Appendix: Analysis log and Splus documentation

Data input and cleaning

Read in the data.

```
> micro <- read.table("prdata.dat")
> names(micro) <- c("age", "sex", "smokstat", "quetelet", "vituse",
                    "calories", "fat", "fiber", "alcohol","cholesterol",
                    "betadiet", "retdiet", "betaplasma", "retplasma")
> dim(micro)
[1] 315  14
```

Check for missing values; there are none.

```
> sum(is.na(micro))
[1] 0
```

Do some recoding.

```
> micro$sex <- micro$sex-1  # Now M=0, F=1
> micro$sex <- as.factor(micro$sex)
> micro$smokstat <- as.factor(micro$smokstat)
> micro$vituse <- as.factor(micro$vituse)
> micro$obesity <- rep(0,315)
> micro$obesity <- ifelse(micro$sex==0&micro$quetelet>28, 1, 0)
> micro$obesity <- ifelse(micro$sex==1&micro$quetelet>27, 1, micro$obesity)
> micro$obesity <- as.factor(micro$obesity)
> micro$percentfat <- 9*micro$fat/micro$calories
```

Check for outliers and extreme observations.

```
> boxplot(micro$age)              # No problems
> summary(micro$sex)              # No apparent miscodings
> summary(micro$smokstat)         # No apparent miscodings
> boxplot(micro$quetelet)         # Lots of high outliers, nothing unreasonable
> summary(micro$vituse)           # No apparent miscodings
> boxplot(micro$calories)         # One really high outlier; who is this?
> micro[micro$calories==max(micro$calories),]
   age sex smokstat quetelet vituse calories   fat fiber alcohol cholesterol
62  65   0        3 23.37617      3   6662.2 164.3  11.3     203         603
   betadiet retdiet betaplasma retplasma obesity percentfat
62     2893    1364         96       317       0  0.2219537
> summary(micro$percentfat)       # Find normal percent fat in diet
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
 0.1630 0.3295  0.3848 0.3823 0.4325  0.6303
> micro$percentfat[62]            # 62's percent fat is low
[1] 0.2219537
> rank(micro$quetelet)[62]/315    # And he's in the lower 50% of BMI
[1] 0.3936508
```

It looks like 62's recorded calories was a mistake; keep an eye on this case. Continuing...

```
> boxplot(micro$fat)               # Some more high outliers
> rank(micro$percentfat)[152]/315  # Highest (152) is also high in percent fat
[1] 0.9301587
> boxplot(micro$percentfat)        # Few outliers, nothing extreme
> boxplot(micro$fiber)             # Lots of high outliers, nothing unreasonable
> boxplot(micro$alcohol)           # One huge outlier -- 203 drinks (29 a day!)
> micro[micro$alcohol==max(micro$alcohol),]  # It's 62 again
> boxplot(micro$cholesterol)       # Lots of high outliers, nothing unreasonable
> boxplot(micro$betadiet)          # Lots of high outliers, nothing unreasonable
> boxplot(micro$retdiet)           # One very high outlier
> micro[micro$retdiet==max(micro$retdiet),]  # This is subject 171; watch her
> boxplot(micro$betaplasma)        # Right-skewed, one subject is ZERO
> boxplot(micro$retplasma)         # Right-skewed, no obvious extreme cases
```

Since subject 62 was an extreme outlier for both calories and alcohol, and because it is difficult to determine which other variables are reliable given this, I dropped him from these analyses. Later I went back and reran the models with and without him (see end of Appendix).

```
micro2 <- micro[-62,]
```

Univariate analyses

- Demographics:
  - Age: fairly symmetric, no outliers

    ```
    > summary(micro2$age)
     Min. 1st Qu. Median Mean 3rd Qu. Max.
     19.0 39.0    47.5   50.1 62.0    83.0
    > stem(micro2$age)
    N = 314   Median = 47.5
    Quartiles = 39, 62

    Decimal point is 1 place to the right of the colon

       1 : 9
       2 : 2234
       2 : 556677789999
       3 : 01111122222233333333344444
       3 : 55555566666666777777777778888888889999999
       4 : 0000000111111111111122222222333333333344444444444
       4 : 55555555666666666666777788888888999999999999
       5 : 0000001112222233333444
       5 : 5555556666666677777888999
    ```

```
6 : 00000112222233444444
6 : 55555566666666667778999999
7 : 0000000111112222333333334444444
7 : 55555677788
8 : 2333
```

- Sex: 41 (13.1%) males and 273 (86.9%) females

  ```
  > summary(micro2$sex)
    0   1
   41 273
  ```

- Health related:

  - Smoking: 157 (50%) never smoked, 115 (36.6%) are former smokers, and 42 (13.4%) are current smokers

    ```
    > summary(micro2$smokstat)
       1   2  3
     157 115 42
    ```

  - Quetelet index: right skewed, with 6 high outliers (subjects 25, 190, 226, 236, 249, and 278); there is a mound of data below 27 or 28 (the cutoff for obesity), but many subjects are higher than this normal range

    ```
    > summary(micro2$quetelet)
      Min. 1st Qu. Median  Mean 3rd Qu.  Max.
     16.33 21.79   24.74  26.17 28.90   50.40
    > stem(micro2$quetelet)
    N = 314   Median = 24.73935
    Quartiles = 21.78854, 28.9498

    Decimal point is at the colon

       16 : 36
       17 :
       18 : 34666899
       19 : 02444677889
       20 : 00111122222234444455666777777889
       21 : 000011112222333555566777777888889
       22 : 000002245555555666677799
       23 : 0011111122333333344555567788999999
       24 : 001112333345567777899
       25 : 00111112222244566677777789999999
       26 : 113334444556788999
       27 : 0233335555889
       28 : 0003344446789
       29 : 000011222236678
    ```

```
    30 : 013334577
    31 : 247778
    32 : 00137
    33 : 001234677
    34 : 1126
    35 : 0234
    36 : 0456
    37 : 1399
    38 : 2
    39 : 2456
    40 : 377
    41 : 679
    42 : 9

High: 44.20619 44.91065 45.85808 46.73763 49.12033 50.40333
```

- Obesity: 213 (67.8%) are not considered obese, 101 (32.2%) are

```
> summary(micro2$obesity)
   0   1
 213 101
```

- Vitamin use: 122 (38.9%) use them fairly often, 82 (26.1%) do, but not often, and 110 (35%) do not use them

```
> summary(micro2$vituse)
   1  2   3
 122 82 110
```

• Diet:

- Calories: slightly right skewed with 2 high outliers

```
> summary(micro2$calories)
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  445.2 1336.0  1665.0 1781.0 2093.0  4374.0
> stem(micro2$calories)
N = 314   Median = 1665.05
Quartiles = 1333.8, 2094.5

Decimal point is 3 places to the right of the colon

   0 : 4
   0 : 6777
   0 : 888888888999
   1 : 000000000011111111111111111
   1 : 222222222222222222233333333333333333
   1 : 44444444444444444444444555555555555555555555
   1 : 66666666666666666666666666677777777777777777777777
```

```
   1 : 8888888888888888888899999999999999
   2 : 000000000000000000111111111111111
   2 : 22222222222233333333333
   2 : 4444445555555
   2 : 6677777777
   2 : 8888889999
   3 : 0011111
   3 : 22233
   3 : 455

High: 3711.0 4373.6
```

– Fat: slightly right skewed with 3 high outliers

```
> summary(micro2$fat)
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  14.40  53.93   72.90  76.76  95.18  235.90
> stem(micro2$fat)
N = 314   Median = 72.9
Quartiles = 53.9, 95.2

Decimal point is 1 place to the right of the colon

    1 : 4
    2 : 02455699
    3 : 001113333344455556677889
    4 : 012233344445555566777788899
    5 : 000001111222223333444444455555566667777777888888999999 9
    6 : 00011111222223333344455555667889
    7 : 011122233333344444555555666666777777788899999
    8 : 001111122222334444445555678999
    9 : 222233444444555556677788889999
   10 : 1134556679
   11 : 000111223333455689
   12 : 0011112345566689
   13 : 023569
   14 : 145
   15 : 5
   16 : 0366
   17 : 13

High: 199.0 202.7 235.9
```

– Percent calories from fat: fairly symmetric, with no outliers

```
> summary(micro2$percentfat)
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
 0.1630 0.3304  0.3849 0.3828 0.4327  0.6303
```

```
> stem(micro2$percentfat)
N = 314   Median = 0.384949
Quartiles = 0.330244, 0.43288

Decimal point is 1 place to the left of the colon

   1 : 67
   1 : 9
   2 : 11
   2 : 23
   2 : 4555
   2 : 66666777777
   2 : 88888899999999
   3 : 00000001111111111111111
   3 : 22222222222222333333333
   3 : 4444444445555555555555555
   3 : 6666666666666666666777777777777777
   3 : 888888888888888899999999999999999999999
   4 : 0000000000000011111111111111
   4 : 222222222222222333333333333333
   4 : 4444444444444444444455555555
   4 : 66666666666777777777777
   4 : 8899999
   5 : 0000011
   5 : 22333
   5 : 5
   5 : 667
   5 :
   6 : 1
   6 : 3
```

- Fiber: right skewed with 5 high outliers

```
> summary(micro2$fiber)
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  3.100  9.125  12.100 12.790 15.600  36.800
> stem(micro2$fiber)
N = 314   Median = 12.1
Quartiles = 9.1, 15.6

Decimal point is at the colon

    3 : 17
    4 : 7799
    5 : 012233345669999
    6 : 000112333356678899
    7 : 0011344566778999
```

```
     8 : 0222334455555677788889
     9 : 00122334455666666688899
    10 : 0122222333444444555666666788888899999
    11 : 111222223444456669
    12 : 00111122333555799999999
    13 : 0112222223333444555666677788899
    14 : 0112222333444466677899999
    15 : 001112566789
    16 : 00111223455668889
    17 : 0113345566777789
    18 : 122448
    19 : 01234579
    20 : 0134568
    21 : 148
    22 : 13569
    23 : 039
    24 : 0
    25 : 2
    26 : 235

High: 28.7 30.2 33.8 35.4 36.8
```

– Alcohol: mode is at zero, very right skewed, many high outliers

```
> summary(micro2$alcohol)
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  0.000  0.000   0.300  2.643  3.200  35.000
> stem(micro2$alcohol)
N = 314   Median = 0.3
Quartiles =  0, 3.2

Decimal point is at the colon

   0 : zzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzx
   0 : 55555555555666777777778899999
   1 : 00000000000122223334
   1 : 55577
   2 : 00000012234
   2 : 569
   3 : 0000122224
   3 : 577
   4 : 0001112
   4 : 5579
   5 : 000002
   5 : 67
   6 : 1224
   6 : 5578
```

```
   7 : 0000011222223
   7 :
   8 : 00034
   8 : 5
   9 : 0

High: 10.0 10.0 10.5 10.5 11.0 11.0 11.0 14.0 14.0 14.1 14.1 14.1 14.2 15.0 15.0
High: 15.5 17.0 18.0 18.0 18.2 20.0 21.0 22.0 35.0 35.0
```

– Cholesterol: right skewed with 5 high outliers

```
> summary(micro2$cholesterol)
  Min. 1st Qu. Median  Mean 3rd Qu.  Max.
  37.7 155.0   206.2  241.3 308.2   900.7
> stem(micro2$cholesterol)
N = 314   Median = 206.2
Quartiles = 154.9, 308.8

Decimal point is 2 places to the right of the colon

   0 : 4
   0 : 566677778888889999999
   1 : 0000000000001111111222222222333333444444444444
   1 : 555555555555666666666666777777777777777777888888888888888889999999999999
   2 : 00000000000000111111111122222222233333333333444444
   2 : 5555555555556666666666777777778888888899
   3 : 0001111112333333333444444
   3 : 555566666666778888889
   4 : 00122223333334444
   4 : 55667779
   5 : 01122
   5 : 557

High: 689.4 718.8 747.5 814.7 900.7
```

– Dietary beta-carotene: right skewed with 8 high outliers

```
> summary(micro2$betadiet)
 Min. 1st Qu. Median Mean 3rd Qu. Max.
  214 1115    1795   2183 2804    9642
> stem(micro2$betadiet)
N = 314   Median = 1795
Quartiles = 1114, 2809

Decimal point is 3 places to the right of the colon

   0 : 223344
   0 : 55555666666666666677777778888888888899999999999
```

```
   1 : 000000000000111111111111111222222222222222222222333333333444444444444444
   1 : 55555555555555666667777777777777777778888899999
   2 : 0000000011111111111122222222233333333444444444
   2 : 55555556666667777777888999999999
   3 : 01111122333333444444
   3 : 555566666777889
   4 : 0001333444
   4 : 5578899
   5 : 0134
   5 : 689
   6 : 01

High:  6321  6558  6943  6948  7026  7362  8046  9642
```

– Dietary retinol: right skewed wtih 6 high outliers

```
> summary(micro2$retdiet)
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
   30.0  479.5   707.0  831.0 1027.0  6901.0
> stem(micro2$retdiet)
N = 314   Median = 707
Quartiles = 479, 1027

Decimal point is 2 places to the right of the colon

    0 : 38
    1 : 1244668
    2 : 112334444678999
    3 : 01111222233455778999
    4 : 0012222223333444444455555555666778888999999
    5 : 0011222223445555558889999
    6 : 00000112333334444455556666667777888899999
    7 : 001111222333344555667888999
    8 : 0011122234445556666667899
    9 : 00111223334445566677778889
   10 : 011335555566778
   11 : 0000122256788
   12 : 001112556689
   13 : 1123334467
   14 : 0456788
   15 : 23346789
   16 : 223
   17 : 06
   18 : 12244
   19 :
   20 :
   21 : 02
```

```
High:  2167  2263  2571  3061  4041  6901
```

- Outcomes:
  - Plasma beta-carotene: very right skewed with 10 high outliers

```
> summary(micro2$betaplasma)
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
    0.0   89.5   140.0  190.2  230.5  1415.0
> stem(micro2$betaplasma)
N = 314   Median = 140
Quartiles = 89, 231

Decimal point is 2 places to the right of the colon

   0 : z1
   0 : 2222223333
   0 : 444444444455555555
   0 : 66666677777777777777
   0 : 8888888888888888889999999999999999999
   1 : 00000000000000000001111111111111111
   1 : 222222222222222222333333333333333
   1 : 44444444445555555555555
   1 : 66666666666666677777777777777
   1 : 888888899999
   2 : 000001111111
   2 : 222222333333
   2 : 44445555
   2 : 6677777
   2 : 8889999999
   3 : 00000
   3 : 2222233333
   3 : 45
   3 : 6777
   3 : 889
   4 : 011
   4 : 222333
   4 : 5
   4 : 67
   4 : 999

High:   613   751   786   800   858   990  1010  1212  1391  1415
```

  - Plasma retinol: fairly symmetric except for 5 high outliers

```
> summary(micro2$retplasma)
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
```

```
  179.0  467.0   566.0  603.7  717.5  1727.0
> stem(micro2$retplasma)
N = 314   Median = 566
Quartiles = 466, 719

Decimal point is 2 places to the right of the colon

    1 : 899
    2 : 23
    2 : 556899
    3 : 0022334
    3 : 56667777888899999
    4 : 0000000111112222222333333333444444
    4 : 5566666777777777888889999999
    5 : 00000000111112222222222222333333334444444
    5 : 5555566666666666667777777888999999999
    6 : 000000111122222222233333444
    6 : 555555666677788888889999
    7 : 0000011112222333333444
    7 : 5556666778888999
    8 : 0000112222222233344
    8 : 55556788
    9 : 0012333
    9 : 55999
   10 : 033
   10 :
   11 : 04
   11 : 9

High:  1249  1262  1443  1517  1727
```

Check for normality on continuous variables.

```
> par(mfrow=c(4,3))
> f.qqenv(micro2$age)
> f.qqenv(micro2$quetelet)
> f.qqenv(micro2$calories)
> f.qqenv(micro2$fat)
> f.qqenv(micro2$percentfat)
> f.qqenv(micro2$fiber)
> f.qqenv(micro2$alcohol)
> f.qqenv(micro2$cholesterol)
> f.qqenv(micro2$betadiet)
> f.qqenv(micro2$retdiet)
> f.qqenv(micro2$betaplasma)
> f.qqenv(micro2$retplasma)
```

Bivariate analyses

Take a quick look at bivariate distributions; use as a reference later.

```
f.mypairs(micro2)
```

Both outcome variables are right skewed; further analyses will work with the log transformed variables.

```
> par(mfrow=c(2,2), mai=c(.7,.7,.7,.2))
> f.qqenv(micro2$retplasma, ylab="Plasma Retinol", cex=1.2)
> f.qqenv(log(micro2$retplasma), ylab="Log Plasma Retinol", cex=1.2)
> f.qqenv(micro2$betaplasma, ylab="Plasma Beta-carotene", cex=1.2)
> f.qqenv(log(micro2$betaplasma), ylab="Log Plasma Beta-carotene", cex=1.2)
> micro2$logretplasma <- log(micro2$retplasma)
> micro2$logbetaplasma <- log(micro2$betaplasma)
```

Retinol

It was suggested that alcohol is related to plasma retinol levels. Is this true?

```
> summary(lm(logretplasma~alcohol, data=micro2))
...
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   6.3097   0.0213   296.5836   0.0000
    alcohol   0.0140   0.0038     3.6999   0.0003

Residual standard error: 0.3324 on 312 degrees of freedom
Multiple R-Squared: 0.04203
F-statistic: 13.69 on 1 and 312 degrees of freedom, the p-value is 0.0002549
...
> plot(micro2$alcohol, micro2$logretplasma)
> abline(lm(logretplasma~alcohol, data=micro2))
> par(mfrow=c(2,3))
> plot(lm(logretplasma~alcohol, data=micro2))
```

Residual plots show major violation of the assumption of constant variance. Since so many subjects do not drink at all, taking the log would not be appropriate. Instead, I categorize their drinking as none, medium, or high. This is effectively throwing away some of the data, but I think it's more clinically relevant anyway.

```
> micro2$alcCat <- rep("00",314)    # No drinks gets category ‘‘00’’
> micro2$alcCat <- ifelse(micro2$alcohol>0 & micro2$alcohol<=10,
                          "Lo",micro2$alcCat)     # 0-10 gets ‘‘Lo’’
> micro2$alcCat <- ifelse(micro2$alcohol>10,"Hi",micro2$alcCat)# >10 gets ‘‘Hi’’
> micro2$alcCat <- as.factor(micro2$alcCat)
> #Creation of dummy variables
```

```
> alc00 <- ifelse(micro2$alcCat=="00", 1, 0)
> alcLo <- ifelse(micro2$alcCat=="Lo", 1, 0)
> alcHi <- ifelse(micro2$alcCat=="Hi", 1, 0)
> summary(lm(logretplasma~alcLo+alcHi, data=micro2))
...
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   6.2855   0.0318   197.7680   0.0000
      alcLo   0.0776   0.0404     1.9199   0.0558
      alcHi   0.2298   0.0767     2.9955   0.0030

Residual standard error: 0.3348 on 311 degrees of freedom
Multiple R-Squared: 0.03106
F-statistic: 4.985 on 2 and 311 degrees of freedom, the p-value is 0.007397
...
> micro2$alcCat <- ordered(micro2$alcCat, levels=c("00", "Lo", "Hi"),
                           labels=c("None", "Moderate", "High"))
> par(mai=c(.7,.7,.7,.2))
> boxplot(split(micro2$logretplasma,micro2$alcCat), bxp.style="old",
          xlab="Alcohol Consumption", ylab="log Plasma Retinol",
          cex=1.4)
```

Drinking heavily actually increases the average log Plasma Retinol levels.

Beta-carotene

It was suggested that BMI, cholesterol, calories, vitamin use, and fiber are related to plasma beta-carotene. Is this true?

- BMI

```
> summary(lm(logbetaplasma~quetelet, data=micro2))
Problem in dimnames(rinv) <- list(cnames, cnames): Component 1 of dimnames has length 2, should be 0
#Problem is that 264 had 0 for betaplasma: omit
> micro3 <- micro2[-256,]
> summary(lm(micro3$logbetaplasma~micro3$quetelet))
...
Coefficients:
                   Value Std. Error  t value Pr(>|t|)
    (Intercept)   5.8737   0.1815    32.3692   0.0000
micro3$quetelet  -0.0349   0.0068    -5.1622   0.0000

Residual standard error: 0.7196 on 311 degrees of freedom
Multiple R-Squared: 0.07892
F-statistic: 26.65 on 1 and 311 degrees of freedom, the p-value is 4.362e-07
...
> plot(micro3$quetelet,micro3$logbetaplasma)
```

```
> abline(lm(micro3$logbetaplasma~micro3$quetelet))
> plot(micro3$quetelet, micro3$betaplasma)
> lines(sort(micro3$quetelet),
        exp(fitted(lm(micro3$logbetaplasma~micro3$quetelet)))
        [order(micro3$quetelet)])
> par(mfrow=c(2,3))
> plot(lm(micro3$logbetaplasma~micro3$quetelet))
```

The residual plots show problems with non-constant variance. Does it improve the model to take the log(BMI)?

```
> par(mfrow=c(1,2))
> f.qqenv(micro3$quetelet)
> f.qqenv(log(micro3$quetelet))
> summary(lm(micro3$logbetaplasma~log(micro3$quetelet)))
...
Coefficients:
                       Value Std. Error  t value Pr(>|t|)
         (Intercept)   8.2508   0.6290    13.1166   0.0000
log(micro3$quetelet)  -1.0152   0.1937    -5.2412   0.0000

Residual standard error: 0.7187 on 311 degrees of freedom
Multiple R-Squared: 0.08116
F-statistic: 27.47 on 1 and 311 degrees of freedom, the p-value is
2.95e-07
...
> plot(log(micro3$quetelet),micro3$logbetaplasma)
> abline(lm(micro3$logbetaplasma~log(micro3$quetelet)))
> par(mai=c(.7,.7,.7,.2))
> plot(micro3$quetelet, micro3$betaplasma, xlab="Quetelet Index",
       ylab="Plasma Beta-carotene (ng/ml)")
> lines(sort(micro3$quetelet), exp(fitted(lm(micro3$logbetaplasma~log(micro3$quetelet))))
        [order(log(micro3$quetelet))], cex=1.4)
> par(mfrow=c(2,3))
> plot(lm(micro3$logbetaplasma~log(micro3$quetelet)))
```

Now the variance in the residual plots looks more homogeneous.

- Cholesterol

```
> summary(lm(micro3$logbetaplasma~micro3$cholesterol))
...
Coefficients:
                     Value Std. Error  t value Pr(>|t|)
       (Intercept)   5.1831   0.0903    57.4029   0.0000
micro3$cholesterol  -0.0009   0.0003    -2.7781   0.0058
```

```
Residual standard error: 0.7406 on 311 degrees of freedom
Multiple R-Squared: 0.02421
F-statistic: 7.718 on 1 and 311 degrees of freedom, the p-value is
0.005801
...
> plot(micro3$cholesterol,micro3$logbetaplasma)
> abline(lm(micro3$logbetaplasma~micro3$cholesterol))
> plot(micro3$cholesterol, micro3$betaplasma)
> lines(sort(micro3$cholesterol),
        exp(fitted(lm(micro3$logbetaplasma~micro3$cholesterol)))
        [order(micro3$cholesterol)])
> par(mfrow=c(2,3))
> plot(lm(micro3$logbetaplasma~micro3$cholesterol))
```

Again, problems with non-constant variance in the residuals. Try taking the log transformation of cholesterol.

```
> par(mfrow=c(1,2))
> f.qqenv(micro3$cholesterol)
> f.qqenv(log(micro3$cholesterol))
> summary(lm(micro3$logbetaplasma~log(micro3$cholesterol)))
...
Coefficients:
                           Value Std. Error  t value Pr(>|t|)
            (Intercept)   5.7863   0.4329    13.3666   0.0000
log(micro3$cholesterol)  -0.1544   0.0806    -1.9161   0.0563

Residual standard error: 0.7454 on 311 degrees of freedom
Multiple R-Squared: 0.01167
F-statistic: 3.671 on 1 and 311 degrees of freedom, the p-value is
0.05627
...
> plot(log(micro3$cholesterol),micro3$logbetaplasma)
> abline(lm(micro3$logbetaplasma~log(micro3$cholesterol)))
> par(mai=c(.7,.7,.7,.2))
> plot(micro3$cholesterol, micro3$betaplasma,
       xlab="Cholesterol consumed (mg/day)",
       ylab="Plasma Beta-carotene (ng/ml)'', cex=1.4))
> lines(sort(micro3$cholesterol),
        exp(fitted(lm(micro3$logbetaplasma~log(micro3$cholesterol))))
        [order(log(micro3$cholesterol))])
```

Fewer problems with heteroscedasticity, but R-squared goes down.

- Calories

```
> summary(lm(logbetaplasma~calories, data=micro3))
...
Coefficients:
                  Value Std. Error  t value Pr(>|t|)
    (Intercept)   5.0729   0.1286    39.4468   0.0000
micro3$calories  -0.0001   0.0001    -0.9232   0.3566

Residual standard error: 0.7487 on 311 degrees of freedom
Multiple R-Squared: 0.002733
F-statistic: 0.8522 on 1 and 311 degrees of freedom, the p-value is
0.3566
...
> plot(micro3$calories,micro3$logbetaplasma)
> abline(lm(logbetaplasma~calories, data=micro3))
```

There doesn't seem to be a relationship here. Maybe it's different for men and women?

```
> summary(lm(logbetaplasma~sex*calories, data=micro3))
...
Coefficients:
               Value Std. Error  t value Pr(>|t|)
 (Intercept)   5.2229   0.2238    23.3419   0.0000
         sex  -0.2349   0.2238    -1.0499   0.2946
    calories  -0.0002   0.0001    -1.6893   0.0922
sex:calories   0.0002   0.0001     1.7498   0.0811

Residual standard error: 0.7416 on 309 degrees of freedom
Multiple R-Squared: 0.02806
F-statistic: 2.973 on 3 and 309 degrees of freedom, the p-value is
0.03197
...
> par(mai=c(.7,.7,.7,.2))
> plot(micro3$calories,micro3$logbetaplasma, type="n",
       xlab="Calories consumed per day",
       ylab="Log Plasma Beta-carotene (log ng/ml)", cex=1.4)
> text(micro3$calories,micro3$logbetaplasma, labels=as.vector(micro3$sex))
> abline(5.2229,-0.0002,lty=1)
> abline(5.2227,0,lty=2)
> legend(locator(1), c("0=Males", "1=Females"), lty=1:2)
> par(mfrow=c(2,3))
> plot(lm(micro3$logbetaplasma~micro3$calories*micro3$sex))
```

The effects are only significant at about .1, but there is a suggestion that calories might matter for males but not females. This would be something to look at in future studies.

- Vitamin use

```
> VitHi <- ifelse(micro3$vituse==1, 1, 0)
> VitMed <- ifelse(micro3$vituse==2, 1, 0)
> VitNo <- ifelse(micro3$vituse==3, 1, 0)
> summary(lm(micro3$logbetaplasma~VitHi+VitMed))
...
Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  4.7178  0.0693    68.0570  0.0000
      VitHi  0.4308  0.0958     4.4978  0.0000
     VitMed  0.2920  0.1061     2.7526  0.0063

Residual standard error: 0.727 on 310 degrees of freedom
Multiple R-Squared: 0.06269
F-statistic: 10.37 on 2 and 310 degrees of freedom, the p-value is
4.384e-05
...
> boxplot(split(micro3$logbetaplasma,micro3$vituse),
          bxp.style="old", xlab="Vitamin Use",
          ylab="log Plasma Beta Carotene")
> par(mfrow=c(2,3))
> plot(lm(micro3$logbetaplasma~VitHi+VitMed))
```

Taking vitamins is associated with a small but significant raising of log plasma beta-carotene levels.

- Fiber

```
> summary(lm(logbetaplasma~fiber, data=micro3))
...
Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  4.5314  0.1068    42.4214  0.0000
      fiber  0.0336  0.0077     4.3564  0.0000

Residual standard error: 0.7279 on 311 degrees of freedom
Multiple R-Squared: 0.05751
F-statistic: 18.98 on 1 and 311 degrees of freedom, the p-value is
1.798e-05
...
> plot(micro3$fiber,micro3$logbetaplasma)
> abline(lm(logbetaplasma~fiber, data=micro3))
> par(mai=c(.7,.7,.7,.2))
> plot(micro3$fiber, micro3$betaplasma,
       xlab="Fiber consumed (g/day)", ylab="Plasma Beta-carotene (ng/ml)",
```

```
       cex=1.4)
> lines(sort(micro3$fiber),
        exp(fitted(lm(logbetaplasma~fiber, data=micro3)))
        [order(micro3$fiber)])
> par(mfrow=c(2,3))
> plot(lm(logbetaplasma~fiber, data=micro3))
```

Fiber is also a significant predictor by itself, tending to increase log beta-carotene.

Model selection

Create data subsets, one for model building and one for model testing.

```
> sample <- sample(314,200)
> modeldata <- micro2[sample,]
> testdata <- micro2[-sample,]
```

Retinol

Begin with what we know, which is that alcohol has an effect, and build from there.

```
> alc00 <- ifelse(modeldata$alcCat=="00", 1, 0)
> alcLo <- ifelse(modeldata$alcCat=="Lo", 1, 0)
> alcHi <- ifelse(modeldata$alcCat=="Hi", 1, 0)
> summary(lm(logretplasma~alcLo+alcHi, data=modeldata))
...
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   6.2505   0.0415   150.4989   0.0000
      alcLo   0.1038   0.0532     1.9505   0.0525
      alcHi   0.2139   0.1068     2.0021   0.0466

Residual standard error: 0.3548 on 197 degrees of freedom
Multiple R-Squared: 0.02953
F-statistic: 2.997 on 2 and 197 degrees of freedom, the p-value is 0.05221
...
```

We will go through variables which might plausibly be related to plasma retinol (based on prior knowledge and the pairs plot) and try adding these terms to the model, testing with nested F-tests.

- Age - add it to the model

```
> anova(lm(logretplasma~alcLo+alcHi+age, data=modeldata),
        lm(logretplasma~alcLo+alcHi,     data=modeldata))
                Terms Resid. Df      RSS Test Df Sum of Sq  F Value
```

```
1 alcLo + alcHi + age       196 22.93300
2       alcLo + alcHi       197 24.80587 -age -1  -1.87287 16.00674

         Pr(F)
1
2 8.942857e-05
> summary(lm(logretplasma~alcLo+alcHi+age, data=modeldata))
...
Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  5.8990  0.0966    61.0898  0.0000
      alcLo  0.1357  0.0519     2.6155  0.0096
      alcHi  0.2371  0.1031     2.2989  0.0226
        age  0.0067  0.0017     4.0008  0.0001

Residual standard error: 0.3421 on 196 degrees of freedom
Multiple R-Squared: 0.1028
F-statistic: 7.486 on 3 and 196 degrees of freedom, the p-value is 9.053e-05
...
```

- Sex - don't add

```
> anova(lm(logretplasma~alcLo+alcHi+age,     data=modeldata),
        lm(logretplasma~alcLo+alcHi+age+sex, data=modeldata))
                      Terms Resid. Df      RSS Test Df  Sum of Sq   F Value
1       alcLo + alcHi + age       196 22.93300
2 alcLo + alcHi + age + sex       195 22.83534 +sex  1 0.09765896 0.8339485

      Pr(F)
1
2 0.3622609
```

- Smoking - don't add

```
> anova(lm(logretplasma~alcLo+alcHi+age,          data=modeldata),
        lm(logretplasma~alcLo+alcHi+age+smokstat, data=modeldata))
                           Terms Resid. Df      RSS      Test Df Sum of Sq
1            alcLo + alcHi + age       196 22.93300
2 alcLo + alcHi + age + smokstat       194 22.60996 +smokstat  2 0.3230357

  F Value    Pr(F)
1
2 1.38587 0.252571
```

- BMI - don't add

```
> anova(lm(logretplasma~alcLo+alcHi+age,          data=modeldata),
        lm(logretplasma~alcLo+alcHi+age+quetelet, data=modeldata))
                         Terms Resid. Df      RSS      Test Df Sum of Sq
1            alcLo + alcHi + age       196 22.93300
2 alcLo + alcHi + age + quetelet       195 22.86551 +quetelet  1 0.0674856

    F Value     Pr(F)
1
2 0.5755258 0.4489874
```

- Vitamin use - don't add

```
> anova(lm(logretplasma~alcLo+alcHi+age,        data=modeldata),
        lm(logretplasma~alcLo+alcHi+age+vituse, data=modeldata))
                       Terms Resid. Df      RSS    Test Df  Sum of Sq
1          alcLo + alcHi + age       196 22.93300
2 alcLo + alcHi + age + vituse       194 22.89306 +vituse  2 0.03993893

    F Value     Pr(F)
1
2 0.1692249 0.8444435
```

- Calories - don't add

```
> anova(lm(logretplasma~alcLo+alcHi+age,          data=modeldata),
        lm(logretplasma~alcLo+alcHi+age+calories, data=modeldata))
                         Terms Resid. Df      RSS      Test Df  Sum of Sq
                         Terms Resid. Df      RSS      Test Df  Sum of Sq
1            alcLo + alcHi + age       196 22.93300
2 alcLo + alcHi + age + calories       195 22.89754 +calories  1 0.03545454

    F Value     Pr(F)
1
2 0.3019378 0.5832989
```

- Fiber - don't add

```
> anova(lm(logretplasma~alcLo+alcHi+age,       data=modeldata),
        lm(logretplasma~alcLo+alcHi+age+fiber, data=modeldata))
                      Terms Resid. Df      RSS   Test Df Sum of Sq   F Value
1         alcLo + alcHi + age       196 22.93300
2 alcLo + alcHi + age + fiber       195 22.82896 +fiber  1 0.1040338 0.8886339

      Pr(F)
1
2 0.3470142
```

- Cholesterol - don't add

```
> anova(lm(logretplasma~alcLo+alcHi+age,             data=modeldata),
        lm(logretplasma~alcLo+alcHi+age+cholesterol, data=modeldata))
                              Terms Resid. Df     RSS         Test Df
1               alcLo + alcHi + age       196 22.9330
2 alcLo + alcHi + age + cholesterol       195 22.8978 +cholesterol  1

   Sum of Sq   F Value     Pr(F)
1
2 0.03519562 0.2997295 0.5846781
```

- Percent fat - don't add

```
> anova(lm(logretplasma~alcLo+alcHi+age,            data=modeldata),
        lm(logretplasma~alcLo+alcHi+age+percentfat, data=modeldata))
                             Terms Resid. Df      RSS        Test Df
1              alcLo + alcHi + age       196 22.93300
2 alcLo + alcHi + age + percentfat       195 22.85317 +percentfat  1

   Sum of Sq   F Value     Pr(F)
1
2 0.07983222 0.6811871 0.4101868
```

- Dietary retinol - don't add

```
> anova(lm(logretplasma~alcLo+alcHi+age,         data=modeldata),
        lm(logretplasma~alcLo+alcHi+age+retdiet, data=modeldata))
                          Terms Resid. Df      RSS     Test Df  Sum of Sq
1           alcLo + alcHi + age       196 22.93300
2 alcLo + alcHi + age + retdiet       195 22.88782 +retdiet  1 0.04518367

    F Value     Pr(F)
1
2 0.3849566 0.535686
```

Check the age and alcohol model against the results of stepwise regression.

```
> modeldata$alcCat <- as.factor(modeldata$alcCat)
> f.sum.step(stepwise(modeldata[,c(1:6,8,10,16,19)],
                      modeldata$logretplasma),
             data=modeldata, y=modeldata$logretplasma)
         BIC   R2a age sex smokstat quetelet vituse calories fiber cholesterol
1(+ 1) -77.2 0.055   1   0        0        0      0        0     0           0
2(+10) -77.8 0.089   1   0        0        0      0        0     0           0
```

```
       percentfat alcCat
1(+ 1)          0      0
2(+10)          0      1
```

The two variable model performs well; on the test data, see how this compares to a model that doesn't include alcCat, only age.

Beta-carotene

Begin with a model based on the bivariate predictors: log BMI, log cholesterol, vitamin use, fiber.

```
> modeldata2 <- modeldata[-47,]      #get rid of infinite value
> modeldata2$logquetelet <- log(modeldata2$quetelet)
> modeldata2$logcholesterol <- log(modeldata2$cholesterol)
> summary(lm(logbetaplasma~logquetelet+logcholesterol+vituse+fiber,
          data=modeldata2))
...
Coefficients:
                 Value Std. Error t value Pr(>|t|)
   (Intercept)  8.4287  0.8674     9.7170  0.0000
   logquetelet -0.9457  0.2314    -4.0865  0.0001
logcholesterol -0.1615  0.0961    -1.6808  0.0944
       vituse1 -0.0101  0.0611    -0.1646  0.8694
       vituse2 -0.0962  0.0351    -2.7416  0.0067
         fiber  0.0309  0.0094     3.2925  0.0012

Residual standard error: 0.6871 on 193 degrees of freedom
Multiple R-Squared: 0.1875
F-statistic: 8.906 on 5 and 193 degrees of freedom, the p-value is 1.257e-07
...
```

Now try adding terms to the model

- Age - don't add

```
> anova(lm(logbetaplasma~logquetelet+logcholesterol+vituse+fiber,
           data=modeldata2),
        lm(logbetaplasma~logquetelet+logcholesterol+vituse+fiber+age,
           data=modeldata2))
                                                 Terms Resid. Df      RSS Test
1       logquetelet + logcholesterol + vituse + fiber       193 91.10786
2 logquetelet + logcholesterol + vituse + fiber + age       192 90.04722 +age

  Df Sum of Sq F Value     Pr(F)
1
2  1  1.060639 2.26151 0.1342678
```

- Smoking - add to model

```
> anova(lm(logbetaplasma~logquetelet+logcholesterol++vituse+fiber,
           data=modeldata2),
        lm(logbetaplasma~logquetelet+logcholesterol+vituse+fiber+smokstat,
           data=modeldata2))
                                                    Terms Resid. Df      RSS
1          logquetelet + logcholesterol +  + vituse + fiber       193 91.10786
2 logquetelet + logcholesterol + vituse + fiber + smokstat       191 87.62320

       Test Df Sum of Sq  F Value      Pr(F)
1
2 +smokstat  2  3.484659 3.797909 0.02412915
```

- Percent fat - don't add

```
> anova(lm(logbetaplasma~logquetelet+logcholesterol+vituse+fiber+smokstat,
        data=modeldata2),
        lm(logbetaplasma~logquetelet+logcholesterol+vituse+fiber+smokstat+percentfat,
        data=modeldata2))
                                                                 Terms
1             logquetelet + logcholesterol + vituse + fiber + smokstat
2 logquetelet + logcholesterol + vituse + fiber + smokstat + percentfat

  Resid. Df      RSS        Test Df   Sum of Sq    F Value     Pr(F)
1       191 87.62320
2       190 87.61369 +percentfat  1 0.009509494 0.02062239 0.8859646
```

- Dietary Beta-carotene - don't add

```
> anova(lm(logbetaplasma~logquetelet+logcholesterol+vituse+fiber+smokstat,
           data=modeldata2),
        lm(logbetaplasma~logquetelet+logcholesterol+vituse+fiber+smokstat+betadiet,
           data=modeldata2))
                                                               Terms
1            logquetelet + logcholesterol + vituse + fiber + smokstat
2 logquetelet + logcholesterol + vituse + fiber + smokstat + betadiet

  Resid. Df      RSS      Test Df  Sum of Sq   F Value     Pr(F)
1       191 87.62320
2       190 87.52895 +betadiet  1 0.09425375 0.2045976 0.6515511
```

- Alcohol category - don't add

```
> modeldata2$alcCat <- as.factor(modeldata2$alcCat)
> anova(lm(logbetaplasma~logquetelet+logcholesterol+vituse+fiber+smokstat,
```

```
              data=modeldata2),
            lm(logbetaplasma~logquetelet+logcholesterol+vituse+fiber+smokstat+alcCat,
              data=modeldata2))
                                                                  Terms Resid. Df
1         logquetelet + logcholesterol + vituse + fiber + smokstat       191
2 logquetelet + logcholesterol + vituse + fiber + smokstat + alcCat       189

       RSS    Test Df Sum of Sq  F Value    Pr(F)
1 87.62320
2 85.94156 +alcCat  2  1.681644 1.849109 0.160213
```

So we now have a model that includes log BMI, log cholesterol, vitamin use, fiber, and smoking status. Let's check this against the restults of stepwise again. We'll look at the top ten models in terms of BIC.

```
> f.sum.step(stepwise(modeldata2[,c(1:3,20,5,6,8,21,11,16,19)],
                      modeldata2$logbetaplasma, method="exhaustive"),
            data=modeldata2, y=modeldata2$logbetaplasma)
          BIC   R2a age sex smokstat logquetelet vituse calories fiber
 2(#2) -218.1 0.132   0   0        0           1      0        0     1
 4(#2) -218.3 0.185   0   0        0           1      1        1     1
 3(#1) -218.6 0.165   0   0        1           1      0        0     1
 4(#3) -218.9 0.180   0   0        1           1      0        1     1
 2(#1) -219.1 0.142   0   0        1           1      0        0     0
 4(#1) -220.3 0.186   0   0        1           1      1        0     1
 1(#1) -220.3 0.093   0   0        0           1      0        0     0
 5(#1) -220.5 0.202   0   0        1           1      1        1     1
 3(#2) -220.7 0.166   0   0        1           1      1        0     0
 5(#2) -221.2 0.197   0   0        0           1      1        1     1
...
       logcholesterol betadiet percentfat alcCat
 2(#2)              0        0          0      0
 4(#2)              0        0          0      0
 3(#1)              0        0          0      0
 4(#3)              0        0          0      0
 2(#1)              0        0          0      0
 4(#1)              0        0          0      0
 1(#1)              0        0          0      0
 5(#1)              0        0          0      0
 3(#2)              0        0          0      0
 5(#2)              0        0          0      1
...
```

- log BMI - shows up in all the models; keep it
- log cholesterol - doesn't show up in any of the models; drop it???
- vitamin use - shows up in half the models; keep it

- fiber - shows up in 7 of the models; keep it
- smoking status - shows up in 6 of the models; keep it

So the only thing stepwise causes us to reconsider is log cholesterol, which wasn't a very good predictor to begin with. Let's examine it again.

```
> anova(lm(logbetaplasma~logquetelet+vituse+fiber+smokstat,
           data=modeldata2),
        lm(logbetaplasma~logquetelet+vituse+fiber+smokstat+cholesterol,
           data=modeldata2))
                                                  Terms Resid. Df      RSS
1               logquetelet + vituse + fiber + smokstat       192 88.46467
2 logquetelet + vituse + fiber + smokstat + cholesterol       191 86.57923

          Test Df Sum of Sq  F Value      Pr(F)
1
2 +cholesterol  1  1.885436 4.159408 0.04278121
> anova(lm(logbetaplasma~logquetelet+vituse+fiber+smokstat,
           data=modeldata2),
        lm(logbetaplasma~logquetelet+vituse+fiber+smokstat+logcholesterol,
           data=modeldata2))
                                                     Terms Resid. Df      RSS
1                  logquetelet + vituse + fiber + smokstat       192 88.46467
2 logquetelet + vituse + fiber + smokstat + logcholesterol       191 87.62320

             Test Df Sum of Sq  F Value     Pr(F)
1
2 +logcholesterol  1 0.8414687 1.834223 0.1772296
```

This is the same issue that came up in the bivariate analyses. Cholesterol is a significant predictor, but log cholesterol is not. Let's look at some partial regression plots.

```
> y.lm <- lm(logbetaplasma~logquetelet+vituse+fiber+smokstat, data=modeldata2)
> x.lm <- lm(cholesterol~logquetelet+vituse+fiber+smokstat, data=modeldata2)
> plot(x.lm$residuals,y.lm$residuals)
> abline(lm(y.lm$residuals~x.lm$residuals))
> y.lm <- lm(logbetaplasma~logquetelet+vituse+fiber+smokstat, data=modeldata2)
> x.lm <- lm(logcholesterol~logquetelet+vituse+fiber+smokstat, data=modeldata2)
> plot(x.lm$residuals,y.lm$residuals,
       ylab="Residuals from logbetaplasma~logquetelet+vituse+fiber+smokstat",
       xlab="Residuals from logcholesterol~logquetelet+vituse+fiber+smokstat")
> abline(lm(y.lm$residuals~x.lm$residuals))
```

Neither one looks like it is adding very much information. My inclination based on this test is to drop it from the model, but I will wait and see how a model including it compares against the baseline model (log BMI, vitamin use, fiber, smoking status) when predicting the test data.

Model validation

Retinol

We return to the question of whether a model including both age and alcohol consumption category is better than a model including age only.

```
> alc00 <- ifelse(modeldata$alcCat=="00", 1, 0)
> alcLo <- ifelse(modeldata$alcCat=="Lo", 1, 0)
> alcHi <- ifelse(modeldata$alcCat=="Hi", 1, 0)
>
> ret1.lm <- lm(logretplasma~alcLo+alcHi+age, data=modeldata)
> ret2.lm <- lm(logretplasma~age,             data=modeldata)
>
> rm(alc00,alcLo,alcHi)
>
> alc00 <- ifelse(testdata$alcCat=="00", 1, 0)
> alcLo <- ifelse(testdata$alcCat=="Lo", 1, 0)
> alcHi <- ifelse(testdata$alcCat=="Hi", 1, 0)
>
> ret.newdata <- as.data.frame(cbind(alcLo,alcHi,testdata$age))
> names(ret.newdata) <- c("alcLo","alcHi","age")
>
> ret1.fitted <- predict(ret1.lm, newdata=ret1.newdata)
> ret2.fitted <- predict(ret2.lm, newdata=ret1.newdata)
>
> par(mfrow=c(1,2), mai=c(.7,.7,.7,.2))
> plot(testdata$logretplasma, ret1.fitted, xlim=c(5.5,7.5),
       ylim=c(5.5,7.5), xlab="Actual Log Plasma Retinol",
       ylab="Predicted Log Plasma Retinol", cex=1.4)
> abline(0,1)
> title(main="Model with age and alcohol")
> plot(testdata$logretplasma, ret2.fitted, xlim=c(5.5,7.5),
       ylim=c(5.5,7.5), xlab="Actual Log Plasma Retinol",
       ylab="Predicted Log Plasma Retinol", cex=1.4)
> abline(0,1)
> title(main="Model with age only")
> par(mfrow=c(1,1))
> legend(locator(1), "Perfect Prediction", lty=1, cex=1.4)
```

Looking at this plot, including alcohol category seems to improve the prediction slightly. Let's look at the residuals from the prediction.

```
> ret1.resid <- ret1.fitted - testdata$logretplasma
> ret2.resid <- ret2.fitted - testdata$logretplasma
>
> par(mfrow=c(2,2))
> boxplot(ret1.resid, bxp.style="old")
```

```
> f.qqenv(ret1.resid)
> plot(ret1.fitted, ret1.resid)
> abline(0,0)
> plot(testdata$logretplasma, ret1.fitted, xlim=c(5.5,7.5), ylim=c(5.5,7.5))
> abline(0,1)
> par(mfrow=c(1,1))
> title(main="Model with age and alcohol")
>
> par(mfrow=c(2,2))
> boxplot(ret2.resid, bxp.style="old")
> f.qqenv(ret2.resid)
> plot(ret2.fitted, ret2.resid)
> abline(0,0)
> plot(testdata$logretplasma, ret2.fitted, xlim=c(5.5,7.5), ylim=c(5.5,7.5))
> abline(0,1)
> par(mfrow=c(1,1))
> title(main="Model with age only")
```

The residuals look ok in both cases; including alcohol category decreases the range. Now we summarize the final model for retinol.

```
> summary(ret1.lm)
...
Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  5.8990  0.0966    61.0898  0.0000
      alcLo  0.1357  0.0519     2.6155  0.0096
      alcHi  0.2371  0.1031     2.2989  0.0226
        age  0.0067  0.0017     4.0008  0.0001

Residual standard error: 0.3421 on 196 degrees of freedom
Multiple R-Squared: 0.1028
F-statistic: 7.486 on 3 and 196 degrees of freedom, the p-value is 9.053e-05
...
> par(mfrow=c(2,3), mai=c(.7,.7,.7,.2))
> plot(ret1.lm)
```

Beta-carotene

Here the question is whether to include cholesterol as a predictor, or to use a model that includes only log BMI, vitamin use, fiber, and smoking status. We will compare the predictions of these two models.

```
> beta1.lm <- lm(logbetaplasma~logquetelet+vituse+fiber+smokstat,
                 data=modeldata2)
> beta2.lm <- lm(logbetaplasma~logquetelet+vituse+fiber+smokstat+cholesterol,
                 data=modeldata2)
```

```
>
> testdata$logquetelet <- log(testdata$quetelet)
> beta1.fitted <- predict(beta1.lm, newdata=testdata)
> beta2.fitted <- predict(beta2.lm, newdata=testdata)
>
> par(mfrow=c(1,2), mai=c(.7,.7,.7,.2))
> plot(testdata$logbetaplasma, beta1.fitted, xlim=c(3.5,7.5),
       ylim=c(3.5,7.5), xlab="Actual Log Plasma Beta-carotene",
       ylab="Predicted Log Plasma Beta-carotene", cex=1.4)
> abline(0,1)
> title(main="Model 1: log BMI, vituse, fiber, smoking")
> plot(testdata$logbetaplasma, beta2.fitted, xlim=c(3.5,7.5),
       ylim=c(3.5,7.5), xlab="Actual Log Plasma Beta-carotene",
       ylab="Predicted Log Plasma Beta-carotene", cex=1.4)
> abline(0,1)
> title(main="Model 1 + cholesterol")
> par(mfrow=c(1,1))
> legend(locator(1), "Perfect Prediction", lty=1, cex=1.4)
```

The predictions are virtually indistinguishable. Let's look at the residuals.

```
> beta1.resid <- beta1.fitted - testdata$logbetaplasma
> beta2.resid <- beta2.fitted - testdata$logbetaplasma
>
> par(mfrow=c(2,2))
> boxplot(beta1.resid, bxp.style="old")
> f.qqenv(beta1.resid)
> plot(beta1.fitted, beta1.resid)
> abline(0,0)
> plot(testdata$logbetaplasma, beta1.fitted, xlim=c(3.5,7.5), ylim=c(3.5,7.5))
> abline(0,1)
> par(mfrow=c(1,1))
> title(main="Model with BMI, smoking, fiber")
>
> par(mfrow=c(2,2))
> boxplot(beta2.resid, bxp.style="old")
> f.qqenv(beta2.resid)
> plot(beta2.fitted, beta2.resid)
> abline(0,0)
> plot(testdata$logbetaplasma, beta2.fitted, xlim=c(3.5,7.5), ylim=c(3.5,7.5))
> abline(0,1)
> par(mfrow=c(1,1))
> title(main="Model with BMI, smoking, fiber, vitamins, and cholesterol")
```

The range of the residuals is very slightly less when we include cholesterol, but I don't think this is a convincing reason to include it in the model. So we take a look at the final model.

```
> vitMod <- ifelse(modeldata2$vituse==2, 1, 0)
> vitHi <- ifelse(modeldata2$vituse==1, 1, 0)
> smokeFormer <- ifelse(modeldata2$smokstat==2, 1, 0)
> smokeCurrent <- ifelse(modeldata2$smokstat==3, 1, 0)
>
> summary(lm(logbetaplasma~logquetelet+vitMod+vitHi+fiber+smokeFormer+
             smokeCurrent, data=modeldata2))
...
Coefficients:
                Value Std. Error  t value Pr(>|t|)
 (Intercept)   8.0089   0.7788    10.2837   0.0000
 logquetelet  -1.0670   0.2286    -4.6664   0.0000
      vitMod   0.2760   0.1256     2.1969   0.0292
       vitHi   0.2756   0.1149     2.3978   0.0175
       fiber   0.0228   0.0094     2.4339   0.0159
 smokeFormer  -0.1825   0.1041    -1.7535   0.0811
smokeCurrent  -0.4366   0.1563    -2.7932   0.0057

Residual standard error: 0.6788 on 192 degrees of freedom
Multiple R-Squared: 0.211
F-statistic: 8.56 on 6 and 192 degrees of freedom, the p-value is 2.996e-08
...
> par(mfrow=c(2,3), mai=c(.7,.7,.7,.2))
> plot(beta1.lm)
```

The same model analyses were run using a dataset that included subject number 62, who was an outlier in terms of both drinks per week and calories per day. This subject had considerable influence when alcohol was one of the independent variables, but did not change the significance of the other findings. The models without this subject are presented in this report because they are easier to interpret.

Bibliography

[1] Junker, Brian. *Determinants of Plasma Retinol and Beta-Carotene Levels*. Available at:

`http://stat.cmu.edu/~brian/707`