Analysis of Personal Characteristics Linked to Plasma Concentrations of Retinol and Beta-carotine

36-707: Applied Regression Analysis

, ,
,,@stat.cmu.edu
Carnegie Mellon University

October 29, 2001

1 Abstract

This paper analyzes the association between personal characteristics and dietary intake on the plasma concentrations of beta-carotine and retinol. Low concentrations of these micronutrients may be associated with increased risk of certain types of cancer. The analysis of cross-sectional data does not suggest that personal characteristics and dietary factors have determining influence on micronutrient concentrations. In fact, the results are insignificant to such a degree that a follow-up study would seem appropriate only if the cross-sectional design is abandoned for an entirely new approach, perhaps longitudinal. We found our models unable to explain even 10 percent of variation in micronutrient concentrations for a separate data set not used for model fitting. Moreover, dietary intake levels of the micronutrients were found to have no significant correlation with concentration levels.

2 Introduction

The analysis in this report is in support of ongoing research of the association between low plasma concentration levels of micronutrients retinol and beta-carotine and the development of certain types of cancer. In particular, we will examine a sub-area of this problem, namely, how personal characteristics and dietary habits influence the aforementioned plasma concentrations. The data at our disposal is from a cross-sectional study design tracking 315 study subjects who had an elective surgical procedure during a three-year period, to biopsy or remove a lesion that was found to be non-cancerous.

This research is important if we are to begin to study what types of behaviors or characteristics may lead to increased risk of cancer. Ultimately, successful discovery of such links could provide researchers with compelling evidence that may be used to offer recommendations to the general public.

Our analysis suggests that we have much ground to cover in order to provide accurate predictions of plasma concentrations. However, we have been able to, in a limited manner, identify some factors whose influence/association can be understood directionally (either an increase or decrease in plasma concentrations). We are quick to add that the magnitude of that influence is a more complicated problem that will require additional study.

3 Description of Data

The data for our study contains 315 observations on 14 variables.

Core Variables:

age: Age (years)
sex: Sex (1=Male, 2=Female)
smokstat: Smoking status (1=Never, 2=Former, 3=Current Smoker)
quetelet: Quetelet index ($weight/(height^2)$); values above 27 kg/m^2 (female) or 28 kg/m^2 (male) indicate obesity
vituse: Vitamin Use (1=Yes, fairly often, 2=Yes, not often, 3=No)
calories: Number of calories consumed per day.
fat: Grams of fat consumed per day.
fiber: Grams of fiber consumed per day.
alcohol: Number of alcoholic drinks consumed per week.
cholesterol: Cholesterol consumed (mg per day).
betadiet: Dietary beta-carotene consumed (mcg per day).
retdiet: Dietary retinol consumed (mcg per day)

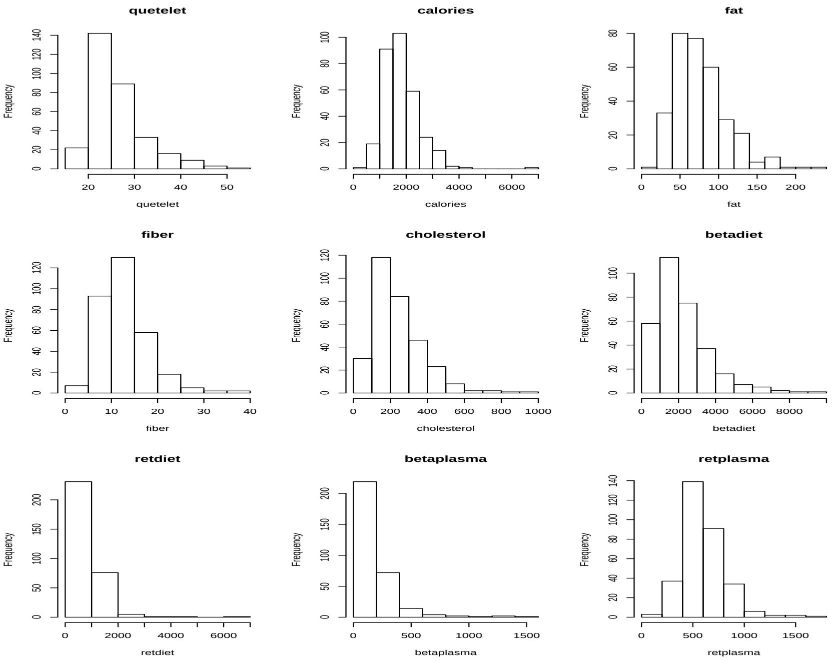

Figure 1: Histograms of several variables

betaplasma: Plasma beta-carotene (ng/ml)
retplasma: Plasma Retinol (ng/ml)

As we can see in Figure 1, several of our quantitative variables demonstrate moderate to severe right skew in their histograms. Among these are both of our response variables, **betaplasma** and **retplasma**. This non-normality motivates a natural log transformation which will tend to make the variables more compatible with our linear regression framework. More on the selection of this transformation can be found in the Technical Appendices.

Notice the histograms of the transformed variables in Figure 2. There is noticeably more symmetry in virtually all cases. Figures 3 and 4 show a comparison of the normal quantile plots before and after the log transformation. It is clear that the transformed variables fit the expected pattern (straight line) for normality much better than the untransformed variables. We will therefore create new variables by using a log transform on this collection of variables.

New variables resulting from natural logarithm transformation:

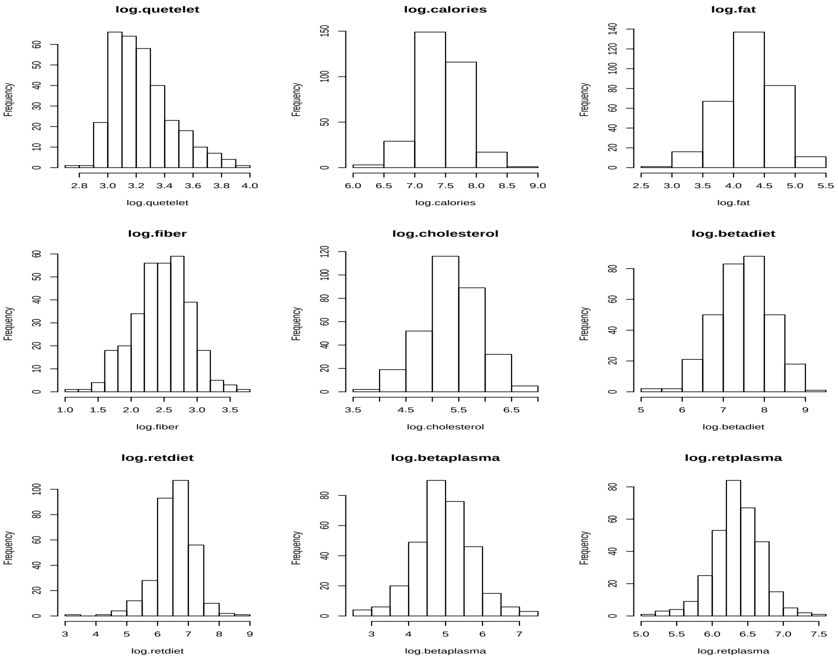

Figure 2: Histograms of transformed variables

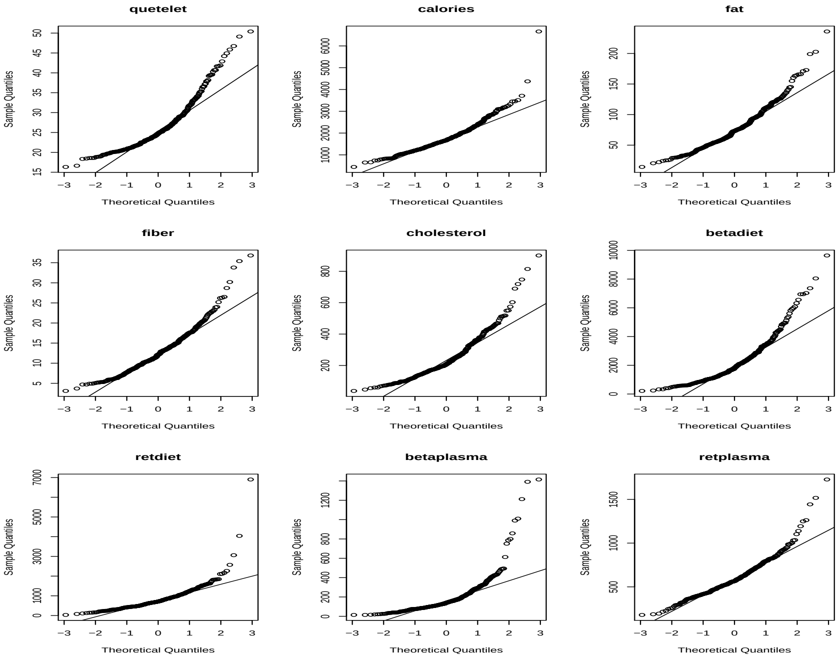

Figure 3: Normal Quantile Plots of untransformed variables

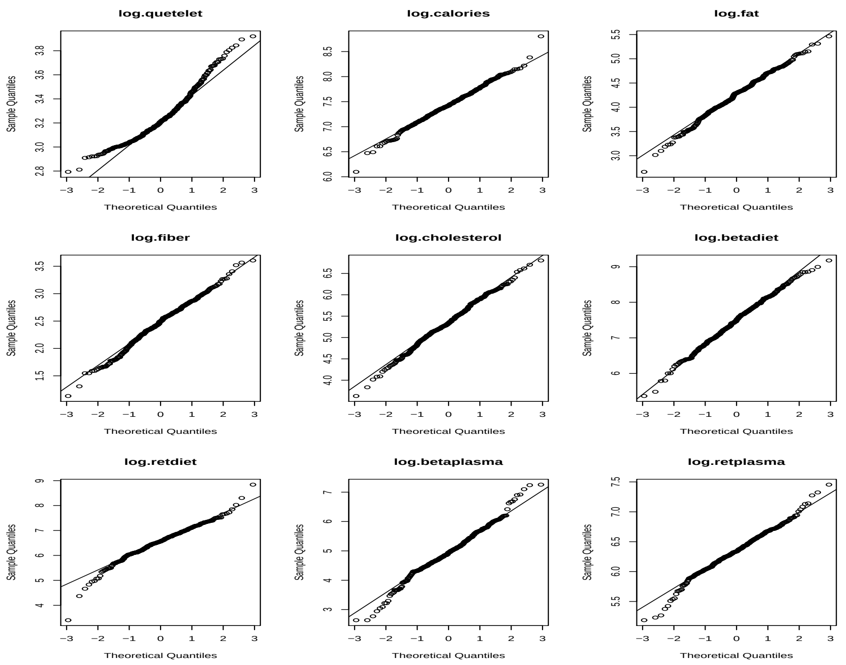

Figure 4: Normal Quantile Plots of transformed variables

log.betadiet
log.betaplasma
log.calories
log.fat
log.fiber
log.cholesterol
log.retdiet
log.retplasma
log.quetelet

In addition to applying a log transformation to many of our variables, it is also necessary to re-code our discrete categorical variables to simplify interpretation. For one quatitative variable, **alcohol**, we have discretized it into disjoint categories and coded the dummy variables, **dummy.alcohol.moderate** and **dummy.alcohol.excess**. Since part of our goal is to establish a set of recommendations to the public, it seems appropriate to examine alcohol consumption in broad categories of drinking considering the high degree of unintentional self-reporting error we might expect. Also, since there were a large percentage of non-drinkers it makes sense to categorize in this way. Note: This re-coding eliminates the problem of one observed value of alcohol consumption that was higher than might seem possible (203 drinks per week!).

New coding for discrete variables:

dummy.male: 1=male, 0=female
dummy.smokstat.current: 1=current smoker, 0=not current smoker
dummy.smokstat.former: 1=former smoker, 0=current or non-smoker
dummy.alcohol.moderate: 1= drink, but no more than 1 per day, 0=else
dummy.alcohol.excess: 1=more than one drink per day, 0=less than one drink per day
dummy.vituse.often: 1=take vitamins fairly often, 0=else
dummy.vituse.notoften: 1=take vitamins, not often, 0=else

Additional information on variables can be found in the Technical Appendices.

4 Analysis and Results

We used the backward elimination method for constructing our regression models for log.betaplasma and log.retplasma. This method consists of starting with a model containing all independent variables and removing the variable with the least significance at every step in the process. At the end of each cycle, we are left with a new, smaller model which can be compared to our original model using nested testing methods described in greater detail in the Technical Appendices. That comparison will tell us whether the bigger model provides a better fit than our new model. We are trying to assess whether it was of much importance that we removed the variable.

For each variable we remove, we compare our new model to the original model. This let's us determine whether we are able to jointly remove all the variables found insignificant to that point. Our goal during this procedure is to construct models that will help predict concentrations of both micronutrients. We want to identify those characteristics that most significantly contribute to the determination of concentration levels.

Here is the final model we constructed for predicting log.betaplasma:

```
Call:
lm(formula = log.betaplasma ~ log.retplasma + dummy.smokstat.current +
           dummy.smokstat.former + log.quetelet + log.fiber +
           dummy.alcohol.moderate +
           dummy.alcohol.excess, data = fit.sample.outliers.removed)
```

```
Residuals:
      Min        1Q    Median        3Q       Max 
-1.966476 -0.384888 -0.008416  0.377853  1.838306 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)              3.4921     1.2431   2.809 0.005503 ** 
log.retplasma            0.4883     0.1504   3.247 0.001384 ** 
dummy.smokstat.current  -0.4733     0.1539  -3.075 0.002427 ** 
dummy.smokstat.former   -0.2567     0.1039  -2.470 0.014408 *  
log.quetelet            -0.8262     0.2133  -3.874 0.000149 ***
log.fiber                0.4211     0.1095   3.845 0.000166 ***
dummy.alcohol.moderate   0.2771     0.1018   2.721 0.007132 ** 
dummy.alcohol.excess    -0.3417     0.1692  -2.020 0.044854 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 0.641 on 184 degrees of freedom
Multiple R-Squared: 0.3041,     Adjusted R-squared: 0.2777 
F-statistic: 11.49 on 7 and 184 DF,  p-value: 4.523e-012 
```

Our model describes a negative association for smoking (current or former), the quetelet index, and drinking in excess of one drink per day. Positive associations include retinol concentration, fiber intake, and moderate levels of drinking. In fact, our base case, non-drinkers, tend to have lower concentration levels than those who drink moderately.

All of our variables were significant at p=.05, while **log.quetelet** and **log.fiber** were significant at p=.001. The diagnostic plots, Figure 5, suggest that our model has normal residuals and a residual cloud with no clear pattern. This final model for log.betaplasma was altered by the removal of several outliers in our independent variables. Our first model contained vitamin use (**dummy.vituse.often** and **dummy.vituse.notoften**), but they were no longer significant after we removed the outliers.

When we tested this model on a subset of data that had been set aside, we found that it has very limited prediction value. The following regression output shows that our predictions were better than the strawman model, but the low R-squared value (.06021) suggests that our model does not explain a high degree of variation in log.betaplasma. Our residuals plot (Figure 6), does not reveal any pattern which might suggest our model

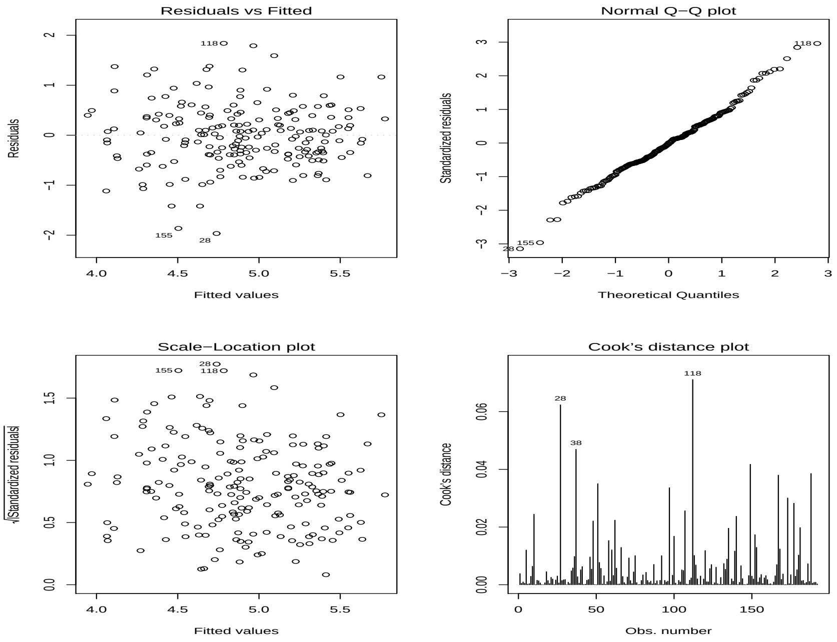

Figure 5: Diagnostics for log.betaplasma Regression

would be better served with the addition of another variable. Since we have tested all of our variables, it would not be possible to add another significant variable in any case.

```
Call:
lm(formula = log.betaplasma ˜ fitted.test)

Residuals:
     Min       1Q   Median       3Q      Max
-1.77554 -0.42227 -0.06862  0.46783  2.14477

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9232     0.7886   3.707 0.000328 ***
fitted.test   0.4321     0.1613   2.679 0.008504 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6746 on 112 degrees of freedom
Multiple R-Squared: 0.06021,    Adjusted R-squared: 0.05182
```

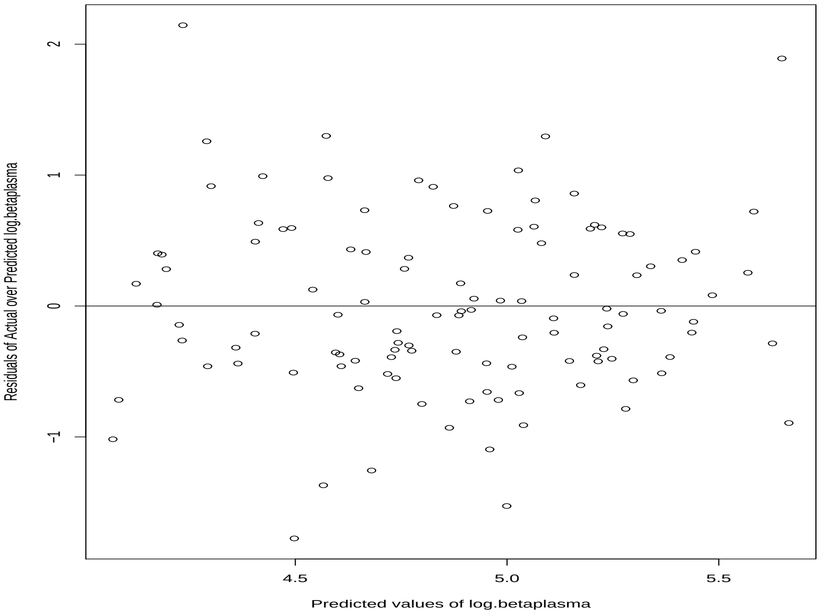

Figure 6: Residuals Plot for log.betaplasma prediction

```
F-statistic: 7.175 on 1 and 112 DF,  p-value: 0.008504
```

Here is the final model we constructed for predicting log.retplasma:

```
Call:
lm(formula = log.retplasma ~ log.betaplasma + age +
     dummy.alcohol.excess +
          dummy.alcohol.moderate, data = fit.sample)

Residuals:
     Min       1Q   Median       3Q      Max
-0.99188 -0.18498 -0.01776  0.19568  1.07843

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            5.754913   0.154968  37.136  < 2e-16 ***
log.betaplasma         0.062754   0.030638   2.048  0.04188 *
age                    0.004094   0.001593   2.570  0.01092 *
dummy.alcohol.excess   0.199205   0.074520   2.673  0.00815 **
dummy.alcohol.moderate 0.046524   0.050510   0.921  0.35814
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3153 on 195 degrees of freedom
Multiple R-Squared: 0.09496,    Adjusted R-squared: 0.0764
F-statistic: 5.115 on 4 and 195 DF,  p-value: 0.0006109
```

Our model for **log.retplasma** was whittled down much further than the **log.betaplasma** model. The final specification includes only three significant positive coefficients. Age, beta-carotine concentration, and alcohol consumption have significant positive relationships with retinol concentration. It seems that excessive alcohol consumption (more than one drink per day) is more highly associated with higher plasma concentrations than moderate drinking or no drinking at all. The coefficient for **dummy.alcohol.excess** is significant at p=.01. At this time we would ask the reader to interpret these results with an appropriate degree of skepticism and not use our **log.retplasma** model as an invitation to embark on a new life of alcoholic abandonment. Only a follow-up study could give a green light on that one.

After removing the outliers among our independent variables, the model for **log.retplasma** did not change. Each of our coefficients remained significant, and none of the previously removed variables achieved a newfound significance. The diagnostic plot in Figure 7 shows no clear indication that we have a systematic departure from normality. Our normal quantiles plot looks to be fairly straight.

When testing our model on a separate sample of data, it showed low predictive value for log.retplasma. The following regression output reveals a low R-squared value (.08983).

```
Call:
lm(formula = log.retplasma ~ fitted.test.ret)

Residuals:
      Min        1Q    Median        3Q       Max
-1.081781 -0.199816  0.005377  0.209117  1.070988

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -0.3416     2.0285  -0.168  0.86655
fitted.test.ret   1.0618     0.3194   3.325  0.00120 **
```

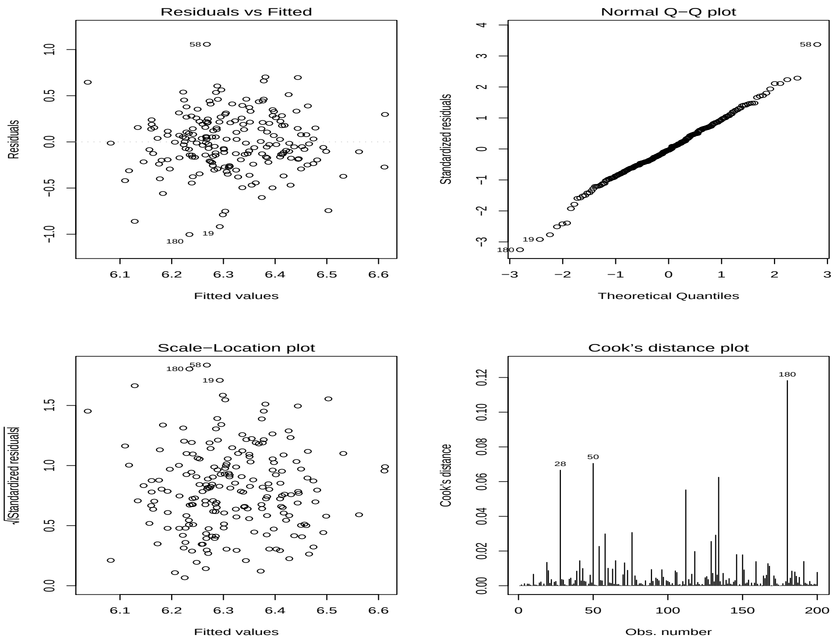

Figure 7: Diagnostics for log.retplasma Regression

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3341 on 112 degrees of freedom
Multiple R-Squared: 0.08983,    Adjusted R-squared: 0.0817
F-statistic: 11.05 on 1 and 112 DF,  p-value: 0.001197
```

We see in Figure 8 that our residuals plot for the prediction test is a pretty good residual cloud with no discernible pattern. Overall, neither of our micronutrient concentration regression models is able to explain very much of the variation in concentration by utilizing our personal characteristics and dietary intake explanatory variables. In fact, dietary intake variables (**log.betadiet**, **log.retdiet**) did not make either of our final models.

5 Discussion/Conclusion

We were unable to accurately predict plasma concentrations of Retinol and Beta-carotine utilizing the cross-sectional data recorded for this study. The standard linear regression techniques did not provide sufficient overall fit to give us confidence that we are able to

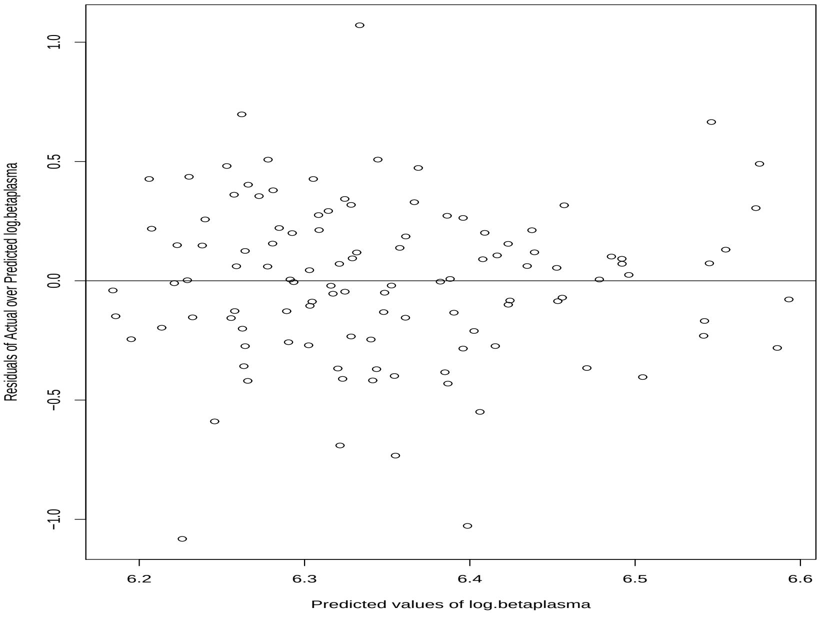

Figure 8: Residuals Plot for log.retplasma prediction

explain variation in concentration in this manner. We were, however, able to identify multiple variables which are associated with either higher or lower concentrations of beta-carotine.

Smoking is associated with lower concentration of betaplasma. This is even more true for current smokers. While high quetelet scores are also associated with lower concentrations. Moderate alcohol consumption, fiber intake, and retinol concentration were all positively associated with levels of betaplasma.

Concentrations of Retinol seems to be less influenced by characteristics of personal behavior. The evidence suggests that concentrations of Retinol tend to be higher among the older subjects. Also, the subjects who averaged more than 7 alcoholic drinks per week tended to have a higher concentration. Finally, the concentration of Beta-carotine is positively associated with retinol concentration.

Clearly the type of data we have collected in this study has been insufficient to help us produce prediction models which explain a majority of the variation in plasma concentrations of beta-carotine and retinol. The cross-sectional nature of the study may have inhibited our ability to construct our prediction models. Perhaps a longitudinal study would be a more capable approach to producing predictions. Tracking changes in personal behaviors and studying the concomitant changes in concentration would, at the very least, better situate us to make recommendations to the public with a sense of the impact that we might expect those behavioral changes to have on an individual basis. This approach could be the correct next step in the research given the unacceptably poor fits of our current models.

Another issue that arose during the analysis was possible sample bias. The data had been collected about individuals, all of whom "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." In a future study, it would be helpful to have a sense how the study subjects, if chosen in the same way, compare to the general population. If, for instance, non-cancerous growths are highly correlated with cancer and low plasma concentrations, then we may have experienced some interference resulting from what appears to be a biased sample. Also, the high percentage of women (87 percent) in our study could have also had an influence on our analysis.

Perhaps the most surprising finding in our analysis was the lack of significant correlation between dietary intake of the micronutrients and their plasma concentration levels. Given this non-correlation, it is quite likely to be very difficult to pin down a reliable prediction of concentration levels. We finish this analysis with a greater appreciation of the complexity of this prediction problem. Clearly, a new approach is needed.

6 Technical Appendices

- Log Transformations

  Log transformations were used to reduce right skew in many variables. All of the standard reasons to transform data are well-known. Our main goal in this analysis was to bring outlying observations closer to the main body of the data. The nature of our data tended to make positive outliers very common. Therefore, the only type of transformation we needed was the natural logarithm.

- Nested F-tests

  We conducted F-tests during every stage in our backward elimination process for model selection. Starting with a model containing all of our independent variables, we removed, during each iteration, the variable with the highest p-value. A test comparing our final models with the big initial models suggested that we could not reject the null hypothesis that the slope coefficients for all excluded variable were zero. Essentially, we employed the Extra-Sum-of-Squares Principle to conclude that our smaller (final) models did not have significantly less explanatory ability than the big models. Occam's razor was our guide in developing the models.

- Outliers

  After conducting univariate analysis of our core variables, several
  potential outliers were identified.

  ```
  List of outliers
  -----------------

  Variable    Value(s)
  -----------------------------------
  retplasma  1443, 1517 and 1727
  betaplasma 0, 1212, 1391 and 1415
  retdiet 4041 and 6901
  betadiet 9642 and 8046
  cholesterol 900.7
  fat 199.0, 202.7, and 235.9
  calories 6662.2
  ```

  Note: The log transformation we implemented would tend to reduce the degree to which these observations are outliers. However, it should not matter if we are a bit conservative and take out some observations which may be on the fence.

The value of 0 for betaplasma in a single case suggests that there may be some kind of coding error at work. We feel that we may confidently remove that observation. The other outliers may have been coded correctly, but we will still need to keep an eye on them to be sure that they do not have too much influence in our models. This can most efficiently be achieved by fitting our models with the full set model-fitting data and compare the result to the models without these outliers. NOTE: These outliers exist somewhere in our full set of data (315 observations). We will not necessarily be removing all 16 from our model-fitting sample since some of them will likely be found in the test-fitting sample (115 observations) which we have set aside to test the prediction ability of our models.

We created a new sample without outliers.

```
> fit.sample.outliers.removed_fit.sample[-c(197,171,115,58,54,50,41,1),]
```

- Summary of Variables

```
      age           alcohol            betadiet   
 Min.   :19.00   Min.   :  0.000   Min.   : 214  
 1st Qu.:39.00   1st Qu.:  0.000   1st Qu.:1116  
 Median :48.00   Median :  0.300   Median :1802  
 Mean   :50.15   Mean   :  3.279   Mean   :2186  
 3rd Qu.:62.50   3rd Qu.:  3.200   3rd Qu.:2836  
 Max.   :83.00   Max.   :203.000   Max.   :9642  

   betaplasma        calories       cholesterol   
 Min.   :  14.0   Min.   : 445.2   Min.   : 37.7  
 1st Qu.:  90.0   1st Qu.:1338.0   1st Qu.:155.0  
 Median : 140.0   Median :1666.8   Median :206.3  
 Mean   : 189.9   Mean   :1796.7   Mean   :242.5  
 3rd Qu.: 230.0   3rd Qu.:2100.4   3rd Qu.:308.9  
 Max.   :1415.0   Max.   :6662.2   Max.   :900.7  

 dummy.alcohol.excess   dummy.alcohol.moderate
 Min.   :0.0000         Min.   :0.0000        
 1st Qu.:0.0000         1st Qu.:0.0000        
 Median :0.0000         Median :1.0000        
 Mean   :0.1302         Mean   :0.5175        
```

```
3rd Qu.:0.0000         3rd Qu.:1.0000
Max.   :1.0000         Max.   :1.0000

  dummy.male     dummy.smokstat.current
Min.   :0.0000   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.0000
Median :0.0000   Median :0.0000
Mean   :0.1333   Mean   :0.1365
3rd Qu.:0.0000   3rd Qu.:0.0000
Max.   :1.0000   Max.   :1.0000

dummy.smokstat.former  dummy.vituse.notoften
Min.   :0.0000          Min.   :0.0000
1st Qu.:0.0000          1st Qu.:0.0000
Median :0.0000          Median :0.0000
Mean   :0.3651          Mean   :0.2603
3rd Qu.:1.0000          3rd Qu.:1.0000
Max.   :1.0000          Max.   :1.0000

dummy.vituse.often      fat             fiber
Min.   :0.0000      Min.   : 14.40   Min.   : 3.10
1st Qu.:0.0000      1st Qu.: 53.95   1st Qu.: 9.15
Median :0.0000      Median : 72.90   Median :12.10
Mean   :0.3873      Mean   : 77.03   Mean   :12.79
3rd Qu.:1.0000      3rd Qu.: 95.25   3rd Qu.:15.60
Max.   :1.0000      Max.   :235.90   Max.   :36.80

   quetelet        retdiet         retplasma
Min.   :16.33   Min.   :  30.0   Min.   : 179.0
1st Qu.:21.80   1st Qu.: 480.0   1st Qu.: 466.0
Median :24.74   Median : 707.0   Median : 566.0
Mean   :26.16   Mean   : 832.7   Mean   : 602.8
3rd Qu.:28.85   3rd Qu.:1037.0   3rd Qu.: 716.0
Max.   :50.40   Max.   :6901.0   Max.   :1727.0
```

- Univariate Analysis of Core Variables

```
``age''
```

```
stemplot

  1 | 9
  2 | 2234
  2 | 556677789999
  3 | 01111122222233333333344444
  3 | 55555566666666777777777778888888888999999
  4 | 0000000111111111111122222222333333333344444444444
  4 | 55555555666666666666777788888888999999999999
  5 | 0000001112222333334444
  5 | 5555556666666667777888999
  6 | 00000112222233444444
  6 | 55555556666666667778999999
  7 | 000000011111222233333334444444
  7 | 55555677788
  8 | 2333
```

Minimal skew is detected in the stemplot.

```
summary(age)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  19.00   39.00   48.00   50.15   62.50   83.00
```

‘‘sex’’

It is clear that the sample is heavily biased towards women.
Of the 315 cases, only 42 are males.

‘‘smokstat’’

```
> table(smokstat)
smokstat
  1   2   3
157 115  43
```

The smoking variable is split almost perfectly even between persons who have never smoked and those who are either current or former smokers.

‘‘quetelet’’

```
> summary(quetelet)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  16.33   21.80   24.74   26.16   28.85   50.40 
```

```
> stem(quetelet)
  16 | 36
  18 | 3466689902444677889
  20 | 0011112222223444445566677777889000011112222233355556677777888889
  22 | 000002245555555666777990011111122333333444555567788999999
  24 | 001112333345567777899001111122222445666777788999999
  26 | 1133344445567889990233335555889
  28 | 000334444678900001122223667 8
  30 | 013334577247778
  32 | 00137001234677
  34 | 11260234
  36 | 04561399
  38 | 22456
  40 | 377679
  42 | 9
  44 | 299
  46 | 7
  48 | 1
  50 | 4
```

‘‘vituse’’

```
> table(vituse)
vituse
  1   2   3 
122  82 111 
```

```
The vitamin use variable is most concentrated among regular
users or
those who don’t take any vitamins.  A smaller number take vitamins
occasionally.
```

```
‘‘calories’’

<INSERT boxplot.calories>

> summary(calories)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  445.2  1338.0  1667.0  1797.0  2100.0  6662.0


> stem(calories)

  The decimal point is 3 digit(s) to the right of the |

  0 | 4
  0 | 6777888888888999
  1 | 00000000001111111111111111122222222222222222223333333333333333344444+6
  1 | 55555555555555555555666666666666666666666666666677777777777777777777+25
  2 | 00000000000000000011111111111111122222222222333333333333444444
  2 | 555555566777777778888889999
  3 | 0011111222334
  3 | 557
  4 | 4
  4 |
  5 |
  5 |
  6 |
  6 | 7


6662.2 is clearly a large positive outlier



‘‘fat’’
```

```
> summary(fat)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.40   53.95   72.90   77.03   95.25  235.90
```

```
> stem(fat)

  The decimal point is 1 digit(s) to the right of the |

   1 | 4
   2 | 02455699
   3 | 001113333344555556778899
   4 | 012333344445555666777788999
   5 | 000001111222223333444455555555566667777777888889999999
   6 | 00111122222222333334444555567788
   7 | 00112222333333444445555555666667777777788899999
   8 | 00111112222233444444555667999
   9 | 0222333444445555556777888889999
  10 | 1134566679
  11 | 000112223333455699
  12 | 0011112345566689
  13 | 023569
  14 | 145
  15 | 5
  16 | 03466
  17 | 13
  18 |
  19 | 9
  20 | 3
  21 |
  22 |
  23 | 6

>
```

```
199.0, 202.7, and 235.9 are all much higher than the rest of the pack.
```

```
``fiber''
```

```
> summary(fiber)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.10    9.15   12.10   12.79   15.60   36.80


> stem(fiber)

  The decimal point is at the |

   2 | 17
   4 | 7799012233345669999
   6 | 0001123333566788990011344566778999
```

‘‘alcohol’’

Since there are so many people who have zero drinks per week, it makes since to construct a dummy variable. The variable dummy.alcohol.excess will track persons who average more than one drink per day. The variable dummmy.alcohol.moderate will be for those who drink, but do not drink more than 7 drinks per week (<= 1 per day).

Here are some of the raw numbers.

```
> table(dummy.alcohol.excess)
dummy.alcohol.excess
```

```
  0   1
274  41

> table(dummy.alcohol.moderate)
dummy.alcohol.moderate
  0   1
152 163
```

The number of non-drinkers in our sample is 111.

NOTE: There is a huge outlier in alcohol variable. Fortunately, our use of the dummy variable coding for alcohol makes this point moot and also makes the most sense for providing a recommendation to the public. There is unlikely to be much attention given to any recommendation that says 3.4 drinks per week is better than 2.7. A dummy variable makes the most sense.

‘‘cholesterol’’

```
> stem(cholesterol)

  The decimal point is 2 digit(s) to the right of the |

  0 | 4
  0 | 566677778888889999999
  1 | 0000000000001111111222222222333333444444444444
  1 | 55555555555666666666666677777777777777788888888888888888889999999999
  2 | 000000000000000111111111112222222233333333334444444
  2 | 5555555555566666666667777777778888888899
  3 | 0001111112333333333444444
  3 | 555566666667778888889
  4 | 00122223333334444
  4 | 55667779
  5 | 01122
  5 | 557
  6 | 0
  6 | 9
  7 | 2
  7 | 5
  8 | 1
  8 |
  9 | 0

>
```

900.7 has achieved some separation from the pack.

‘‘betadiet’’

```
> stem(betadiet)

  The decimal point is 3 digit(s) to the right of the |

  0 | 223344
  0 | 55556666666666666677777778888888888899999999999
  1 | 0000000000001111111111111111222222222222222333333333444444444444444
  1 | 55555555555556666677777777777777777788889999999
  2 | 0000000011111111111112222222233333333444444444
  2 | 555555566666677777778889999999999
  3 | 01111122333333444444
  3 | 555566666777889
  4 | 0001333444
  4 | 5578899
  5 | 0134
  5 | 689
  6 | 013
  6 | 699
  7 | 04
  7 |
  8 | 0
  8 |
  9 |
  9 | 6
```

```
9642 and 8046 are considerably higher than the rest.
```

```
‘‘retdiet’’

> stem(retdiet)

  The decimal point is 3 digit(s) to the right of the |

  0 | 0111112222222222223333333333333333333344444444444444444444444444
  0 | 55555555555555555555555555555555555555666666666666666666666666666666667+75
  1 | 00000000000000000000111111111111111122222222223333333333333344444
  1 | 55555555566666667888888
  2 | 1123
  2 | 6
  3 | 1
  3 |
  4 | 0
  4 |
  5 |
  5 |
  6 |
  6 | 9
```

```
4041 and 6901 are far from the other values
```

```
‘‘betaplasma’’

> stem(betaplasma)

  The decimal point is 2 digit(s) to the right of the |

   0 | 11222233333444444444445555555566666677777777777888888888888888888889+4
   1 | 00000000000000000000111111111111111111122222222222222223333333333333+52
   2 | 00000011111112222333333344444555567777778899999999
   3 | 000012222233333456777899
```

```
    4 | 011223334567999
    5 |
    6 | 1
    7 | 59
    8 | 06
    9 | 9
   10 | 1
   11 |
   12 | 1
   13 | 9
   14 | 2
```

1212,1391 and 1415 are all higher than the rest.

‘‘retplasma’’

```
> stem(retplasma)

  The decimal point is 2 digit(s) to the right of the |

    1 | 899
    2 | 23556899
    3 | 0022233456667777888899999
    4 | 0000000111112222222333333334444444556666677777777888889999999
    5 | 00000000011111222222222223333333334444444455555666666666666677777778
    6 | 000000111222222222333333444555555666777788888889999
    7 | 00000111122233333334445556666778888999
    8 | 000001122222233334455556888
    9 | 00223345599
   10 | 0034
   11 | 049
   12 | 56
   13 |
   14 | 4
   15 | 2
   16 |
   17 | 3
```

```
1443, 1517 and 1727 are contributing to a right skew.
```

- No correlation between dietary intake and plasma concentration

  One of the surprising findings in our analysis is that there is no significant relationship among the plasma concentrations and the dietary intake for either micronutrient. We present the simple regression models which reveal as much:

```
Call:
lm(formula = log.betaplasma ~ log.betadiet)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.94072 -0.49732 -0.04325  0.49920  2.11696 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   4.31112    0.76052   5.669 1.14e-07 ***
log.betadiet  0.09489    0.10017   0.947    0.346    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 0.6931 on 112 degrees of freedom
Multiple R-Squared: 0.007948,   Adjusted R-squared: -0.0009091 
F-statistic: 0.8974 on 1 and 112 DF,  p-value: 0.3455 
```

```
Call:
lm(formula = log.retplasma ~ log.retdiet)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.23012 -0.23896  0.02739  0.24537  1.02904 
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.77497    0.33376  20.299   <2e-16 ***
log.retdiet -0.05737    0.05105  -1.124    0.263
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3482 on 112 degrees of freedom
Multiple R-Squared: 0.01115,    Adjusted R-squared: 0.002322
F-statistic: 1.263 on 1 and 112 DF,  p-value: 0.2635
```

7 Bibliography and Credits

I conferred with Brian Junker on the analysis found in this report and used his class handouts and notes extensively. No additional sources were used.