Protein Consumption

Introduction

Food consumption is of broad interest in the modern world. Patterns of food consumption are related to issue in overall health, obesity, nutrition, economics and the environment. This analysis is concerned with the relationship between meat consumption and consumption in other food groups in 25 European countries. It is of interest to determine how meat consumption is related to consumption in the other food groups. The data include measures of red meat, white meat, eggs, milk, fish, cereals, starchy foods, nuts (which includes pulses and oil-seeds), fruits & vegetables. The units of the variables are not reported, but are assumed to be of the form (unit weight)/(unit time).

Methods

A principal component factor analysis was performed to ascertain the relatedness of consumption across the different food groups. An preliminary factor analysis was performed to assess the correlation structure and the sampling adequacy. SAS reports the partial correlations, which were all small to moderate, indicating a structure conducive to factor analysis. The variables fish and fruits & vegetables have low Kaiser’s sampling adequacy: 0.3704 and 0.3838, respectively. While these measures of sampling adequacy are considered low and make the variables candidates for deletion, the variables were retained because of high interest in these food categories. We used the eigenvalue criterion for choosing the number of factors, cross checked by the proportion of variation explained by the factors. Both the quartimax and varimax rotation methods were performed, in order to choose the rotation that gave the highest interpretability. All analyses were performed in SAS.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Red Meat</th>
<th>White Meat</th>
<th>Eggs</th>
<th>Milk</th>
<th>Fish</th>
<th>Cereals</th>
<th>Starch</th>
<th>Nuts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red Meat</td>
<td>1</td>
<td>-0.38485</td>
<td>0.4247</td>
<td>0.03148</td>
<td>-0.20075</td>
<td>-0.19904</td>
<td>-0.18009</td>
<td>-0.1564</td>
</tr>
<tr>
<td>White Meat</td>
<td>-0.38485</td>
<td>1</td>
<td>-0.4905</td>
<td>-0.35694</td>
<td>-0.43608</td>
<td>-0.17071</td>
<td>-0.06826</td>
<td>-0.56192</td>
</tr>
<tr>
<td>Eggs</td>
<td>0.4247</td>
<td>0.4905</td>
<td>1</td>
<td>0.30681</td>
<td>-0.1712</td>
<td>-0.38474</td>
<td>0.31877</td>
<td>0.34529</td>
</tr>
<tr>
<td>Milk</td>
<td>0.03148</td>
<td>-0.35694</td>
<td>0.30681</td>
<td>1</td>
<td>-0.11749</td>
<td>-0.15863</td>
<td>-0.22915</td>
<td>-0.50885</td>
</tr>
<tr>
<td>Fish</td>
<td>-0.20075</td>
<td>-0.43608</td>
<td>-0.1712</td>
<td>-0.11749</td>
<td>1</td>
<td>-0.66229</td>
<td>0.26044</td>
<td>0.01814</td>
</tr>
<tr>
<td>Cereals</td>
<td>-0.19904</td>
<td>-0.17071</td>
<td>-0.38474</td>
<td>-0.15863</td>
<td>-0.66229</td>
<td>1</td>
<td>-0.01474</td>
<td>0.21437</td>
</tr>
<tr>
<td>Starch</td>
<td>-0.18009</td>
<td>-0.06826</td>
<td>0.31877</td>
<td>-0.22915</td>
<td>0.26044</td>
<td>-0.01474</td>
<td>1</td>
<td>-0.30275</td>
</tr>
<tr>
<td>Nuts</td>
<td>-0.1564</td>
<td>-0.56192</td>
<td>0.34529</td>
<td>-0.50885</td>
<td>-0.01814</td>
<td>0.21437</td>
<td>-0.30275</td>
<td>1</td>
</tr>
</tbody>
</table>
Results and Conclusions

The initial factor procedure identified 3 factors with eigenvalues greater than 1. These three factors explained about 81% of the variation, with the first factor explaining about 50% and the next two explaining about 18% and 13% respectively. There were some cross loadings in the initial factor rotation. The quartimax and varimax rotation loadings are shown below, with the factors loadings grouped for interpretation. The different rotation lead to different groupings and hence, possibly, different interpretations. Both rotations included cross loadings making interpretation more difficult. In the quartimax rotation, the factors could be labeled the red meat, white meat and fish factors, respectively, with red meat being highly associated with milk, eggs and cereals; white meat associated with nuts, and fish associated with starches. Though other interpretations are possible. For the varimax rotation, the factors could be labeled the staple proteins, gourmet proteins and alternative proteins respectively.

The sample size is not quite large enough for this analysis. Perhaps a more interesting study would include non-European countries to look at how protein consumption is related across cultural settings.

Point Allocation

From a total of 40 points possible:

- Not in journal article format: -7 points
- Not reporting partial correlations: -4
- Not reporting sampling adequacy: -4
- Not stating criterion for inclusion/exclusion: -4
- Not discussing cross loadings: -4
- Not showing final factors with variables grouped: -4
- Not reporting the proportion variance explained: -4
- Not interpreting the factors: -4
- Poor grammar or spelling, or confusing language: -5
Wages

Introduction

What determines a person’s hourly wage is an important topic in a modern economy. Poverty among working heads-of-household is becoming common in the United States, and continues to increase. The minimum wage varies from state to state and is hotly debated among politicians and economists. Issues of fairness with respect to race, gender, education and age also dominate discussions in the United States and other multi-cultural societies with complex economies.

This analysis seeks to explore the variables that predict wage. In this analysis we explore the relationship between earned hourly wage and personal characteristics using a data set containing 534 observations on 11 variables sampled from the United States’ Current Population Survey of 1985.

The variables include:

1. EDUCATION: Number of years of education.
2. SOUTH: Indicator variable for Southern Region (1=Person lives in South, 0=Person lives elsewhere).
3. SEX: Indicator variable for sex (1=Female, 0=Male).
4. EXPERIENCE: Number of years of work experience.
5. UNION: Indicator variable for union membership (1=Union member, 0=Not union member).
6. WAGE: Wage (dollars per hour).
7. AGE: Age (years).
8. RACE: Race (1=Other, 2=Hispanic, 3=White).
9. OCCUPATION: job category (1=Management, 2=Sales, 3=Clerical, 4=Service, 5=Professional, 6=Other).
10. SECTOR: Sector (0=Other, 1=Manufacturing, 2=Construction).
11. MARR: Marital Status (0=Unmarried, 1=Married)

The variables OCCUPATION and SECTOR were dropped from the analysis, in the interest of simplicity.

Methods

<table>
<thead>
<tr>
<th>Race</th>
<th>race1</th>
<th>race2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hispanic</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>White</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Dummy variables were created for the race variable as shown in the table. No interactions or higher order terms were considered. A matrix scatterplot of the continuous predictor variables against wages was created to assess linearity and homogeneity of variance, so that any corrective measures could be taken. The plots were messy, even after the response variable was log transformed. This could be due to the several categorical variables. Rather than look at scatterplots by each category separately to assess the appropriateness of the linear regression model, the full model was fit and the residuals from the full model were analyzed for normality and equal variance. These plots are shown. The log transformation improved the problem with the variance, though the relationships appear to be weak.
A histogram of the residuals from the full model looked very nice, with a good overlay of the normal pdf and the histogram. All of the tests for normality reported by SAS were non-significant at the 95% confidence limit. A scatterplot of the residuals versus the predicted showed some concern for increasing variance, but was difficult to ascertain because of the low number of observations at the lower range for predicted wage. There was some indication of possible outliers, with cook’s D values lying distant from the others, but since the actual values were small and we have no reason to suspect the quality of the data, so these were retained.

A stepwise multiple regression and all exploratory and validation procedures were performed in SAS.

Results and Conclusions
The stepwise procedure resulted in a model that included education, experience, sex, union membership, south, marital status and the second race dummy variable. The latter being significant at the 90% level, with a p-value of 0.0656. The exclusion of age as a predictor is somewhat surprising, but inspection of the scatterplots reveals that age and experience are highly and tightly correlated, so only one of these variables is needed.

The race dummy variable was recoded as a binary variable of white/non-white and the regression refit including the variables selected by the stepwise procedure. The white/non-white variable had a p-value of 0.1312 and so was dropped.

The final model included education, experience, sex, south, and union membership. The parameter estimates are shown below.
Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>41.74690</td>
<td>10.43673</td>
<td>51.74</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>529</td>
<td>106.69992</td>
<td>0.20170</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>533</td>
<td>148.44682</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE: 0.44911  R-Square: 0.2812  Dependent Mean: 2.05919  Adj R-Sq: 0.2758  Coeff Var: 21.81012

Parameter Estimates

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|-------|----|-------------------|----------------|---------|------|-----|
| Intercept| Intercept | 1  | 0.76449           | 0.12353        | 6.19   | <.0001 |
| Ed       | Ed    | 1  | 0.09443           | 0.00805        | 11.74  | <.0001 |
| Exper    | Exper | 1  | 0.01237           | 0.00169        | 7.33   | <.0001 |
| Sex      | Sex   | 1  | -0.25781          | 0.03914        | -6.59  | <.0001 |
| SOUTH    | SOUTH | 1  | -0.12638          | 0.04325        | -2.92  | 0.0036 |

The adjusted $R^2$ of the model is only about 0.2758, indicating that many other variables not included in the data set are needed to predict wage with a reasonable degree of accuracy.

Two data points with unknown wages were fed into model to get a predicted value for wage. The two individuals had the following characteristics, in order:

1. EDUCATION: 18, 5
2. SOUTH: yes, no
3. SEX: female, female
4. EXPERIENCE: 2 years, 3 years
5. UNION: no, yes
6. AGE: 23, 33
7. RACE: hispanic, other
8. OCCUPATION: professional, professional
9. SECTOR: Manufacturing, other
10. MARR: married, unmarried

The predictions and the confidence intervals are shown below.

<table>
<thead>
<tr>
<th>Obs</th>
<th>Variable</th>
<th>Value</th>
<th>Predicted Mean</th>
<th>Std Error</th>
<th>95% CL Mean</th>
<th>95% CL Predict</th>
</tr>
</thead>
<tbody>
<tr>
<td>535</td>
<td></td>
<td>2.1048</td>
<td>0.0605</td>
<td>1.9860</td>
<td>2.2363</td>
<td>2.9960</td>
</tr>
<tr>
<td>536</td>
<td></td>
<td>1.0160</td>
<td>0.0858</td>
<td>0.8473</td>
<td>1.1846</td>
<td>1.9142</td>
</tr>
</tbody>
</table>

The predicted values in the table are the natural logs. Back transforming to find the predicted value in the original units (dollars), we have: $8.21/hour for the first individual and $2.76/hour for the second individual, with corresponding confidence intervals: ($3.37,$19.98) and ($1.13,$6.97)
respectively.

**Interpretation**

Because the model has $\ln wages$ as the response variable, interpretation requires back transformation. Let $W$ denote wages, $Ed$ denote education, $Ex$ denote experience, $Sex$ denote sex and $So$ denote South. We have:

$$\ln W = 0.76449 + 0.09443 \times Ed + 0.01237 \times Ex - 0.25781 \times Sex - 0.12638 \times So$$

$$\rightarrow e^{\ln W} = e^{0.76449+0.09443\times Ed+0.01237\times Ex-0.25781\times Sex-0.12638\times So}$$

In 1985, for a fixed level of education and experience, a person living in the South would expect to make $e^{-0.12638} = 0.88$ of what his or her counterpart would expect to make in other regions of the country.

For a fixed level of education and experience a woman living in the same region of the country would expect to make $e^{-0.2581} = 0.7725$ of what her male counterpart would expect.

For a fixed gender, region, and experience, for every extra year of education a person would expect his or her wage to increase by $e^{0.09443} = 1.099$ times, or about 9.9%.

For a fixed gender, region, and education, for every extra year of experience, a person would expect his or her wage to increase by $e^{0.01237} = 1.0125$ times, or about 1.25%.

The regression model shows that there remained some gender bias in wages in 1985. Some might also argue that experience ought to count more than education, at least in some sectors. An analysis that included market sector in the model might be interesting to look for interactions between sector and education, and sector and experience.

**Point Allocation**

From a total of 60 points possible:

- Not in journal article format: -9 points
- Mis-coding of dummy variables: -4
- No exploratory residual analysis: -4
- No discussion of remedial measures: -4
- Not testing for a binary race variable: -4
- Not reporting final model: -4
- Not discussing or showing final residual analysis: -4
- Not reporting the proportion variance explained: -4
- Not reporting confidence intervals for prediction: -4; if reported but incorrect: -2
- No interpretation: -8
- Not back transforming if made transformations: -4
- Poor grammar or spelling, or confusing language: -5