Statistics 600 Problem Set 5

Due Friday, December 3rd at midnight

1. Suppose we have two matrices X_a and X_b, where X_a is $n \times p+1$, X_b is $n \times p+q$, and $\text{col}(X_a) \subset \text{col}(X_b)$. We observe $y = Ey + \epsilon$, where $Ey \in \text{col}(X_a)$. Suppose we compare the models $Ey \in X_a$ and $Ey \in X_b$ using BIC. Get an exact expression for the probability that X_b is favored by BIC over X_a for a given sample size n, in the setting where the errors ϵ are Gaussian. Then give conditions for this probability to go to zero in a more general (non-Gaussian) setting.

2. Suppose we have n clusters of size n_i, for a total sample size of $N = \sum_{i=1}^{n} n_i$. Further, suppose that within each cluster the observations are exchangeable (given X), so that within cluster i, the covariance of $y_i = (y_{i1}, \ldots, y_{in_i})$ is equal to $\sigma^2 I_{n_i \times n_i} + \tau^2 1_{n_i} 1'_{n_i}$. We fit a linear model using these data, yielding an estimate $\hat{\beta}_{\text{gls}} \in \mathcal{R}^{p+1}$. (i) Derive an analytic expression for $\text{cov}(\hat{\beta}_{\text{gls}})$. (ii) Derive a simplified version of this covariance matrix that holds when all covariates except the intercept are centered within each cluster. (iii) Use the Sherman-Morrison-Woodbury identity in the centered setting of (ii) to determine the relationship between $\text{cov}(\hat{\beta}_{\text{gls}})$ for a general τ^2, and for the setting where $\tau^2 = 0$. (iv) Numerically evaluate your results from (i)-(iii) in a few scenarios and briefly explain your findings.

3. Continuing with #2 above, suppose the true value of τ^2 is positive, and we use a working model in which the observations are taken to be uncorrelated (i.e. we estimate β using OLS). (i) What is $\text{cov}(\hat{\beta})$ in this case? (ii) Specialize your answer from part (i) to the setting where $X'X = N I_{p+1 \times p+1}$. (iii) Numerically evaluate your results from (i)-(ii) in a few scenarios and briefly explain your findings.

4. Using the NHANES data from the last assignment, create a cluster variable based on all possible distinct combinations of `SDMVSTRA` and `SDMVPSU`. Then take your best model for explaining blood pressure variation in terms of age, sex, and BMI (from the last problem set) and use its residuals to estimate the ICC for these clusters. Then, report the standard errors for all model parameters using an ICC of zero, and the estimated ICC.

5. Suppose we have a design matrix $X \in \mathcal{R}^{n \times p}$ and the data are partioned into q clusters. Consider a fixed effects analysis of an outcome $y \in \mathcal{R}^n$. Derive a formula for the expected value of $\widehat{\text{var}}(\hat{\theta}_1, \ldots, \hat{\theta}_q)$. Now suppose that the θ_j are actually random variables with variance τ^2. Describe how τ^2 can be estimated in this analysis. Assess these ideas using the same model and data from exercise #4.

6. Suppose we have a primary variable of interest $x \in \mathcal{R}$, a confounder $z \in \mathcal{R}$, and the data are partitioned into clusters. The mean structure is $E[y|x, z] = \beta_0 + \beta_1 x + \beta_2 z$. Consider situations where the ICC of z with respect to the clusters is large. Conduct a simulation study to assess the coverage probability of the 95% confidence interval for β_1 for GLS and fixed effects analyses.

7. Consider the estimated ICC based on applying the method of moments to the OLS residuals, as given in the course notes. In contrast, we can fit a mixed effects model using maximum likelihood (or REML). Conduct a simulation study to assess the bias and variance of the estimated ICC using these two techniques.