



# Evaluating the Main Predictors of Human Beta-Carotene Levels

Luyang Zhang and Christopher Li



# Background

- Prior literature has shown a negative relationship between beta-carotene levels and risk of cancer
- Studying the determinants of beta-carotene levels can inform avenues for cancer prevention
- Past studies have shown several variables to be good predictors (carotene intake, gender, etc.)
- This study was designed to evaluate the effect of multiple predictors on both beta-carotene and retinol
- Our analysis focuses on beta-carotene: studies have shown this to have a stronger association with cancer - more important to study predictors of beta-carotene
- Goal will be to derive a model that is good at prediction and accurately quantifies relationships

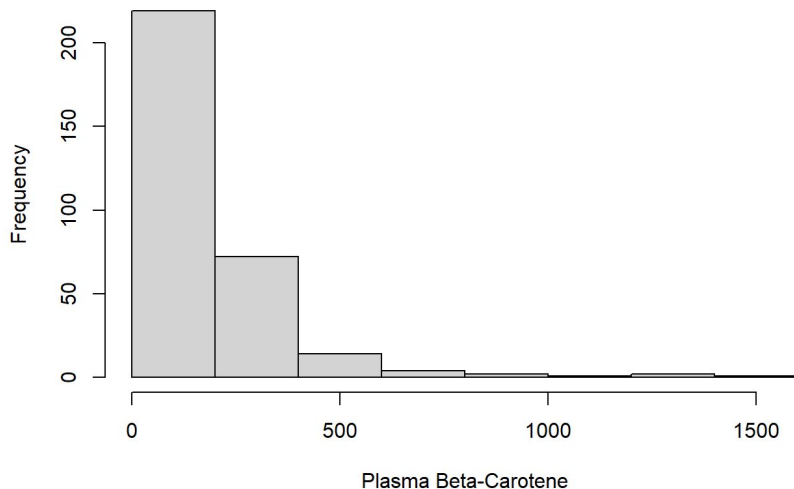
# Dataset

- 315 patients who had an elective surgical procedure during a three-year period to biopsy or remove a lesion of the lung, colon, breast, skin, ovary or uterus that was found to be non-cancerous.
  - Important that these patients did not have cancer: so beta-carotene levels are more representative of population
- Variables to note:
  - Quetelet - weight / height<sup>2</sup> (body mass index) - approximate measure of body fat
  - Beta-carotene levels (ng/ml): removed one observation where value was zero (measurement error)
  - Beta-carotene intake (mcg per day): approximate measure, based on frequency of consumption of certain foods
  - Three levels of smoking status, three levels of vitamin use (shown later)
- Mainly females, people of all ages (19 - 83)

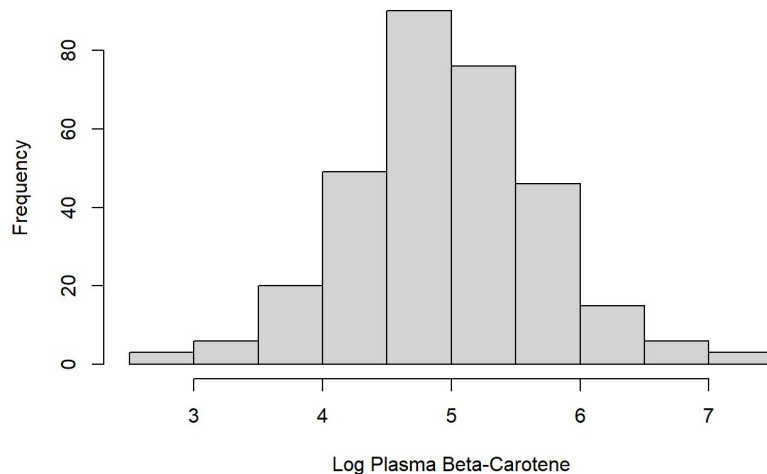
# Exploratory Data Analysis: Response Variable

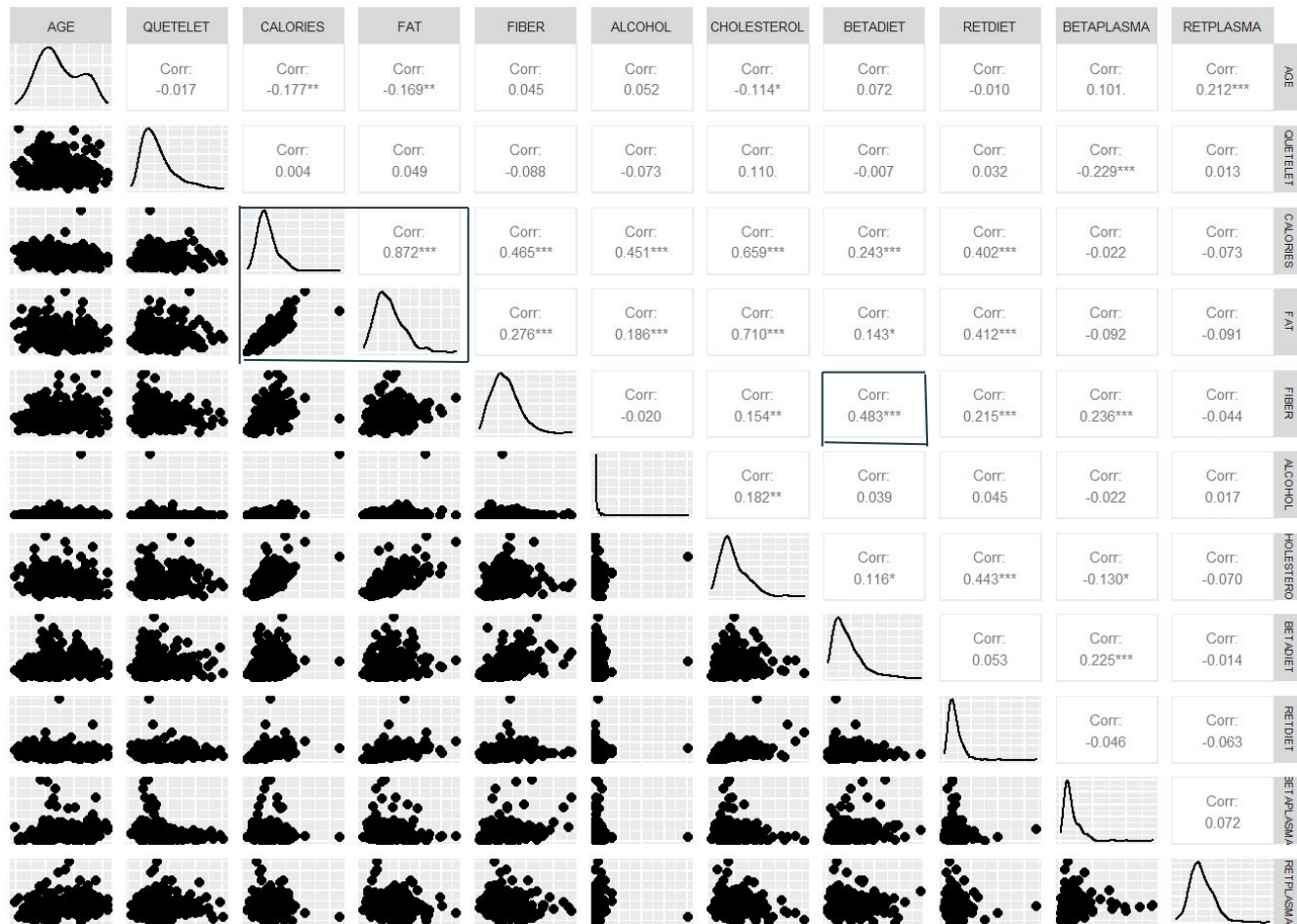
- Response variable is very right-skewed, so log transformation makes sense for normal error model

Histogram of Plasma Beta-Carotene

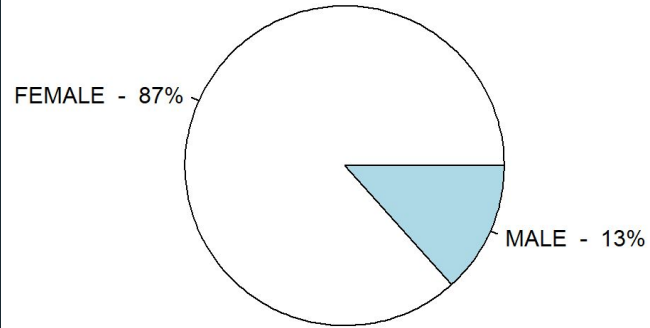


Histogram of Log Plasma Beta-Carotene

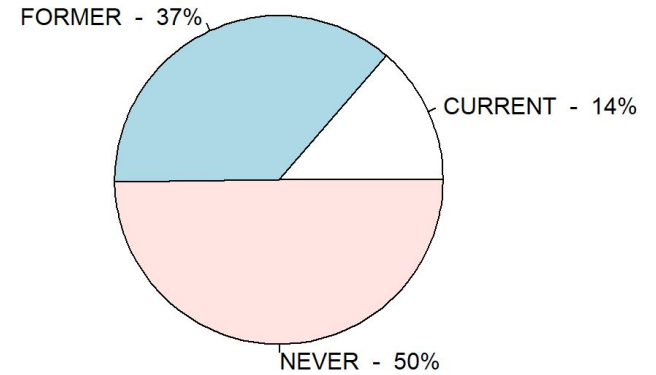




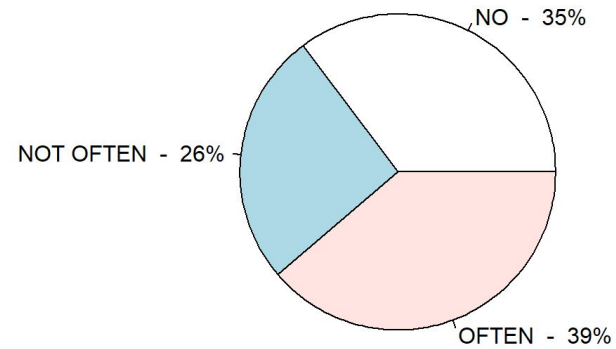
**SEX: pie chart**

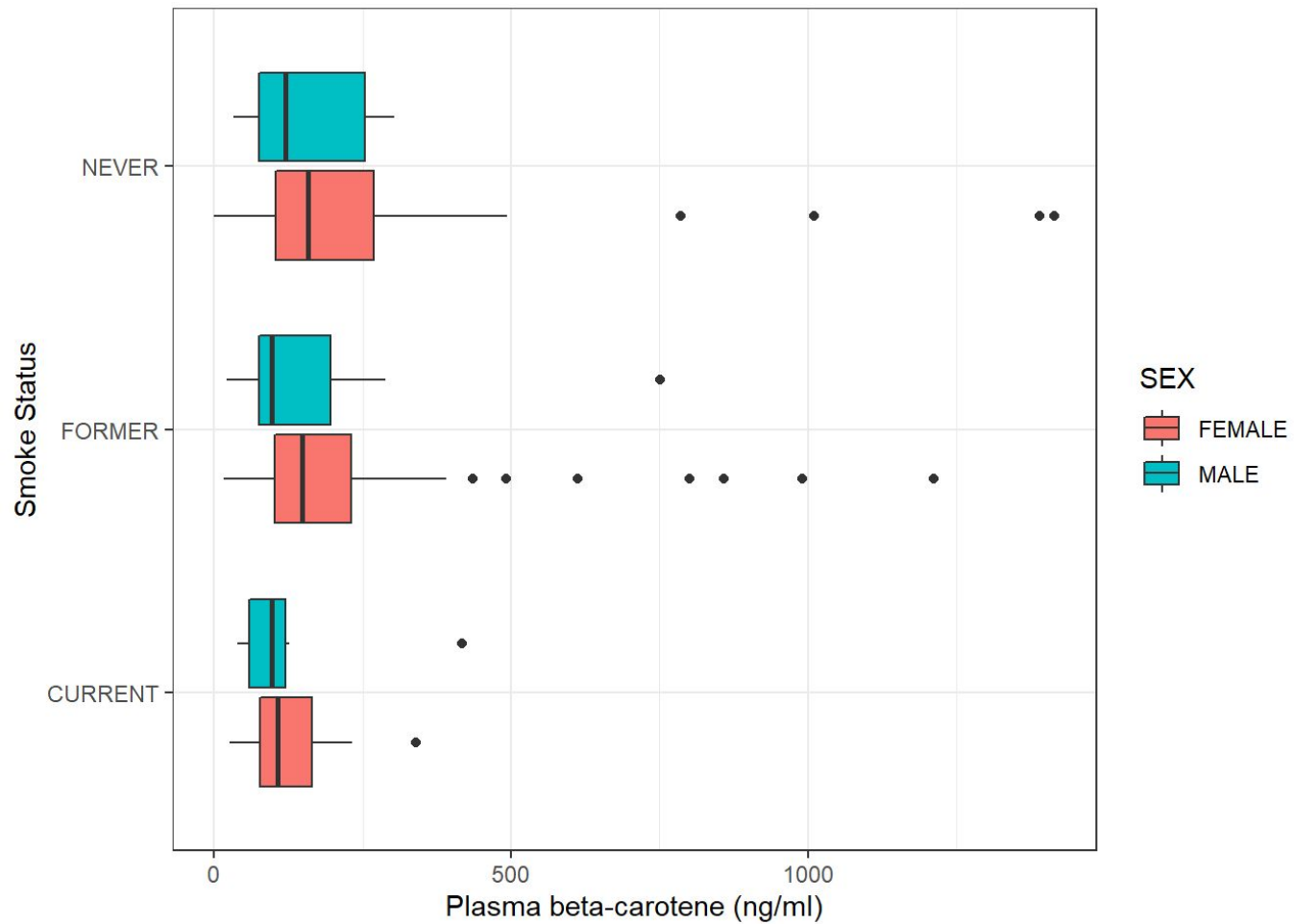


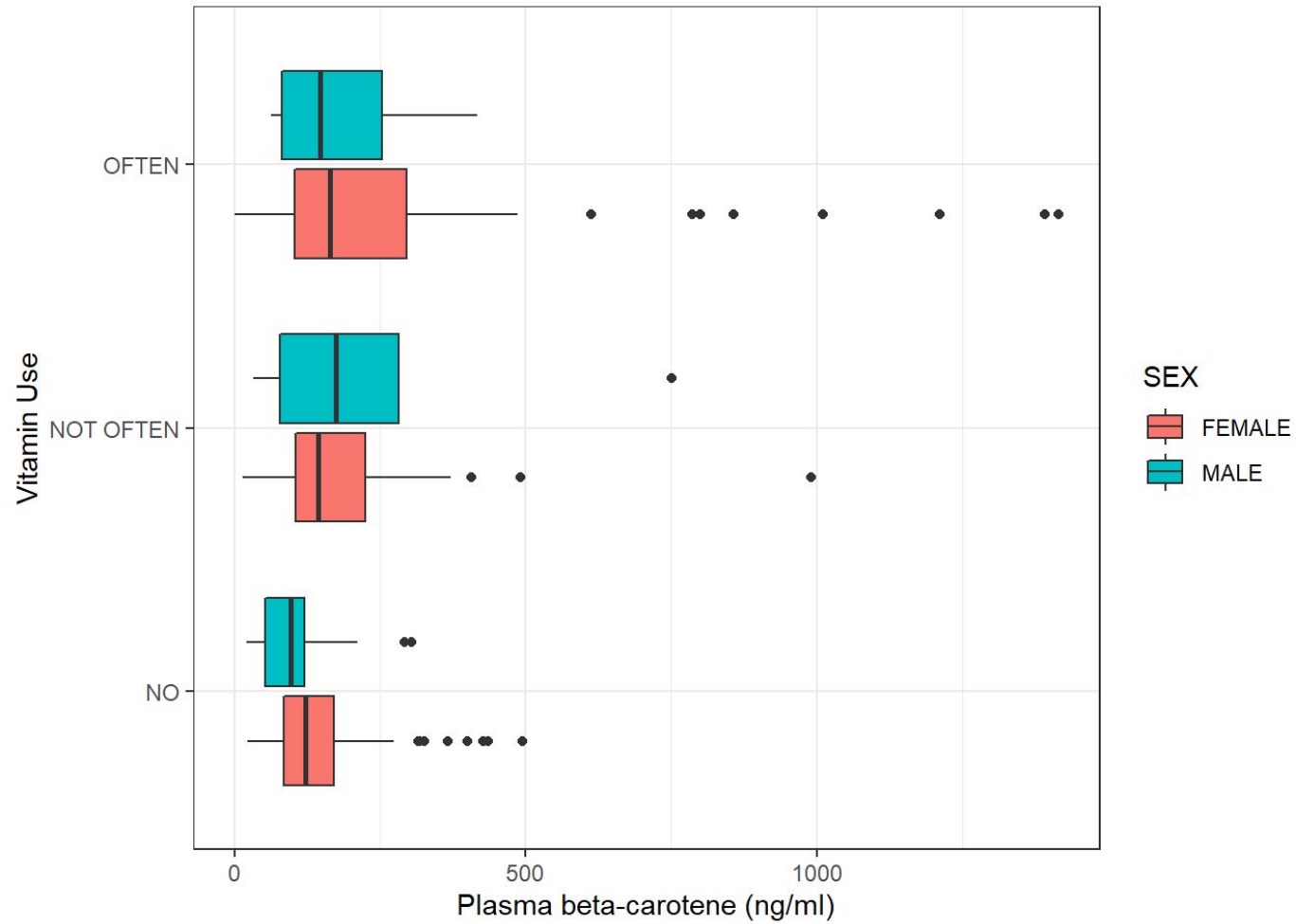
**SMOKSTAT: pie chart**



**VITUSE: pie chart**



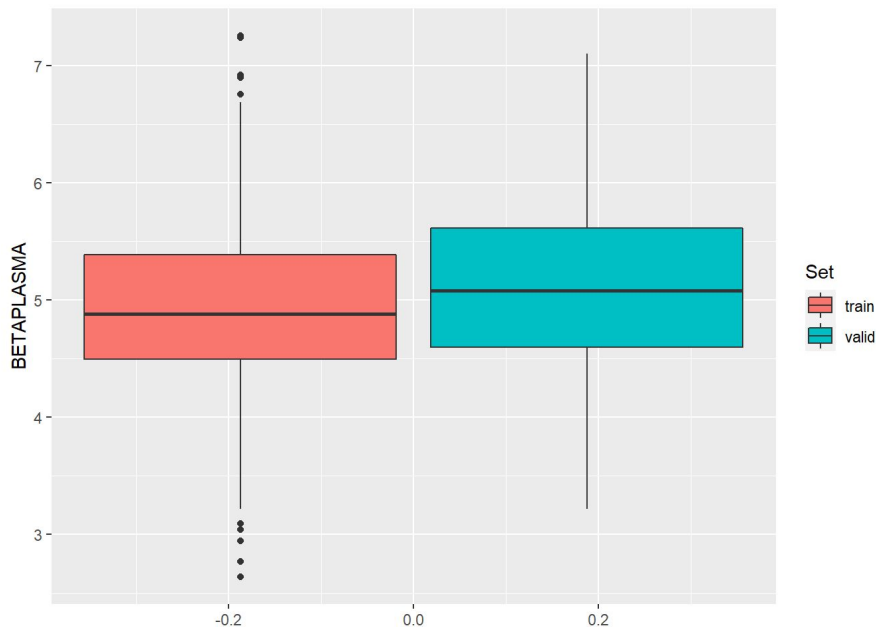






# Splitting Data into Training and Validation

- Split dataset 80/20
- Training dataset 250
- ~11 possible explanatory variables (14 possible coefficients)
- Many more possible interactions
- Since we have a smaller sample, we choose to use a “larger” training set
- Confirm that training and validation are similarly distributed



# Model Selection Procedure

- Remove retinol variables from consideration
- Stepwise Regression - additive model
- Consider adding/removing additional variables based on literature and exploratory analysis
- Stepwise Regression - interaction model
- Decide if these interactions and/or other interactions should be added
- Land on final model and execute model diagnostics.

# Stepwise Regression: Additive Model

- Choose to use AIC as criterion
- Variables consistent with literature
- Others worth considering for model

```
BETAPLASMA ~ QUETELET + FIBER + CALORIES + VITUSE + FAT + BETADIET
```

	Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
1				249	141.4577	-140.3651
2	+ QUETELET	1	11.2686428	248	130.1891	-159.1183
3	+ FIBER	1	7.8677107	247	122.3213	-172.7023
4	+ CALORIES	1	5.6565390	246	116.6648	-182.5390
5	+ VITUSE	2	5.0691770	244	111.5956	-189.6448
6	+ FAT	1	1.4723217	243	110.1233	-190.9651
7	+ BETADIET	1	0.9336995	242	109.1896	-191.0938

# Alter Stepwise Model Based on Other Evidence

- Keep beta-carotene intake, quetelet, vitamin use: literature
- Also keep calories and fiber, very significant
- Stepwise not necessarily best model
  - May only be local optimum, and makes decisions based on very marginal differences in AIC
- Add gender and smoking: based on literature, exploratory analysis, and best subset selection
- Remove fat, high collinearity with calories, stepwise chose calories first
- Preferred additive model: initial results

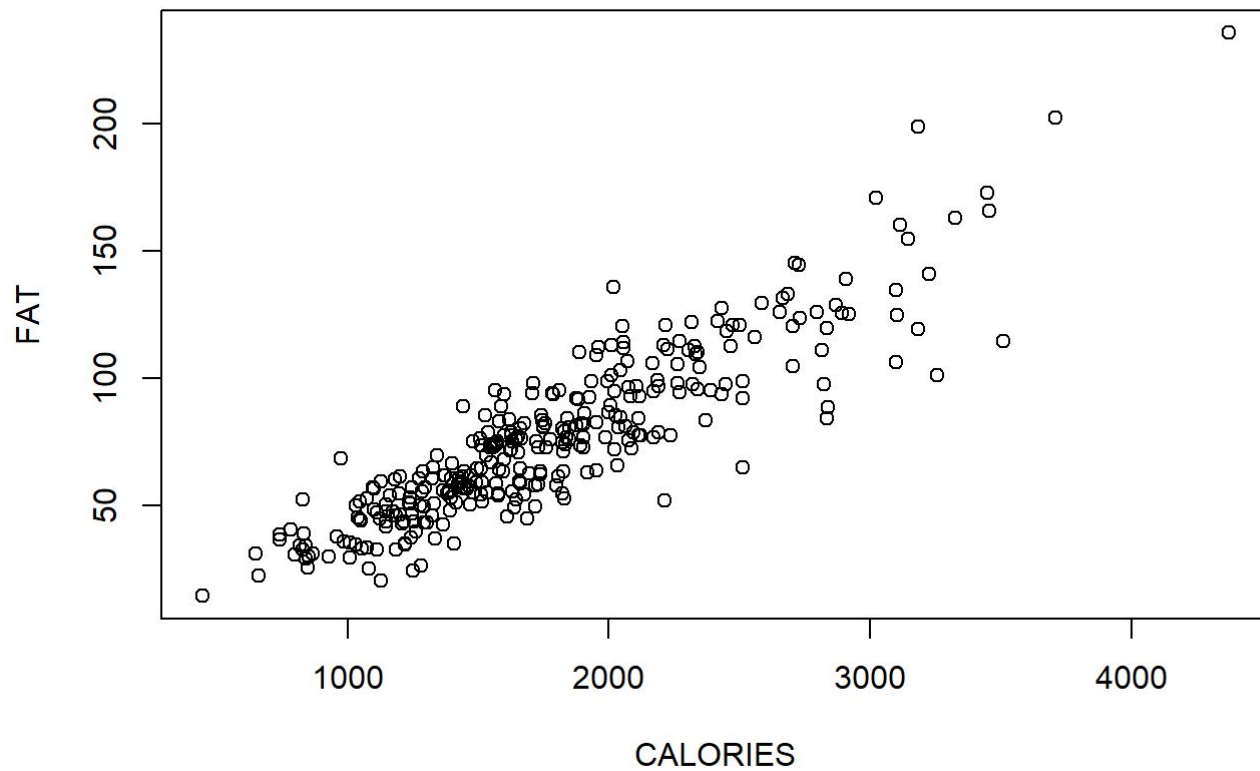
## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	5.301e+00	2.499e-01	21.209	< 2e-16	***
QUETELET	-3.061e-02	6.989e-03	-4.380	1.78e-05	***
FIBER	3.793e-02	1.054e-02	3.599	0.000388	***
CALORIES	-2.599e-04	8.783e-05	-2.959	0.003393	**
VITUSENOT OFTEN	2.269e-01	1.126e-01	2.015	0.045054	*
VITUSEOFTEN	2.797e-01	1.030e-01	2.716	0.007080	**
BETADIET	4.764e-05	3.413e-05	1.396	0.164038	
SEXMALE	-1.536e-01	1.443e-01	-1.064	0.288211	
SMOKSTATFORMER	1.723e-01	1.354e-01	1.272	0.204597	
SMOKSTATNEVER	2.113e-01	1.341e-01	1.576	0.116404	

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6734 on 240 degrees of freedom  
Multiple R-squared: 0.2306, Adjusted R-squared: 0.2017  
F-statistic: 7.992 on 9 and 240 DF, p-value: 2.458e-10

Fat vs. Calories: Correlation Coefficient = 0.897



# Stepwise Regression: Interaction Model

- Allow all two-way interactions
- Interaction between smoking status and vitamin use not justified by literature
- Interaction between sex and carotene intake could make sense, given past literature - propose adding to model
- Conduct further investigation on interaction between smoking status and fiber

Initial Model:

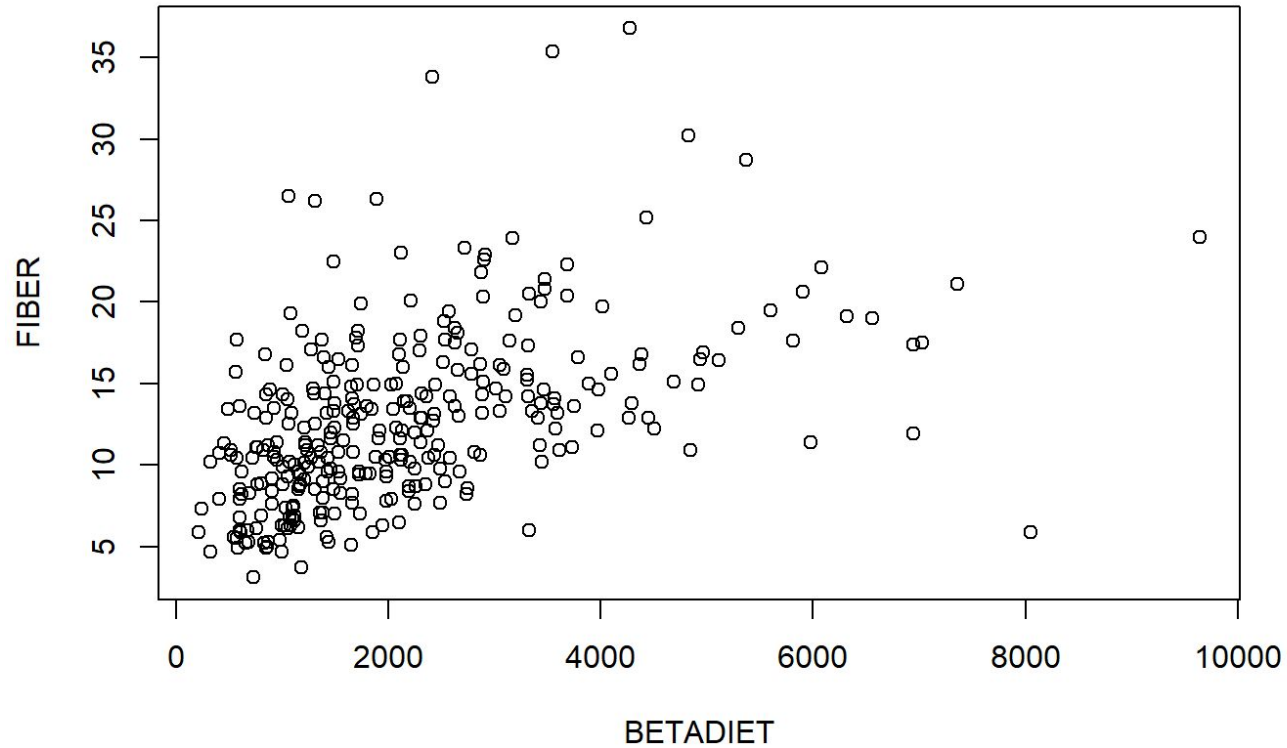
```
BETAPLASMA ~ QUETELET + FIBER + CALORIES + VITUSE + BETADIET +  
SEX + SMOKSTAT
```

Final Model:

```
BETAPLASMA ~ QUETELET + FIBER + CALORIES + VITUSE + BETADIET +  
SEX + SMOKSTAT + VITUSE:SMOKSTAT + BETADIET:SEX + FIBER:SMOKSTAT
```

	Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
1				240	108.83739	-187.9015
2	+ SMOKSTAT:VITUSE	4	6.003025	236	102.83436	-194.0853
3	+ SEX:BETADIET	1	2.752223	235	100.08214	-198.8674
4	+ SMOKSTAT:FIBER	2	2.281181	233	97.80096	-200.6316

Fiber vs. Beta-Carotene Intake: Correlation Coefficient = 0.485



# Final Model With Interactions

- Confirms findings from prior literature
  - Weight/height<sup>2</sup>
  - Vitamin use
- Confirms interaction between carotene intake and smoking status
- Suggests interaction effect between gender and carotene intake
- Calories: important for model
- Effect of beta-carotene intake confounded by fiber and interaction

```
lm(formula = BETAPLASMA ~ QUETELET + FIBER + CALORIES + VITUSE +  
    BETADIET + SEX + SMOKSTAT + BETADIET:SMOKSTAT + BETADIET:SEX,  
    data = train)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.08246	-0.36923	0.00345	0.37187	1.94988

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.473e+00	2.765e-01	19.791	< 2e-16 ***
QUETELET	-2.864e-02	6.914e-03	-4.143	4.77e-05 ***
FIBER	3.907e-02	1.043e-02	3.746	0.000226 ***
CALORIES	-2.634e-04	8.699e-05	-3.028	0.002733 **
VITUSENOT OFTEN	2.044e-01	1.110e-01	1.842	0.066666 .
VITUSEOFTEN	2.725e-01	1.014e-01	2.686	0.007737 **
BETADIET	-8.357e-05	9.113e-05	-0.917	0.360024
SEXMALE	5.309e-01	3.016e-01	1.760	0.079674 .
SMOKSTATFORMER	-8.318e-02	2.267e-01	-0.367	0.713946
SMOKSTATNEVER	-1.341e-01	2.192e-01	-0.612	0.541339
BETADIET:SMOKSTATFORMER	1.432e-04	1.009e-04	1.420	0.156983
BETADIET:SMOKSTATNEVER	1.847e-04	9.789e-05	1.887	0.060408 .
BETADIET:SEXMALE	-3.263e-04	1.295e-04	-2.520	0.012405 *
---				

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6621 on 237 degrees of freedom

Multiple R-squared: 0.2655, Adjusted R-squared: 0.2283

F-statistic: 7.138 on 12 and 237 DF, p-value: 4.339e-11



# Validation

## Internal Validation:

- Press P measures predictive ability of model using LOOCV
- Press P = **119.68**, SSE = **103.90**

## External Validation:

- Mean Squared Prediction Error measures predictive ability of model using hold out data (20% of initial dataset, 63 observations)
- MSPE = **0.445**, SSE (training data) / 250 = **0.416**

# Refit Model on Full Dataset

- Larger sample size decreases standard errors
- Magnitude and sign of coefficient estimates pretty similar to model fit on training data
- Some small differences
  - Vitamin use more significant
  - Gender less of an effect
  - Smoking status and carotene intake interaction more significant

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.564e+00	2.529e-01	22.003	< 2e-16 ***
QUETELET	-3.076e-02	6.339e-03	-4.852	1.96e-06 ***
FIBER	3.383e-02	9.306e-03	3.636	0.000326 ***
CALORIES	-2.322e-04	7.310e-05	-3.177	0.001645 **
VITUSENOT OFTEN	2.281e-01	9.792e-02	2.329	0.020514 *
VITUSEOFTEN	2.922e-01	8.924e-02	3.274	0.001185 **
BETADIET	-1.451e-04	8.365e-05	-1.734	0.083913 .
SEXMALE	1.518e-01	2.488e-01	0.610	0.542338
SMOKSTATFORMER	-1.402e-01	2.055e-01	-0.682	0.495571
SMOKSTATNEVER	-9.499e-02	2.016e-01	-0.471	0.637854
BETADIET:SMOKSTATFORMER	2.162e-04	9.033e-05	2.393	0.017325 *
BETADIET:SMOKSTATNEVER	2.343e-04	9.005e-05	2.602	0.009726 **
BETADIET:SEXMALE	-1.384e-04	1.006e-04	-1.377	0.169657

---

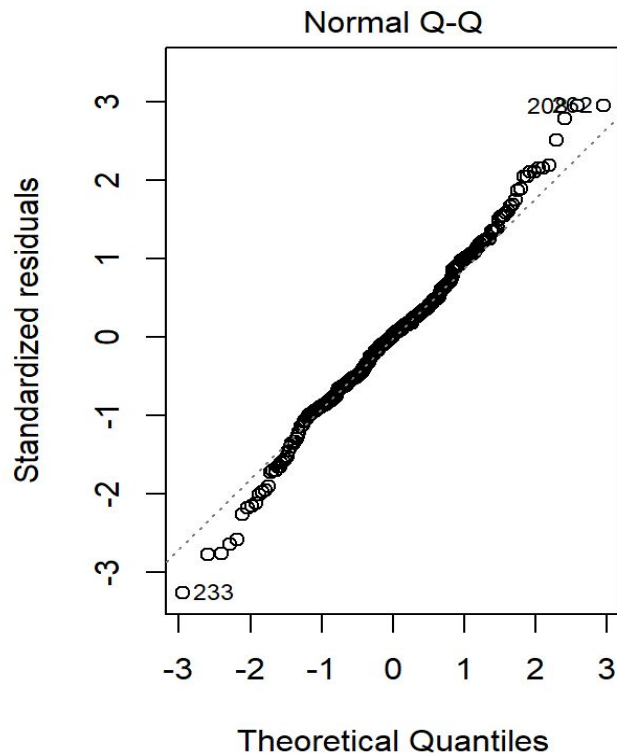
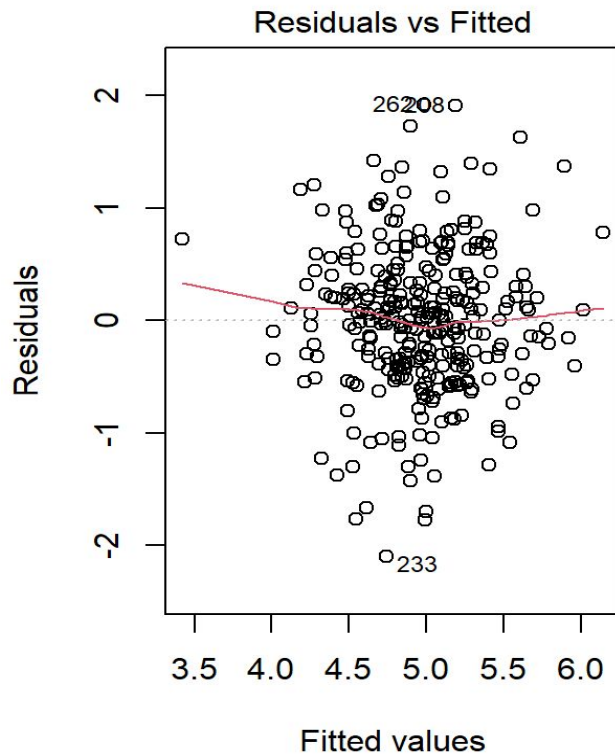
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6565 on 300 degrees of freedom

Multiple R-squared: 0.2604, Adjusted R-squared: 0.2308

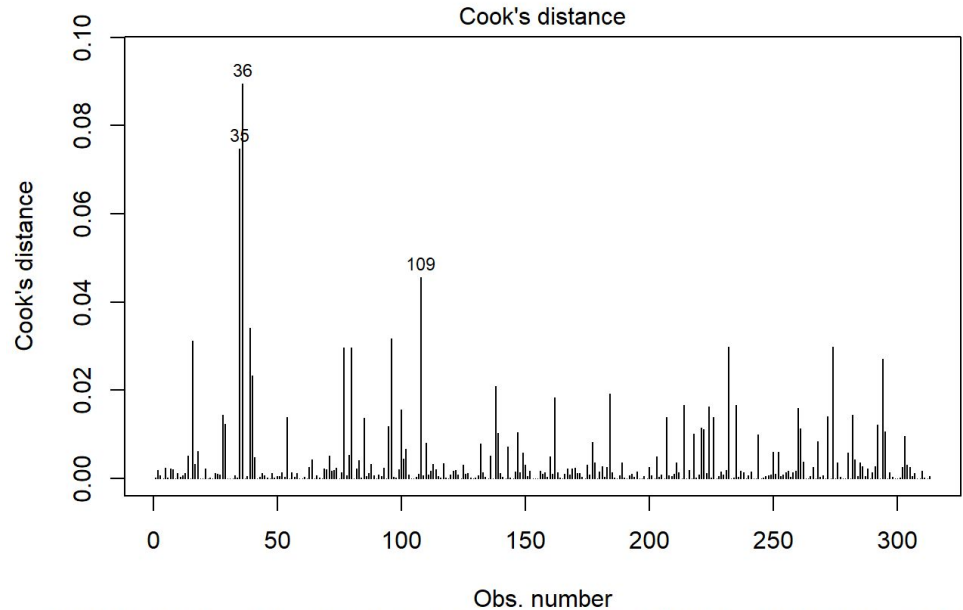
F-statistic: 8.803 on 12 and 300 DF, p-value: 2.052e-14

# Model Diagnostics



# Model Diagnostics: Cont'd

- Several influential cases, but most influential are obs 35 and 36
- 35 has very low carotene intake but high carotene levels
- 36 has very high carotene intake but low carotene levels



# Discussion

## Findings

- Negatively associated with BMI, smoking, and calorie intake
- Positively associated with taking vitamins, fiber intake, and carotene intake (assuming not smoking)

## Limitations

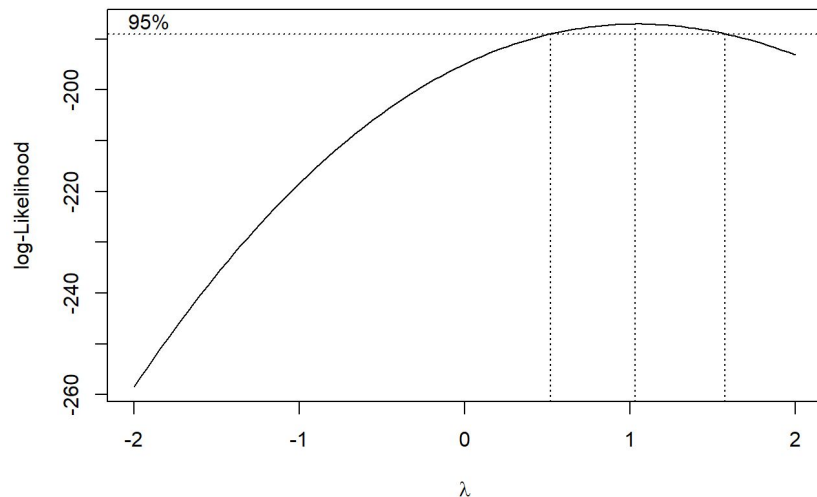
- Inaccuracy of beta carotene intake measurements - cause influential cases and may reduce accuracy of coefficients
- Unclear what vitamins are being taken
- Multicollinearity between fat and calories as well as between fiber and carotene intake
- Small sample size, so coefficient estimates may be sensitive to sampling variability
  - e.g. gender no longer statistically significant after running full model (compared to model ran on training data)

# Conclusion

- External validity seems decent, diverse age, BMI, and habits
- Results are consistent with literature
- This can inform action: what can patients do to reduce cancer risk
- This analysis cannot conclude causation, just association
- Predict carotene without directly measuring it (which could be more tedious/costly)
  - Instead, can predict beta-carotene levels using more accessible information (mass and height, daily calorie intake, etc.)
- Further areas of research
  - Scientific mechanism that some of these factors affect carotene levels
  - Smoking, calories vs. fat, fiber, vitamins
  - Larger studies

# Fit Preliminary Model

- All possible variables except for retinol levels and retinol intake, don't want to confound effects



Coefficients:

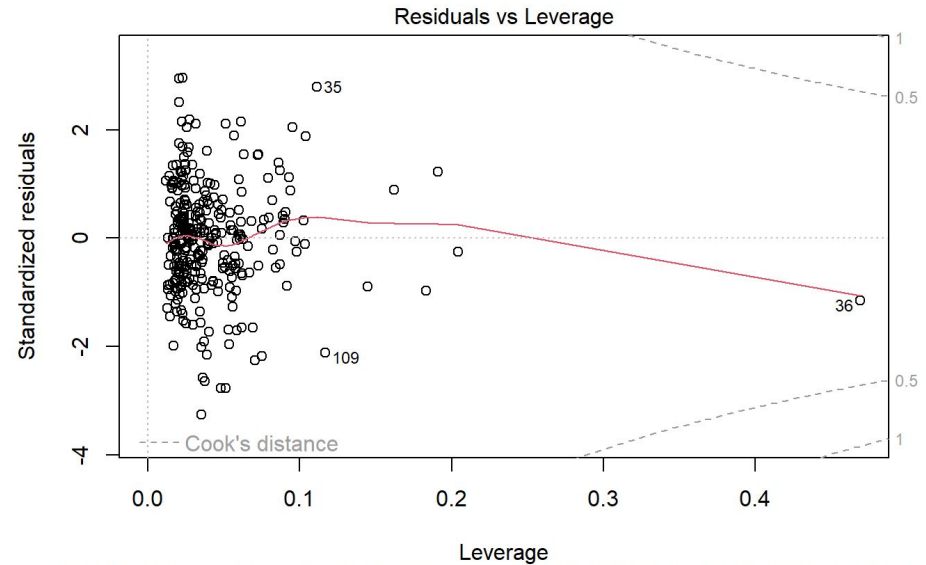
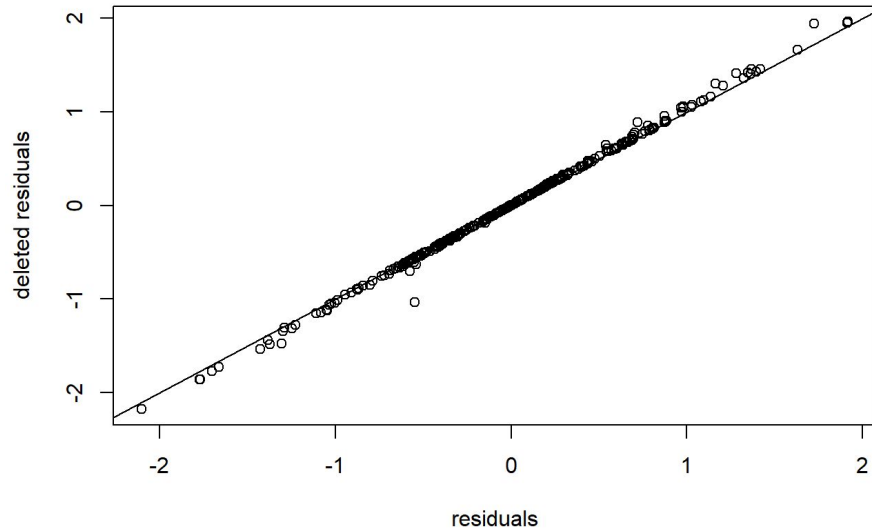
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.102e+00	3.012e-01	16.939	< 2e-16 ***
AGE	3.920e-03	3.341e-03	1.174	0.241772
SEXMALE	-1.889e-01	1.565e-01	-1.207	0.228522
SMOKSTATFORMER	1.522e-01	1.362e-01	1.117	0.264955
SMOKSTATNEVER	1.976e-01	1.345e-01	1.469	0.143111
QUETELET	-2.977e-02	7.134e-03	-4.173	4.23e-05 ***
VITUSENOT OFTEN	2.640e-01	1.133e-01	2.331	0.020610 *
VITUSEOFTEN	3.116e-01	1.056e-01	2.951	0.003488 **
CALORIES	-5.628e-04	2.520e-04	-2.233	0.026468 *
FAT	6.720e-03	3.811e-03	1.763	0.079181 .
FIBER	4.605e-02	1.316e-02	3.499	0.000558 ***
ALCOHOL	5.192e-03	9.955e-03	0.522	0.602470
CHOLESTEROL	-4.441e-04	4.979e-04	-0.892	0.373386
BETADIET	4.312e-05	3.449e-05	1.250	0.212416

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

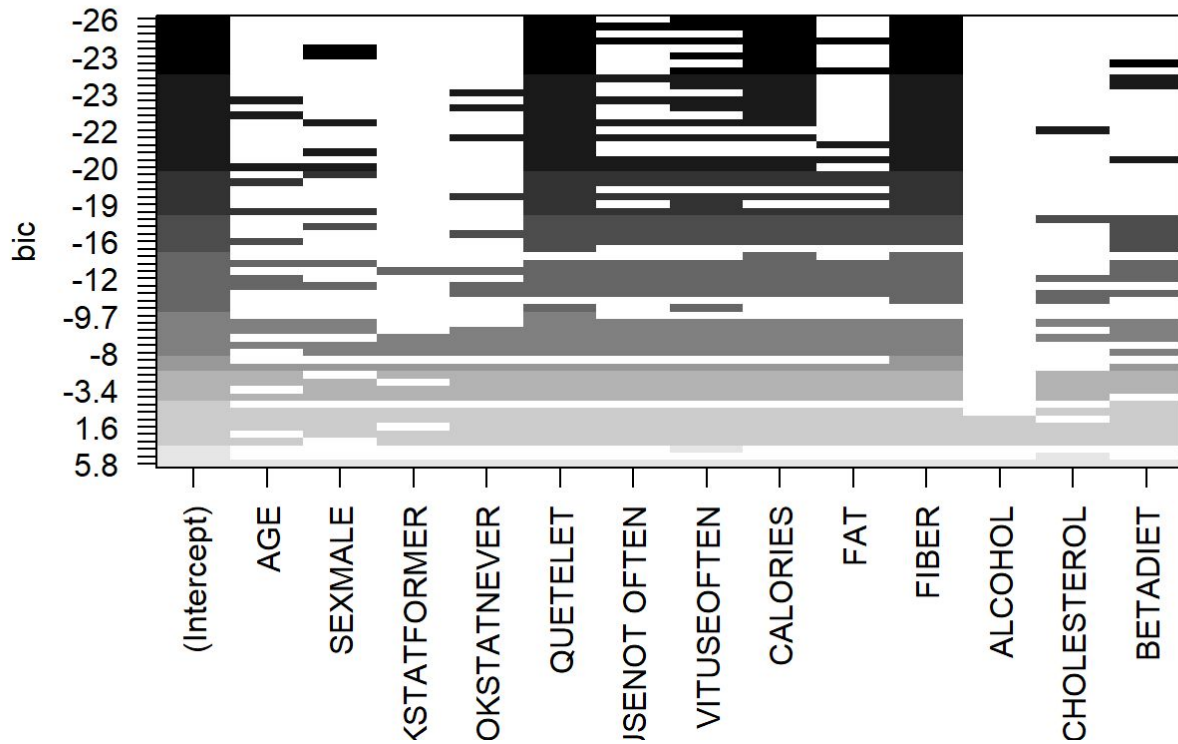
Residual standard error: 0.671 on 236 degrees of freedom  
 Multiple R-squared: 0.2489, Adjusted R-squared: 0.2075  
 F-statistic: 6.015 on 13 and 236 DF, p-value: 1.193e-09

# Model Diagnostics: Cont'd





## Drop in BIC Compared to Intercept-Only Model



# EDA: Beta-Carotene vs. Retinol

- Not very correlated, so our analysis does not suggest much about predictors of retinol

