Biost/Stat 533
Emerson, Spring 09
Homework 4
May 6, 2009

The problems of this homework is a midterm examination I have previously given. As a midterm, it was closed book, closed notes, and it was administered in 50 minutes.

As a homework exercise, you are free to use books and notes, but you might find it advantageous to try it without it. The homework should be handed in on Monday, May 11.

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.

1. Consider a regression model in which response variables $Y_i, i = 1, \dots, n$ satisfy

$$Y_i = \beta_0 + Z_i\beta_1 + W_i\beta_2 + \epsilon_i$$

with the ϵ_i's independent and identically distributed according to $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$. We consider the situation in which we are interested in making inference about β_1, and we are trying to decide whether to regress $\vec{Y}$ on the full model including $\vec{Z}$ and $\vec{W}$, or whether to regress $\vec{Y}$ on a reduced model including only $\vec{Z}$ as a predictor (in both cases we will include an intercept). Notationally, the reduced model is

$$Y_i = \gamma_0 + Z_i\gamma_1 + \epsilon_i^*$$

and let $\mathbf{X} = (\vec{1}_n \quad \vec{Z} \quad \vec{W})$ and $\mathbf{U} = (\vec{1}_n \quad \vec{Z})$ be the design matrices for the full and reduced models, respectively, with $\widehat{\vec{\beta}}$ and $\widehat{\vec{\gamma}}$ be the ordinary least squares estimates from the corresponding regression models.

a. Without loss of generality, we may assume $\sum_{i=1}^n Z_i = 0$ and $\sum_{i=1}^n W_i = 0$. Why?

b. Under what conditions does $\hat{\gamma}_1 = \hat{\beta}_1$?

c. What are the expectations of $\widehat{\vec{\beta}}$ and $\widehat{\vec{\gamma}}$? Under what conditions is $\hat{\gamma}_1$ unbiased for β_1?

d. What is the variance of $\widehat{\vec{\beta}}$ and $\widehat{\vec{\gamma}}$? Under what conditions is $Var(\hat{\gamma}_1) = Var(\hat{\beta}_1)$?

e. Suppose $\beta_2 = 0$ and $\sum_{i=1}^n Z_iW_i = 0$. What are the relative advantages and disadvantages of choosing the reduced model over the full model?

f. Suppose $\beta_2 = 0$ and $\sum_{i=1}^n Z_iW_i \neq 0$. What are the relative advantages and disadvantages of choosing the reduced model over the full model?

g. Suppose $\beta_2 \neq 0$ and $\sum_{i=1}^n Z_i W_i = 0$. What are the relative advantages and disadvantages of choosing the reduced model over the full model?

h. Suppose $\beta_2 \neq 0$ and $\sum_{i=1}^n Z_i W_i \neq 0$. What are the relative advantages and disadvantages of choosing the reduced model over the full model?

2. Consider an "error in the variables" model in which there is a true relationship between response Y and predictor W given by $Y_i = \beta_1 + \beta_2 W_i + \epsilon_i$ with $\epsilon_i \sim (0, \sigma^2 > 0)$ totally independent. Suppose that W is unobserved, and we instead have X, an imprecise measurement of W which follows the relation $X_i = \alpha_1 + \alpha_2 W_i + \delta_i$, with $\delta_i \sim (0, \tau^2 > 0)$ totally independent of each other and the ϵ's. We thus fit a regression model $Y_i = \gamma_1 + \gamma_2 X_i + \eta_i$.

   a. Show that the ordinary least squares estimate $\widehat{\vec{\gamma}}$ is generally biased as an estimator of $\vec{\beta}$, even when $\alpha_1 = 0$ and $\alpha_2 = 1$. Under what conditions is it unbiased? (Hint: What must $\mathbf{X}^T\mathbf{W}$ equal if $\widehat{\vec{\gamma}}$ is to be unbiased for $\vec{\beta}$? What does it equal?)

3. Suppose independent response variables $Y_i \sim \mathcal{E}(\lambda_i)$, $\lambda_i > 0$, for $i = 1, \ldots, n$ are distributed according to an exponential distribution with

$$\begin{aligned} \text{density} \quad f_i(y_i) &= \frac{1}{\lambda_i} e^{-y_i/\lambda_i} \\ \text{cdf} \quad F_i(y_i) &= 1 - e^{-y_i/\lambda_i} \\ \text{mean} \quad E[Y_i] &= \lambda_i \\ \text{variance} \quad Var(Y_i) &= \lambda_i^2 \end{aligned}$$

Recall that in the exponential, λ is a scale parameter such that if $Y \sim \mathcal{E}(\lambda)$ then for $c > 0$, $cY \sim \mathcal{E}(c\lambda)$.

   a. Consider a linear regression model with $\lambda_i = \vec{x}_i^T \vec{\beta}$ for known predictor vectors $\vec{x}_i$. Is inference based on the asymptotic normality of least squares estimators of $\vec{\beta}$ valid in this setting? Justify your answer. If it is not valid, briefly describe a regression analysis that would provide asymptotically valid inference for this model.

   b. Suppose $Z_i = \mu_i + \delta_i$ where μ_i is an unknown parameter and $e^{\delta_i} \sim \mathcal{E}(1)$ are independent. What is the distribution of e^{Z_i}?

   c. For independent response variables Y_i as above, consider a linear regression model

$$\log(Y_i) = \vec{x}_i^T \vec{\gamma} + \epsilon_i$$

   Is inference based on the asymptotic normality of least squares estimators of $\vec{\gamma}$ valid in this setting? Justify your answer. If it is not valid, briefly describe a regression analysis that would provide asymptotically valid inference for this model.

4. Consider response variables $Y_i, i = 1, \ldots, n$ and known predictors x_i. Let $x_i^* = x_i - \overline{x}$ be transformed predictors obtained by centering the x_i's about their mean. Consider linear regression models

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$Y_i = \gamma_0 + \gamma_1 x_i^* + \epsilon_i$$

with independent identically distributed errors $\epsilon_i \sim (0, \sigma^2)$. Let $\widehat{\vec{\beta}}$ and $\widehat{\vec{\gamma}}$ be OLSE from the corresponding models.

a. How does $Var(\hat{\beta}_1)$ compare to $Var(\hat{\gamma}_1)$?

b. How does $Var(\hat{\beta}_0)$ compare to $Var(\hat{\gamma}_0)$?

c. Suppose we want to make inference about the average response when $x = x_0$. Specifically, we wish to test that $E[Y \mid x = x_0] = c$. Describe a hypothesis test based on $\widehat{\vec{\beta}}$ that is asymptotically valid for this setting. (I want formulas, but matrix notation is fine, providing you have defined your notation.)

d. Suppose we were to also construct the hypothesis test of part (c) using $\widehat{\vec{\gamma}}$. Which test is more efficient to make this inference?