Biost/Stat 533
Emerson, Spring 14
Midterm Examination
May 7, 2014

This examination is closed book, closed notes. You have 80 minutes to complete the exam. Concise answers are greatly to be preferred, however, in all cases you should provide enough detail to make clear the reasoning behind your answer. Each problem is worth 25 points.

If there are any problems that you believe are not solvable without making additional assumptions, state clearly the (reasonable) assumptions you made in order to solve the problem.

For all problems you may adopt the following notation.

- $\vec{1}_n$ is the n-vector having every element equal to 1.
- $\mathbf{I}_n$ is the n dimensional identity matrix.
- For random n-vector $\vec{X} = (X_1, \ldots, X_n)^T$:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \qquad S_{XX} = \sum_{i=1}^n (X_i - \overline{X}_n)^2 \qquad v_X^2 = \frac{1}{n} S_{XX} \qquad s_X^2 = \frac{n}{n-1} v_X^2$$

- Given two random n-vectors $\vec{X}$ and $\vec{Y}$:

$$S_{XY} = \sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n) \qquad r_{XY} = \frac{\frac{1}{n} S_{XY}}{\sqrt{v_X^2\, v_Y^2}}$$

- Analogous notation can be used for random vectors $\vec{W}$, $\vec{Z}$, etc.
- Let $\mathbf{A}$ be a $p \times p$ invertible matrix that is partitioned into $r \times r$ matrix $\mathbf{R}$, $s \times s$ matrix $\mathbf{S}$, and $r \times s$ matrix $\mathbf{T}$:

$$\mathbf{A} = \begin{bmatrix} \mathbf{R} & \mathbf{T} \\ \mathbf{T}^T & \mathbf{S} \end{bmatrix}.$$

Then the inverse of $\mathbf{A}$ can be found as either

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{R}^{-1} + \mathbf{R}^{-1}\mathbf{T}(\mathbf{S} - \mathbf{T}^T\mathbf{R}^{-1}\mathbf{T})^{-1}\mathbf{T}^T\mathbf{R}^{-1} & -\mathbf{R}^{-1}\mathbf{T}(\mathbf{S} - \mathbf{T}^T\mathbf{R}^{-1}\mathbf{T})^{-1} \\ -(\mathbf{S} - \mathbf{T}^T\mathbf{R}^{-1}\mathbf{T})^{-1}\mathbf{T}^T\mathbf{R}^{-1} & (\mathbf{S} - \mathbf{T}^T\mathbf{R}^{-1}\mathbf{T})^{-1} \end{bmatrix}$$

or

$$\mathbf{A}^{-1} = \begin{bmatrix} (\mathbf{R} - \mathbf{T}\mathbf{S}^{-1}\mathbf{T}^T)^{-1} & -(\mathbf{R} - \mathbf{T}\mathbf{S}^{-1}\mathbf{T}^T)^{-1}\mathbf{T}\mathbf{S}^{-1} \\ -\mathbf{S}^{-1}\mathbf{T}^T(\mathbf{R} - \mathbf{T}\mathbf{S}^{-1}\mathbf{T}^T)^{-1} & \mathbf{S}^{-1} + \mathbf{S}^{-1}\mathbf{T}^T(\mathbf{R} - \mathbf{T}\mathbf{S}^{-1}\mathbf{T}^T)^{-1}\mathbf{T}\mathbf{S}^{-1} \end{bmatrix}.$$

1. Consider a regression model in which response variables $\vec{Y} = (Y_1, \ldots, Y_n)^T$ satisfy

$$\vec{Y} \mid \mathbf{X} \sim (\mathbf{X}\vec{\beta}, \Sigma),$$

where $\mathbf{X}$ is a $n \times p$ design matrix.

   a. Provide a formula for the ordinary least squares estimator $\widehat{\vec{\beta}_{OLS}}$ and a formula for its standard error.

   b. Provide a formula for the generalized least squares estimator $\widehat{\vec{\beta}_{GLS}}$ that has the minimal variance among LS estimators. Provide a formula for its standard error.

   c. Supposing $\Sigma = \sigma^2 \mathbf{V}$ for some known form of $\mathbf{V}$ with σ^2 unknown, provide an estimator of σ^2 and describe its optimality properties.

2. Consider a regression model in which response variables $\vec{Y} = (Y_1, \ldots, Y_n)^T$ satisfy

$$\vec{Y} \mid \mathbf{X} \sim (\mathbf{X}\vec{\beta}, \Sigma),$$

where $\mathbf{X}$ is a $n \times p$ design matrix.

   a. Provide a definition for an estimable function.

   b. State and prove the Gauss-Markov Theorem in the setting of a $\Sigma = \sigma^2 \mathbf{I}_n$.

   c. Extend the result in part b to an arbitrary covariance matrix Σ.

3. Let response $\vec{Y}$ and covariates $\vec{X}$ and $\vec{W}$ be measurements made on n independent subjects. Consider linear regression models

$$(Y_i \mid X_i = x_i) = \beta_0 + x_i\beta_x + \epsilon_i^*$$
$$(Y_i \mid X_i = x_i, W_i = w_i) = \gamma_0 + x_i\gamma_x + w_i\gamma_w + \epsilon_i$$
$$(W_i \mid X_i = x_i) = \alpha_0 + x_i\alpha_x + \epsilon_i^{**}$$

with OLSE $\widehat{\vec{\beta}}$, $\widehat{\vec{\gamma}}$, and, $\widehat{\vec{\alpha}}$, respectively. Let $\vec{U}$ be the residuals from the regression of $\vec{Y}$ on $\vec{X}$, and let $\vec{V}$ be the residuals from the regression of $\vec{W}$ on $\vec{X}$.

   a. Show that the OLSE slope from regressing $\vec{U}$ on $\vec{V}$ is equal to $\hat{\gamma}_w$.

   b. Show that the OLSE slope from regressing $\vec{Y}$ on $\vec{V}$ is equal to $\hat{\gamma}_w$.

   c. Show that the estimated standard errors for the slopes from the regressions in parts a and b are not necessarily equal. Prove that the estimated standard error for the slope in part a is equal to the estimated standard error for $\hat{\gamma}_w$.

4. We consider a two stage sampling scheme:

   - In the first stage we observe m_1 vector $\vec{Z}$ and perform a regression analysis based on $m_1 \times p$ design matrix $\mathbf{W}$ (with $rank(\mathbf{W}) = p < m_1$) using regression model:
   $$\vec{Z} = \mathbf{W}\vec{\beta} + \vec{\epsilon}^{*}$$
   with $\vec{\epsilon}^{*} \sim (0_{m_1}, \sigma^2 \mathbf{I}_{m_1})$. Let $\widehat{\vec{\beta}^{(1)}}$ be the best linear unbiased estimator of $\vec{\beta}$ from the first stage analysis.

   - In the second stage we observe m_2 vector $\vec{U}$ and perform a regression analysis based on $m_2 \times p$ design matrix $\mathbf{V}$ (with $rank(\mathbf{V}) = p < m_2$) using regression model:
   $$\vec{U} = \mathbf{V}\vec{\beta} + \vec{\epsilon}^{**}$$
   with $\vec{\epsilon}^{*} \sim (0_{m_2}, \sigma^2 \mathbf{I}_{m_2})$. Let $\widehat{\vec{\beta}^{(1)}}$ be the best linear unbiased estimator of $\vec{\beta}$ from the second stage analysis.

   - In a combined analysis we consider regression model
   $$\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$$
   where
   $$\vec{Y} = \begin{pmatrix} \vec{Z} \\ \vec{U} \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} \mathbf{W} \\ \mathbf{V} \end{pmatrix}$$
   and $\widehat{\vec{\beta}^{(C)}}$ is the best linear unbiased estimator of $\vec{\beta}$ from an analysis using the combined data.

   a. Suppose $p = 2$ show that the slope from $\widehat{\vec{\beta}^{(C)}}$ is a weighted average of the slopes from $\widehat{\vec{\beta}^{(1)}}$ and $\widehat{\vec{\beta}^{(2)}}$. Explicitly state the weights.

   b. Suppose $p = 2$ show that the intercept from $\widehat{\vec{\beta}^{(C)}}$ is a weighted average of the intercepts from $\widehat{\vec{\beta}^{(1)}}$ and $\widehat{\vec{\beta}^{(2)}}$. Explicitly state the weights.

   c. Again supposing $p = 2$ show the parameter estimate vector $\widehat{\vec{\beta}^{(C)}}$ is a weighted average of the parameter estimate vectors $\widehat{\vec{\beta}^{(1)}}$ and $\widehat{\vec{\beta}^{(2)}}$. Explicitly state the weights.

   d. Prove that the results of part c can be generalized to arbitrary p.

5. Outline a proof of a central limit theorem for OLSE from simple linear regression. Be sure to specify the context in which your CLT holds.

6. Consider a linear regression model of response n-vector $\vec{Y}$ on the constant vector $\vec{1}_n$ and covariate n-vector $\vec{X}$, where each element x_i of $\vec{X}$ is either 0 or 1. Show that when all elements of $\vec{Y}$ are independent, inference about the slope from the simple linear regression model is asymptotically equivalent to the t test that presumes equal variances.