

Masters Comprehensive Exam
Stat 505-506 August 2012

1. Use the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{V})$.
For this problem, assume \mathbf{X} is of full column rank.
 - (a) State the Gauss Markov Theorem. (10 pts)
 - (b) What is the (sampling) distribution of $\mathbf{X}\hat{\boldsymbol{\beta}}$? ($\hat{\boldsymbol{\beta}}$ is the Gauss-Markov estimate of $\boldsymbol{\beta}$) (10 pts)

2. Suppose we are fitting a four-level treatment and a linear continuous covariate, x , allowing slopes to change with treatment. Assume errors are iid $N(0, \sigma^2)$ and that we have five observations per treatment group with a reasonable range of x values within each of the groups.
 - (a) Write out the “treatment effects” model using Greek letters for all parameters. Use β_0 for overall intercept, τ_i 's for treatment effects, β_1 for overall slope, and γ_i 's for the interaction terms. (5 pts)
 - (b) Give an example of a non-estimable function involving a γ_i . Explain why it is not estimable.? (5 pts)
 - (c) How do we work around the estimability problem to answer questions of interest about the slopes? (10 pts)

3. Suppose that we have a multilevel model which we represent as:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_ib_i + \boldsymbol{\varepsilon}_i, i = 1, 2, \dots, 12$$

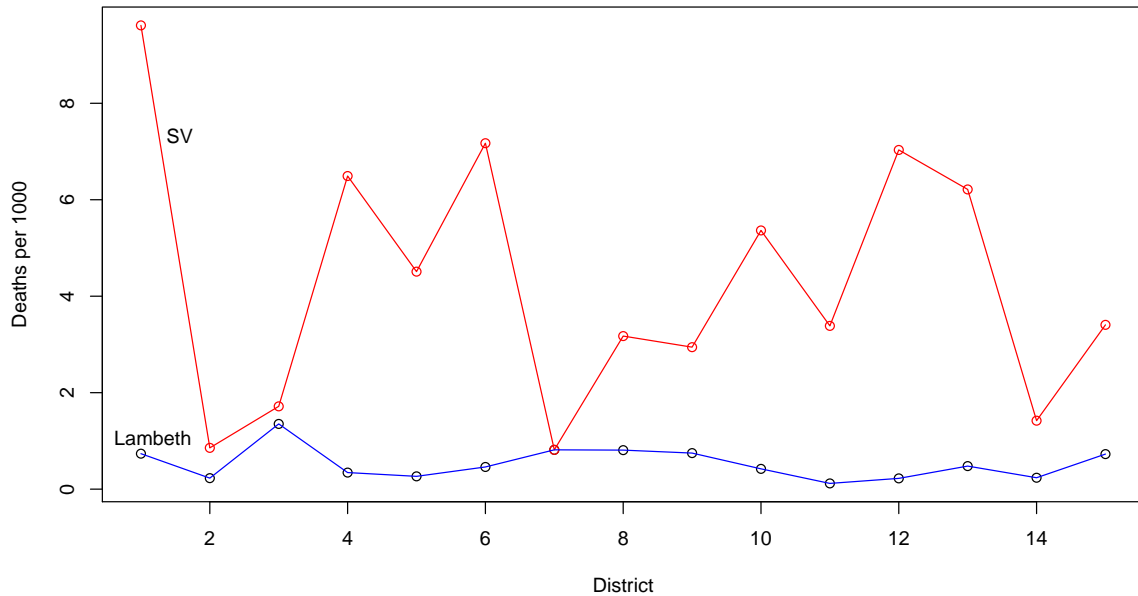
where $\boldsymbol{\varepsilon}_i \sim iid N(0, \sigma^2\mathbf{I})$ independent of $b_i \sim iid N(0, \sigma_b^2)$ and each response vector \mathbf{y}_i contains 4 values, \mathbf{X}_i 's are 4 by 2, and each \mathbf{Z}_i is a vector of 4 ones.

- (a) What is the correlation between two responses in the same \mathbf{y}_i vector? (5 pts)
- (b) What is the distribution of \mathbf{y}_i ? (5 pts)
- (c) What is the MLE of $\boldsymbol{\beta}$? (5 pts)

4. John Snow collected data on cholera deaths in London in 1849 and traced the source of water for houses in which a death occurred to get the following table where “population” is an estimate of the number of people in the district who are supplied water by South & Vauxhall or by Lambeth water companies.

District	Water Source			
	South & Vauxhall		Lambeth	
	population	deaths	population	deaths
Christchurch	9613	11	6409	13
Kent-road	10876	52	7250	5
Borough-road	9517	61	6345	7
London-road	10702	21	7134	8
Trinity	12553	52	8369	6
St. Peter Walworth	17917	84	11944	4
St. Mary	8420	19	5613	1
Waterloo-road 1st	8453	8	5635	2
Waterloo-road 2nd	11009	25	7339	8
Lambeth Church 1st	11045	6	7364	9
Lambeth Church 2nd	16070	34	10714	13
Kennington 1st	14557	63	9704	5
Kennington 2nd	11309	34	7539	3
Brixton	8766	5	5844	2
St. George	9509	34	6340	4

The Lambeth company drew water from the Thames above the points where raw London sewage was discharged into the river, but S&V company’s intake was lower. The companies both served each of these districts. Death rates seem quite different:



The data were stacked into a new data frame containing district, water source, population and deaths (4 columns) and the following model was fit:

```
grandDF$surv <- grandDF$pop - grandDF$deaths
grand.fit2 <- lmer(cbind(deaths, surv) ~ water + (1|dist),
                  data = grandDF, family=binomial)
```

and these output were obtained:

```
Generalized linear mixed model fit by the Laplace approximation
Formula: cbind(deaths, surv ) ~ water + (1 | dist)
Data: grandDF
AIC    BIC logLik deviance
125.4 129.7 -59.72    119.4
Random effects:
Groups Name          Variance Std.Dev.
dist  (Intercept) 0.26261  0.51245
Number of obs: 30, groups: dist, 15

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.7001     0.1710  -45.03  <2e-16
waterSV       2.1428     0.1145   18.72  <2e-16
```

- What effect does `family = binomial` have? Explain the assumed model. (5 pts)
- Interpret the `waterSV` fixed effect. What does this estimate mean about the odds of cholera death? (5 pts)
- Interpret the district effects. (5 pts)
- Snow was trying to convince the medical establishment (before germ theory was proposed) that cholera was a contagion spread by drinking contaminated water. What assumptions are needed to provide causal inference in this case? (10 pts)