Linear Regression Models
P8111

Lecture 21

Jeff Goldsmith
April 12, 2016

Today's Lecture

- LDA review
- Model interpretation
- Model estimation
- Example (pig data!)

Recall the setting

- We observe data y_{ij}, x_{ij} for subjects $i = 1, \ldots I$ at visits $j = 1, \ldots, J_i$

- Overall, we pose the model

$$y = X\beta + \epsilon$$

where $\text{Var}(\epsilon) = \sigma^2 V$ and

$$V = \begin{bmatrix} V_1 & 0 & \ldots & 0 \\ 0 & V_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & & V_I \end{bmatrix}$$

General ideas

- We discussed two approaches:
  - Random (mixed) effects models introduce *random* subject effects and assume uncorrelated errors
  - Marginal models include only fixed effects and assume correlated subject-level errors
- In some cases these introduce equivalent correlation structures; an example is that a random intercept model induces uniform within-subject correlations
- Marginal models separate the mean structure and correlation; random effect models induce correlation by introducing random quantities into the mean

Some comparisons

- Marginal models may be less immediately obvious or interpretable
- However, marginal models can be more robust to misspecification through robust SE estimates

Marginal model

The marginal model formulation is

$$y = X\beta + \epsilon$$

where

- $\epsilon \sim \mathrm{N}\left[0, \nu^2 V\right]$

- This approach focuses on the *marginal* distribution of y, rather than on a subject-level *conditional* distribution

- Coefficients have a marginal interpretation – compare subjects based only on covariate values

- Interpretation is analogous to a cross-sectional model

Remember GLS

Given the model

$$y = X\beta + \epsilon$$

where $\epsilon \sim N(0, \sigma^2 V)$ with V known, we are essentially assuming

$$y \sim N(X\beta, \sigma^2 V)$$

Using MLE, we find that $\hat{\beta}_{GLS} = (X^T V^{-1} X)^{-1} X^T V^{-1}$

Estimation – marginal model

- If we can use MLE when V is known, maybe we can use MLE to estimate V as well
- Our log likelihood function is

$$l(\boldsymbol{\beta}, \sigma^2, V; \boldsymbol{y}, \boldsymbol{X}) = -\frac{1}{2}\left[n\log(\sigma^2) + \log(|V|) + \frac{1}{\sigma^2}(\boldsymbol{y} - \boldsymbol{X\beta})^T V^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})\right]$$

- Using profile likelihood, we find that for any V_0

$$\hat{\boldsymbol{\beta}}(V_0) = (\boldsymbol{X}^T V_0^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T V_0^{-1}$$

Estimation – marginal model

- Estimation of V and σ is done through restricted maximum likelihood
  - Standard MLE produces biased variance estimates; REML adjusts for the number of fixed effects components that are estimated
- Often V is structured parametrically to ease estimation and computation
- We won't worry about how this is done

Random intercept model

A random intercept model with one covariate is given by

$$y_{ij} = \beta_0 + b_i + \beta_1 x_{ij} + \epsilon_{ij}$$

where

- $b_i \sim \mathrm{N}\left[0, \tau^2\right]$
- $\epsilon_{ij} \sim \mathrm{N}\left[0, \nu^2\right]$

Random intercept model

More compactly, we write

$$y = X\beta + Zb + \epsilon$$

where

- $b \sim \mathrm{N}\left[0, \tau^2 I_I\right]$
- $\epsilon \sim \mathrm{N}\left[0, \nu^2 I_n\right]$

Random intercept model

In the model

$$y = X\beta + Zb + \epsilon$$

why might we assume b are random rather than estimating them as fixed effects?

Random intercept model

In the model

$$y = X\beta + Zb + \epsilon$$

why might we assume b are fixed rather than considering them to be random?

Random intercept model interpretation

Random effect models have a conditional interpretation.

- Mean conditions on subject effects: $E(y_{ij}|x_{ij}, \boldsymbol{\beta}, b_i)$
- To derive a marginal (averaged across subjects) mean, one can use iterated expectations

- (For an identity link, the coefficients also have a marginal interpretation – this is not true for generalized RE models)

Estimation – random effect model

Still done using MLE, but now we include random effects; our model is

$$y = X\beta + Zb + \epsilon$$

where

- $b \sim \mathrm{N}\left[0, \tau^2 I_I\right]$
- $\epsilon \sim \mathrm{N}\left[0, \nu^2 I_n\right]$

Estimation – random effect model

Estimation – random effect model

Estimation – BLUPs

- Our estimate for fixed and random effects are

$$\left[\begin{array}{c} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{b}} \end{array} \right] = \left(\boldsymbol{C}^T \boldsymbol{C} + \frac{\nu^2}{\tau^2} R \right)^{-1} \boldsymbol{C}^T \boldsymbol{y}$$

- These are referred to as "BLUPs" (the "P" is for "predictions")
- The variances ν^2 and τ^2 are estimated via REML
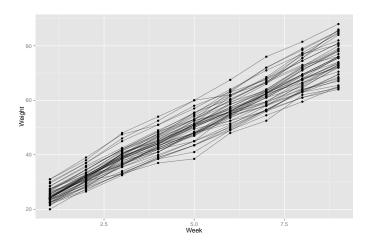
BLUPs

BLUPs are a lot like BLUEs

- They are the best linear unbiased predictors when one has both fixed and random effects
- We derived them using Normal distributions, but even without distributional assumptions these are BLUP
- The Normal distributions help with the assumptions are satisfied, in that one can get distribution-based inference
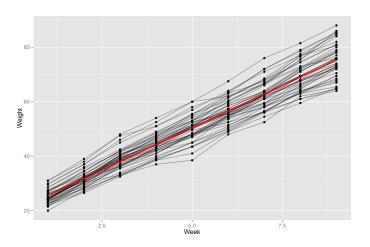
Penalized spline regression

Note that BLUPs look an awful lot like penalized regression estimates

- The "mixed model framework" is commonly used for penalized spline estimation
- The "tuning parameter" is a ratio of variances, and can be estimated via REML (rather than CV)
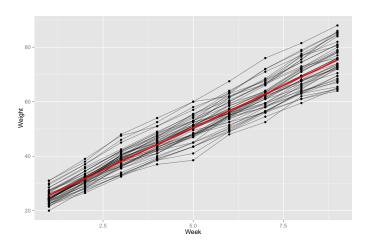- This approach provides a method for inference
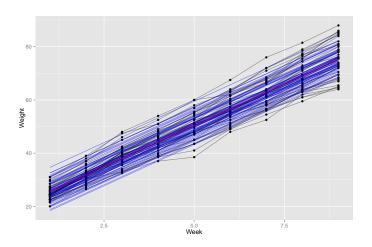
Pig data

Pig data

OLS fit for pig data

Pig data

OLS code

```
>   lin.mod = lm(weight ~ num.weeks, data = pig.weights)
>   summary(lin.mod)

Residuals:
    Min      1Q   Median      3Q     Max
-11.9051  -2.5348  -0.1952   2.5949  13.1751

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.35561    0.46054   42.03   <2e-16 ***
num.weeks    6.20990    0.08184   75.88   <2e-16 ***

Residual standard error: 4.392 on 430 degrees of freedom
```

Pig data

Marginal model fit for pig data

Pig data

Marginal model code

```
> library(gee)
> marg.mod = gee(weight ~ num.weeks, id = id.num, corstr = "exchangeable",
                 data = pig.weights)
> summary(marg.mod)

Model:
 Link:                      Identity
 Variance to Mean Relation: Gaussian
 Correlation Structure:     Exchangeable

Summary of Residuals:
       Min          1Q      Median          3Q         Max
-11.9050926  -2.5347801  -0.1951968   2.5949074  13.1751157

Coefficients:
             Estimate Naive S.E.   Naive z Robust S.E. Robust z
(Intercept) 19.355613  0.5983680  32.34734  0.39963854 48.43280
num.weeks    6.209896  0.0393321 157.88366  0.09107443 68.18485

Working Correlation
          [,1]      [,2]      [,3]      [,4] ...
 [1,] 1.0000000 0.7690313 0.7690313 0.7690313 ...
 [2,] 0.7690313 1.0000000 0.7690313 0.7690313 ...
 [3,] 0.7690313 0.7690313 1.0000000 0.7690313 ...
 [4,] 0.7690313 0.7690313 0.7690313 1.0000000 ...
 ...
```
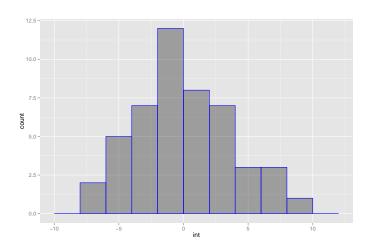
Pig data

Mixed effect model fit for pig data

Pig data

Mixed effect model code

```
> library(lme4)
> ranef.mod = lmer(weight ~ (1 | id.num) + num.weeks, data = pig.weights)
> summary(ranef.mod)
Linear mixed model fit by REML
Formula: weight ~ (1 | id.num) + num.weeks
   Data: pig.weights
  AIC  BIC logLik deviance REMLdev
 2042 2058  -1017     2030    2034
Random effects:
 Groups   Name        Variance Std.Dev.
 id.num   (Intercept) 15.1418  3.8913
 Residual              4.3947  2.0964
Number of obs: 432, groups: id.num, 48

Fixed effects:
            Estimate Std. Error t value
(Intercept) 19.35561    0.60311   32.09
num.weeks    6.20990    0.03906  158.97

> (15.1418) / (15.1418 + 4.3947)
[1] 0.7750518
```

Pig data

Histogram of estimated random intercepts

Today's big ideas

- Estimation in LDA
- Example + code

Potential reading on mixed effects models – Semiparametric Regression (Ruppert, Wand, Carroll) Ch 4