Predicting Caloric Intake from Fat and Fibre Consumption

Author: Leigh West

Exploratory Analysis

Figure 1 (below) depicts the pairs plot of age, calories consumed in a day, fat and fibre against each other. In evaluating the relationship between our response variable (calories) and our predictor variables, there is no apparent association between caloric intake and age. A strong (r = 0.87), positive, linear association exists between fat and caloric intake, and a moderate (r = 0.47), positive linear association exists between fibre consumption and caloric intake. The pairs plot suggests that a suitable model for predicting caloric intake may include fibre and fat consumption, but not age.

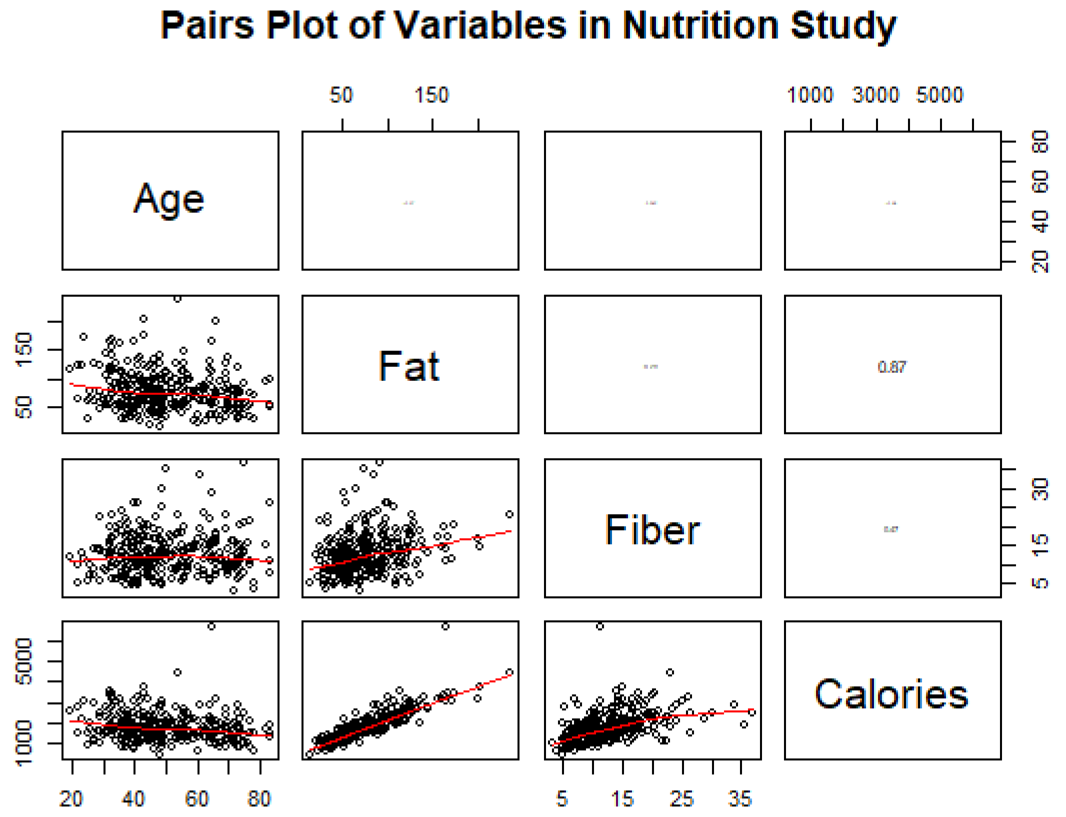

Figure 1: A pairs plot of the nutrition study variables: age, fat (grams), fibre (grams) and caloric intake per day.

With respect to the relationships between the predictor variables, there is no apparent association between age and fat, and age and fibre intake. A weak (r = 0.28), positive relationship exists between fat and fibre intake.

Using R Studio, a main effects model fitted to the form Calories = $\mu + \beta_{1\ \mathrm{Age}} + \beta_{2\ \mathrm{Fat}} + \beta_{3\ \mathrm{Fibre}} + \varepsilon$ is as follows:

Expected (Calories) = 285.81 – 2.49 x Age +15.97 x Fat + 31.69 x Fibre.

Assessing Evidence of Multicollinearity

The model is assessed for multicollinearity using the following indicators:

1. Significant correlations between pairs of independent variables in the model;
2. Nonsignificant t-tests for the majority (or all) of the predictor variables when the F-test for the model usability is significant
3. Negative signs in the estimated parameters where the opposite is expected (or vice versa); and
4. A variance inflation factor (VIF) for a predictor greater than 10.

From the pairs plot (Figure 1), a weak (r = 0.28) relationship exists between fat and fibre consumption. There are very weak to no associations between age and fat (r = -0.17) and age and fibre (r= 0.05).

We consider the following hypotheses with respect to the global F-test:

H_0: The true slope coefficients of each of the predictor variables is zero ($\beta_{1\ \mathrm{Age}} = \beta_{2\ \mathrm{Fat}} = \beta_{3\ \mathrm{Fibre}} = 0$).

H_a: The true slope coefficient of at least one of the predictor variables is different to zero.

From the Table of Regression Coefficients (Table 2), the F-statistic is 464 on 3 and 311 degrees of freedom, p-value = $2.2\text{x}10^{-16} < 0.05$. Hence, we have strong evidence to reject the null hypothesis in favour of the alternative, and conclude that at least one of our predictor variables is useful for predicting caloric intake. The adjusted R^2 is 0.8156, which suggests that 81.56% of the variability in caloric intake can be explained by this main effects model using three predictors. The t-tests of all three predictor variables are significant, with p = 0.03<0.05 for age, and $p < 2\text{x}10^{-16} < 0.05$ for both fat and fibre.

Table 1: Analysis of Variance Table (Main Effects Model)

```
Analysis of Variance Table

Response: Calories
           Df    Sum Sq   Mean Sq   F value     Pr(>F)
Age         1   4541563   4541563    53.214  2.502e-12
Fat         1 106059935 106059935  1242.722  < 2.2e-16
Fiber       1   8198249   8198249    96.060  < 2.2e-16
Residuals 311  26542259     85345
```

The signs of the estimated parameters (Table 2) are consistent with the signs of the correlation coefficients in the pairs plot (Figure 1). Table 2 also depicts the variance inflation factors (VIF) for the predictor variables. A VIF of greater than 10 is indicative of multicollinearity, however all three predictors have a VIF of approximately 1.

In summary, given there is only a weak correlation between two of our predictor variables (fat and fibre consumption), an absence of nonsignificant t-tests in the presence of a large F-statistic, that the signs of the estimated parameters are consistent with the respective signs of the correlation coefficients and the VIF for the predictor values are considerably smaller than 10, the model does not appear to have a multicollinearity problem.

Table 2: Table of Regression Coefficients (Main Effects Model)

```
Coefficients:
              Estimate  Std. Error  t value   Pr(>|t|)   Variance
                                                         Inflation
(Intercept)   285.8099     79.7921    3.582   0.000396          0
Age            -2.4886      1.1531   -2.158   0.031688   1.039318
Fat            15.9713      0.5165   30.925    < 2e-16   1.123079
Fiber          31.6932      3.2337    9.801    < 2e-16   1.093019

Residual standard error: 292.1 on 311 degrees of freedom
Multiple R-squared:  0.8174,   Adjusted R-squared:  0.8156
F-statistic:   464 on 3 and 311 DF,  p-value: < 2.2e-16
```

Variable Screening and Model Selection

Variable screening was performed in R using forward stepwise regression, commencing with the null model and the upper model set to include all main effects and possible interactions. Per Table 3, the null model produced an Akaike's Information Criteria (AIC) of 4,110, which was reduced to 3,579 in the final step.

Table 3: Forward Stepwise Regression (First vs Final Step)

```
First Step
Start:  AIC=4110.24
Calories ~ 1
        Df  Sum of Sq         RSS    AIC
+ Fat    1  110475565    34866442   3663
+ Fiber  1   31491594   113850412   4035
+ Age    1    4541563   140800444   4102
<none>                  145342006   4110
```

```
Final Step
Step:  AIC=3577.73
Calories ~ Fat + Fiber + Age + Fat:Fiber
           Df Sum of Sq        RSS    AIC
<none>                    26132365   3578
+Age:Fat    1     90622   26041743   3579
+Age:Fiber  1     83099   26049266   3579
```

Table 4: Analysis of Variance Table (Final Stepwise Model)

```
Analysis of Variance Table

Response: Calories
             Df    Sum Sq   Mean Sq   F value      Pr(>F)
Fat           1 110475565 110475565  1310.537      <2e-16
Fiber         1   7926709   7926709    94.032      <2e-16
Age           1    397473    397473     4.715      0.0307
Fat:Fiber     1    409894    409894     4.862      0.0282
Residuals   310  26132365
```

Table 5: Table of Regression Coefficients and Confidence Intervals (Final Stepwise Model)

```
Coefficients:
              Estimate  Std. Error  t value   Pr(>|t|)       2.5%    97.5%
(Intercept)    64.0944    128.0561     0.50      0.617   -187.875  316.063
Fat            18.9977      1.4653    12.97     <2e-16     16.115   21.881
Fiber          48.0486      8.0834     5.94    7.5e-09     32.143   63.953
Age            -2.4423      1.1462    -2.13      0.034     -4.698   -0.187
Fat:Fiber      -0.2154      0.0977    -2.21      0.028     -0.408   -0.023

Residual standard error: 290 on 310 degrees of freedom
Multiple R-squared:  0.82,     Adjusted R-squared:  0.818
F-statistic:  354 on 4 and 310 DF,  p-value: <2e-16
```

The model produced in the final step of the forward stepwise regression is:

Expected (Calories) = 64.09 – 2.44 x Age + 19.00 x Fat + 48.05 x Fibre – 0.22 x Fat:Fibre

Per the analysis of variance of the final stepwise model (Table 4), both fat and then fibre produced large F-statistics (p-value < $2\text{x}10^{-16}$, respectively), and age and the interaction between fat and fibre were also regarded as statistically significant predictors when added to the model (both approximately 0.03 < 0.05). From Table 5, the adjusted R^2 is 0.818, which suggests that 81.80% of the variability in caloric intake can be explained by the final stepwise model using three main effects predictors and the interaction term fat:fibre. This is only a marginal improvement of 0.06% when compared to the main effects model.

Example Prediction Using the Model

The estimated fat coefficient when a person eats 25 grams of fibre per day:

= 19 + -0.22(25)

= 13.5

In other words, it is estimated that the caloric intake of a person who consumes 25 grams of fibre per day will increase by 13.5 calories for every additional gram of fat, holding age constant.

Concluding Remarks

This paper has proposed and evaluated an interaction model designed to predict caloric intake using an individual's age, fat and fibre consumption. Consideration of a pairs plot indicated that age would not be a statistically useful predictor of caloric intake, and a weak (r = 0.28) relationship existed between our fat and fibre predictor variables. However, formal evaluation of the main effects model:

Expected (Calories) = 285.81 – 2.49 x Age +15.97 x Fat + 31.69 x Fibre

identified that age was a statistically significant predictor of caloric intake and the model did not appear to have a multicollinearity problem. The adjusted R^2 of the main effects model was 0.8156, which suggests that 81.56% of the variability in caloric intake can be explained using three predictors.

Forward stepwise regression was subsequently utilised to consider the inclusion of interaction terms and the following model was produced:

$$\text{Expected (Calories)} = 64.09 – 2.44 \times \text{Age} + 19.00 \times \text{Fat} + 48.05 \times \text{Fibre} – 0.22 \times \text{Fat:Fibre}.$$

Statistical analysis of this interaction model returned an adjusted R^2 value of 0.818, which suggests that 81.80% of the variability in caloric intake can be explained by the final stepwise model using three main effects predictors and the interaction term fat:fibre. Although this is only a marginal improvement of 0.06% when compared to the main effects model, the addition of an interaction term does not require the collection of any additional data when compared to the main effects model. Accordingly, the author proposes this interaction model be utilised in predicting expected caloric intake.