Multilevel pooled radon model

Names: (signatures only please, printed names will not be counted)

1.) 4.)

2.) 5.)

3.) 6.)

Overview

Last time we examined the completely pooled model for the Minnesota radon data. This time we will look at the unpooled model and, finally, the partially pooled or hierarchical model.

The pooled model is:

$$y = a + \beta \cdot x + e$$

Where:

- y is the measured radon concentration
- a is the intercept of the regression line
- β is the coefficient of the floor number (0=basement, 1=first)
- e is the noise term, assumed to be independent normal $(0, \sigma_y)$

The unpooled model is:

$$y = a_i + \beta \cdot x + e$$

Where:

- y is the measured radon concentration
- a_i is the intercept of the regression line for county i
- β is the coefficient of the floor number (0=basement, 1=first)
- e is the noise term, assumed to be independent normal $(0, \sigma_y)$

The partially pooled model is:

$$a_i = mu_a + e_a$$

$$y = a_i + \beta \cdot x + e$$

This model has two levels and is called a multilevel or hierarchical model. The first level is:

- a_i is the intercept for county i, i=$1, 2, \ldots, 85$
- μ_a is the (hypothetical) mean of the county intercepts
- e is the noise term for the county intercept, assumed to be independent normal $(0, \sigma_a)$

and the second level is:

- y is the measured radon concentration
- a_i is the estimated intercept for county i obtained from the model above, i=$1, 2, \ldots, 85$
- β is the coefficient of the floor number (0=basement, 1=first)
- e is the noise term, assumed to be independent normal $(0, \sigma_y)$

The data elements are:

- `county` County number
- `floor_measure` Floor - 0 or 1
- `log_radon` natural log of radon concentration

While this model is nominally a simple regression, because x takes only values zero and one, it is computationally identical to a single-factor ANOVA with two levels.

The difference is that we might use this model to predict radon on the second floor by coding $x = 2$, something that would have no meaning in an ANOVA model.

The line between regression and ANOVA can be rather blurry. This model would fall under a class of models known as ”regression on dummy variables” where ”dummy variables” refers to predictors that only assume the values zero and one. This terminology is common in many fields, including economics.

Like the single-factor ANOVA with two levels, there are only two fitted values, which correspond to the mean radon level for basements and first floors.

The .Rnw files are:

- `Multilevel_unpooled.Rnw`
- `Multilevel_partially_pooled.Rnw`

Questions:

1) In the unpooled model postprocessing, we computed a point estimate of the standard deviation of the county intercepts `sigma_a` using the R `apply` function. What was its value?

2) The `print(stanfit)` output for the unpooled model contains a point extimate of `sigma_y`. What is that value?

3) What is the point estimate of the standard deviation of the county intercepts `sigma_a` in the partially pooled model (it should be in the `print(stanfit)` output)?

4) The `print(stanfit)` output for the partially pooled model contains a point extimate of `sigma_y`. What is that value?

5) How much difference is there between the unpooled and partially pooled model's estimates of `sigma_a`? Of `sigma_y`?

6) Negative radon concentrations do not make sense. Since the county intercept is the same as the predicted mean basement concentration, a model that produces a lot of negative values for this is not as plausible as one that rarely produces negative values. Compare the boxplots from the unpooled and partially pooled models. Do they indicate much difference in the likelihood of negative values for a_i?