Biost/Stat 533
Emerson, Spr 14
Homework #3
April 18, 2014

Written problems to be handed in Monday, April 28.

Under no circumstances may you refer to a homework key from this or other classes. While you may work with other students to derive a solution, when you write up your final solution, you may not refer to any other source. You must be able to develop your answer as if it were being done in a closed book, closed notes examination. You must provide a signed pledge to that effect:

On my honor I have neither given nor received unauthorized aid on the completion of this homework.

1. Consider a linear regression model $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$ where the first column of the design matrix is filled with 1's (so we are fitting an intercept). Assume that $var(\vec{\epsilon}) = \sigma^2\mathbf{I}_n$.

   a. Suppose $E[\epsilon_i] = \theta$ for $i = 1, \ldots, n$. What is the distribution for the OLS estimator $\widehat{\vec{\beta}}$? In particular, how does the assumption of nonzero mean for the errors alter the interpretation and distribution of the slope parameters $\beta_1, \ldots, \beta_{p-1}$?

   b. Let $\mathbf{X}^*$ be a design matrix derived from $\mathbf{X}$ by subtracting the corresponding column means from the elements in columns 2 through p. That is $X^*_{ij} = X_{ij} - \sum_{i=1}^n X_{ij}/n$. If we fit the regression model $\vec{Y} = \mathbf{X}^*\vec{\beta}^* + \vec{\epsilon}$, how does the OLS estimator $\widehat{\vec{\beta}}^*$ relate to $\widehat{\vec{\beta}}$ from the original problem. In particular, how does the interpretation and distribution of of each of the regression parameters change?

2. Consider a linear regression model $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$ where the first column of the design matrix is filled with 1's (so we are fitting an intercept). Consider adding an additional predictor $\vec{Z}$ to the model where, for some fixed j, $Z_i = 1$ if $i = j$ and $Z_i = 0$ otherwise. Let $\mathbf{X}^*$ be the augmented matrix in which the $(p+1)$th column is $\vec{Z}$, and consider fitting the regression model $\vec{Y} = \mathbf{X}^*\vec{\gamma} + \vec{\epsilon}^*$

   a. How do the parameter estimates $\hat{\gamma}_0, \ldots, \hat{\gamma}_{p-1}$ differ from $\widehat{\vec{\beta}}$?

   b. How do the parameter estimates $\hat{\gamma}_0, \ldots, \hat{\gamma}_{p-1}$ differ from the estimates obtained by fitting the first model with the jth case deleted?

3. Suppose $Y_i \sim (\mu_0, \sigma^2)$ for $i = 1, \ldots, n_0$ and $Y_i \sim (\mu_1, \sigma^2)$ for $i = n_0 + 1, \ldots, n = n_0 + n_1$, with $Cov(Y_i, Y_j) = 0$ for $i \neq j$. We are interested in estimating $\mu_1 - \mu_0$. For notational convenience, let $\vec{w}$ be an n-vector such that $w_i = 1$ for $1 \leq i \leq n_0$ and $w_i = 0$ otherwise, and let $\vec{z} = \vec{1}_n - \vec{w}$. (In all parts of this problem please provide formulas in terms of simple statistics, not matrix notation.)

   a. Using design matrix $\mathbf{X} = (\vec{1}_n \quad \vec{w})$, find the ordinary least squares estimator $\widehat{\vec{\beta}}$ for regression model $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$. Find vector $\vec{a}$ such that estimable function $\vec{a}^T\vec{\beta} = \mu_1 - \mu_0$, and provide the formula and mean and variance for $\vec{a}^T\widehat{\vec{\beta}}$.

   b. Using design matrix $\mathbf{X} = (\vec{1}_n \quad \vec{z})$, find the ordinary least squares estimator $\widehat{\vec{\beta}}$ for regression model $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$. Find vector $\vec{a}$ such that estimable function $\vec{a}^T\vec{\beta} = \mu_1 - \mu_0$, and provide the formula and mean and variance for $\vec{a}^T\widehat{\vec{\beta}}$.

   c. Using design matrix $\mathbf{X} = (\vec{w} \quad \vec{z})$, find the ordinary least squares estimator $\widehat{\vec{\beta}}$ for regression model $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$. Find vector $\vec{a}$ such that estimable function $\vec{a}^T\vec{\beta} = \mu_1 - \mu_0$, and provide the formula

and mean and variance for $\vec{a}^T\widehat{\vec{\beta}}$.

d. Using design matrix $\mathbf{X} = (\vec{1}_n \quad \vec{w} \quad \vec{z})$, find the ordinary least squares estimator $\widehat{\vec{\beta}}$ for regression model $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$. Find vector $\vec{a}$ such that estimable function $\vec{a}^T\vec{\beta} = \mu_1 - \mu_0$, and provide the formula and mean and variance for $\vec{a}^T\widehat{\vec{\beta}}$.

4. Let $\mathbf{X}$ (dimension $n \times p$) and $\mathbf{W}$ (dimension $n \times r$) be design matrices with the same range spaces (so $\mathcal{R}[\mathbf{X}] = \mathcal{R}[\mathbf{W}]$, where $\mathcal{R}[\mathbf{X}] = \{\vec{y} : \vec{y} = \mathbf{X}\vec{a},\ \vec{a} \in \mathcal{R}^p\}$ and $\mathcal{R}[\mathbf{W}] = \{\vec{y} : \vec{y} = \mathbf{W}\vec{a},\ \vec{a} \in \mathcal{R}^r\}$). Show that regression models $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$ and $\vec{Y} = \mathbf{W}\vec{\gamma} + \vec{\epsilon}$ are alternative parameterizations of each other. Furthermore show that if $\vec{a}^T\vec{\beta}$ is an estimable function, then there exists an estimable function $\vec{b}^T\vec{\gamma}$ such that estimates $\vec{a}^T\widehat{\vec{\beta}}$ and $\vec{b}^T\widehat{\vec{\gamma}}$ are equal for all $\vec{Y} \in \mathcal{R}^n$ and all least squares estimators $\widehat{\vec{\beta}}$ and $\widehat{\vec{\gamma}}$.

5. Suppose n-vector $\vec{\epsilon}$ has $E[\vec{\epsilon}] = \vec{0}$ and $Cov[\vec{\epsilon}] = \mathbf{V}$ with $rank(\mathbf{V}) = n$. Let $\widehat{\vec{\beta}} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\vec{Y}$ be the ordinary least squares estimator of $\vec{\beta}$ and $\widehat{\vec{\beta}}_G = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}^T\mathbf{V}^{-1}\vec{Y}$ be the generalized least squares estimator of $\vec{\beta}$ in regression model $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$.

a. Find the mean and variance of estimators $\vec{a}^T\widehat{\vec{\beta}}$ and $\vec{a}^T\widehat{\vec{\beta}}_G$ of estimable function $\vec{a}^T\beta$.

b. Show that a best linear unbiased estimator of estimable function $\vec{a}^T\vec{\beta}$ is $\vec{a}^T\widehat{\vec{\beta}}_G$.

6. Consider the simple linear regression model $Y_i = \beta_0 + x_i\beta_1 + \epsilon_i$ for $i = 1, \ldots, n$, with x_i known predictors, $\vec{\beta} = (\beta_0, \beta_1)^T$ an unknown parameter vector to be estimated and/or tested, and $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$. Without loss of generality, we will assume that $\sum_{i=1}^n x_i = 0$. Let $\sigma_i^2 = \alpha_i + x_i\gamma > 0$ with γ and $\vec{\alpha}$ unknown nuisance parameters subject to $\vec{\alpha}^T\vec{x} = 0$. Let $\widehat{\vec{\beta}}$ be the ordinary least squares estimate of $\vec{\beta}$.

a. What is the mean and variance of $\hat{\beta}_1$?

b. Under what conditions will the estimated variance of $\hat{\beta}_1$ based on the ordinary least squares regression analysis be consistent for the true variance of $\hat{\beta}_1$.

c. What restrictions on the problem would be necessary for $\widehat{\vec{\beta}}$ to be asymptotically normally distributed? (You need not rigorously derive an asymptotic distribution, instead just briefly discuss the ways that this setting differs from the assumptions under which we derived the asymptotic distribution in class, and what general requirements might address those problems.)

d. What would be the effect of using the asymptotic results for ordinary least squares regression analysis on tests of $H_0 : \hat{\beta}_1 = 0$? Consider the effect that the variance of the α_is and the value of γ has on your answer.

e. What would be the effect of using the asymptotic results for ordinary least squares regression analysis on confidence intervals for β_1? Consider the effect that the variance of the α_is and the value of γ has on your answer.

f. How do the above results compare to results for the t tests as considered in homework # 1?