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

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

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

• Overall, we are maximizing

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

• Equivalently, we are minimizing the negative log-likelihood

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

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

\[ \text{const} + \frac{\sum_j (\alpha_j - \mu_\alpha)^2}{2\sigma_\alpha^2} + \frac{\sum_i (y_i - \alpha_{j[i]})^2}{2\sigma_y^2} \]

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

Trade-off between making the first term small by setting all \( \alpha_j \) close to \( \mu_\alpha \), and making the \( \alpha_{j[i]} \) close to the measurements \( y_i \) in each county

The measurements can’t be too far away from the county means
Imagine we have lots of counties

- Consider one individual county $j$ with few measurements, and imagine we are only trying to estimate $\alpha_j$ (imagine the means for the other counties and the $\sigma$s have all been already set)

- Minimize $\frac{(\alpha_j - \mu_\alpha)^2}{2\sigma_\alpha^2} + \frac{\sum_{y \in \text{county } j} (y - \alpha_j)^2}{2\sigma_y^2}$

- $\sigma_\alpha^2$ large $\Rightarrow \alpha_j$ is close to the mean of the measurements
  - Possible overfitting for county $j$ b/c of small sample size

- $\sigma_\alpha^2$ small $\Rightarrow \alpha_j$ is close to $\mu_\alpha$
  - 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 $\beta$’s that are close to 0
- $\lambda$ 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
- $\lambda$ 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