Midterm Review STAT 505 Fall 2014

Linear model: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$; $\boldsymbol{\epsilon} \sim (0, \sigma^2\boldsymbol{V})$

1. What condition on the $\boldsymbol{X}$ matrix assures us that all individual elements of $\boldsymbol{\beta}$ are estimable? In what common situations does that condition hold?

2. What assumptions are used for the Gauss-Markov Theorem? What is the "punch line"? Some people use $\boldsymbol{\epsilon} \sim (0, \sigma^2\boldsymbol{I})$, others use $\boldsymbol{\epsilon} \sim (0, \sigma^2\boldsymbol{V})$. Be able to show equivalence.

3. The t-tests printed in a coefficient summary table are conditional on what? With several predictors in a model, the output of an R `anova(model)` command prints F tests which are conditional on what?

4. How does the presence of a strong interaction effect change interpretation in a two-way ANOVA or in an ANCOVA model?

5. Why might we not remove a main effect in a two-way ANOVA model when it has a large p-value?

6. When can we draw causal inference between a predictor and a response?

7. How do we determine the scope of inference of a study? (There are 2 questions to answer.)

8. How does R parametrize a factor with 4 levels? SAS? When will the two packages give different estimates of an estimable contrast after fitting a linear model? How does this relate to the idea of a generalized matrix inverse?

9. How do we diagnose a problem with nonconstant variance? Does nonconstant variance bias coefficient estimates? Explain two ways to handle the problem of nonconstant variance.

10. Explain the components of the mixed model:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + b_i\mathbf{1} + \boldsymbol{\epsilon}_i$$

for a randomized block design with 4 treatments each applied once in each of 5 blocks using a mixed model. Under the usual assumptions of iid Gaussian errors and random effects, derive the correlation between two measurements within the same block, and for two measurements in different blocks.

11. When does a semivariogram plot indicate that a spatial correlation structure is needed? Be able to estimate range and nugget.

12. What plot is used to look for temporal autocorrelation? What patterns indicate that AR1 structure is needed? compound symmetric? How would you test to see that independence is violated?

13. If we have 4 observations on each individual, equally spaced in time, and a CS correlation estimated as 0.40, what is the estimate of the within-individual correlation matrix $\mathbf{R}_i$? Similarly know how symmetric correlations are structured.

14. Again with 4 observations per individual, assume the variance is proportional to time (t_i) to power 2α. Given times 1, 3, 5, and 7 days, and $\alpha = 0.3$, write out $\widehat{\boldsymbol{D}_i}$, diagonal matrix of standard deviations.

15. Given both Compound Symmetric correlations and variance proportional to $x_i^{2\alpha}$ (where x_i's are observed predictors) in (14), write out the individual's estimated variance-covariance matrix, $\widehat{\boldsymbol{V}}_i$.

16. Non-constant Variance
    Transformations: The Box-Cox approach makes residuals more normal, a Δ-method approach can stabilize variance.

$$\text{If } \text{var}(y) \propto \mu_y^{2\alpha},\ (\alpha \neq 0) \text{ then } \text{var}(y^{-\alpha+1}) \approx \text{var}(y)[(-\alpha+1)\mu_y^{-\alpha}]^2 \propto \mu_y^{2\alpha-2\alpha} = 1$$

    Weighting: Find weights proportional to inverse variance – why is iteration required? (Also true for correlation models).

17. Given the output of a linear model fit with R, be able to compute the variance of any estimable linear combination, $\boldsymbol{c}^T\boldsymbol{\beta}$ and build a confidence interval for it. Know it's distribution when residuals are normally distributed, and for "large" samples.

18. Diagnostics
    What are the "usual" assumptions for a linear model?

    (a) Plot residuals versus fits. What problems might show up here and in the scale-location plot?

    (b) qqnorm plots on residuals. When do we need to worry about non-normality? Should 48 residual points all be right on the line? When does CLT kick in?

    (c) leverage and Cook's Distance plots to identify influential points.

    When should we worry about one point being too influential?

19. What approaches are available for cases when the normality assumption is not appropriate?

20. LRT: What are the degrees of freedom for an "Extra Sums of Squares" (ESS) F test for nested models. Null and alternative hypotheses? When do we reject? If we fail to reject, what is the conclusion? What assumptions are needed?

21. Be able to set up a test of contrasts to be equivalent to an ESS F test. What assumptions are needed?