Biost/Stat 533
Emerson, Spr 09
Homework #3
April 20, 2009

Written problems to be handed in Wednesday, April 29.

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 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?

3. Let $Y_i \sim Bernoulli(p_i), i = 1, \ldots, n$ be independent random variables with $p_i = \vec{x}_i^T\vec{\beta}$ for known predictor vector $\vec{x}_i$.

   a. Is inference about $\vec{\beta}$ using ordinary least squares regression analysis asymptotically valid for this problem? If so, provide justification. If not, are there any situations in which it might be approximately valid?

   b. Describe an iterative approach in which weighted least squares might be used to address this problem. What undesirable small sample behavior with respect to the range of estimates $\hat{p}_i$ might persist under this analysis scheme?

4. Consider a linear regression model relating response $\vec{Y}$ to an intercept and two predictor vectors $\vec{W}$ and $\vec{Z}$ (so design matrix $\mathbf{X} = (\vec{1}_n \quad \vec{W} \quad \vec{Z})$ has $X_{i1} \equiv 1$ for $i = 1, \ldots, n$ and $X_{i2} = W_i$ and $X_{i3} = Z_i$ and

$\vec{\beta} = (\beta_0, \beta_1, \beta_2)^T$). Assume $E[\vec{\epsilon}] = \vec{0}$ and $var(\vec{\epsilon}) = \sigma^2 \mathbf{I}_n$.

a. Show that the correlation between OLS estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ is opposite in sign to the sample correlation between $\vec{W}$ and $\vec{Z}$ and that the two slope estimates are uncorrelated if the sample correlation between $\vec{W}$ and $\vec{Z}$ is zero.

b. Suppose we hold $S_{WW} = (\vec{W} - E[\vec{W}])^T(\vec{W} - E[\vec{W}])$, S_{ZZ}, and σ^2 constant, but we may freely vary $S_{WZ} = (\vec{W} - E[\vec{W}])^T(\vec{Z} - E[\vec{Z}])$. For what value of S_{WZ} do we minimize the variance of $\hat{\beta}_1$ and $\hat{\beta}_2$? What does this suggest about our ability to test for an association between Y and W adjusting for Z when W and Z are correlated?

5. Consider again the linear regression model in Problem 4 in which we will assume the true model is

$$\vec{Y} = \beta_0 + \vec{W}\beta_1 + \vec{Z}\beta_2 + \vec{\epsilon}$$

but we want to also consider fitting a model

$$\vec{Y} = \gamma_0 + \vec{W}\gamma_1 + \vec{\epsilon}^*$$

.

a. Under what conditions is the OLS estimate $\hat{\beta}_1$ equal to the OLS estimate $\hat{\gamma}_1$?

b. Under what conditions is the standard error of $\hat{\beta}_1$ equal to the standard error of $\hat{\gamma}_1$.

c. Under what conditions is the <u>estimated</u> standard error of $\hat{\beta}_1$ equal to the <u>estimated</u> standard error of $\hat{\gamma}_1$.

d. Under what conditions is $\hat{\gamma}_1$ unbiased for β_1?

e. Under what conditions is $\hat{\gamma}_1$ BLUE for β_1?

f. Suppose in particular that $\beta_1 = 0$ and $\beta_2 \neq 0$. What is the impact of this situation on the distribution of $\hat{\gamma}_1$, and how would $\hat{\gamma}_1$ compare to $\hat{\beta}_1$ from the full model? Compare this situation to the setting in which $\beta_2 = 0$ and $\beta_1 \neq 0$.

6. 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?