

This examination is closed book, closed notes. You have 110 minutes to complete the exam. Concise answers are greatly to be preferred. Unless otherwise stated, facts known from memory may be stated without proof. When you are asked to provide justification, it is okay to only provide enough detail to make clear the reasoning behind your answer.

Each problem is worth 25 points. You should work

- Problem 1.
- Problem 2.
- Problem 3 or problem 4. (You can work both and get extra credit.)
- Problem 5 or problem 6. (You can work both and get extra credit.)
- Problem 7a or problem 7b. (You can work both and get extra credit.)

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 two regression models in which response variables $Y_i, i = 1, \dots, n$ are presumed to satisfy

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

$$Y_i = \gamma_0 + Z_i\gamma_1 + W_i\gamma_2 + \delta_i$$

Suppose that the ϵ_i 's are independent and identically distributed according to $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$, and the δ_i 's are independent and identically distributed according to $E(\delta_i) = 0$, $Var(\delta_i) = \tau^2$. Let $\hat{\beta}$ and $\hat{\gamma}$ be the OLSE from the two regression models.

- a. Under what conditions is β_1 equal to γ_1 ?

Ans: If $\gamma_2 = 0$ or if $r_{WZ} = 0$ (or if $E[W | Z]$ is equal to some constant independent of Z).

(Note first that the above conditions are somewhat redundant. If $E[W | Z]$ is equal to some constant independent of Z , then we must have $r_{WZ} = 0$, but not necessarily vice-versa (think of a U shaped curve) I accepted the last condition in lieu of the assumption about the correlation, but I think it better to give the more general condition.

The condition based on γ_2 is obvious. To see the restriction on the correlation, it is easiest to consider the unbiasedness of OLSE. We know $\hat{\beta}_1 = \hat{\gamma}_1$ if $r_{WZ} = 0$, and because these estimators are unbiased, we must have $\beta_1 = \gamma_1$.

Of course, the above condition does not guarantee all of the stipulations of the problem. Hence, if we consider the setting in which $r_{WZ} = 0$ and the assumptions of the two models hold, then we get the last condition as the way in which $r_{WZ} = 0$ in this problem. Note that if we condition on Z_i , then in order for $\beta_1 = \gamma_1$ we must have that $\beta_0 + \epsilon_i = \gamma_0 + W_i\gamma_2 + \delta_i$. The left hand side are independent random variables having conditional (on Z_i) mean β_0 and conditional (again, on Z_i) variance σ^2 . The same must be true of the right hand side. Now by taking the conditional mean of the RHS, we know $\beta_0 = \gamma_0 + \gamma_2 E[W_i | Z_i]$. So we would have to have $E[W_i | Z_i]$ a constant independent of the value of Z_i or $\gamma_2 = 0$ or both. Similarly, by considering the conditional variance of the RHS, we know $\sigma^2 = \gamma_2^2 Var(W_i | Z_i) + \tau^2$, and we must have $Var(W_i | Z_i)$ a constant independent of Z_i or $\gamma_2 = 0$ or both.

Had I not demanded that all the δ_i 's be identically distributed, then I could consider other cases in which there were no linear association between W_i and Z_i in the sample.)

- b. Under what conditions is $\hat{\beta}_1$ equal to $\hat{\gamma}_1$?

Ans: If $r_{WZ} = 0$.

(Technically, we could also have $\hat{\beta}_1 = \hat{\gamma}_1$ if $r_{WZ} = r_{YZ}/r_{YW}$ in a particular sample. However, this can not be guaranteed by experimental design, so we are fairly uninterested in this case. In any case, it is immaterial whether $\gamma_2 = 0$. See the key to problem 5 of Homework 3.)

c. Under what conditions is $E[\hat{\beta}_1 | \vec{Z}]$ equal to $E[\hat{\gamma}_1 | \vec{Z}, \vec{W}]$?

Ans: Same as part a. (We know OLSE are unbiased.)

d. Under what conditions is $Var(\hat{\beta}_1 | \vec{Z})$ equal to $Var(\hat{\gamma}_1 | \vec{Z}, \vec{W})$?

Ans: If $r_{WZ} = 0$ and either $\gamma_2 = 0$ or $Var(W|Z) = 0$.

(We know

$$Var(\hat{\beta}_1 | \vec{Z}) = \frac{Var(Y | Z)}{nVar(Z)} \quad \text{and} \quad Var(\hat{\gamma}_1 | \vec{Z}, \vec{W}) = \frac{Var(Y | Z, W)}{nVar(Z)(1 - r_{WZ}^2)}.$$

Furthermore, we know

$$Var(Y | Z) = E_W[Var(Y | Z, W)] + Var(E[Y | Z, W]) = \sigma^2 + \gamma_2^2 Var(W|Z).$$

The above restrictions that $r_{XZ} = 0$ and $\gamma^2 = 0$ are sufficient. Similarly, the restrictions that $r_{XZ} = 0$ and $Var(W|z) = 0$ are sufficient, though they are decidedly less interesting, as they would violate the conditions of parts a and c. Technically, we could also have $r_{WX} \neq 0$ and $\gamma_2 \neq 0$, with

$$Var(W|Z) = \frac{\sigma^2 r_{XW}^2}{\gamma_2^2 (1 - r_{WX}^2)},$$

but this possibility is uninteresting, because it cannot be guaranteed through experimental design: we do not know the value of γ_2 .

e. Suppose we are scientifically interested in testing hypotheses about γ_1 . What are the relative advantages of using the first model which regresses only on \vec{Z} versus the second model that regresses on both \vec{Z} and \vec{W} ? That is, describe the settings that one should prefer the first model and the settings that one should prefer the second model. **Very briefly** justify your answer.

Ans: First suppose that we believe that there is some chance $\gamma_2 \neq 0$. Then if $r_{XW} \neq 0$, we need to use the adjusted model in order to avoid confounding. Even if $r_{XW} = 0$, we would generally gain precision by using the adjusted model, because we would want to estimate our standard error using $Var(Y | Z, W)$, rather than $Var(Y | Z)$. Now if we are certain that $\gamma_2 = 0$, then if $r_{WZ} \neq 0$, we would definitely want to avoid the “variance inflation” that would come with the use of the adjusted model, so we would use the unadjusted model. Even if $r_{XW} = 0$, we would probably still want to use the unadjusted model to allow one more degree of freedom in estimating our residual variance.

f. Suppose you are able to design an experiment in order to make inference about γ_1 . What statistical design would provide the greatest precision? Be sure to specify the analysis model you will use.

Ans: 1:1 randomization between the lowest and highest possible values of Z . If I believed $\gamma_2 \neq 0$, I would include only subjects who have some single specified value of W . Analysis would be by the unadjusted model

The 1:1 randomization was chosen because of homoscedasticity, and I could limit sampling to a single value of W and two values of Z because the linear model was stipulated. In real life, we would never believe that assumption. If we want to be able to investigate a common treatment response across groups having different values of W , we would instead typically use randomization stratified by W in order to have $r_{WZ} = 0$ and use the adjusted analysis. Furthermore, Koch's postulates (an early treatise on how to establish cause and effect in infectious diseases) would argue that sampling multiple values of Z would help establish a dose-response. The randomization ratio across dose groups would generally be chosen to optimize either the detection of nonlinearity of dose response or to allow collection of secondary outcomes at each dose group.

g. I did not ask you this, but: Under what conditions is β_0 equal to γ_0 ?

Ans: If $\gamma_2 = 0$ or if $E[W|Z = 0] = 0$.

h. I did not ask you this, but: Under what conditions is $\hat{\beta}_0$ equal to $\hat{\gamma}_0$?

Ans: Sufficient conditions are that $r_{WX} = 0$ and $\overline{W} = 0$.

2. Consider the adjusted model of problem 1, but suppose now we are interested in testing whether the distribution of Y is affected by either Z or W (i.e., we test $H_0 : \gamma_1 = \gamma_2 = 0$).
- a. Briefly describe a testing procedure for the case where the distribution of δ_i is arbitrary. Very, very briefly justify the distributional assumptions for your test.

Ans: Our null hypothesis can be expressed in matrix form as $H_0 : \mathbf{A}\hat{\gamma} = \vec{0}$, where matrix

$$A = \begin{pmatrix} 0, 1, 0 \\ 0, 0, 1 \end{pmatrix}$$

has rank 2. In problem 1, we have $\delta_i \sim (0, \tau^2)$ independently, but the exact form of the distribution is unspecified. We let $\mathbf{W} = \begin{pmatrix} \vec{1} & \vec{Z} & \vec{W} \end{pmatrix}$. From the CLT for regression, we know that as the smallest eigenvalues of $\mathbf{W}^T \mathbf{W}$ go to infinity,

$$\hat{\gamma} \sim \mathcal{N}_3(\vec{\gamma}, \tau^2(\mathbf{W}^T \mathbf{W})^{-1}).$$

Hence, in that asymptotic setting and under H_0 , the quadratic form

$$Q = \frac{\hat{\gamma}^T \mathbf{A}^T (\mathbf{A}(\mathbf{W}^T \mathbf{W})^{-1} \mathbf{A}^T)^{-1} \mathbf{A} \hat{\gamma}}{\tau^2} \sim \chi_2^2.$$

Because $\hat{\tau}^2 = (\vec{Y} - \mathbf{W}\hat{\gamma})^T (\vec{Y} - \mathbf{W}\hat{\gamma}) / (n-3) \rightarrow_p \tau^2$, Slutsky's theorem allows us to define $Q^* = Q\tau^2/\hat{\tau}^2$ as also having an approximate χ_2^2 null distribution. Thus we can define a level α hypothesis test based on

$$\text{reject } H_0 \iff Q^* > \chi_2^2(1 - \alpha),$$

where $\chi_2^2(1 - \alpha)$ is the $1 - \alpha$ quantile of the chi square distribution having 2 degrees of freedom. In the setting of linear regression and analysis of variance, we most often instead use

$$\text{reject } H_0 \iff Q^* > F_{2, n-3}(1 - \alpha),$$

where $F_{2, n-3}(1 - \alpha)$ is the $1 - \alpha$ quantile of the F distribution having 2 and $n - 3$ degrees of freedom. Such an approach is in keeping with the exact test specified in part b, and it is statistically valid, because as $n \rightarrow \infty$, $F_{2, n-3}(1 - \alpha) \rightarrow \chi_2^2(1 - \alpha)$. In either formulation, the hypothesis test is consistent.

- b. How does the justification for your test and the distributional assumptions change when we presume $\delta_i \sim \mathcal{N}(0, \sigma^2)$?

Ans: If $\delta_i \sim \mathcal{N}(0, \tau^2)$ independently, then $Q \sim \chi_2^2$ and $(n-3)\hat{\tau}^2/\tau^2 \sim \chi_{n-3}^2$ are independent random variables. Hence, $Q^* \sim F_{2, n-3}$ exactly. This test is efficient.

Work either problem 3 or problem 4 (both earns extra credit).

3. Consider a regression model for response variables Y_{ij} where $j = 1, \dots, n_i$ notes the measurements made on each of $i = 1, \dots, m$ subjects. We model the distribution of Y_{ij} as

$$Y_{ij} = \beta_0 + X_{ij}\beta_1 + \epsilon_{ij},$$

where our experimental design is such that $X_{ij} = x_i$ for all $j = 1, \dots, n_i$. We assume that $\epsilon_{ij} \sim (0, \sigma^2)$ and that measurements made on different subjects are uncorrelated (so $\text{corr}(\epsilon_{ij}, \epsilon_{k\ell}) = 0$ if $i \neq k$), but measurements made on the same subject have correlation ρ (so $\text{corr}(\epsilon_{ij}, \epsilon_{i\ell}) = \rho$ if $j \neq \ell$).

- a. Suppose we combine all data into a single sample (hence ignoring the subjects) and use standard linear regression software to compute OLSE $\hat{\vec{\beta}}$. Notationally, let $\vec{Y}_i = (Y_{i1}, \dots, Y_{in_i})^T$ and $\vec{Y} = (\vec{Y}_1^T, \dots, \vec{Y}_m^T)^T$. Let \vec{X} be the corresponding values of the predictors. Our OLS regression model can be represented as $\vec{Y} | \vec{X} \sim (\vec{\mu}, \Sigma)$. What is the asymptotic distribution of $\hat{\vec{\beta}}$? What optimality criteria does it satisfy?

Ans: The OLSE has approximate distribution

$$\hat{\vec{\beta}} \sim \mathcal{N}_2 \left(\vec{\beta}, (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \Sigma \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \right).$$

The OLSE is unbiased and consistent, but it is not the BLUE unless $\rho = 0$ or n_i is the same for all individuals.

- b. Consider again the setting of part a. How does the estimated standard error of $\hat{\beta}_1$ as returned by the OLS software compare to the true standard error? (A qualitative answer with the reasoning behind it will suffice.)

Ans: The standard linear regression software will have presumed that $\Sigma = \sigma^2 \mathbf{I}_n$. Because the covariate value is constant among the correlated observations made on the same individual, the estimated standard error will be too small if $\rho > 0$ and too large if $\rho < 0$.

(Note: It is straightforward to establish the above. We know that the OLSE for β_1 is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Owing to the structure of the repeated measurements made on each subject, we can write this OLSE as a sum over individuals by defining $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$ and writing the estimator as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^m (X_{i1} - \bar{X}) n_i \bar{Y}_i}{\sum_{i=1}^m n_i (X_{i1} - \bar{X})^2}.$$

Because the individuals are independent of one another, the true standard error is

$$\text{Var}(\hat{\beta}_1 | \vec{X}) = \frac{\sum_{i=1}^m (X_{i1} - \bar{X})^2 n_i^2 \text{Var}(\bar{Y}_i | X_{i1})}{\left(\sum_{i=1}^m n_i (X_{i1} - \bar{X})^2\right)^2}.$$

$\text{Var}(\bar{Y}_i | X_{i1})$ is found by noting that

$$\vec{Y}_i | X_{i1} \sim \left((\beta_0 + \beta_1 X_{i1}) \vec{1}_{n_i}, \sigma^2 \left((1 - \rho) \mathbf{I}_{n_i} + \rho \vec{1}_{n_i} \vec{1}_{n_i}^T \right) \right).$$

Hence, because

$$\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij} / n_i = \vec{1}_{n_i}^T \vec{Y}_i / n_i,$$

we have

$$\begin{aligned} \bar{Y}_i | X_{i1} &\sim \left((\beta_0 + \beta_1 X_{i1}), \frac{\sigma^2}{n_i^2} \vec{1}_{n_i}^T \left((1 - \rho) \mathbf{I}_{n_i} + \rho \vec{1}_{n_i} \vec{1}_{n_i}^T \right) \vec{1}_{n_i} \right) \\ &\sim \left((\beta_0 + \beta_1 X_{i1}), \frac{\sigma^2}{n_i} (1 + (n_i - 1)\rho) \right). \end{aligned}$$

Substituting the formula for the variance in the earlier equation yields

$$\text{Var}(\hat{\beta}_1 | \vec{X}) = \frac{\sum_{i=1}^m (X_{i1} - \bar{X})^2 n_i \sigma^2 (1 + (n_i - 1)\rho)}{\left(\sum_{i=1}^m n_i (X_{i1} - \bar{X})^2\right)^2}.$$

Note that if any $n_i > 1$, this expression for the variance is monotonically increasing in ρ . Standard OLSE linear regression packages presume that $\rho = 0$, and thus they will overestimate the true standard error when $\rho < 0$, and they will underestimate the true standard error when $\rho > 0$. The degree to which the estimated standard error is wrong will depend upon the n_i 's and the value of ρ .

- c. How could you improve on the statistical analysis performed in part a? What optimality criteria would that analysis satisfy? Briefly describe the barriers to that analysis and any iterative approaches that might allow more optimal, valid statistical inference.

Ans: The BLUE would be the generalized least squares estimate

$$\hat{\beta}_G = (\mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Sigma^{-1} \vec{Y}.$$

It would have the lowest variance of all unbiased linear estimators. It would also be consistent. The problem is that we do not know the value of ρ , and that is crucial to being able to form the properly weighted analysis (the value of σ^2 is not as crucial, because it is a scale factor common to all measurements). We could, however, find an estimate of ρ as the correlation

among the residuals on the same individual from an OLSE regression, and then construct an estimate of Σ/σ^2 to use in a GLSE analysis.

(There are a wide variety of estimates of ρ that could be used. In a balanced design (where n_i is the same for all i), the easiest approach would be the intraclass correlation. In an unbalanced design there are many alternative formulations for computing the intraclass correlation. In GEE, we generally use some method of moments estimate. One simple (simplistic?) approach might be to estimate a within cluster covariance matrix V where $V_{jk} = \text{corr}(Y_{ij}, Y_{ik})$ is estimated among those individuals having $n_i \geq j$ and $n_i \geq k$. We could then use

$$\hat{\rho} = \frac{\sum_{j < k} w_{jk} \hat{V}_{jk}}{\sum_{j < k} w_{jk}},$$

a weighted average of the individual estimates. An intuitive weighting might be based on the number of pairs contributing to the estimate \hat{V}_{jk} .)

- d. Consider now an alternative analysis in which we measure the mean response on each subject $M_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$. What are the mean and variance of M_i ?

Ans: Because $X_{ij} = X_{i1}$ for all j , $\text{Var}(M_i | X_{i1})$ is found by noting

$$\vec{Y}_i | X_{i1} \sim \left((\beta_0 + \beta_1 X_{i1}) \vec{1}_{n_i}, \sigma^2 \left((1 - \rho) \mathbf{I}_{n_i} + \rho \vec{1}_{n_i} \vec{1}_{n_i}^T \right) \right).$$

Hence, because

$$M_i = \sum_{j=1}^{n_i} Y_{ij}/n_i = \vec{1}_{n_i}^T \vec{Y}_i / n_i,$$

we have

$$\begin{aligned} M_i | X_{i1} &\sim \left((\beta_0 + \beta_1 X_{i1}), \frac{\sigma^2}{n_i^2} \vec{1}_{n_i}^T \left((1 - \rho) \mathbf{I}_{n_i} + \rho \vec{1}_{n_i} \vec{1}_{n_i}^T \right) \vec{1}_{n_i} \right) \\ &\sim \left((\beta_0 + \beta_1 X_{i1}), \frac{\sigma^2}{n_i} (1 + (n_i - 1)\rho) \right). \end{aligned}$$

- e. Briefly describe an optimal, valid regression analysis to estimate β_1 using \vec{M} . How does this analysis compare to that in part c?

Ans: We can use the regression model $E[M_i | W_i] = \beta_0 + \beta_1 W_i$. If all the n_i 's are the same, then we have homoscedasticity and we can use OLSE. Such an analysis with standard statistical software will provide valid inference, and the estimated residual standard error estimates $\sigma \sqrt{(1 + (n_i - 1)\rho)/n_i}$. In the general case in which the n_i 's are not necessarily the same, we can find the GLSE

$$\hat{\beta}_G = (\mathbf{W}^T \Sigma^{*-1} \mathbf{W})^{-1} \mathbf{W}^T \Sigma^{*-1} \vec{M},$$

where $W_i = X_{i1}$ and $\Sigma_{ii}^* = \sigma^2(1 + (n_i - 1)\rho)/n_i$ and $\Sigma_{ij}^* = 0$ if $i \neq j$. Because $X_{ij} = W_i$ for all j , this estimate would be exactly the same as that in part **c**, if we use the same values for $\hat{\rho}$ and $\hat{\sigma}^2$ in estimating the weight matrices Σ and Σ^* . In either the balanced case or the unbalanced case, the analysis would differ only in the degrees of freedom used in the critical value for an F statistic. This is because each observation made on the same subject is weighted equally in either analysis, and the total weights assigned to each subject is the same in either analysis.

(Note that in the unbalanced case we must use the full data to get an estimate of ρ and σ^2 .)

Work either problem 3 or problem 4 (both earns extra credit).

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

$$Y_i = \beta_0 + X_i\beta_1 + \epsilon_i$$

with ϵ_i 's independent and identically distributed according to $E(\epsilon_i) = 0$ and $Var[\epsilon_i] = \alpha_0 + \alpha_1 X_i$, with α_0 and α_1 unknown.

- a. What is the impact of performing standard OLSE on this model? (Brevity is sufficient and greatly admired.)

Ans: If the distribution of the X_i 's is symmetric (no skewness) or if $\alpha_1 = 0$, there is no impact on the validity of an OLSE analysis. If the distribution of the X_i 's is positively skewed, then OLSE based inference is anti-conservative when $\alpha_1 > 0$ and conservative when $\alpha_1 < 0$. If the distribution of the X_i 's is negatively skewed, then OLSE based inference is anti-conservative when $\alpha_1 < 0$ and conservative when $\alpha_1 > 0$.

- b. How might you use the residuals from OLSE to estimate α_0 and α_1 ?

Ans: OLSE are unbiased, hence so long as the linear model is correct, the OLSE residuals are unbiased (but slightly correlated) estimates of the errors. Now because error $\epsilon_i \sim (0, \alpha_0 + \alpha_1 X_i)$, we know $E[\epsilon_i^2 | X_i] = \alpha_0 + \alpha_1 X_i$. Letting $\hat{\epsilon}_i$ be the i th residual, we could fit regression model

$$E[\hat{\epsilon}_i^2 | X_i] = \alpha_0 + \alpha_1 X_i.$$

(A key point to remember is that if we have a zero mean random variable, then the expectation of the squared random variable is the variance. I note that the squared estimated residuals are slightly correlated and quite likely heteroscedastic. However OLSE will still provide unbiased estimates of $\vec{\alpha}$.)

- c. Describe a way in which the linear regression methods we have discussed this quarter might be used to provide more optimal, asymptotically valid inference about β_1 .

Ans: First perform OLSE to obtain estimates of the residual errors. Then perform an OLSE regression of the squared residuals as in part b. Then construct an estimated covariance matrix $\hat{\Sigma} = \text{diag}(\hat{\alpha}_0 + \hat{\alpha}_1 X_1, \dots, \hat{\alpha}_0 + \hat{\alpha}_1 X_n)$, and find GLSE

$$\hat{\beta}_G = (\mathbf{X}^T \hat{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\Sigma}^{-1} \vec{Y}.$$

We can then iterate this process using the residuals from the GLSE regression to find more precise estimates of $\vec{\alpha}$ and then better estimates of $\hat{\Sigma}$ to use in a GLSE regression.

Work either problem 5 or problem 6 (both earns extra credit).

5. Consider a probability model in which $Y_i | X_i \sim (\mu_i, \sigma^2)$ independently, where the μ_i are assumed to depend on the value of X_i in some unspecified manner (not necessarily linear in X). Suppose, however, that we fit a linear regression model using standard software in order to obtain OLSE $\hat{\beta}$. (This problem is often referred to as “model misspecification”.)

- a. Provide a formula for the resulting OLSE $\hat{\beta}_1$.

Ans:

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

- b. What is $E[\hat{\beta}_1 | \bar{X}]$?

Ans:

$$E[\hat{\beta}_1 | \bar{X}] = \frac{\sum_{i=1}^n (X_i - \bar{X})\mu_i}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

- c. Under what conditions will $\hat{\beta}_1$ consistently estimate some population parameter β_1 , and what is β_1 ? (You need not be rigorous here, just give the general conditions.)

Ans: Supposing that as $n \rightarrow \infty$ the relative proportion of the observations sampled at each value of X_i stays constant, $\hat{\beta}_1$ consistently estimates a contrast across the individual group means

$$\hat{\beta}_1 \rightarrow_p \beta_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})\mu_i}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

- d. Under what conditions will the estimated standard error of $\hat{\beta}_1$ (as returned by standard linear regression software) be consistent for the true standard error? Will hypothesis tests of $H_0 : \beta_1 = 0$ be valid (where β_1 meets the conditions in part c)?

Ans: The squared standard error of $\hat{\beta}_1$ is

$$\text{Var}(\hat{\beta}_1 | \bar{X}) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 \text{Var}(Y_i | X_i)}{(\sum_{i=1}^n (X_i - \bar{X})^2)^2} = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

The estimated standard error will use the squared estimated residual standard error

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\beta}_0 + \hat{\beta}_1 X_i)^2,$$

which is consistent for σ^2 only if $\mu_i = \beta_0 + \beta_1 X_i$ for all i . Otherwise, the estimated standard error is too large, as it reflects both the random error $Y_i - \mu_i$ and the systematic error $\mu_i - \beta_0 - \beta_1 X_i$, where β_1 is defined as in part c and β_0 is consistently estimated by $\bar{Y} - \beta_1 \bar{X}$ (presuming that the sampling of the X_i s is as in part c). Note that under the strong null $H_0 : \mu_i = \mu, i = 1, \dots, n$, we do have $\beta_0 = \mu, \beta_1 = 0$, and $\mu_i = \beta_0 + \beta_1 X_i$, so the tests using standard OLSE software are valid for tests of the strong null hypothesis. However, they will not in general be valid for tests of the weak null that there is no linear trend in the μ_i 's.

Work either problem 5 or problem 6 (both earns extra credit).

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

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

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, considering separately the case of testing a hypothesis $H_0 : \beta_1 = 0$ and computing a confidence interval for β_1 . 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. This condition is satisfied under H_0 , but it is not satisfied under alternatives. Hence, hypothesis testing is valid under a strong null hypothesis (but not under a weak null hypothesis only testing for linear trends). Confidence intervals will not be valid due to the heteroscedasticity under the alternatives.

One approach around this problem would be to iteratively use weighted least squares with the current estimate of $\vec{\beta}$ at each iteration used to estimate the covariance matrix for \vec{Y} . When using such an approach with confidence intervals, either the sample size must be large enough to minimize the role of the mean-variance relationship, or the confidence intervals must be computed by inverting the score or likelihood ratio tests.

An equivalent formulation of this approach for this exponential family distribution with an identity link would be to use likelihood based regression methods.

- 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.

(This model can be viewed as a multiplicative model of the geometric mean (whether or not the ϵ_i 's are known to be identically distributed) or a multiplicative model of any quantile (given the presumption of i.i.d. ϵ_i 's). On the other hand, a likelihood based model of λ_i using a log link is a multiplicative model of the mean. So these are very different analysis approaches.)

7. 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: Consider simple linear regression in which (Y_i, x_i) are pairs of response R.V.'s and known predictors. Y_i 's are independently distributed $Y_i | x_i \sim (\mu_i, \sigma^2)$ with $\sigma^2 < \infty$ known, $\mu_i = \beta_0 + \beta_1(x_i - \bar{x})$, and $(Y_i - \mu_i) \sim_{iid} (0, \sigma^2)$. Then for $\mathbf{X} = (\vec{1}_n \bar{x} - \bar{x})$, $\vec{\beta} = (\beta_0 \beta_1)^T$, and OLSE $\hat{\vec{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X} \vec{Y}$,

$$\begin{aligned} \vec{Z}_n &= (\mathbf{X}^T \mathbf{X})^{1/2} (\hat{\vec{\beta}} - \vec{\beta}) \\ &= \left(\begin{array}{c} \sqrt{n}(\hat{\beta}_0 - \beta_0) \\ \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} (\hat{\beta}_1 - \beta_1) \end{array} \right) \rightarrow_d \mathcal{N}_2(0, \sigma^2 \mathbf{I}_2) \end{aligned}$$

Pf: Basic idea:

1. Consider arbitrary $a, b \in \mathcal{R}$ and $W_n = aZ_{n1} + bZ_{n2}$ so that $E[W_n] = 0$ and $Var[W_n] = (a^2 + b^2)\sigma^2$.
2. Notice that $W_n = \sum_{i=1}^n w_{ni}$ with the w_{ni} 's independent and $w_{ni} = k_i \epsilon_i$.
3. Use L-F CLT to show $W_n \rightarrow_d \mathcal{N}(0, (a^2 + b^2)\sigma^2)$ by showing that the Lindeberg condition holds if

$$\max \left\{ \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2} \right\} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

4. Then apply the Cramer-Wold device to show the multivariate normal asymptotic distribution for \vec{Z}_n .

(More detail is provided below. For \vec{Z}_n and $\hat{\vec{\beta}}$ defined as above, straightforward matrix multiplication yields

$$\begin{aligned} \hat{\beta}_0 &= \bar{Y} \\ \hat{\beta}_1 &= \frac{S_{XY}}{S_{XX}} = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{j=1}^n (x_j - \bar{x})^2}. \end{aligned}$$

Further,

$$\begin{aligned}
 \hat{\beta}_0 - \beta_0 &= \frac{1}{n} \sum_{i=1}^n (Y_i - \beta_0) \\
 &= \frac{1}{n} \sum_{i=1}^n [(x_i - \bar{x})\beta_1 + \epsilon_i] \\
 &= \bar{\epsilon} \\
 \hat{\beta}_1 - \beta_1 &= \frac{1}{\sum_{j=1}^n (x_j - \bar{x})^2} \sum_{i=1}^n (x_i - \bar{x})(Y_i - (x_i - \bar{x})\beta_1) \\
 &= \frac{1}{\sum_{j=1}^n (x_j - \bar{x})^2} \sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \epsilon_i) \\
 &= \frac{\sum_{i=1}^n (x_i - \bar{x})\epsilon_i}{\sum_{j=1}^n (x_j - \bar{x})^2}.
 \end{aligned}$$

Hence $Z_{n1} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i$ and $Z_{n2} = \frac{\sum_{i=1}^n (x_i - \bar{x})\epsilon_i}{\sqrt{\sum_{j=1}^n (x_j - \bar{x})^2}}$. Now consider $W_n = aZ_{n1} + bZ_{n2}$.

$$W_n = \sum_{i=1}^n \left(\frac{a}{\sqrt{n}} + \frac{b(x_i - \bar{x})}{\sqrt{\sum_{j=1}^n (x_j - \bar{x})^2}} \right) \epsilon_i \equiv \sum_{i=1}^n w_{ni},$$

so that

$$\begin{aligned}
 E[w_{ni}] &= \sum_{i=1}^n \left(\frac{a}{\sqrt{n}} + \frac{b(x_i - \bar{x})}{\sqrt{\sum_{j=1}^n (x_j - \bar{x})^2}} \right) E[\epsilon_i] = 0 \\
 \text{Var}[w_{n,i}] &= \left(\frac{a}{\sqrt{n}} + \frac{b(x_i - \bar{x})}{\sqrt{\sum_{j=1}^n (x_j - \bar{x})^2}} \right)^2 \sigma^2 \\
 V_n &\equiv \sum_{i=1}^n \text{Var}[w_{ni}] = (a^2 + b^2)\sigma^2.
 \end{aligned}$$

We now want to use the L-F CLT to show that $W_n/V_n \rightarrow_d \mathcal{N}(0, 1)$, so we must show that

$$\frac{1}{V_n} \sum_{i=1}^n E[w_{n,i}^2 \times 1_{\{|w_{ni}| > M\sqrt{V_n}\}}] \rightarrow 0 \quad \forall M > 0.$$

To this end, let $c_{ni} = \frac{a}{\sqrt{n}} + \frac{b(x_i - \bar{x})}{\sqrt{\sum_{j=1}^n (x_j - \bar{x})^2}}$ and $c_n^* = \max\{c_{ni}\}$ and suppose that

$$\max \left\{ \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2} \right\} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

$$\Rightarrow c_n^* \rightarrow 0 \text{ as } n \rightarrow \infty$$

So,

$$\begin{aligned} \frac{1}{V_n} \sum_{i=1}^n E[c_{ni}^2 \epsilon_i^2 \times 1_{[|c_{ni}\epsilon_i| > M\sqrt{V_n}]}] &\leq \frac{1}{V_n} \sum_{i=1}^n E[c_{ni}^2 \epsilon_i^2 \times 1_{[|\epsilon_{ni}| > M\sqrt{V_n}/c_n^*]}] \\ &= \frac{1}{V_n} E[\epsilon_i^2 \times 1_{[|\epsilon_{ni}| > M\sqrt{V_n}/c_n^*]}] \sum_{i=1}^n c_{ni}^2 \\ &\quad (\text{since } \epsilon_i \text{ are iid}) \\ &= \frac{1}{(a^2 + b^2)\sigma^2} E[\epsilon_i^2 \times 1_{[|\epsilon_{ni}| > M\sqrt{V_n}/c_n^*]}] \times (a^2 + b^2) \\ &= \frac{1}{\sigma^2} E[\epsilon_i^2 \times 1_{[|\epsilon_{ni}| > M\sqrt{V_n}/c_n^*]}] \rightarrow 0 \\ &\quad (\text{since } M\sqrt{V_n}/c_n^* \rightarrow \infty \text{ since } c_n^* \rightarrow 0 \text{ as } n \rightarrow \infty) \end{aligned}$$

So by the dominated convergence theorem the Lindeberg condition holds and $aZ_{n1} + bZ_{n2} \rightarrow_d \mathcal{N}(0, 1)$. Finally, by appealing Cramer-Wold we have the desired result.

The restriction placed on the x_i 's in the above proof

$$\max \left\{ \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2} \right\} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

can have two notable violations: 1) If there is only one value of x_i sampled for $i > n^*$, or 2) if additional x_i 's are always chosen to be influential. More generally we can show that the OLSE of an estimable $\vec{\beta}$ (when $\vec{\beta}$ estimable we have a design matrix of full rank) from linear regression model $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$ with ϵ_i iid $(0, \sigma^2)$ are consistent as long as the smallest eigenvalue of $\mathbf{X}^T \mathbf{X}$ approach ∞ as $n \rightarrow \infty$. This guarantees we're not just sampling one point, but instead are gaining information with each observation.)

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

- b. Provide a brief outline of a proof that parametric likelihood based regression provides a basis for optimal inference in regular problems. Be sure to indicate where the key assumptions for regular problems play a role and the sense in which the inference is optimal.

Ans: Consider a probability model in which $Y_i \sim f(y; \theta_i)$ independently where $\theta_i = h(\beta, \vec{x}_i)$ for known covariate vector \vec{x}_i and

- a. (identifiability) the form of pdf (or pmf) $f(y; \theta)$ is known up to the unknown parameter θ , and $\theta_i \neq \theta_j$ implies cdfs $F(y; \theta_i)$ and $F(y; \theta_j)$ are not identical;
- b. (common support) the set $\{y : f(y; \theta_i) > 0\}$ is independent of θ_i ;
- c. (interchange of differentiation and integration) the pdf (or pmf) $f(y; \theta)$ can be twice differentiated with respect to θ under the integration with respect to y ;
- d. (boundedness) the third derivative of the pdf $f(y; \theta)$ with respect to θ is bounded in some neighborhood of the true value of θ ; and
- e. (information growth) as $n \rightarrow \infty$, the rank of the Fisher's information (negative expectation of the second derivative of the density of \vec{Y} with respect to β) is of constant rank and has eigenvalues all going to ∞ .

Then

1. From the boundedness of the third derivatives, we know that there exists a root $\hat{\vec{\beta}}$ of the likelihood equations $\vec{U}(\hat{\vec{\beta}}) = \frac{\partial}{\partial \vec{\beta}} \log f(\vec{y}; \theta) \Big|_{\vec{\beta}=\hat{\vec{\beta}}} = 0$ such that $\hat{\vec{\beta}} \rightarrow_p \vec{\beta}$.
2. From the common support and the ability to differentiate the density under the integral sign, we know that the i th observation's contribution to the score equations has $E[\vec{U}_i(\vec{\beta})] = \vec{0}$ and $Var(\vec{U}_i(\vec{\beta})) = \mathcal{I}_i(\vec{\beta})$, Fisher's information.
3. From the independence of the observations and conditions on the information growth, application of the Lindeberg-Feller CLT to the $\vec{U}_i(\vec{\beta})$ shows that under the true value of β

$$\vec{U}(\vec{\beta}) \sim \mathcal{N}_p(0, \mathcal{I}(\vec{\beta})).$$

4. A first order Taylor expansion of $\vec{U}(\hat{\beta})$ around $\vec{\beta}$ yields

$$\hat{\beta} \sim \mathcal{N}_p(\vec{\beta}, \mathcal{I}^{-1}(\vec{\beta})).$$

5. A second order Taylor expansion of the log likelihood (log density) $\mathcal{L}(\vec{\beta})$ about $\hat{\beta}$ yields

$$-2(\mathcal{L}(\vec{\beta}) - \mathcal{L}(\hat{\beta})) \sim \chi_p^2.$$

(For more detailed sketch of the proof, see the handout on asymptotic likelihood theory.)