Exercise #6
Due: February 24, 2010

Biostat/Stat 571
Winter 2010
P. Heagerty

Reading: • Verbeke & Molenberghs, Chapters 5 & 6

**** GIVEN THE SHORT WEEK – PLEASE CHOOSE ONE OF THESE QUESTIONS FOR THE EXERCISE THIS WEEK *****

Pre-post Designs

1. A clinical investigation randomizes individuals to receive either active treatment (`TX=1`) or to control (`TX=0`) after first recording a baseline measurement, Y_{i0}. Subsequent follow-up records an outcome at (at least one) follow-up time for each participant, Y_{i1}. The goal of the study is to assess whether these is an impact on Y due to treatment. There are a number of potential analyses that can be proposed for this study design. Let `TX` denote the treatment assignment and let `post` be an indicator for the follow-up time. Assume that we have m subjects in each group. Assume that $\sigma^2 = \text{var}(Y_{i0}) = \text{var}(Y_{i1})$ (although for certain MLE estimators we will consider the relaxation to $\sigma_0^2 = \text{var}(Y_{i0})$, $\sigma_1^2 = \text{var}(Y_{i1})$).

(a) Consider the regression model:

$$E(Y_{ij} \mid \boldsymbol{X}_i) = \beta_0 + \beta_1 \cdot \texttt{TX}_i + \beta_2 \cdot \texttt{post}_{ij} + \gamma \cdot \texttt{TX}_i \cdot \texttt{post}_{ij} \quad .$$

Please provide an interpretation of the parameter γ appropriate for a general scientific audience.

(b) Consider a repeated measures model for $\boldsymbol{Y}_i = (Y_{i0}, Y_{i1})$, that assumes $\text{var}(Y_{ij}) = \sigma_j^2$ and $\text{cov}(Y_{i0}, Y_{i1}) = \sigma_0\sigma_1\rho$, and assumes multivariate normality. Show that the MLE for γ is given by the difference of the differences: $\widehat{\gamma}^{(1)}$ equals the mean of $Y_{i1} - Y_{i0}$ for the treatment minus the mean of $Y_{i1} - Y_{i0}$ for the control group.

(c) Calculate the variance of $\widehat{\gamma}^{(1)}$ under the assumption that $\sigma_0^2 = \sigma_1^2$.

(d) Another model uses the fact that groups were randomized to constrain the means at baseline, $E(Y_{i0} \mid \texttt{TX}_i = 0) = E(Y_{i0} \mid \texttt{TX}_i = 1)$. This model can be written as

$$E(Y_{ij} \mid \boldsymbol{X}_i) = \beta_0 + \beta_2 \cdot \texttt{post}_{ij} + \gamma \cdot \texttt{TX}_i \cdot \texttt{post}_{ij} \quad .$$

Derive the MLE for γ in the constrained model (with covariance assumptions as in 1(b) above). Denote this estimator as $\widehat{\gamma}^{(2)}$. HINT: factor the likelihood $f(\boldsymbol{Y}_i) = f(Y_{i0})f(Y_{i1} \mid Y_{i0})$.

(e) Calculate the variance of $\widehat{\gamma}^{(2)}$ under the assumption that $\sigma_0^2 = \sigma_1^2$, and assuming σ_0, σ_1, ρ are fixed.

(f) The estimators $\widehat{\gamma}^{(k)}$ are special cases of the general estimator

$$\widehat{\gamma}(\alpha) = \overline{Y_{i1}(\texttt{TX}=1) - \alpha Y_{i0}(\texttt{TX}=1)} - \overline{Y_{i1}(\texttt{TX}=0) - \alpha Y_{i0}(\texttt{TX}=0)}$$

Another common proposed estimator is $\widehat{\gamma}(0)$ which simply compares the means at the follow-up time and ignores the baseline. Calculate the expectation of $\widehat{\gamma}(\alpha)$ for any fixed α assuming the model in (d) holds. What is the variance of $\widehat{\gamma}(\alpha)$? When is $\widehat{\gamma}(1)$ more precise than $\widehat{\gamma}(0)$? What is the optimal choice of α?

(g) Given the information from (a)-(f) suggest an appropriate analysis of "change" when there are multiple follow-up measurements.

Data Analysis

2. **MACS Data – CD4 and Viral Load**: In the last exercise we considered data from the Multicenter Aids Cohort Study (MACS). Scientific interest is in whether the rate of decline in CD4 is associated with the baseline viral load measurement. In this exercise we will use regression methods to answer this question.

(a) Use appropriate estimation methods to fit the model:

$$E(Y_{ij} \mid \boldsymbol{X}_i) = \beta_0 + \beta_1 \cdot \texttt{month}_{ij} + \beta_2 \cdot X_i + \beta_3 \cdot \texttt{month}_{ij} \cdot X_i$$

where X_i is the baseline viral load measurement (or a suitable transformation). Interpret your results. First give an interpretation of the parameters in this regression model. Second, comment on whether there is a significant association between the baseline viral load measurement and the rate of decline in CD4 based on your estimates for this model.

(b) The model in (a) makes some strong assumptions about how the rate of decline (slope for month) differs with the baseline viral load measurement. As a way of checking this model consider using viral load after categorizing this variable (for example create a factor based on the quartiles of viral load). Your regression model will take the form:

$$E(Y_{ij} \mid \boldsymbol{X}_i) = \beta_0 + \beta_1 \cdot \texttt{month}_{ij} + \sum_{k=2}^{C} \beta_{2,k} \cdot X_i(k) + \sum_{k=2}^{C} \beta_{3,k} \cdot \texttt{month}_{ij} \cdot X_i(k)$$

where $X_i(k)$ is a dummy variable that indicates the category for viral load (with C total categories). Interpret your results. First give an interpretation of the parameters in this regression model. Second, comment on whether there is a significant association between the baseline viral load measurement and the rate of decline in CD4 based on your estimates for this model.

(c) Each of the models considered above can be viewed as examples of a "varying coefficient" model:

$$E(Y_{ij} \mid \boldsymbol{X}_i) = \gamma_0(X_i) + \gamma_1(X_i) \cdot \texttt{month}_{ij}$$

Note that in this model the slope for time (month) depends on the value of the covariate X_i. In model (a) we have assumed that $\gamma_1(X_i) = (\beta_1 + \beta_3 \cdot X_i)$ while in model (b) we use a discrete function where $\gamma_1(X_i) = [\beta_1 + \sum_{k=2}^{C} \beta_{3,k} \cdot X_i(k)]$. Can we exploit the advantages of models (a) and (b): in model (a) we use viral load in its continuous form, but make strong functional assumptions; while in model (b) we use a categorical version of viral load but make no functional assumptions as to how the rate of decline differs for the different viral load categories.

If we let $\gamma_0(X_i)$ and $\gamma_1(X_i)$ take richer functional forms than the linear form used in model (a) then we can provide a flexible description of how the rate of decline differs for different values of baseline viral load. For example, we may allow a quadratic function for the coefficients:

$$\begin{aligned} \gamma_0(X_i) &= \gamma_{0,0} + \gamma_{0,1} \cdot X_i + \gamma_{0,2} \cdot X_i^2 \\ \gamma_1(X_i) &= \gamma_{1,0} + \gamma_{1,1} \cdot X_i + \gamma_{1,2} \cdot X_i^2 \end{aligned}$$

This model assumes that the rate of decline in CD4 will vary with viral load according to a quadratic curve as given by the function $\gamma_1(X_i)$.

This model can be fit using standard correlated data regression methods as it corresponds to a standard linear mean model:

$$\begin{aligned} E(Y_{ij} \mid \boldsymbol{X}_i) &= \gamma_0(X_i) + \\ &\quad \gamma_1(X_i) \cdot \texttt{month}_{ij} \\ &= \gamma_{0,0} + \gamma_{0,1} \cdot X_i + \gamma_{0,2} \cdot X_i^2 + \\ &\quad \left(\gamma_{1,0} + \gamma_{1,1} \cdot X_i + \gamma_{1,2} \cdot X_i^2\right) \cdot \texttt{month}_{ij} \end{aligned}$$

This shows that the model can be written in terms of the "basis" elements for the functions $\gamma_0(X_i)$ and $\gamma_1(X_i)$, including product terms with `month`.

Use a varying coefficient model for the rate of decline in CD4 that characterizes how the rate of decline depends on the value of the baseline viral load. I recommend that you use natural splines for the coefficient functions and simply choose 2 knots – but you are welcome to choose a different parametric spline method if you desire.

Plot the estimated coefficient function $\gamma_1(X_i)$ with point-wise 95% confidence bands, and interpret specific values. Does this plot suggest that the model given in (a) is adequate?

(d) [**optional**] Consider use of a cubic spline for $\gamma_1(X_i)$ – plot your estimated function with 95% pointwise confidence intervals and compare results to (a)–(c).

(e) [**optional**] Compare the inference for model (a) based on WLS or LMM using different weight/covariance matrices. Are results/conclusions sensitive to the choice of weight matrix?