ST344 Lab 5: A model answer

David Firth

(Actually there is rather more here than would be needed in a 'model answer' for the Lab 5 assignment. I also take the opportunity here to explain in some more detail how to work with multiplicative models, i.e., models that are additive on the *log* scale. DF)

1. Background, and methods used

We analyse data from the American Time Use Survey for 2017. The aim here is to describe the relationship between just one time-use category — top-level category 03, "Caring for and helping household members" — and respondent age and sex.

We fit a sequence of log-linear models for mean of the time-use variable (Y_i, say, for respondent i), as a flexible function of age (a_i) and sex (s_i with values m and f). The specific models all have the form

$$\log E(Y_i) = \alpha_{s_i} + \boldsymbol{\beta}'_{s_i} \mathbf{n}_i(a_i)$$

where each $\mathbf{n}_i$ is respondent i's vector of basis values for a natural cubic spline function of age, with spline knots placed arbitrarily at ages 20, 40, 60 and 80.

The three models fitted are:

- Model 1: no dependence on sex. Here $\alpha_m = \alpha_f$ and $\boldsymbol{\beta}_m = \boldsymbol{\beta}_f$.
- Model 2: the dependence on age has the same *shape* for both sexes. This is as in Model 1, but now α_m and α_f are allowed to be different. The value of $\exp(\alpha_m - \alpha_f)$ is the ratio between mean time use of males and females, which in this model is taken to be the *same ratio at all ages*.
- Model 3: dependence on age may have different shapes for the two sexes. This is the most general model, with no constraint on the parameters α_m, α_f, $\boldsymbol{\beta}_m$ and $\boldsymbol{\beta}_f$.

The models can all be fitted in *R*, by using function `ns()` from the *splines* package to generate the basis for a natural cubic spline, and `glm()` to fit the log-linear model with variance taken to be proportional to the square of the mean.

We assess the relative support for Models 1–3 through the analysis of variance; and we will report our conclusions based on the *simplest* of those models that is not clearly rejected on the basis of the data.

2. Analysis of variance table

	Residual DF	Residual deviance	DF drop	Deviance drop	F	p-value
Model 1	10217	35953				
Model 2	10216	35469	1	483.8	22.46	$< 10^{-5}$
Model 3	10211	35368	5	101.3	0.94	0.45

(That actually is *analysis of deviance* rather than analysis of variance, since the deviances for multiplicative-error models are not exactly sums of squares. But the procedure is the same as analysis of variance; and the use of the F distribution in this more general (non-Normal) context remains justified, as an asymptotic approximation.)

From the table, it is clear that Model 1 should be rejected in favour of Model 2. But Model 3 does not improve enough upon Model 2 to justify the 5 additional parameters that are needed to fit separate spline

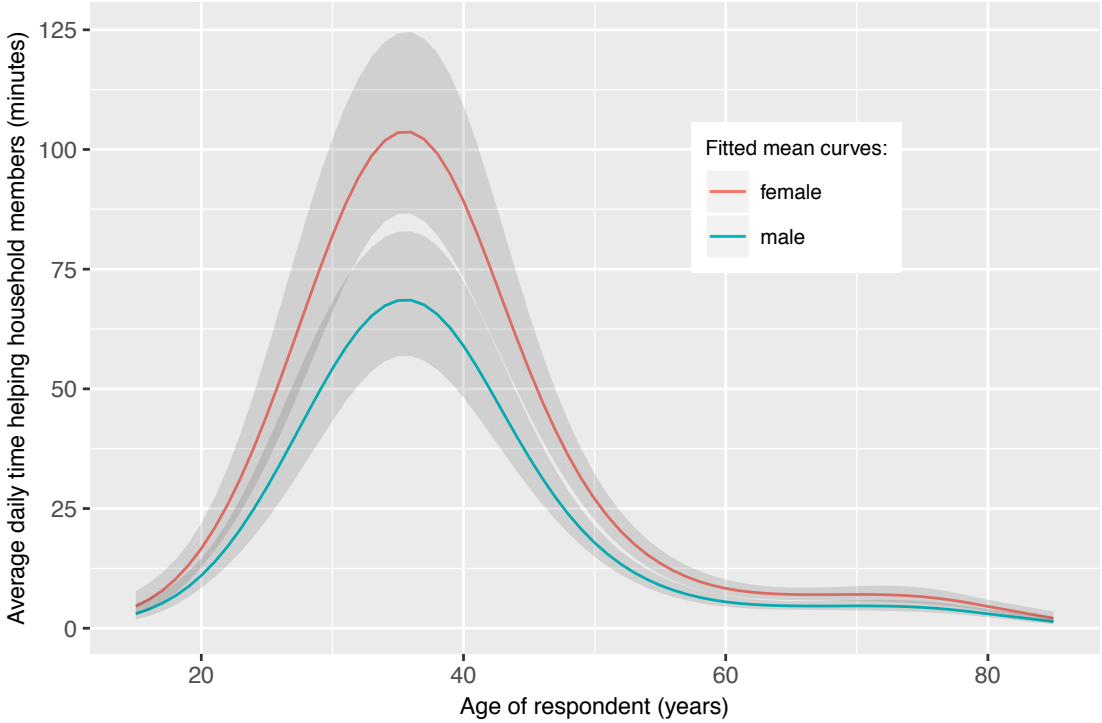
functions for male and female respondents. So we base our conclusions on Model 2 (after doing some basic checks on the residuals from Model 2, which are clearly non-normal but otherwise show no problems).

3. Conclusions

A good summary of the dependence of the average time spent on "Caring for and helping household members" upon *age* and *sex* is as follows.

- The shape of the relationship with age does not differ appreciably between males and females.
- The very youngest and the very oldest respondents report rather little time spent in this category, on average. The peak time spent (averaging around 100 minutes per day for females) is reached at around age 35.
- From the fit of Model 2: The average time spent by males is estimated to be $\exp(\hat{\alpha}_m - \hat{\alpha}_f) = \exp(-0.41) = 66\%$ of the corresponding time spent by females, at every age. An approximate 95% confidence interval for that effect on the log scale is $[-0.414-(1.96\times0.091),\ -0.414+(1.96\times0.091)] = [-0.591,\ -0.235]$. From that we can obtain a 95% confidence interval for the amount of time spent by males, as a percentage of the time spent by females, by exponentiating the endpoints of the interval that was calculated on the log scale. The resultant interval, around the already-calculated point estimate of 66%, is thus $[\exp(-0.591),\ \exp(-0.235)] = [55\%,\ 79\%]$.

The relationship with age — for males and for females — is illustrated graphically below, with 95% confidence bands around the *Model 2* fitted mean curves.



4. Limitations of this analysis

In this analysis I did not use the survey weights from the ATUS dataset. It would be interesting to see if that makes a difference to the conclusions.