Predicting Beta-Carotene Concentration: Ordinal vs. Multinomial Logistic Regression

Nancy Truong, Xinxu Yao

2025

Project Overview

- Goal: Model the probability of beta-carotene concentration categories.
- Identify the best-fitting model for accurate prediction.
- Methods: Ordinal Logistic Regression and Multinomial Logistic Regression.

Beta-Carotene Categories

- Continous beta-carotene measurements were categorized into:
  - Deficient: < 50 mg/ml
  - Low: 50–225 mg/ml
  - High: > 225 mg/ml
- These thresholds were chosen based on nutritional guidelines.

Deficient	Low	High
23	212	80

Table: Sample table for 3 categories

Ordinal Logistic Regression: Modeling Ordered Categories

- Ordered categories: a logical order between the categories (Deficient $<$ Low $<$ High).
- Model predicts cumulative probabilities of being in a category or below.
- Coefficients reflect effect on odds of being in a category or below.
- Covariates: bmi, age, calories, fat, cholesterol, fiber, alcohol, betadiet, smokstat, sex, vituse.

Model Selection for Ordinal Logistic Regression

- Selection tools: Backward elimination, Forward selection, and Stepwise regression with criterion AIC/BIC, LRT.
- Due to sample size limitations, a full interaction model was not feasible.
- A linear model was used to identify interactions significantly affecting beta-carotene levels (betaplasma).
- Interactions not significantly influencing the continuous response were assumed unlikely to impact the ordinal categories.

Model Selection for Ordinal Logistic Regression

- **Model inter**:
  plasma_123 $\sim$ bmi + age + fat + cholesterol + fiber + alcohol + betadiet + smokstat + sex + vituse + age:betadiet + fiber:vituse + alcohol:vituse + smokstat:vituse
- Backward elimination + stepwise regression (BIC).
  Model interstep: plasma_123 $\sim$ bmi + age + cholesterol + betadiet + vituse
- vituse is a categorical variable.
  - vituse_new = "yes" or "no".
  - LRT: p-value = 0.6616888
- **Model interstep_red**: plasma_123 $\sim$ bmi + age + cholesterol + betadiet + vituse_new

Model Selection for Ordinal Logistic Regression

- Another model?:
  polr(betaplasma_123 $\sim$ all predictors)

Variable	**GVIF**	**Df**	**GVIF$^{1/(2\cdot \mathbf{Df})}$**
bmi	1.065938	1	1.032443
age	1.269467	1	1.126706
calories	12.765374	1	3.572866
fat	7.985909	1	2.825935
cholesterol	2.097929	1	1.448423
fiber	2.369830	1	1.539425
alcohol	2.401768	1	1.549764
betadiet	1.290810	1	1.136138
smokstat	1.152737	2	1.036174
sex	1.272761	1	1.128167
vituse	1.133150	2	1.031744

Exclude calories $\rightarrow$ **Model excl**

$\rightarrow$ backward elimination + stepwise regression (BIC)

$\rightarrow$ **Model inter**.

Models from Project 2

- Best AIC Model
  plasma_123 $\sim$ bmi + age + fat + cholesterol + fiber + betadiet+ fat:betadiet + cholesterol:betadiet + fiber:betadiet + bmi:betadiet
- Best BIC Model
  plasma_123 $\sim$ betadiet + bmi + vituse_new + age + betadiet:bmi
- Stepwise Regression:
  - **Model p2aic**: plasma_123 $\sim$ bmi + fat + cholesterol + fiber + betadiet + fat:betadiet + cholesterol:betadiet + fiber:betadiet
  - **Model p2bic**: plasma_123 $\sim$ betadiet + bmi + vituse_new + age

Model Comparison Table

Table: Ordinal Model Comparison

Model	df	AIC	BIC	R^2D	Adj. R^2D
model.excl	13	460.069	508.852	0.145	0.123
model.inter	23	461.896	548.205	0.181	0.169
model.interstep	8	459.583	489.604	0.126	0.114
model.interstep_red	7	457.775	**484.043**	0.126	0.114
model.p2aic	12	**457.179**	502.210	0.147	0.127
model.p2bic	7	461.239	487.507	-446.239	0.109

Model Comparison Table

Table: Best AIC vs Best BIC

Model	**model.p2aic** (best AIC)	**model.interstep_red** (best BIC)
Accuracy	0.7238	0.7079
P-value[Acc>NIR]	0.03	0.1028
Cohen's κ	0.2606	0.2165
P-value[McNemar's Test]	4.074e-12	7.161e-12

Model Performance Metrics

Table: Goodness of fit for model.p2aic

Metric	Deficient	Low	High
Sensitivity	0.00000	0.9575	0.31250
Specificity	1.00000	0.2524	0.95745
Pos Pred Value	NaN	0.7250	0.71429
Neg Pred Value	0.92698	0.7429	0.80357
Prevalence	0.07302	0.6730	0.25397
Detection Rate	0.00000	0.6444	0.07937
Detection Prevalence	0.00000	0.8889	0.11111
Balanced Accuracy	0.50000	0.6050	0.63497

Table: Goodness of fit for model.interstep_red

Metric	Deficient	Low	High
Sensitivity	0.043478	0.9481	0.26250
Specificity	1.000000	0.2233	0.94894
Pos Pred Value	1.000000	0.7153	0.63636
Neg Pred Value	0.929936	0.6765	0.79078
Prevalence	0.073016	0.6730	0.25397
Detection Rate	0.003175	0.6381	0.06667
Detection Prevalence	0.003175	0.8921	0.10476
Balanced Accuracy	0.521739	0.5857	0.60572

Multinomial Logistic Regression: Modeling Unordered Categories

- Assumption: No logical order among categories.
- Model estimates probability of each category independently.
- Coefficients compare each category to the reference. And reference is the largest category "Low".

Model Selection for Multinomial Logistic Regression

- **Model inter**: plasma_123 $\sim$ bmi + age + fat + cholesterol + fiber + alcohol + betadiet + smokstat + sex + vituse + age:betadiet + fiber:vituse + alcohol:vituse + smokstat:vituse
- **Model backstep**: plasma_123 $\sim$ bmi + age + cholesterol + alcohol + betadiet + smokstat + vituse + alcohol:vituse + smokstat:vituse
- **Model null**: plasma_123 $\sim$ 1
- **Model forstep**: plasma_123 $\sim$ vituse + betadiet + bmi + age + cholesterol + smokstat

Model Comparison Table

Table: Multinomial Model Comparison

Model	**df**	**AIC**	**BIC**	**R^2D**	**Adj. R^2D**
model.inter	44	476.378	641.491	0.235	0.231
model.null	2	511.571	519.076	0.000	-0.004
model.backstep	32	459.367	579.449	0.221	0.217
model.forstep	26	458.695	556.262	0.199	0.195
model.p2aic	22	**451.973**	534.529	0.196	0.192
model.p2bic	14	459.857	**512.393**	0.149	0.145
model.p2aic$_{red}$	22	**451.973**	534.529	0.196	0.192

Model Comparison Table

Table: Best AIC vs Best BIC

Model	**model.p2aic**(best AIC)	**model.p2bic**(best BIC)
Accuracy	0.7397	0.7206
P-value[Acc $>$ NIR]	0.0062	0.0394
Cohen's κ	0.3244	0.2778
P-value[McNemar's Test]	1.621e-10	1.464e-09

Model Performance Metrics

Table: Goodness of fit for model.p2bic

Metric	Low	Deficient	High
Sensitivity	0.9528	0.086957	0.36250
Specificity	0.3204	1.000000	0.94894
Pos Pred Value	0.7426	1.000000	0.70732
Neg Pred Value	0.7674	0.932907	0.81387
Prevalence	0.6730	0.073016	0.25397
Detection Rate	0.6413	0.006349	0.09206
Detection Prevalence	0.8635	0.006349	0.13016
Balanced Accuracy	0.6366	0.543478	0.65572

Table: Goodness of fit for model.p2bic

Metric	Low	Deficient	High
Sensitivity	0.9340	0.00000	0.36250
Specificity	0.3010	1.00000	0.93191
Pos Pred Value	0.7333	NaN	0.64444
Neg Pred Value	0.6889	0.92698	0.81111
Prevalence	0.6730	0.07302	0.25397
Detection Rate	0.6286	0.00000	0.09206
Detection Prevalence	0.8571	0.00000	0.14286
Balanced Accuracy	0.6175	0.50000	0.64721

Comparing Ordinal vs. Multinomial Logistic Regression

Table: Parameter Estimates(Ordinal)

Parameter	**Estimate**	**2.5%**	**97.5%**
bmi	-0.017	-0.098	0.063
age	0.021	0.003	0.039
fat	0.043	0.018	0.068
cholesterol	-0.011	-0.017	-0.005
fiber	-0.044	-0.154	0.060
betadiet	0.874	-0.198	1.963
fat:betadiet	-0.022	-0.034	-0.012
cholesterol:betadiet	0.004	0.002	0.007
fiber:betadiet	0.046	0.008	0.089
bmi:betadiet	-0.023	-0.053	0.008
Intercept(Low/Deficient)	-1.873	-2.202	-1.544
Intercept(High/Low)	-0.023	-0.053	0.008

Comparing Ordinal vs. Multinomial Logistic Regression

Table: Parameter Estimates(multinomial) Deficient

Parameter	**Estimate**	**2.5%**	**97.5%**
intercept	-5.956	-6.350	-5.563
bmi	0.110	0.037	0.183
age	0.018	-0.012	0.049
fat	0.009	-0.033	0.052
cholesterol	0.008	-0.002	0.017
fiber	-0.293	-0.496	-0.089
betadiet	0.875	-0.298	2.047
fat:betadiet	0.000	-0.0197	0.0205
cholesterol:betadiet	-0.003	-0.008	0.002
fiber:betadiet	0.072	0.001	0.142
bmi:betadiet	-0.034	-0.070	0.001

Comparing Ordinal vs. Multinomial Logistic Regression

Table: Parameter Estimates(multinomial) High

Parameter	**Estimate**	**2.5%**	**97.5%**
intercept	-4.589	-4.940	-4.238
bmi	0.052	-0.008	0.111
age	0.031	0.011	0.050
fat	0.058	0.026	0.090
cholesterol	-0.010	-0.019	-0.001
fiber	-0.170	-0.305	-0.035
betadiet	2.054	1.155	2.954
fat:betadiet	-0.029	-0.043	-0.015
cholesterol:betadiet	0.004	0.000	0.007
fiber:betadiet	0.089	0.031	0.147
bmi:betadiet	-0.066	-0.100	-0.032

Comparing Ordinal vs. Multinomial Logistic Regression

Table: Ordinal Model vs Multinomial Model

Model	**model.p2aic**(ordinal)	**model.p2aic**(multinomial)
Accuracy	0.7238	0.7397
P-value[Acc $>$ NIR]	0.03	0.0062
Cohen's κ	0.2606	0.3244
P-value[McNemar's Test]	4.074e-12	1.621e-10
df	**12**	**22**

Comparing Ordinal vs. Multinomial Logistic Regression

- Key Differences:
  - Ordinal model assumes category order.
  - Multinomial model treats categories as unrelated.
- Proportional Odds Assumption:
  - The central assumption of ordinal logistic regression is that the effect of the explanatory variable on different categories of the response variable is proportional. That is, the coefficients of the explanatory variables are the same in all cumulative probabilities.

Conclusion and Recommendations

- Summary: Both models work depending on the data.
- Recommendation: Use ordinal regression when there is logical order in the categories of response variable and you want a smaller model.
- Interesting question: Can we find an optimal threshold to satisfy the proportional odds assumption?