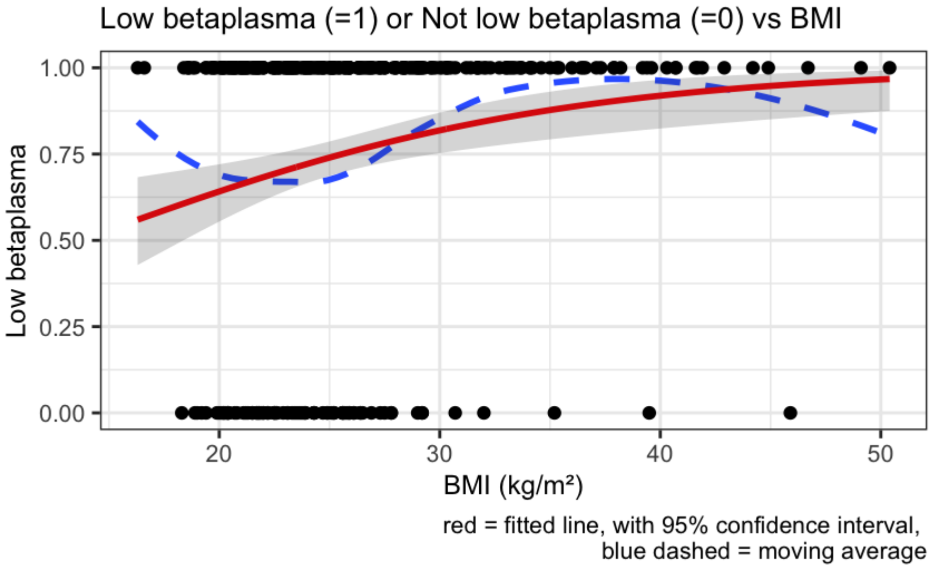

Figure 1: Plasma b-carotene by BMI

The probability of having low plasma β-carotene concentration increases as BMI rises from 20kg/m^2 to 35kg/m^2 (refer to Figure 1, indicated by the blue dashed moving average line). However, for BMI that is below 20kg/m^2 and above 35kg/m^2, the probability seems to decrease. Additionally, this supports our findings from Project 1, which showed that a lower plasma β-carotene level corresponded to a higher BMI.

We now fit a logistic regression for lowplasma_01 as the response variable and bmi as the explanatory variable (Model 2.a).

Table 5: β-estimates, the e^{β}-estimates and their profile likelihood based 95% confidence intervals

Variable	Parameter	Estimate (β_i)	95% CI for β_i	Estimate ($\exp(\beta_i)$)	95% CI for $\exp(\beta)$
(Intercept)	β_0	-1.27	(-2.71, 0.06)	0.28	(0.07, 1.07)
bmi	β_1	0.09	[0.04, 0.15]	1.10	[1.04, 1.16]

Some interpretations from Figure 1:

- As BMI rises to 35kg/m^2, the predicted probability of having low plasma β-carotene increases steadily.
- For BMI above 35kg/m^2, the probability seems to have a flatter increasing trend.
- The blue dashed line (moving average) follows a similar trend to the predicted probability, but has a higher variability at lower and higher BMI.

Let x =0bmi, then our model is

$$\text{log odds} = \beta_0 + \beta_1 x \iff \text{odds} = e^{\beta_0+\beta_1 x} \tag{2.a}$$

Let new bmi $= t_0 + \Delta t$, then we have the odds ratio:

$$\text{odd ratio} = \text{OR} = \frac{e^{\beta_0+\beta_1(x+\Delta x)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1 \Delta x}$$

We get $\hat{\beta_1} = 0.09246156$,

Table 6: Odds ratio (OR), with 95% confidence interval

Δx	OR	2.5%	97.5%
1	1.0969	1.0411	1.1631
-1	0.9117	0.9605	0.8598
-10	0.3967	0.6682	0.2208

2.b Test

We use both a Wald test and an LR-test to determine if the probability of having low plasma β-carotene changes with BMI.

We have the hypotheses,

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

For Wald test for β_1 (when the number of observations n is very large), we have the test statistic, which is asymptotic normal when the null hypothesis H$_0$,

$$Z = \frac{\hat{\beta_1} - 0}{d(\beta_1)} \sim N(0, 1)$$

.

The observed value of the test statistic is 3.282548 and the p-value is 0.001028734. Since the p-value is less than the significance level 0.05, we can reject the null hypothesis and conclude that BMI has a significant effect on the probability of success.

For LR-test, the test statistic, and its asymptotic distribution when H_0 is true,

$$D_0 - D \sim \chi^2_\alpha(1).$$

The observed test statistic is 13.19563 and the p-value is 0.0002806021. Since the p-value is substantially less than $\alpha = 0.05$, we reject the null hypothesis and conclude that BMI has a significant effect on the probability of having low plasma β-carotene.

We see that both tests lead to the same conclusion that BMI is statistically significant and the p-value of LR test is smaller than Wald test's p-value. Since $n = 315$ being medium size dataset, LR-test is preferred to use in this case.

2.c Leverage

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.

The leverage values for Model 2.a and those from a simple linear regression with BMI as covariate are plotted in Figure 2.

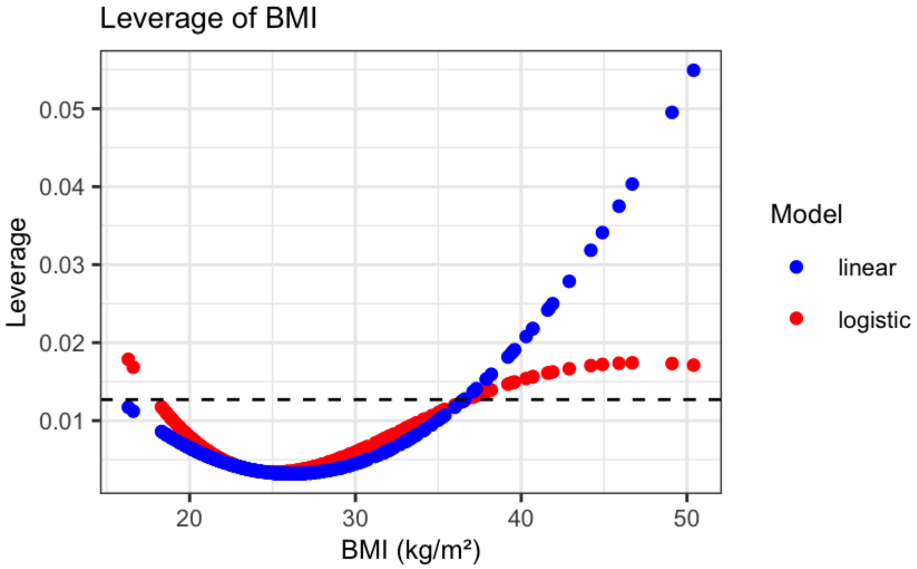

Figure 2: Leverage

For both linear regression model (blue dots in Figure 2) and logistic regression model (red dot), the leverage curves have a U-shape, which is lowest when BMI is approximately 25kg/m^2. However, the leverage value for linear regression increases significantly as BMI increases beyond approximately 35kg/m^2. In contrast, the leverage for logistic regression shows a slight increase up to a BMI of 45, after which it tends to decline. This is because the leverage for logistic now depend on X (0bmi) and Y (0lowplasma_01) and, as such, are no longer indicators of outliers with respect to X. Therefore, it is less sensitive to extremes of 0bmi.

2.d Standardized deviance residuals

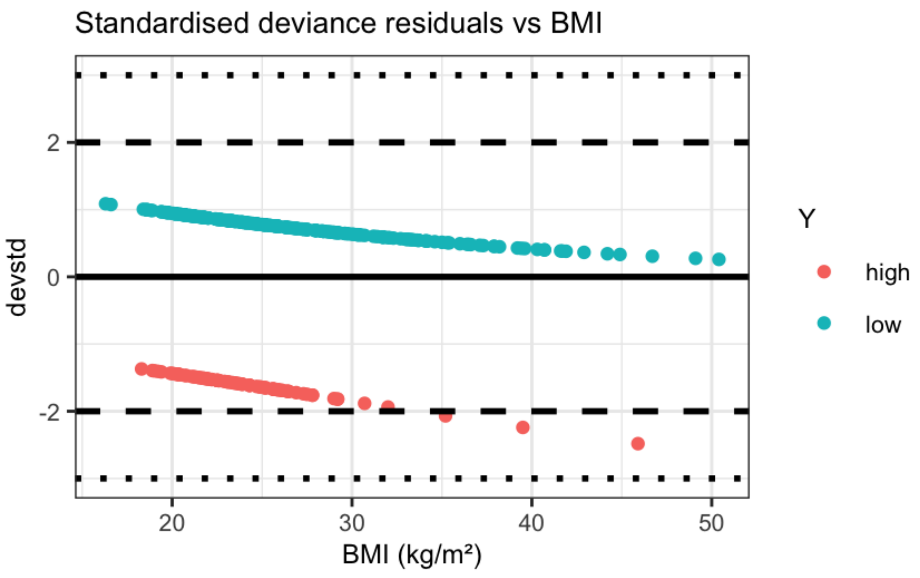

Figure 3: Standardized deviance residuals

From Figure 3 that shows the standardized deviance residuals for model 2.a against BMI, for observations with low plasma β-carotene concentrations, the residuals generally decrease as BMI increases. Conversely, for other observations with high plasma β-carotene concentrations, the residuals tend to increase with increasing BMI.

There are three observations with residuals outside ±2. These correspond to observations with high concentrations of β-carotene and high BMI values. They stand out as outliers because they go against the general trend of high BMI being associated with low plasma β-carotene concentration (see Figure 1).

2.e Cook's distance and observations with high influence

Cook's Distance for Model 2.a is shown in figure 4.

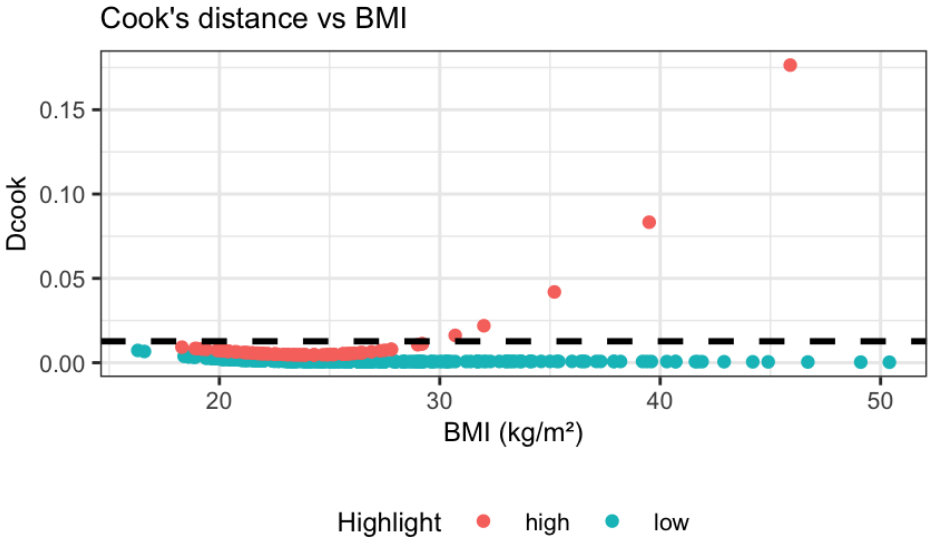

Figure 4: Cook's distance vs. BMI

And the observation that has the highest Cook's Distance is

$$\text{Age} = 43, \quad \text{Sex} = \text{female}, \quad \text{BMI} = 45.9, \quad \text{betaplasma} = 241, \quad \text{lowplasma\_01} = 0,$$

$$\text{phat} = 0.9515330 \quad \text{stddevresid} = \text{-2.482072}, \quad \text{D} = 0.17647045.$$

The observation that has the second highest Cook's Distance is

$$\text{Age} = 49, \quad \text{Sex} = \text{female}, \quad \text{BMI} = 39.5, \quad \text{betaplasma} = 249, \quad \text{lowplasma\_01} = 0,$$

$$\text{phat} = 0.9157100 \quad \text{stddevresid} = \text{-2.240923}, \quad \text{D} = 0.08332085.$$

This limited amount of data in the high BMI, high plasma value makes the model less stable in that region. These two observations have large residuals, indicating a poor fit relative to the model's prediction. As seen in Figure 1, there are relatively few observations with very high BMI values (especially when BMI is greater than 40). In particular, we only have one observation at bmi = 39.5 and bmi = 45.9 with high plasma β-carotene concentrations. The majority of observations with BMI greater than 40 have low plasma β-carotene concentration.

Therefore, these individual observations have more influence on the model fit, leading to higher Cook's Distance.

3 Multiple logistic regression and model selection

3.a Full model with all variables

We fit a multiple model using all 11 x-variables, including calories. Then we calculate the values of VIF (variance inflation factor) or GIVF (generalized variance 3 inflation factor) for each variable (Table 7).

Table 7: VIF/GVIF of full model

Variable	GVIF	DF	GVIF$^{(1/(2*DF))}$
bmi	1.074668	1	1.036662
age	1.261831	1	1.123312
calories	10.623520	1	3.259374
fat	7.269588	1	2.696217
cholesterol	1.828183	1	1.352103
fiber	2.576775	1	1.605234
alcohol	1.323134	1	1.150276
betadiet	1.311511	1	1.145212
smokstat	1.104874	2	1.025246
sex	1.274001	1	1.128717
vituse	1.132064	2	1.031496

Despite the change in model from linear regression in Project 1.3 (c) to logistic regression, the GVIF values remain very similar. This is because GVIF measures multicollinearity among the x variables, which is independent of the type of regression model used.

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. In all further analyses, we exclude the variable calories.

3.b Variable selection

We now fit a full model with all the x-variables (except calories), and their two-way interactions (Full Model):

$$\begin{aligned}\text{lowplasma\_01} \sim (&\text{bmi} + \text{age} + \text{fat} + \text{cholesterol} + \text{fiber}\\ &+ \text{alcohol} + \text{betadiet} + \text{smokstat} + \text{sex} + \text{vituse})^2\end{aligned} \tag{Full}$$

We use Bayesian Information Criterion (BIC) as the model selection criterion:

$$\text{BIC}(p+1) = \ln n \cdot (p+1) - 2\ln L(\hat{\beta}),$$

where n is the sample size, $p+1$ is the number of parameters, and $L(\hat{\beta})$ is the maximized log-likelihood. The "best" model is the one with THE smallest BIC.

Case 1. Backward elimination and stepwise regression, BIC as criterion

We perform Backward elimination, starting with the Full Model with BIC = 613.61. We get the Backward Model with BIC = 340.56:

$$\begin{aligned}\text{lowplasma\_01} \sim\ & \text{bmi} + \text{age} + \text{fat} + \text{cholesterol} + \text{fiber} + \text{betadiet} \\ & + \text{fat:betadiet} + \text{cholesterol:betadiet} + \text{fiber:betadiet}\end{aligned} \tag{Backward}$$

We then continue to perform stepwise regression, using the resulting model (Backward Model) as the starting point.

The final model (BackStep Model) with BIC = 339.4 is:

$$\begin{aligned}\text{lowplasma\_01} \sim\ & \text{bmi} + \text{age} + \text{fat} + \text{cholesterol} + \text{fiber} + \text{betadiet} \\ & + \text{fat:betadiet} + \text{cholesterol:betadiet} + \text{fiber:betadiet} + \text{bmi:betadiet}\end{aligned} \tag{BackStep}$$

Case 2. Forwards selection and Stepwise regression, BIC as criterion

The original model (Null Model) has BIC = 362.74:

$$\text{lowplasma\_01} \sim 1 \tag{Null}$$

We perform Forward selection, starting with the Null Model, we get the Forward Model, with BIC = 327.9 is as follows:

$$\text{lowplasma\_01} \sim \text{bmi} + \text{age} + \text{betadiet} + \text{vituse} + \text{bmi:betadiet} \tag{Forward}$$

We then continue to perform stepwise regression, using the resulting model (Forward Model) as the starting point, it results in the starting model. This is because neither adding nor removing any variables leads to a further improvement in the BIC.

Among the two stepwise regression models, the "Forward selection then Stepwise regression" model (which is the same as the Forward Model) includes the categorical variable vituse, which has three levels. To test whether simplifying this variable improves model fit, we will test whether the number of levels can be reduced from three to two using a likelihood-ratio test.

We perform a likelihood ratio test comparing the "full" Forward Model against a reduced model in which the categorical variable vituse is simplified from three levels to two.

Firstly, we assume that the "Not often" category of vituse does not have a significant effect on the probability of success and fit a reduced model for that, in which "yes" and "no" are two levels of vituse:

$$\text{log odd} = \beta_0 + \beta_1 \cdot \text{vituse}_{\text{yes}} + \beta_2 \cdot \text{vituse}_{\text{no}} + \beta_3 \cdot \text{bmi} + \beta_4 \cdot \text{age} + \beta_5 \cdot \text{betadiet} + \beta_6 \cdot \text{bmi:betadiet} \tag{Reduced}$$

Therefore, we want to test

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

, where β_j is the coefficient corresponding to the "Not often" category of vituse.

If H_0 is true, then, asymptotically

$$D_{\text{red}} - D_{\text{full}} \sim \chi^2(1)$$

The observed value of the test statistic, $D_{\text{red}} - D_{\text{full}} = 0.3499975$ and the p-value is 0.5541146. Since the p-value is greater than the significance level 0.05, we do not reject the null hypothesis and conclude that the category "Not often" of categorical variable vituse does not have a significant effect on the probability of success.

The following Table 8 presents the estimated β-parameters of the four models:

Table 8: Estimated β-parameters of Backward, BackStep, Forward and Reduced Models

	Variable	**Backward**	**BackStep**	**Forward**	**Reduced**
1	(Intercept)	0.8543	4.4788	4.6012	4.7398
2	bmi	0.0870	-0.0407	-0.0565	-0.0567
3	age	-0.0281	-0.0293	-0.0353	-0.0363
4	fat	-0.0554	-0.0569	–	–
5	cholesterol	0.0112	0.0109	–	–
6	fiber	0.1444	0.1303	–	–
7	betadiet	-0.2434	-2.0428	-2.1308	-2.1606
8	fat:betadiet	0.0280	0.0291	–	–
9	cholesterol:betadiet	-0.0045	-0.0044	–	–
10	fiber:betadiet	-0.0822	-0.0745	–	–
11	bmi:betadiet	–	0.0627	0.0723	0.0734
12	vitusenotoften	–	–	0.2083	–
13	vituseno	–	–	1.4599	–
14	vituse_newno	–	–	–	1.3901

One interesting fact is that the results are relatively similar among the four models but not the same. What do we know about these results is that they are all local minimums, but not necessarily global minimum.

The BackStep Model is derived by applying a stepwise regression procedure to the Backward Model, which results in the inclusion of only one additional (interaction) term: bmi:betadiet. This explains why the coefficient estimates remain similar across these two models, while the coefficients for bmi and betadiet both show notable differences.

3.c Model comparison

For model comparison, we use several metrics, including AIC, BIC and the adjusted McFadden's R^2. AIC and BIC were previously used in Project 1 to compare models based on their goodness of fit and complexity. In this project, we additionally introduce the adjusted McFadden's R^2, which is defined as:

$$\text{R2McF.adj} = R^2_{\text{McF,adj}} = 1 - \frac{\mathrm{ln}L(\beta) \hat{-} p/2}{\mathrm{ln}L(\hat{\beta}_0)} = 1 - \frac{D+p}{D_0},$$

where ln $L(\hat{\beta})$ is the log-likelihood of the fitted model, ln $L(\hat{\beta}_0)$ is the log-likelihood of the null model, and p is the number of parameters. The "best" model is the simplest one with a high $R^2_{\text{McF,adj}}$.

The following Table 9 presents the number of the McFadden's adjusted pseudo $R^2_{\text{McF,adj}}$, AIC, and BIC for the four models:

Table 9: Model Performance Metrics

Model	**Degree of freedom**	**AIC**	**BIC**	**R2McF.adj**
Backward	10	303.0346	340.5603	0.1820
BackStep	11	298.1199	339.3981	0.1985
Forward	7	301.6360	327.9040	0.1775
Reduced	6	299.9860	322.5014	0.1793

From Table 9, with different criterion function, the best model is:

- AIC: BackStep Model.
- BIC: Reduced Model.
- Fadden's adjusted pseudo R^2: BackStep Model.

3.d Residual analysis

We calculate the standardized deviance residuals for the model with the best AIC (BackStep Model) and the model with the best BIC (Reduced Model).

Firstly, we calculate the deviance residuals, d_i as follows:

$$d_i = \begin{cases} -\sqrt{2\ln\left(\frac{1}{1-\hat{p}_i}\right)} & \text{if } Y_i = 0 \\ +\sqrt{2\ln\left(\frac{1}{\hat{p}_i}\right)} & \text{if } Y_i = 1 \end{cases}$$

, where d_i is the deviance residual for observation i, Y_i is the obversed response for observation i, and $\hat{p}_i$ is the predicted probability of $Y_i = 1$ of the logistic model.

Then the deviance residual will be standardised as:

$$\frac{d_i}{\sqrt{1 - v_{ii}}} \sim N(0, 1)$$

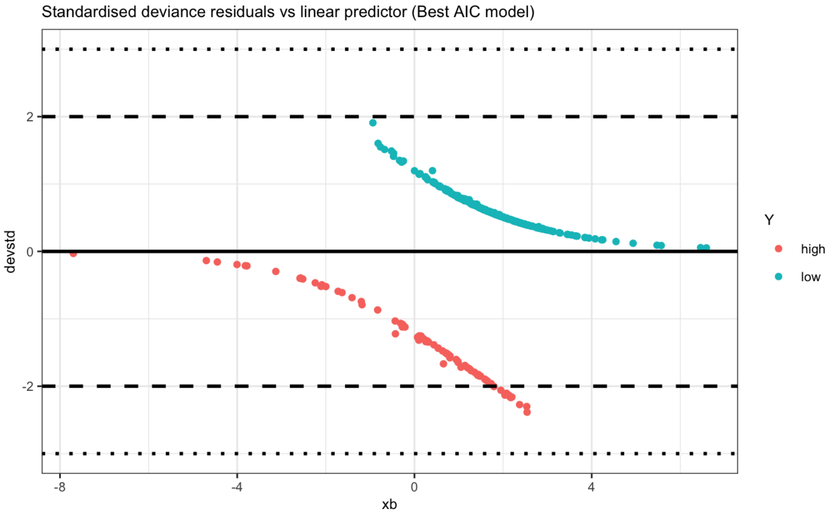

Figure 5: Standardised deviance residuals vs linear predictor (Best AIC model)

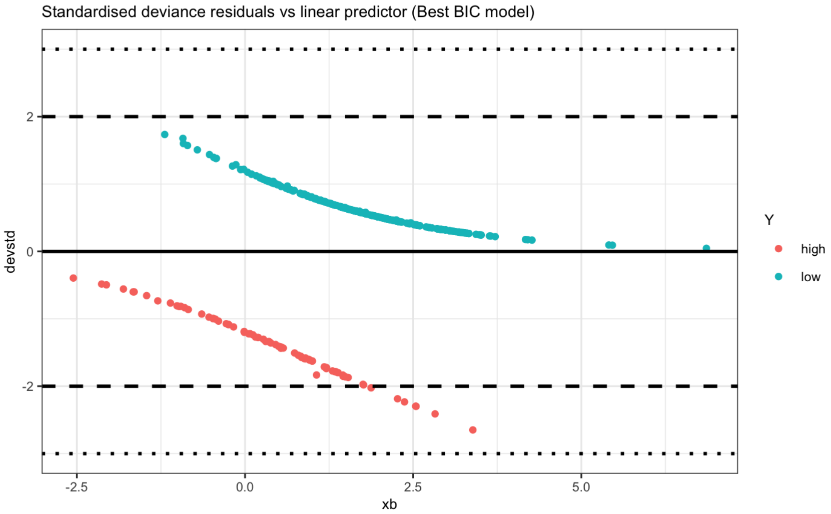

Figure 6: Standardised deviance residuals vs linear predictor (Best BIC model)

Figure 5 and Figure 6 show the plot of standardized deviance residuals against the linear predictor for the models selected by AIC and BIC, respectively. The reference lines at ±2 are also added to help identify unusually large standardized deviance residuals.

The BIC model (aka. Reduced Model) appears to have minimally better-behaved residuals. Using the reference lines at ±2, the BIC model has 7 standardized deviance residuals that exceed this threshold, while the AIC model has 11. This means that the BIC model provides a better fit, with fewer observations exhibiting unusually large residuals.

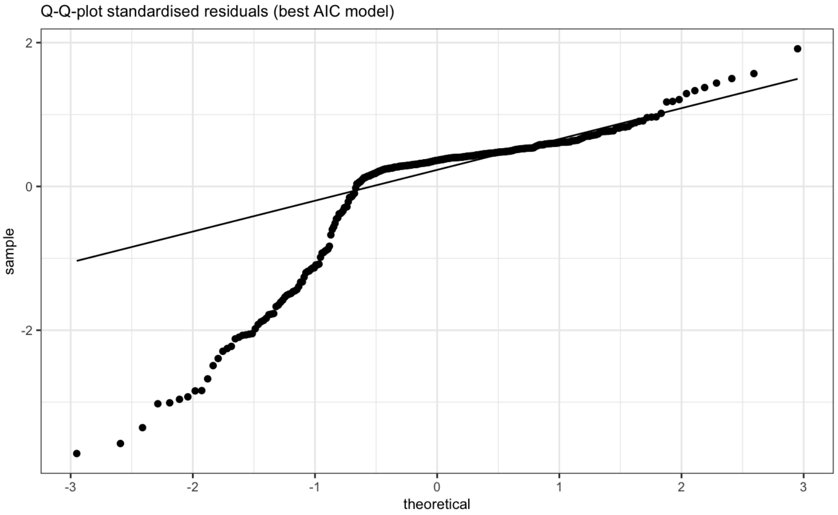

Figure 7: Q-Q-plot standardised residuals (best AIC model)

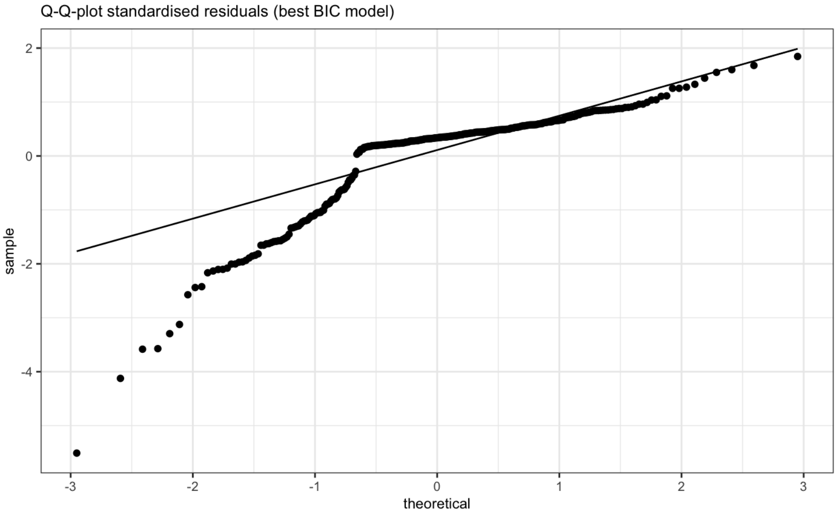

Figure 8: Q-Q-plot standardised residuals (best BIC model)

Both of the QQ-plot of the standardised residuals for the two models (see Figure 8 and Figure 7 show mild non-normality, especially in the lower tail (left side). However, there are more observations with residuals that deviate from the theoretical normal distribution in the AIC model (see the left tail in Figure 7). In contrast, the BIC model shows fewer overall deviations, but it has one particularly extreme residual (at approximately -5.7 on the sample axis), suggesting a significantly bad fit for a specific observation.

4 Goodness-of-fit

From Table 9, we know that the model with the least AIC is BackStep Model, while the model with least BIC is Reduced Model.

4.a Goodness-of-fit with threshold 0.5

We now make a prediction with a cut-off value of 0.5, then we have the confusion matrices for both models in Table 10 and Table 11.

Table 10: Confusion Matrices of BackStep Model with Threshold 0.5

	Reference	
Prediction	high	low
high	27	10
low	53	225

Table 11: Confusion Matrices of Reduced Model with Threshold 0.5

	Reference	
Prediction	high	low
high	26	13
low	54	222

And the goodness-of-fit is in Table 12. We can see from the table that in both models the accuracy is significantly higher than NIR, but Cohen's κ shows that the prediction is better than random classification but still far worse than the reference. In addition, McNemar's Test shows that the False Positives and False Negatives are significantly different in both models, and the sensitivities are far above specificities in both model, so we need to consider trading sensitivity for some specificity. Overall, the BackStep model is slightly better than the Reduced model because it has higher accuracy and Cohen's κ, but there are some same problems in both models.

Table 12: Model Comparison (Threshold = 0.5)

Model	BackStep	Reduced
Accuracy	0.8	0.7873
P-value[Acc>NIR]	0.01468	0.05081
Cohen's κ	0.3585	0.3245
P-value[McNemar's Test]	1.213e-07	1.025e-06
Sensitivity	0.9574	0.9447
Specificity	0.3375	0.3250

4.b ROC-curves and AUC-values

Let the threshold varies in [0,1], then we get the ROC-curves for both models in Figure 9.

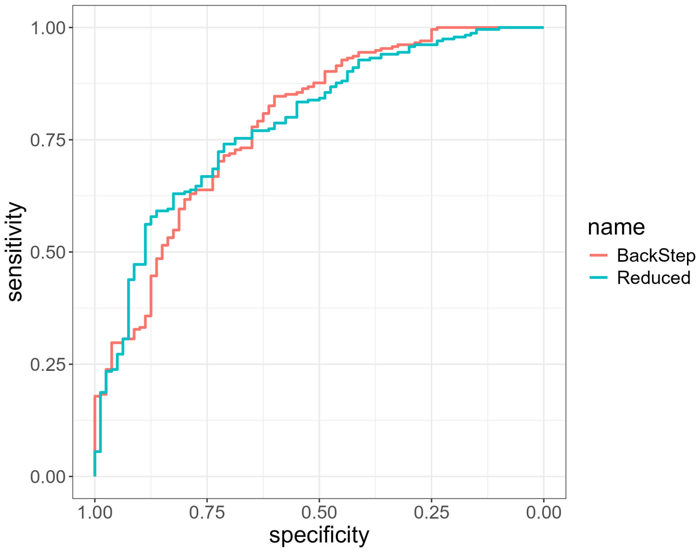

Figure 9: ROC Curves for BackStep and Reduced Models

With the threshold 0.5, the sensitivity is about 0.9, and we can see in that part, BackStep model is slightly better than the Reduced model. However, the Reduced model seems better than the BackStep model in the region which is close to the topleft, so we calculate the AUC-values and the confidence intervals in Table 13. And it seems the two models are quite close. So we perform a DeLong's Test for these two correlated ROC-curves. The result is p-value is 0.9714, which means there is no significant difference between the two AUC-values.

Table 13: AUC Values with 95% Confidence Intervals

Model	AUC	Lower CI	Upper CI
BackStep	0.7874468	0.7285045	0.8463892
Reduced	0.7882979	0.7316104	0.8449853

4.c Goodness-of-fit with optimal threshold

Now we want to trade sensitivity for some specificity, and reach the closest point to the topleft in the ROC-curves and then make a comparison.

Table 14: Optimal Threshold Values

Model	Optimal Threshold	Distance to Topleft
BackStep	0.779499	0.4048974
Reduced	0.7340273	0.3873437

From Table 14, we can see with optimal threshold, the Reduced model is closer to the topleft. And the confusion matrices are presented in Table 15 and 16.

Table 15: Confusion Matrices of BackStep Model with Optimal Threshold

	Reference	
Prediction	high	low
high	57	67
low	23	168

Table 16: Confusion Matrices of Reduced Model with Optimal Threshold

	Reference	
Prediction	high	low
high	57	61
low	23	174

And the goodness-of-fit is presented in the Table 17. We can see the accuracy is reduced in both model, and it is not significantly higher than the NIR anymore. In addition, the Cohen's κ increased a little and there is still a significant difference between False Positive and False Negative. In comparison, the Reduced model is better than the BackStep since it has higher accuracy and Cohen's κ which is contrary to the conclusion of Table 12.

Table 17: Model Comparison (Optimal Threshold)

Model	BackStep	Reduced
Accuracy	0.7143	0.7333
P-value[Acc>NIR]	0.9116	0.7222
Cohen's κ	0.3618	0.3916
P-value[McNemar's Test]	5.826e-06	5.413e-05
Sensitivity	0.7149	0.7404
Specificity	0.7125	0.7125

4.d Conclusion

Taking all the results into account, we think Reduced model is the "best" model between the two models. Although with threshold 0.5, the BackStep model has higher accuracy and Cohen's κ, the Reduced model performs better than the BackStep with optimal threshold. From Figure 9, we can see that in the region that is close to the topleft, the Reduced model is locally better than the BackStep model, and this is the case with more application where there is a balance of sensitivity and specificity.

In conclusion, the best model is as follows,

$$lowplasma\_01 \sim betadiet + bmi + vituseNo + age + betadiet : bmi.$$

The e^{β}-estimates and their corresponding 95% confidence intervals are presented in Table 18.

Table 18: Best Model: Reduced Model

Model	e^{β}-estimates	95% CI
(Intercept)	114.4160	(5.6275, 2366.4457)
Betadiet	0.1153	(0.0303, 0.3766)
BMI	0.9449	(0.8498, 1.0568)
VituseNo	4.0154	(2.0485, 8.4163)
Age	0.9644	(0.9452, 0.9832)
Betadiet:BMI	1.0762	(1.0263, 1.1360)

The variables betadiet and age have significant negative effect on the probability of Low plasma β-carotene, while vituseNo and betadiet:bmi have significant positive effect on the probability of low plasma β-carotene. Betadiet is diet consuming for plasma β-carotene, the more you take, the less possible you have low plasma β-carotene. As for variable age, the elders may have different dietary preferences. They may tend to consume more vegetables and fruits which are the biggest source of plasma β-carotene. And vituseNo means no specific vitamin consuming, and some kind of vitamin can be transformed to plasma β-carotene in the body. So if you don't consume vitamin, you have higher probability to have low plasma β-carotene. As for the interaction betadiet:bmi, this seems counterintuitive, but there might be some metabolic problems caused by obesity.

5 AI use clarification

We used AI for the possible reason for the effect of variables on the probability of having a low plasma β-carotene concentration.

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

6 Author Contributions

We complete the code individually and then discuss the results.

- **Nancy Truong** wrote Part 1, Part 2 and Part 3 of the report.

- **Xinxu Yao** Wrote the rest parts of the report and is responsible for checking and modifying.

References

[1] National Center for Biotechnology Information. *PubChem Compound Summary for CID 5280489, Beta-Carotene.* 2025. URL: `https://pubchem.ncbi.nlm.nih.gov/compound/Beta-Carotene` (visited on 04/17/2025).