Biost/Stat 533
Emerson, Spring 99
Final Examination Key
June 25, 1999

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

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

For each of the following scenarios **very** briefly state

- any optimality properties provided by the method of characterizing the relationship between response Y and predictor Z
- whether the stated analysis would provide valid asymptotic statistical inference about the relationship between response Y and predictor Z in the sense that hypothesis tests would have the correct size and interval estimates would have correct coverage probabilities
- if the statistical inference is invalid, under what conditions the inference would tend to be conservative (size of tests smaller than nominal or coverage of intervals too high) or anti-conservative.

a. The ϵ_i's are independent and identically distributed according to $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$. We use OLS regression of response Y on predictors Z and W to compute a confidence interval for a linear relationship between Y and Z.

Ans: This is the situation in which OLS are BLUE (not UMVUE unless the errors are normally distributed) for estimating the association between Y and Z adjusted for W. If adjustment for W was clearly part of the scientific question, then this model is in that sense optimal. If this is a designed experiment, failure to adjust for W would cause our estimates of $var(\hat{\beta}_1)$ to be too large, hence in such a case we should always adjust for W. If adjustment for W is somewhat arbitrary, then one must consider whether there is truly an association between W and Y. If not (i.e., $\beta_2 = 0$), inclusion of W might lead to variance inflation in that if $cov(Z_i, W_i) \neq 0$, then $var(\hat{\beta}_1)$ will be higher than the variance of the slope might be in a simple linear regression model. However, in this setting, the asymptotic inference is valid: Any variance inflation that occurs is estimated correctly by this full model. The problem with this sort of variance inflation is that we could have found an estimator that had less variability, providing we were willing to change our question to be the association between Y and Z when not adjusting for W.

b. The ϵ_i's are identically distributed according to $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$. The data are measured on rats and are collected on pairs of litter mates, where the $(2i)$th and

$(2i+1)$th measurements are from the same litter. We use OLS regression of response Y on predictors Z and W to compute a confidence interval for a linear relationship between Y and Z.

Ans: The OLS are still unbiased in this setting, but the potential correlation between errors for litter mates means that asymptotic inference will not in general be valid. If the correlation between errors for litter mates is in the same direction as the correlation between the Z_is for litter mates, the inference will tend to be anti-conservative. If the correlation between errors for litter mates is in the opposite direction as the correlation between the Z_is for litter mates, the inference will tend to be conservative.

c. The ϵ_i's are identically distributed according to $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$. The data are measured on rats and are collected on pairs of litter mates, where the $(2i)$th and $(2i+1)$th measurements are from the same litter. We use generalized least squares regression of response Y on predictors Z and W to compute a confidence interval for a linear relationship between Y and Z. In the GLS, we assume a correlation structure in which $Var(\vec{\epsilon}) = \sigma^2\mathbf{V}$ where $V_{ii} = 1$, $V_{2i,2i+1} = \rho$ (known), and $V_{ij} = 0$ otherwise.

Ans: This is the situation in which GLS are BLUE (not UMVUE unless the errors are normally distributed) for estimating the association between Y and Z adjusted for W. If adjustment for W was clearly part of the scientific question, then this model is in that sense optimal. All of the discussion in part (a) pertains to this situation as well.

d. n_i measurements are made on the ith individual and averaged to obtain Y_i (each of the n_i measurements have the same mean and variance). We use OLS regression of response Y on predictors Z and W to compute a confidence interval for a linear relationship between Y and Z.

Ans: The OLS are unbiased in this situation, but there is quite likely to be heteroscedasticity of the observations. That is, let Y_{ij} be the jth measurement on the ith individual and ϵ_{ij} be the error associated with that measurement. If $var(\epsilon_{ij}) = \tau_i^2$, then the variance associated with the measurement of Y_i (the mean of the Y_{ij}'s for $j = 1, \ldots, n_i$) is $\sigma_i^2 = \tau_i^2/n_i$. If we managed to choose n_i such that $\sigma_i^2 = \sigma^2$ was constant, then we have that the answer to part (a) applies. However, under the more typical assumption that $tau_i^2 = \tau^2$ is constant, we would have unequal error variance and the OLS would not be BLUE. Asymptotic variance might be invalid. In the absence of a definite relationship between σ_i^2 and Z_i, when the larger values of σ_i^2 are associated with more extreme values of Z_i the inference will tend to be anti-conservative. If smaller values of σ_i^2 are associated with more extreme values of Z_i in that setting, the inference will tend to be conservative. If the distribution of n_i is independent of Z_i, then inference would be valid. If there is a predictor variance relationship, positively skewed Z_i values will tend to cause anti-conservative inference, while negatively skewed Z_i distribution will tend to cause conservative

inference. If the skewness of the Z_i distribution is zero, inference would be valid. (The above statements consider the setting of Homework 2, problem 2.)

e. n_i measurements are made on the ith individual and averaged to obtain Y_i (each of the n_i measurements have the same mean and variance). We use weighted least squares regression of response Y on predictors Z and W to compute a confidence interval for a linear relationship between Y and Z. In the WLS, we assume a correlation structure in which $Var(\vec{\epsilon}) = \sigma^2\mathbf{V}$ where $V_{ii} = 1/n_i$, and $V_{ij} = 0$ for $i \neq j$.

Ans: Under the notation of the answer to part (d), the WLS is BLUE (not UMVUE unless the errors are normally distributed) for estimating the association between Y and Z adjusted for W when $\tau_i^2 = \tau^2$ is constant. In that case, the discussion of part (a) is relevant. If τ_i^2 is not the same for all individuals, then the discussion in part(d) is relevant.

f. The ϵ_i's are independent and identically distributed according to $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$. We use OLS regression of response Y on predictor Z to compute a confidence interval for a linear relationship between Y and Z.

Ans: This is the situation in which OLS are BLUE (not UMVUE unless the errors are normally distributed) for estimating the association between Y and Z unadjusted for W. If this is a designed experiment (i.e., the sampling of Z_i and W_i is determined by design), our estimates of $var(\hat{\beta}_1)$ will be too large, however, and our statistical inference will be conservative. If this is not a designed experiment, then we must consider whether the more relevant scientific question is one which adjusts or does not adjust for W. If adjustment for W is somewhat arbitrary, then one must consider whether there is truly an association between W and Y. If not (i.e., $\beta_2 = 0$), inclusion of W might lead to variance inflation in that if $cov(Z_i, W_i) \neq 0$, then $var(\hat{\beta}_1)$ will be higher than the variance of the slope might be in a simple linear regression model. If $\beta_2 \neq 0$, then we can gain more power by adjusting for W, and when $r_{ZW} \neq 0$, we will also be avoiding confounding the relationship between Y and Z with the relationship between Y an W.

g. The ϵ_i's are independent and identically distributed according to $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$. We use OLS regression of response Y on predictors Z and W to compute a prediction interval for an individual observation of Y when $Z = z_0$ and $W = w_0$ using

$$\hat{Y}_0 \pm z_{1-\alpha/2}\hat{\sigma}\sqrt{1 + \vec{x}_0^t(\mathbf{X}^T\mathbf{X})^{-1}\vec{x}_0}$$

where $\hat{Y}_0 = \vec{x}_0^T\widehat{\vec{\beta}}$, $\vec{x}_0^T = (1\ z_0\ w_0)$ and $\mathbf{X} = (\vec{1}_n\ \vec{Z}\ \vec{W})$.

Ans: This situation is that of part (a), but now we are also interested in prediction intervals. The inference associated with these prediction intervals will only be valid if the errors are normally distributed (we are using the quantiles of the standard normal distribution). Whether the inference is conservative or anti-conservative

will depend on the exact distribution of the errors and the value of α– no general statement can be made.

2. Consider again the setting of problem 1c in which observations made on litter mates are assumed to be correlated. But now suppose that the correlation between observations is not known. Describe a way in which the linear regression methods we have discussed this quarter might be used to provide asymptotically valid inference when ρ is unknown.

   Ans: The problem here is that if ρ is known, then we know that GLS are optimal and provide valid asymptotic inference. If we do not know the value of ρ, then we need to estimate it. One approach to do this would be to estimate the errors for each observation, and then use the sample correlation between errors on litter mates to construct the variance matrix $\mathbf{V}$. Now, as noted in 1(b), the OLS are unbiased, so if we estimate $\widehat{\vec{\beta}}$ using OLS, then $\hat{\epsilon}_i = Y_i - \hat{\beta}_0 - Z_i\hat{\beta}_1 - W_i\hat{\beta}_2$ is an unbiased estimate of ϵ_i. We then find the correlation of the $\hat{\epsilon}_i$'s between litter mates, and use that value in a GLS analysis. (I was looking for you to discuss the fact that OLS could be used to find reasonable estimates of the errors, that the sample correlation of those error estimates could be easily computed, and that GLS could then be used.)

3. Consider the regression model of problem 1, and assume that the ϵ_i's are independent with $E[\epsilon_i] = 0$ and $Var[\epsilon_i] = \alpha_0 + \alpha_1 Z_i$, with α_0 and α_1 unknown. Describe a way in which the linear regression methods we have discussed this quarter might be used to provide asymptotically valid inference about β_1.

   Ans: The problem here is that if α_0 and α_1 are known, then we know that GLS are optimal and provide valid asymptotic inference. If we do not know the value of α_0 and α_1, then we need to estimate them. One approach to do this would be to estimate the errors for each observation, and then use regression on the squared errors to estimate the α_0 and α_1, and then construct the variance matrix $\mathbf{V}$. Now, as noted in 1(b), the OLS are unbiased, so if we estimate $\widehat{\vec{\beta}}$ using OLS, then $\hat{\epsilon}_i = Y_i - \hat{\beta}_0 - Z_i\hat{\beta}_1 - W_i\hat{\beta}_2$ is an unbiased estimate of ϵ_i. And because $Var[\epsilon_i] = \alpha_0 + \alpha_1 Z_i$, we have that $E[\epsilon_i^2] = \alpha_0 + \alpha_1 Z_i$. Hence, we regress the $\hat{\epsilon}_i$'s on the Z_is to obtain OLS $\hat{\alpha}_0$ and $\hat{\alpha}_1$, and use $V_{ii} = \hat{\alpha}_0 + \hat{\alpha}_1 Z_i$ in a GLS (WLS) analysis. (I was looking for you to discuss the fact that OLS could be used to find reasonable estimates of the errors, that the regression of the squared error estimates on Z would allow estimation of the true variance for each observation, and that GLS could then be used.)

4. 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.

Ans: Because there is a mean variance relationship, OLS based inference would only be valid if the sampling of the predictors and the value of $\vec{\beta}$ were such that $\vec{x}_i^T \vec{\beta}$ were the same for all individuals.

One approach around this problem would be to iteratively use weighted least squares with the current estimate of $widehat\vec{\beta}$ at each iteration used to estimate the covariance matrix for $\vec{Y}$. (see homework #2)

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}?

Ans: $e^{Z_i} = e^{\mu_i} e^{\delta_i}$ so $e^{Z_i} \sim \mathcal{E}(e^{\mu_i})$, a scaled exponential random variable.

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.

Ans: Using the result from part (b), we see that $Z_i = \log(Y_i)$ can be written as $Z_i = \log(\lambda_i) + \epsilon_i$ where the ϵ_i's are independent and identically distributed. This suggests that asymptotic inference for $\vec{\beta}$ based on ordinary least squares estimates would be valid. It should be noted that $E[\epsilon_i] = 1 \neq 0$, so the LSE of the intercept is biased, but that will not affect the distribution of the estimates for the slopes.

5. Do either part a or part b (answering both will get extra credit).

a. Provide a brief outline of a proof of the asymptotic normality of OLS regression estimates in simple linear regression. Clearly state the assumptions required for your proof.

Ans: The key points I was looking for was that we would consider all linear combinations of the normalized version of $\hat{\beta}_0$ and $\hat{\beta}_1$. Such linear combinations turned out to be of the form $\sum W_i$, where $W_i = A_i \epsilon_i$ (where the A_is depend on the predictor X_i and the coefficients from the linear combination being considered), and hence we have independent, but non-identically distributed W_is. The Lindeberg-Feller central limit theorem can be used to show the asymptotic normal distribution of

$\sum W_i$, providing the Lindeberg condition holds. That condition was found to hold so long as the ϵ_is were i.i.d. and the $max((x_i - \overline{x})^2 / \sum_{j=1}^{n}(x_j - \overline{x})^2$ tended toward 0 as n became infinite. Then we could argue the asymptotic bivariate normality of $\widehat{\vec{\beta}}$ based on the Cramér-Wold device.

b. Provide a brief outline of a proof that OLS regression provides a basis for optimal hypothesis tests when $\vec{\epsilon} \sim \mathcal{N}_n(\vec{0}_n, \sigma^2 \mathbf{I}_n)$.

Ans: Here I was looking for you to explicitly show that the likelihood ratio was a function of $(RSS_H - RSS)/RSS$.

6. Suppose we wish to investigate a relationship between blood pressure and race (white, black, asian, other). Suppose further that we perform an analysis modelling blood pressure as a function of race, sex, and a possible race-sex interaction.

a. Describe an analysis model you would use to address this problem, clearly stating the predictors modelled and their corresponding model parameters.

Ans: Either ANOVA or linear regression formulations of the model were acceptable, providing that race was modeled with three dummy variables in the regression model. Preferring regression model notation myself, I would define indicator variables $BLACK$, $ASIAN$, $OTHER$, and $FEMALE$ and interaction terms $B.F = BLACK \times FEMALE$, $A.F = ASIAN \times FEMALE$, and $O.F = OTHER \times FEMALE$ (in this parameterization my intercept will correspond to white males) and use regression model

$$E[Y] = \beta_0 + \beta_1 \times FEMALE + \beta_2 \times BLACK + \beta_3 \times ASIAN + \beta_4 \times OTHER + \beta_5 \times B.F + \beta_6 \times A.F + \beta_7 \times O.F$$

b. Suppose we wish to test for the existence of a race effect on blood pressure. State the hypotheses you would test in terms of your model parameters, the form of the test statistic you would use to test those hypotheses, and the methods you would use to determine a critical region for the test. I want explicit formulas, though matrix notation is fine so long as the matrices are adequately defined.

Ans: If there is to be no difference in blood pressure by race, then the regression parameter for every term that involves race in some way must be 0. Hence our null hypothesis needs to be $H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$. Using OLS $\widehat{\vec{\beta}}$ we can compute quadratic form

$$Q = \frac{(\mathbf{A}\widehat{\vec{\beta}})^T (\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T)^{-1} \mathbf{A}\widehat{\vec{\beta}}}{\hat{\sigma}^2}$$

where $\mathbf{A}$ contains the bottom 6 rows of an 8 dimensional identity matrix. Asymptotically, Q has a chi squared distribution with 6 degrees of freedom under the null hypothesis. I would actually use the F distribution with 6 and $n - 8$ degrees of

freedom, rejecting H_0 if $Q > F_{1-\alpha,6,n-8}$. I note that we could also get this same statistic by considering that

$$Q = \frac{(RSS_H - RSS)/6}{RSS/(n-8)}$$

where $RSS = (n-8)\hat{\sigma}^2$ is computed from the full model and RSS_H is computed from a simple linear regression of SBP on $FEMALE$.

c. Suppose instead that we merely want to determine whether a race-sex interaction on blood pressure exists. State the hypotheses you would test in terms of your model parameters, the form of the test statistic you would use to test those hypotheses, and the methods you would use to determine a critical region for the test. I want explicit formulas, though matrix notation is fine so long as the matrices are adequately defined.

Ans: If there is to be no interaction between sex and race on blood pressure, then the regression parameter for every interaction term must be 0. Hence our null hypothesis needs to be $H_0 : \beta_5 = \beta_6 = \beta_7 = 0$. Using OLS $\widehat{\vec{\beta}}$ we can compute quadratic form

$$Q = \frac{(\mathbf{A}\widehat{\vec{\beta}})^T(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T)^{-1}\mathbf{A}\widehat{\vec{\beta}}}{\hat{\sigma}^2}$$

where $\mathbf{A}$ contains the bottom 3 rows of an 8 dimensional identity matrix. Asymptotically, Q has a chi squared distribution with 3 degrees of freedom under the null hypothesis. I would actually use the F distribution with 3 and $n-8$ degrees of freedom, rejecting H_0 if $Q > F_{1-\alpha,3,n-8}$. I note that we could also get this same statistic by considering that

$$Q = \frac{(RSS_H - RSS)/3}{RSS/(n-8)}$$

where $RSS = (n-8)\hat{\sigma}^2$ is computed from the full model and RSS_H is computed from a linear regression of SBP on $FEMALE$, $BLACK$, $ASIAN$, and $OTHER$.