Multilevel/Hierarchical Models, Overfitting, and Ridge Regression: A Connection

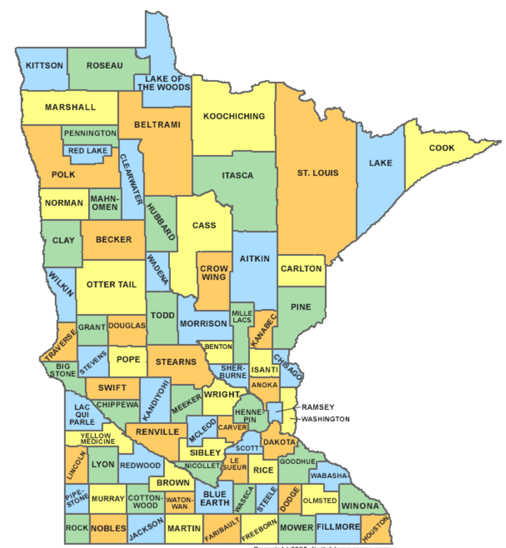
STA303/STA1002: Methods of Data Analysis II, Summer 2016

Michael Guerzhoy

Reminder: Radon Levels in Minnesota

- Radon is a radioactive gas that is known to cause lung cancer, and is responsible for several thousand of lung cancer deaths per year in the US
- Radon levels vary in different homes, and also vary in different counties

Minnesota



Minnesota counties

Partial Pooling

i-th measurement

j[i]: the county where the i-th measurement is taken

$$y_i \sim N\left(\alpha_{j[i]}, \sigma_y^2\right)$$
$$\alpha_{j[i]} \sim N(\mu_\alpha, \sigma_\alpha^2)$$

$\alpha_{j[i]}$: the mean for the county where the i-th measurement is taken

- Let $f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$
- (Approximate) Likelihood used by lme in R:
$$P\left(y_1, y_2, \dots, y_n \middle| \mu_\alpha, \sigma_y^2, \sigma_\alpha^2\right)$$
$$= (\Pi_j f\left(\alpha_j \middle| \mu_\alpha, \sigma_\alpha^2\right)) \left(\Pi_i f\left(y_i \middle| \alpha_{j[i]}, \sigma_y^2\right)\right)$$
- lme finds the α_j, σ_y^2, $\mu_\alpha, \sigma_\alpha^2$ which maximize the likelihood
- Can now look at the different α_j

Estimating the $\alpha_{j[i]}$

- Overall, we are maximizing

$$P\left(y_1, y_2, \ldots, y_n \middle| \mu_\alpha, \sigma_y^2, \sigma_\alpha^2\right)$$
$$= \left(\Pi_j f\left(\alpha_j \middle| \mu_\alpha, \sigma_\alpha^2\right)\right) \left(\Pi_i f\left(y_i \middle| \alpha_{j[i]}, \sigma_y^2\right)\right)$$

- Equivalently, we are minimizing the negative log-likelihood

$$const + \frac{\Sigma_j\left(\alpha_j - \mu_\alpha\right)^2}{2\sigma_\alpha^2} + \frac{\Sigma_i\left(y_i - \alpha_{j[i]}\right)^2}{2\sigma_y^2}$$

The county means can't be too far away from overall mean

The measurements can't be too far away from the county means

Trade-off

- We are minimizing the negative log-likelihood

$$const + \frac{\Sigma_j\left(\alpha_j - \mu_\alpha\right)^2}{2\sigma_\alpha^2} + \frac{\Sigma_i\left(y_i - \alpha_{j[i]}\right)^2}{2\sigma_y^2}$$

The county means can't be too far away from overall mean

The measurements can't be too far away from the county means

Trade-off between making the first term small by setting all α_j close to μ_α, and making the $\alpha_{j[i]}$ close to the measurements y_i in each county

Imagine we have lots of counties

- Consider one individual county j with few measurements, and imagine we are only trying to estimate α_j (imagine the means for the other counties and the σs have all been already set)
  - Minimize $\frac{(\alpha_j - \mu_\alpha)^2}{2\sigma_\alpha^2} + \frac{\Sigma_{y \in county_j}(y - \alpha_j)^2}{2\sigma_y^2}$
  - σ_α^2 large $\Rightarrow$ α_j is close to the mean of the measurements
    - Possible overfitting for county j b/c of small sample size
  - σ_α^2 small $\Rightarrow$ α_j is close to μ_α
    - The estimate for county j doesn't reflect the data for county j

Reminder: ridge logistic regression

- Minimize $NLL + \lambda(\beta_1^2 + \beta_2^2 + \cdots + \beta_k^2)$
  - NLL smaller when the model fits the training data better
- Tradeoff between small NLL and β's that are close to 0
- λ small $\Rightarrow$ $\beta's$ are close to being the $\beta's$ that make the NLL the smallest
  - Possible overfitting for county j b/c of small sample size
- λ large $\Rightarrow$ $\beta's$ are close to 0
  - The estimates for $\beta's$ are less influenced by the actual sample

Ridge Regression and Partial Pooling

- In both cases, we want to avoid overfitting due to small sample sizes
  - With partial pooling, “pull” the county means towards the overall mean
  - With ridge logistic regression, “pull” the coefficients towards 0