Practical on the Health Survey of England Dataset

Considering BMI in other formats

We have so far considered BMI as a continuous quantity and looked at fitting Normal response models. BMI is however often considered in terms of categorisation based on ranges along the BMI spectrum. One has to also consider that although for many people reducing their BMI will improve their health that such a policy of reducing BMI does not work for every individual. In fact BMI has an optimal range below which individuals will become more susceptible to other health problems. BMI is therefore often characterised in terms of 5 categorical ranges (under-weight, ideal, over-weight, obese, morbidly obese). The fact that there are 3 levels of over-weight category and only one under-weight is due to the greater number of people in this range. Replacing the raw BMI values with a categorical predictor has a disadvantage as we do lose precise information for each individual but in practice medical interventions will often target particular risk groups and so it is interesting to look at what factors will result in individuals belonging to particular BMI groups.

The standard (UK) BMI classifications are given in the table below:

<table>
<thead>
<tr>
<th>Category</th>
<th>BMI Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Underweight</td>
<td>Below 18.5</td>
</tr>
<tr>
<td>Ideal Weight</td>
<td>18.5 &lt; Weight &lt; 25</td>
</tr>
<tr>
<td>Over weight</td>
<td>25 &lt; Weight &lt; 30</td>
</tr>
<tr>
<td>Obese</td>
<td>30 &lt; Weight &lt; 40</td>
</tr>
<tr>
<td>Morbidly Obese</td>
<td>Above 40</td>
</tr>
</tbody>
</table>

If we load up the worksheet *binary1.ws* we can then construct this categorical variable. Note that this worksheet should be almost identical to the worksheets used earlier aside from potentially the category names of the various categorical variables. Having loaded up the worksheet we now need to do the following:

Choose *recode by range* from the Data Manipulation menu

Select 0 and 18.5 for values in the range boxes

Select 0 for the new value 0

Select bmival for the input column

Click on the Free Columns button (c47 is chosen)

Click on the Add to action list button.

Next select 18.5 and 25 in the range boxes and 1 in the new value box

The window should look as follows:
Continue by doing the following

Clicking on the **Add to action list** button

Input the remaining ranges clicking on the **Add to action list** button each time.

First **range** 25 and 30 with **new value** 2

Next **range** 30 and 40 with **new value** 3

Finally **range** 40 and 80 with **new value** 4

When you have finished the window should look as follows:
Clicking on the **Execute** button will result in the categorical variable being constructed in c47. We next want to create category names and name the variable.

Choose **Names** from the **Data Manipulation** menu.

Scroll down to column c47.

Click on the **Name** button under column and change c47 to ‘bmicat’

Click on the **Toggle Categorical** button to tell MLwiN the variable is categorical.

Click on the **View** button under categories to see that default labels have been assigned.

Replace these labels using the **edit** button and choose ‘under’ for 0, ‘ideal’ for 1, ‘over’ for 2, ‘obese’ for 3 and ‘m_obese’ for 4

Click on the **OK** button when finished.

We next want to look at the data created:

Choose **View or edit data** from the **Data Manipulation** menu.

By default all columns are shown so click on the **view** button.

We next need to select columns so choose ['year', 'pserial', 'hserial', 'area', 'sex', 'age', 'ethnic', 'bmival', 'bmicat'] – note to select multiple columns use the ctrl button when selecting columns after the first.

Click **OK** when done

The data will then look as shown in the window overleaf. Here we see that the first record is for an overweight white female aged 59 and the second for an obese white male aged 24.
We would next like to look at the numbers of people and proportions that fall into each weight category and we can do this via the Tabulate window:

Select **Tabulate** from the **Basic Statistics** menu

Choose ‘bmicat’ as the **Columns** indicator

Select the **percentage of row totals** tick box

The **Tabulate** window will look as follows:

Pressing the **Tabulate** button will then give the output shown on the next page:
Here we see that in our sample over half the people in the dataset (32.3% + 17.9% + 1.6% = 51.8%) are overweight whilst only 12.5% are underweight and the remaining 35.7% fit in the ideal weight category.

Next we can look at tabulations by gender and ethnic group by clicking in the rows box and choosing the rows to be ‘sex’ or ‘ethnic’ respectively:

For gender we see more women than men in the survey with a bigger proportion of men being overweight but more women in the highest ‘m_obese’ category.

By Ethnicity we see:

Here it is firstly worth noting that the bulk of the data is for the ‘white’ ethnicity category so we shouldn’t overstate differences here. The ‘mixed’, ‘asian’ and ‘other’ categories appear to have less overweight people than the ‘white’ and ‘black’ categories. The ‘mixed’ category has a surprisingly high number of underweight individuals but the sample size is small.

For age this variable is continuous and so we can adapt the Tabulate window to look at the average age per category by selecting options so that the window looks as follows:
Pressing Tabulate then gives the following results:

<table>
<thead>
<tr>
<th></th>
<th>under</th>
<th>ideal</th>
<th>over</th>
<th>obese</th>
<th>m_obese</th>
<th>TOTALS</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>3829</td>
<td>10922</td>
<td>9899</td>
<td>5477</td>
<td>497</td>
<td>30624</td>
</tr>
<tr>
<td>MEANS</td>
<td>9.66</td>
<td>37.7</td>
<td>49.2</td>
<td>50.6</td>
<td>46.6</td>
<td>40.4</td>
</tr>
<tr>
<td>SD'S</td>
<td>11.5</td>
<td>20.7</td>
<td>18.0</td>
<td>16.7</td>
<td>15.1</td>
<td>18.1</td>
</tr>
</tbody>
</table>

Interestingly we see that the underweight group has a much smaller mean age than the other groups. This might call into question the usefulness of the BMI measure for children or at least whether we should use them in our analysis. For this document we will keep them in the analysis but will supply additional sub-datasets so that you can test the models on different age categories in further practical work.

We also have a categorical age variable (‘agecat’) that splits people in 4 groups (under 16s, 17-34, 35-59, 60 plus). We can use this variable in a similar way as we did gender and ethnicity.

Looking at the distributions for each age category we see the following:

<table>
<thead>
<tr>
<th></th>
<th>under</th>
<th>ideal</th>
<th>over</th>
<th>obese</th>
<th>m_obese</th>
<th>TOTALS</th>
</tr>
</thead>
<tbody>
<tr>
<td>under16</td>
<td>N</td>
<td>3485</td>
<td>2084</td>
<td>353</td>
<td>114</td>
<td>5</td>
</tr>
<tr>
<td>ROW %</td>
<td>57.7</td>
<td>34.5</td>
<td>5.8</td>
<td>1.9</td>
<td>0.1</td>
<td>100.0</td>
</tr>
<tr>
<td>17to34</td>
<td>N</td>
<td>193</td>
<td>3173</td>
<td>1909</td>
<td>890</td>
<td>104</td>
</tr>
<tr>
<td>ROW %</td>
<td>3.1</td>
<td>50.6</td>
<td>30.5</td>
<td>14.2</td>
<td>1.7</td>
<td>100.0</td>
</tr>
<tr>
<td>35to59</td>
<td>N</td>
<td>84</td>
<td>3781</td>
<td>4590</td>
<td>2705</td>
<td>276</td>
</tr>
<tr>
<td>ROW %</td>
<td>0.7</td>
<td>33.1</td>
<td>40.1</td>
<td>23.7</td>
<td>2.4</td>
<td>100.0</td>
</tr>
<tr>
<td>over60</td>
<td>N</td>
<td>67</td>
<td>1884</td>
<td>3047</td>
<td>1768</td>
<td>112</td>
</tr>
<tr>
<td>ROW %</td>
<td>1.0</td>
<td>27.4</td>
<td>44.3</td>
<td>25.7</td>
<td>1.6</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Here we see a gradual progression of people moving from underweight in the younger age category across to overweight and obese in later categories.

We can plot the data for BMI against age as follows:
The graph will look as follows (note we have added labels by clicking on the graph and using the window that is brought up)

Here we see the clump of young children with lower BMI. Interestingly the really outlying BMI readings tend to be at younger ages - which appears to mask the overall trend for older people being more overweight. It is possible that the large number of younger outliers may be due to the problem of obesity increasing over time or it might be that the life expectancy of such individuals is lower and hence less individuals in this category appear at later ages. This is however all rather speculative. We also see the 58.8kg 3 year old which one would guess is an error but given the dataset size will not sway conclusions.

Having worked our way through the relationships between the various categories and seen the influence of several demographic factors on BMI we will now consider a specific binary categorisation: What factors are associated with being over-weight?

Here we will create a binary variable that encompasses the difference between the three overweight categories (overweight, obese and morbidly obese) and the two other categories (underweight, ideal).

The easiest way of doing this is to choose the Command interface window from Data Manipulations and type in the commands:

Select Customised Graphs from the Graphs menu

Select ‘bmival’ as the y variable

Select ‘age’ as the x variable

Click on the Apply button.
Looking at risk factors related with being ‘overweight’

We now have our response variable ‘overweight’ and so before we start it is worth tabulating this variable using the Tabulate window as we did previously with the 5 category ‘bmicat’ variable which should give the following:

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>TOTALS</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>14751</td>
<td>15873</td>
<td>30624</td>
</tr>
<tr>
<td>%</td>
<td>48.2</td>
<td>51.8</td>
<td>100.0</td>
</tr>
</tbody>
</table>

This tabulation is simply the earlier 5 category tabulation collapsed into two columns and we could repeat the other tabulations as before, for example the tabulation of the overweight variable against gender is as follows:

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>TOTALS</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>N</td>
<td>6363</td>
<td>7816</td>
</tr>
<tr>
<td>ROW %</td>
<td>44.9</td>
<td>55.1</td>
<td>100.0</td>
</tr>
<tr>
<td>female</td>
<td>N</td>
<td>8388</td>
<td>8057</td>
</tr>
<tr>
<td>ROW %</td>
<td>51.0</td>
<td>49.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Here we see more overweight males than females as we suggested earlier.

Next the tabulation against ethnicity is as follows:

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>TOTALS</th>
</tr>
</thead>
<tbody>
<tr>
<td>white</td>
<td>N</td>
<td>13242</td>
<td>14793</td>
</tr>
<tr>
<td>ROW %</td>
<td>47.2</td>
<td>52.8</td>
<td>100.0</td>
</tr>
<tr>
<td>mixed</td>
<td>N</td>
<td>256</td>
<td>89</td>
</tr>
<tr>
<td>ROW %</td>
<td>74.2</td>
<td>25.8</td>
<td>100.0</td>
</tr>
<tr>
<td>asian</td>
<td>N</td>
<td>789</td>
<td>565</td>
</tr>
<tr>
<td>ROW %</td>
<td>58.3</td>
<td>41.7</td>
<td>100.0</td>
</tr>
<tr>
<td>black</td>
<td>N</td>
<td>268</td>
<td>291</td>
</tr>
<tr>
<td>ROW %</td>
<td>47.9</td>
<td>52.1</td>
<td>100.0</td>
</tr>
<tr>
<td>other</td>
<td>N</td>
<td>176</td>
<td>112</td>
</tr>
<tr>
<td>ROW %</td>
<td>61.1</td>
<td>38.9</td>
<td>100.0</td>
</tr>
</tbody>
</table>

This shows more clearly how being overweight is less common in the ‘mixed’, ‘asian’ and ‘other’ groups in the sample.

Calc c49 = ‘bmicat’ > 1
Name c49 ‘overweight’
Single level logistic regression models

We will first set up a simple logistic regression model ignoring the underlying nested structure of the data in our modelling (although we will specify the structure).

Select the **Equations** window from the **Model** menu

Click on the red y and select ‘overweight’ from the y list.

Select ‘4 ijk’ from the **number of levels** list.

Select ‘area’ for **level 4(l)**

Select ‘year’ for **level 3(k)**

Select ‘hserial’ for **level 2(j)**

Select ‘pserial’ for **level 1(i)**

Click on the **Done** button

Click on the N (for Normal) and select Binomial from the list that appears

Click on the **Done** button

Click on the red $n_{ijkl}$ and select ‘cons’ from the list that appears and click **done**.

Click on the red $x_{o}$ and select ‘cons’ from the list that appears and click **done**.

The **Equations** window will then look as follows:

![Equations window](image)

If we now click on the Start button and click on the Estimates button twice the model will be run and the estimates displayed:
Here we are fitting a basic logistic regression with no predictor variables. This model gives us one coefficient (and its standard error) and the value 0.073 is not immediately that interpretable.

This value is on the logit scale as we are fitting a regression to the logit of the underlying probability as we discussed in the lecture. It is therefore useful to convert this to a probability and this we do by reversing the logit transformation which is done via the anti-logit function.

This can be done in MLwiN in the Command Interface window (available from Data Manipulation menu) if you bring up this window and type the following command:

CALC ALOG 0.073

Then in the Output window the value 0.51824 will be displayed. This is the probability of being overweight and corresponds exactly to the tabulation we did previously. We do actually know something about values on the logit scale – a value of 0 corresponds to 0.5 on the probability scale so a small positive value will correspond to a probability a little bigger than 0.5 as we see here.

Next we can consider adding a predictor to our regression. Let us consider gender first:

On the Equations window click on the Add term button

Choose ‘sex’ on the window that appears.

Leave ‘male’ as a base category and click on the Done button.

Click the Start button.

The Equations window will now look as follows:

overweight_{ijkl} \sim \text{Binomial}(\text{cons}_{ijkl}, \pi_{ijkl})
\logit(\pi_{ijkl}) = 0.206(0.017)\text{cons} + -0.246(0.023)\text{female}_{ijkl}
var(overweight_{ijkl}|\pi_{ijkl}) = \pi_{ijkl}(1 - \pi_{ijkl})/\text{cons}_{ijkl}
For two categories we therefore get two estimates (plus standard errors)

To convert these values to probabilities we again use the Command Interface window and type:

CALC ALOG 0.206

For the males and

CALC ALOG (0.206-0.246)

for the females. This will give values of 0.55132 and 0.49000 respectively again as seen in the earlier tabulation.

Odds Ratios

As mentioned in the lecture the odds ratio is often used as a useful statistic for describing the effect of a parameter. As the logit transform (log(p/(1-p))) is essentially the log odds of the probability we can calculate an odds ratio related to a particular parameter by taking the exponential of the coefficient.

In the Command Interface window if we type

CALC EXPO (-0.246)

We will get the value 0.782 (which is less than 1 showing women have less chance than men of being overweight)

If we instead type

CALC EXPO 0.246

We get the value 1.2789 (= 1/0.782) which means that men are 1.2789 times as likely to be overweight as women.

Model Comparison

We might consider how to test for significant sex differences. The standard approach would be to perform a chi-squared test against the null hypothesis of no sex differences.

Under Ho our expected counts would be as follows:

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>TOTALS</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>female</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6830</td>
<td>7349</td>
<td>14179</td>
</tr>
<tr>
<td></td>
<td>7921</td>
<td>8524</td>
<td>16445</td>
</tr>
</tbody>
</table>

Here the proportions in each gender group are equal and this results in a chi-squared statistic \( \frac{\text{sum}(O-E)^2}{E} \) of around 114.7 on 1df which is incredibly significant (Note the Tabulate window will calculate this chi-squared statistic). From a logistic regression perspective we would usually test the predictor associated with female using an approximate Wald test which can be done in the Intervals and Tests window:
The window should look as follows:

This rather pleasingly gives almost the same chi-squared statistic (the differences are possibly due to rounding when I performed the test by hand) as the standard test done by hand. To look up the significance of this value we can choose **Tail Areas** from the **Basic Statistics** menu and fill in the window as follows:
Clicking on the **Calculate** button results in a tiny P value that appears in the Output window:

\[
\text{Probability 114.458 1 } \\
1.0343 \times 10^{-26}
\]

This means we have highly significant gender differences (partly due to the size of the dataset). We can next repeat our analysis with the ethnicity variable.

In the **Equations** window click on the **female** term and click on **Delete Term** in the window that appears.

Click on **Add Term** and select ‘ethnic’ from the list that appears leaving ‘white’ as the reference category.

Click on the **Done** button and then the **Start** button.

The **Equations** window will then appear as follows:

\[
\begin{align*}
\text{overweight}_{jk} & \sim \text{Binomial} (\text{cons}_{jkl}, \pi_{jkl}) \\
\logit(\pi_{jkl}) &= 0.111(0.012)\text{cons} + -1.167(0.124)\text{mixed}_{jkl} + 0.445(0.056)\text{asian}_{jkl} + \\
&\quad -0.028(0.086)\text{black}_{jkl} + -0.563(0.121)\text{other}_{jkl} \\
\text{var}(\text{overweight}_{jkl} | \pi_{jkl}) &= \pi_{jkl}(1 - \pi_{jkl})/\text{cons}_{jkl}
\end{align*}
\]

Now we have 5 terms for the 5 categories and once again we can convert them to probabilities in the **command interface** window:

For example:

CALC ALOG 0.111 for the white group corresponds to a probability of 0.52772 and CALC ALOG (0.111-1.167) for the mixed group corresponds to a probability of 0.25807. These probabilities are the same as seen in the earlier tabulation.

To check for significant differences between the categories we can again use the **Intervals and Tests** window which needs to be setup as follows:
Pressing the **Calculate** button gives the value 167.609 on 4 degrees of freedom which can be looked up using the **Tail Areas** window:

The P value is so small it will give a value of 0 in the output window meaning that we have incredibly significant effects of ethnicity (again mainly due to the size of the dataset). The coefficients give an indication of the relative effects and so compared to the reference category (white) all coefficients are negative, meaning there are less overweight people in the other groups. Only the black group being close to 0 and having a coefficient that is smaller than twice its standard error (which is a rough test of significance) has a similar probability of being overweight.
Including a continuous predictor

We will next consider the ‘age’ predictor. Here we will focus on the actual age rather than the categorical predictor.

In the **Equations** window click on ‘mixed’

Click on **Delete Term** in the window that appears and click on **yes** to delete all 4 terms

Click on **Add Term** and select ‘age’ from the list of variables.

Click on the **Done** button and then the **Start** button

The **Equations** window will then appear as follows:

\[
\text{overweight}_{yjk} \sim \text{Binomial}(\text{cons}_{jkt}, \pi_{yjk})
\]

\[
\logit(\pi_{yjk}) = -1.734(0.028)\text{cons} + 0.045(0.001)\text{age}_{yjk}
\]

\[
\var(\text{overweight}_{yjk}|\pi_{yjk}) = \pi_{yjk}(1 - \pi_{yjk})/\text{cons}_{yjk}
\]

Here we see a single term which represents the effect of being 1 year older

We can immediately see that this coefficient is very much bigger than its standard error and so is highly significant (you can check with **Intervals and Tests** to confirm this.)

We can then do an odds ratio

\[
\text{CALC EXP} 0.045 = 1.046
\]

This means that for each year older a person is, they are 1.046 times as likely to be overweight – so for a person 10 years older they would be \(\text{EXP} 0.045 \times 10 = 1.5683\) times more likely to be overweight.

We can plot a prediction of how the probability varies with age as follows:

Select **Predictions** from the **Model** menu

Click on **fixed** and select **include all fixed effects**

Choose c60 in the **output from predictions** to box
The window will look as follows:

\[
\text{logit}(\text{overweight}_{ijt}) = \hat{\beta}_0 \text{cons} + \hat{\beta}_1 \text{age}_{ijt}
\]

variable cons age_{ijt}  
fixed $\beta_0$ $\beta_1$  
level 1

Clicking on the \textbf{Calc} button will put predictions for the linear predictor in column \textit{c60}.

To convert these to the probability scale we need to use the \textbf{Command Interface} window to convert all predictions between scales. This is done by typing:

\texttt{CALC c60 = alog(c60)}

We can now plot these predictions against age:

- Select \textbf{Customised graphs} from the \textbf{Graphs} menu
- Select \textit{c60} as \textit{y} variable
- Select \textit{age} as \textit{x} variable
- Select line as \textbf{plot type}

The window should now look as follows:
Click **Apply** and the following graph will appear:

![Graph](image)

Note: Here we have changed axes labels and limits by clicking on the graph and changing settings in the window that appears.

We have assumed a linear relationship with age (on the logit scale). We can of course add additional polynomial terms (to do this use the **add term** button and change the order), for example if we go to the fourth power in age (all terms being significant) we get the following plot:

![Graph](image)

Here we see a steeper rise at the start, a more gradual rise until about 75 and then a tailing off (here we might speculate that possibly we are seeing the more overweight people dying earlier!)

We can of course create a model with several predictors together and so please try out fitting the predictors together by adding terms and also including interactions.
For example if we fit all three main effects for gender, age and ethnicity we get the following in the **Equations** window:

\[
\begin{align*}
\text{overweight}_{ijkl} & \sim \text{Binomial}(\text{cons}_{ijkl}, \pi_{ijkl}) \\
\text{logit}(\pi_{ijkl}) &= -1.550(0.032)\text{cons} + 0.046(0.001)\text{age}_{ijkl} - 0.380(0.026)\text{female}_{ijkl} + \\
&\quad -0.346(0.132)\text{mixed}_{ijkl} + -0.037(0.061)\text{asian}_{ijkl} + 0.320(0.093)\text{black}_{ijkl} + \\
&\quad -0.236(0.129)\text{other}_{ijkl} \\
\text{var}(\text{overweight}_{ijkl} | \pi_{ijkl}) &= \pi_{ijkl}(1 - \pi_{ijkl})/\text{cons}_{ijkl}
\end{align*}
\]

This is interesting as although the ‘age’ main effect is similar to when it is fitted in isolation, the ‘female’ effect is larger in magnitude meaning that controlling for age and ethnicity there is a larger difference in obesity rates between males and females. Also two of the ethnicity effects have changed: the ‘Asian’ effect is now very small and so the earlier lower probability of being overweight for the Asian group was mainly due to a different age/sex mix in this group in the sample. Similarly for the ‘black’ group we now a significantly higher rate of obesity, so once again this was masked by differing age/sex makeup for this group.

We can also investigate interactions. In fact all two-way interactions and three-way interactions are significant and we end up with the complicated model:

\[
\begin{align*}
\text{overweight}_{ijkl} & \sim \text{Binomial}(\text{cons}_{ijkl}, \pi_{ijkl}) \\
\text{logit}(\pi_{ijkl}) &= -1.758(0.044)\text{cons} + 0.052(0.001)\text{age}_{ijkl} + 0.076(0.060)\text{female}_{ijkl} - 1.946(0.508)\text{mixed}_{ijkl} + \\
&\quad -0.377(0.205)\text{asian}_{ijkl} - 0.137(0.316)\text{black}_{ijkl} - 0.669(0.505)\text{other}_{ijkl} + \\
&\quad -0.012(0.001)\text{age female}_{ijkl} + 0.068(0.019)\text{age mixed}_{ijkl} + 0.007(0.006)\text{age asian}_{ijkl} + \\
&\quad 0.003(0.008)\text{age black}_{ijkl} + 0.026(0.015)\text{age other}_{ijkl} + 1.271(0.621)\text{female mixed}_{ijkl} + \\
&\quad -0.186(0.287)\text{female asian}_{ijkl} + 0.098(0.439)\text{female black}_{ijkl} + -0.281(0.730)\text{female other}_{ijkl} + \\
&\quad -0.054(0.022)\text{female age mixed}_{ijkl} + 0.012(0.008)\text{female age asian}_{ijkl} + \\
&\quad 0.022(0.012)\text{female age black}_{ijkl} + -0.018(0.020)\text{female age other}_{ijkl} \\
\text{var}(\text{overweight}_{ijkl} | \pi_{ijkl}) &= \pi_{ijkl}(1 - \pi_{ijkl})/\text{cons}_{ijkl}
\end{align*}
\]

We essentially have 10 groups (2 genders * 5 ethnicities) and we can plot the age curves for each of these as shown in the graph below.
We will leave single level logistic regression models here and now focus on multilevel modelling of our dataset. You might like to consider some of the other predictor variables in your modelling including higher level predictors.

**Multilevel modelling of the data**

All the modelling that we have done thus far in this practical has assumed no structure to the data and involved simple logistic regressions. We will now consider the four levels of nesting that occur in the dataset with people ($p_{serial}$) nested with households ($h_{serial}$) nested within years ($year$) nested within areas ($area$). We will begin by removing all the fixed effects in the model except the intercept (‘cons’)

Select **Equations** from the **Model** menu.

Click on all terms (apart from ‘cons’) and click on **Delete term** and **yes** on the window that appears.

Click on ‘cons’.

Select the tick boxes for ‘l(area)’, ‘k(year)’ and ‘j (hserial)’ leaving the **fixed parameter** box selected.

Click on the **Done** button and then the **Start** button

The model will then run and you should see results as follows:
It is noticeable immediately that the between year within area variance has been estimated as zero. If we select the **Hierarchy Viewer** from the **Model** menu we see the following:

Here we see that there are only 2053 year cohorts within the 1833 areas – in other words for the majority of areas there is only one year cohort present. We will therefore remove the year level from our data (in particular as it causes estimation problems later) and refit the model.

Select **Equations** from the **Model** menu

Click on *overweight* and in the window that follows change **n-levels** to 3-ijk

Change **level 3(k)** to **area**

Click on the **Done** button and the **Start** button.
The model will give the same results but without the zero estimate. This model has been run using the default methods in MLwiN — 1\textsuperscript{st} order MQL estimation. This method is a form of quasi-likelihood estimation and for multilevel discrete response models there are alternative quasi-likelihood methods.

Click on the \textbf{Nonlinear} button on the \textbf{Equations} window

On the window that appears choose \textbf{2\textsuperscript{nd} order} and \textbf{PQL}

Click on the \textbf{Done} button and the \textbf{Start} button

The model will then be run using 2\textsuperscript{nd} order PQL estimation with the following results:

\[
\text{overweight}_{ijk} \sim \text{Binomial}(\text{cons}_{ijk}, \pi_{ijk})
\]

\[
\logit(\pi_{ijk}) = \beta_{0jk} \text{cons} + \nu_{0k} + \tau_{0jk}
\]

\[
\begin{bmatrix}
\nu_{0k} \\
\tau_{0jk}
\end{bmatrix} \sim N(0, \Omega_{\nu}) : \Omega_{\nu} = \begin{bmatrix}
0.033 (0.010)
\end{bmatrix}
\]

\[
\begin{bmatrix}
\nu_{0k} \\
\nu_{0jk}
\end{bmatrix} \sim N(0, \Omega_{\nu}) : \Omega_{\nu} = \begin{bmatrix}
0.308 (0.027)
\end{bmatrix}
\]

\[
\text{var}(\text{overweight}_{ijk} | \pi_{ijk}) = \pi_{ijk}(1 - \pi_{ijk})/\text{cons}_{ijk}
\]

Note: PQL is a better approximation than MQL though it sometimes has more problems converging (in fact if we leave the \textit{year} level in then PQL doesn’t converge).

In the lecture notes we discussed interpreting the coefficients given in the model. We will first see that the intercept has changed slightly and that 0.108 corresponds to a probability of 0.526 (as opposed to 0.518) in the one level model. This slight change is a combination of the estimate being subject specific rather than population average and an adjustment for the lack of balance in sample size across areas and households. The standard error has also increased from 0.011 to 0.013 due to this model controlling for the lack of independence in the dataset.

\textbf{Variance Partition Coefficients}

As we saw in the lecture notes calculating the proportion of variance explained by higher levels in a Binomial model is more difficult. We can directly compare the two higher level variances and see that the between household (within area) variance is almost ten times the
between area variance. If we use the latent variable interpretation then we assume that the level 1 variance = 3.291 and so we have

\[ VPC(\text{area}) = \frac{0.033}{0.033 + 0.305 + 3.291} = 0.009 \]

And

\[ VPC(\text{household}) = \frac{0.305}{0.033 + 0.305 + 3.291} = 0.084 \]

So approximately 8.4% of variation is due to differences between household (within area) and 0.9% is due to between area differences. Note one could also say that 8.4+0.9 = 9.3% of variation is due to between household differences by combining the two effects.

**Looking at the influence of risk factors**

As with the simple logistic regression modelling we can next consider the influence of predictors whilst this time accounting for the structure:

1. Select **Equations** window from the **Model** menu
2. Click on **Add term** and select **sex** from the list
3. Click on the **Done** button and then on the **Start** button

The model will then be fitted as shown below:

\[
\text{overweight}_{ijk} \sim \text{Binomial}(\text{cons}_{ijk}, \pi_{ijk})
\]

\[
\text{logit}(\pi_{ijk}) = \beta_{0jk} \text{cons} + 0.268(0.024)\text{female}_{ijk}
\]

\[
\beta_{0jk} = 0.254(0.019) + \nu_{0jk} + \nu_{0jk}
\]

\[
\begin{bmatrix}
\nu_{0jk} \\
\nu_{0jk}
\end{bmatrix} \sim \text{N}(0, \Omega_{\nu}) : \quad \Omega_{\nu} = \begin{bmatrix}
0.034 & 0.010 \\
0.010 & 0.034
\end{bmatrix}
\]

\[
\begin{bmatrix}
\nu_{0jk} \\
\nu_{0jk}
\end{bmatrix} \sim \text{N}(0, \Omega_{\nu}) : \quad \Omega_{\nu} = \begin{bmatrix}
0.315 & 0.027 \\
0.027 & 0.315
\end{bmatrix}
\]

\[
\text{var}(\text{overweight}_{ijk} | \pi_{ijk}) = \pi_{ijk}(1 - \pi_{ijk}) / \text{cons}_{ijk}
\]

We see that there is a negative coefficient for female as we saw for the single level model. The magnitude has increased but recall from the lecture notes the subject specific interpretation of this coefficient which may explain this increase. We can also test fitting ethnicity in the multilevel framework but instead we will move on to looking at age.
Select **Equations** window from the **Model** menu

Click on **female** and select **Delete term**.

Click on **Add term** and select **age** from the list.

Click on the **grand mean** button.

Click on the **Done** button and then on the **Start** button

The **Equations** window will look as follows:

\[
\text{overweight}_{ijk} \sim \text{Binomial}(\text{cons}_{ijk}, \pi_{ijk})
\]

\[
\logit(\pi_{ijk}) = \beta_{gjk} \text{ cons} + 0.050(0.001)(\text{age-gm})_{ijk}
\]

\[\beta_{gjk} = 0.087(0.015) + \gamma_{0k} + \zeta_{gjk}\]

\[
\begin{bmatrix}
\gamma_{0k}
\end{bmatrix} \sim \mathcal{N}(0, \Omega_{\gamma}) : \hspace{1cm} \Omega_{\gamma} = \begin{bmatrix}
0.038 & 0.012
\end{bmatrix}
\]

\[
\begin{bmatrix}
\zeta_{gjk}
\end{bmatrix} \sim \mathcal{N}(0, \Omega_{\zeta}) : \hspace{1cm} \Omega_{\zeta} = \begin{bmatrix}
0.356 & 0.033
\end{bmatrix}
\]

\[
\text{var(overweight}_{ijk}|\pi_{ijk}) = \pi_{ijk}(1 - \pi_{ijk})/\text{cons}_{ijk}
\]

Again we see a strong effect of age with the probability of being overweight increasing with age. Age has been centred around it’s grand mean value (40.336) and so now the value 0.087 and it’s corresponding probability (ALOG(0.087) = 0.52174) corresponds to an average aged i.e. 40 year old person. We can next ask whether this strong positive effect varies from area to area.

**Fitting Random slopes for age**

We next want to fit random slopes for age:

In the **Equations** window we click on the (age-gm) term .

Select the tick box marked (k)area.

Select the **Done** button and the **More** button

Choose **Numbers** from the **Options** Menu

Change **#digits after decimal point** to 5 and click **Apply**
We then see the following model estimates:

\[
\begin{align*}
\text{overweight}_{jk} & \sim \text{Binomial}(\text{cons}_{jk}, \pi_{jk}) \\
\text{logit}(\pi_{jk}) & = \beta_{0jk} \text{cons} + \beta_{1jk} (\text{age-gm})_{jk} \\
\beta_{0jk} & = 0.13093(0.01523) + \nu_{0k} + \omega_{0k} \\
\beta_{1jk} & = 0.05098(0.00074) + \nu_{1k}
\end{align*}
\]

\[
\begin{bmatrix}
\nu_{0k} \\
\nu_{1k}
\end{bmatrix} \sim \mathcal{N}(0, \Omega_{\nu}) : \Omega_{\nu} =
\begin{bmatrix}
0.04472(0.01265) \\
0.00435(0.00043) & 0.00088(0.00003)
\end{bmatrix}
\]

\[
\begin{bmatrix}
\omega_{0k}
\end{bmatrix} \sim \mathcal{N}(0, \Omega_{\omega}) : \Omega_{\omega} =
\begin{bmatrix}
0.38808(0.03369)
\end{bmatrix}
\]

\[
\text{var} (\text{overweight}_{jk} | \pi_{jk}) = \pi_{jk}(1-\pi_{jk}) / \text{cons}_{jk}
\]

We can test (approximately) whether random slopes results in a better model via the **Interval and Tests** window, which we set up as follows:

Here the value 109.34 corresponds to a hugely significant effect so random slopes improves the model. We can also graph the model via the predictions and graphs windows:
Select **Predictions** from the **Model** menu

Click on **fixed** and select **include all fixed effects**

Click on **level 3** and select **include all level 3 random effects**

Select **c60** for the **output from predictions to** column

The window will then look as follows:

![Predictions window](image)

We next need to calculate the residuals and graph them:

Click on the **Calc** button

Select **Command Interface** from the **Data Manipulation** window

Type the command `CALC c60 = ALOG(c60)`

Select **Customised Graphs** from the **Graph** menu

Select **c60** for the **y variable**

Select **age** for the **x variable**

Select **area** for the **group variable**

The window will look as follows:
Click on the **Apply** button to get the following graph.

Note this graph has a couple of thousand curves so it may be slow to appear.

We can see the level of variability across the different areas with regard the relationship between age and probability of being overweight.
We can also allow age to vary across households. This however will cause convergence problems for the 2nd order PQL method. If we were to switch to 1st order MQL the model will converge and give the following results:

\[
\begin{align*}
\text{overweight}_{ik} & \sim \text{Binomial}(\text{cons}_{ik}, \pi_{ik}) \\
\logit(\pi_{ik}) &= \beta_{0ik} + \beta_{1ik}(\text{age-20})_{ik} \\
\beta_{0ik} &= 0.04560(0.01527) + \nu_{0i} + \mu_{0ik} \\
\beta_{1ik} &= 0.05054(0.00071) + \nu_{1i} + \mu_{1ik}
\end{align*}
\]

At the household level there is also a significant variability in the effect of age on the probability of being overweight. Of course plotting lines for each of the 15,000 households would be impractical but we see that the relationship is complicated and that the influence of age depends on both the household and the area in which it is situated.

**Investigating risk factors related to different weight characterisations**

We have so far considered only a binary response variable. An alternative approach to investigating a categorical response is to consider fitting an ordered multinomial model. In MLwiN by default cases that include missing data are ignored when model fitting (effectively list-wise deleted). This means that when we fitted the first model to the BMI variable it was fitted to the 30,624 individuals with a BMI measurement (out of 36,389 records in the database). These same 30,624 individuals were used in all models aside from those involving ethnicity where only 30,581 individuals were used as 43 people with BMI scores did not have ethnicity recorded.

When we consider multinomial modelling MLwiN no longer has the facility to automatically ignore cases with missing data, and therefore we will use a different worksheet `ordered1.wsz` that has only the 30,581 individuals who have the response and three predictors that we have thus far considered recorded. To create this worksheet we used the Select/Omit Cases window available from the Data Manipulation menu as we will do for the worksheets in the ‘further practical ideas’ section.

We will firstly load up the worksheet and set up a simple ordered response model:
Select **Open Worksheet** from the **File** menu.

Select worksheet ‘ordered1.ws’ from the list and Click on the **Open** button.

Select the **Equations** window from the **Model** menu.

Click on the red \( y \) and Select \( bmicat \) as the \( y \) variable

Select 1-I for **N levels** and \( pserial \) as **level 1(i)**

Click on the \( N \) and choose **Multinomial**

Select **Ordered Proportional Odds** from choices of type of multinomial

Change the reference category to ‘m-obese’

Click on the **Done** button

The **Equations** window should now look as follows:

\[
\begin{align*}
\text{resp}_q & \sim \text{Ordered Multinomial} (\pi_0, \pi_3) \\
\gamma_0 & = \pi_0 \\
\gamma_1 & = \pi_0 + \pi_1 \\
\gamma_2 & = \pi_0 + \pi_1 + \pi_2 \\
\gamma_3 & = \pi_0 + \pi_1 + \pi_2 + \pi_3 \\
\gamma_4 & = 1 \\
\text{logit}(\gamma_0) & = \\
\text{logit}(\gamma_1) & = \\
\text{logit}(\gamma_2) & = \\
\text{logit}(\gamma_3) & = 
\end{align*}
\]

We next need to set up the denominator and the baseline model.

Click on the red \( n \) and select **cons**

Click on the **Done** button.

Click on the **Add Term** button.

Select ‘cons’ and click on the **Add separate coefficients** button.

Click on the **Start** button and the **Estimates** button twice.
The **Equations** window will then look as follows:

\[
\text{resp}_y \sim \text{Ordered Multinomial}(\text{cons}, \pi_y)
\]

\[
\gamma_1 = \pi_1 y, \quad \gamma_2 = \pi_2 + \pi_1 y, \quad \gamma_3 = \pi_3 + \pi_2 y, \quad \gamma_4 = \pi_4 + \pi_3 y, \quad \gamma_5 = 1
\]

\[
\logit(\gamma_1) = -1.946(0.017)\text{cons}(\leqslant \text{under})_y
\]

\[
\logit(\gamma_2) = -0.073(0.011)\text{cons}(\leqslant \text{ideal})_y
\]

\[
\logit(\gamma_3) = 1.417(0.014)\text{cons}(\leqslant \text{over})_y
\]

\[
\logit(\gamma_4) = 4.102(0.044)\text{cons}(\leqslant \text{obese})_y
\]

\[
\text{cov}(\gamma_{ij}, \gamma_{ij'}) = \gamma_{ij'}(1 - \gamma_{ij'})/\text{cons}_j \quad \forall \leqslant i
\]

Here we get a parameter for each threshold and so we can calculate the probabilities of being in each category. Note here these probabilities are based on the data with the 43 people of unknown ethnicity excluded.

The antilogit of -1.946 is 0.12499 which is the probability of being underweight. The antilogit of -0.073 is 0.48176 which is the probability of being either underweight or ideal weight. By subtraction we can then calculate the probability of being in the ideal weight category (0.48176 − 0.12499 = 0.35677). We can in fact recover the probabilities of being in each of the 5 categories from these 4 coefficients and they will correspond exactly to the observed counts. These models are sometimes called cumulative logit models as we are modelling the cumulative probabilities rather than the specific category probabilities. When modelling the cumulative probabilities we only require that the probabilities of being in each cumulative group of categories increases as the number of categories increases i.e. the threshold parameters are ordered and increasing. The difficulty with modelling the category probabilities directly is that they need to be constrained to sum to 1 which is harder to do. For unordered categories this is overcome by looking at several relative probabilities i.e. the probabilities of being in each category relative to a base category.

**Adding in predictor variables**

As with the binary logistic regression model we can next include predictors into the model. The convention in these cumulative models is to include one coefficient for the predictor that is common to all threshold boundaries. This is known as a proportional odds assumption and again simplifies the model and avoids the problem of cumulative probabilities losing their strictly increasing property for certain predictor values.

We will explain this by including sex in the model:
Choose the **Equations** window from the **Model** menu.

Click on the **Add Term** button and select ‘sex’ from the variable list.

Click on the **Add common coefficient** button.

On the window that appears click on the **Include All** button.

Click on the **Done** button and then the **More** button.

The **Equations** window will then look as follows:

Here we see a coefficient of 0.070 for *female* with the 0123 referring to the fact that this term is common to all four thresholds.

We can use the Command Interface window once again to look at probabilities from this model:

\[
\text{CALC ALOG(-1.984)} = 0.12089 \\
\text{CALC ALOG(-1.984 + 0.070)} = 0.12853
\]

Thus our model estimates the probability of being underweight for males to be 0.12089 and for females 0.12853.

In fact we see from the raw data we see the observed probabilities of being underweight are reversed with 13.7% of males being underweight and 11.5% of females being underweight.
This apparent discrepancy is due to the fact that we are assuming a proportional odds assumption. We are thinking of there being an underlying continuous BMI measure (which there is) but we are assuming that underlying risk factors will explain movements from category to category in the same linear way. This makes much more logical sense if for example the data are exam grades from education where one could imagine the process of learning is going to move you from grade to grade with an optimal category at one end of the scale.

Of course we could say that putting on weight will move you from category to category however the difference here is the optimum is not an end category. We can therefore look at fitting separate gender effects for each category:

Click on the ‘female.0123’ terms and click on the **Delete Term** button.

Click on the **Add term** button and choose ‘sex’

This time choose the **Add separate coefficients** button.

Click on the **More** button.

The **Equations** window then looks as follows:
Here we see what appear to be several significant effects of gender (by using the crude method of comparing estimates to their standard errors). These effects are recovering the table of probabilities we have already seen with more underweight males than females but also more overweight males than females until we reach the morbidly obese category where females dominate. This suggests that it might be better in this situation to consider as we have already just one threshold and do a simpler binary logistic regression. In the further practicals we suggest looking at the risk factors for being underweight and looking at this threshold in isolation.

**Multilevel Ordered Response Models**

For purely illustrate purposes we will briefly show a multilevel ordered response model. We will begin as always with a model with no predictors:

```
Select Equations from the Model menu
Click on each of the 4 female related terms in turn and choose to delete term with each of them.
Click on resp$_{ij}$ and on the screen that appears change levels to 5-ijklm
Select area as level 5(m)
Select year as level 4(l)
Select hserial as level 3(k)
Click on Done.
Click on Add Term and select cons
Click on Add Common coeffient and click on the Include all and Done buttons.
Click on the cons.0123 in the Equations window.
Click in the tickboxes for m(area long), l (year long) and k(hserial long) and remove the tick for fixed effects before pressing Done.
Click on the More button.
```

The Equations window will then appear as follows:
Here we see, as with the binary response model earlier, that there is larger variability between households (0.263), some variability between areas (0.026) and no variability between years within households. This will again be due to the lack of repetition of years within areas making it hard to estimate additional effects.

Additional predictor variables

As we saw with the binary regression we can now add predictors to this model and look at their effect so for example continuing with the gender predictor:

Select Equations window from the Model menu

Click on the Add Term button and select ‘sex’ from the variable list

Click on the Add common coefficient button.

On the window that appears click on the Include All button.

Click on the Done button and then the More button.
The Equations window will now look as follows:

```plaintext
The Equations window will now look as follows:

\[ g_{3w} \sim N(0, \Omega_g) : \Omega_g = \begin{bmatrix} 0.027(0.008) \end{bmatrix} \]

\[ f_{5lw} \sim N(0, \Omega_f) : \Omega_f = \begin{bmatrix} 0.000(0.000) \end{bmatrix} \]

\[ v_{5lw} \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} 0.261(0.022) \end{bmatrix} \]

\[ \text{cov}(y_{ijklm}, y_{ijklm}) = \frac{\gamma_{ijklm}(1 - \gamma_{ijklm})}{\gamma_{ijklm}} \]
```

Here we see a fairly similar overall effect for gender as we saw in the single level model with a significant positive effect of being female. This means that the model is saying that women are more likely than men to be in the higher overweight categories. The same issues about using one term to influence all thresholds persist.

Although we could now continue our modelling to include random slopes, other predictor variables etc. we have already seen that in this dataset it might be better to simply consider specific thresholds and binary response models.

We finish this practical by offering further practical ideas that you could try (given time) to confirm your understanding of the material.

**Further practical ideas**

If you wish to test how well you have understood the practical thus far you might like to try the following:

1. Risk factors for being underweight

In the practical we showed how to construct a binary variable that compared overweight people with those that were under or ideal weight. We might also be interested in those that are underweight and construct a variable that takes value 1 for underweight and 0
otherwise. Alternatively we might want to consider the subsample of people who are underweight or ideal weight and remove overweight people from the analysis (use worksheet under1.ws)

2. Risk factors for being under/overweight in the youngest age category

The risk factors in the 0-16 subgroup might be different and so we might like to redo the analysis with this group only (using worksheet young.ws).

3. Risk factors for the working age population

The risk factors in the two age groups 17-34 and 35-59 might differ from other age groups and so we might consider a separate analysis of these data (using worksheet middle.ws).

4. Risk factors for the older population

Finally we might look at risk factors in the oldest age group (60+) and see how these vary from those in the overall analysis (using worksheet old.ws)

5. Using MCMC estimation

In the lectures we mentioned that MCMC estimation can be better for these models but our practical has not included MCMC estimation and if you are interested you could try the models using MCMC. See the MCMC manual for details.