ST 762, TEST SOLUTIONS, FALL 2009

Please sign the following pledge certifying that the work on this test is your own:

"I have neither given nor received aid on this test."

Signature: ____

Printed Name: ____

This test covers material in Chapters 1 – 12 of the class notes.

For each part of each question, please write your answers in the space provided. If you need more space, continue on the back of the page, and indicate clearly that you have continued on the back.

Points for each part of each problem are given in the left margin. TOTAL POINTS = 100.

You are allowed ONE 8 1/2 inch × 11 inch sheet of HANDWRITTEN NOTES, one side.

[10 points] 1. (a) A friend is analyzing data from an experiment he has conducted that are in the form of independent pairs (Y_j, x_j), $j = 1, \ldots, n$, where $n = 200$, the x_j are one-dimensional and positive, and the Y_j are continuous responses. He is very confident that

$$E(Y_j|x_j) = f(x_j, \boldsymbol{\beta}); \quad \text{var}(Y_j|x_j) = \sigma^2 x_j^{2\theta}, \quad j = 1, \ldots, n \tag{1}$$

for some function f that is increasing in x, where $\theta \geq 0$ is a scalar whose value your friend does *not* know. Your friend is also confident that it is reasonable to assume that $\epsilon_j = \{Y_j - f(x_j, \boldsymbol{\beta}_0)\}/\{\sigma_0 x_j^{\theta_0}\}$ are independent and identically distributed (iid), where $(\boldsymbol{\beta}_0, \sigma_0, \theta_0)$ are the true values of $(\boldsymbol{\beta}, \sigma, \theta)$, but he is not confident enough to specify the distribution of the ϵ_j beyond the fact that it is probably *symmetric*.

Your friend would like to estimate $\boldsymbol{\beta}$ by generalized least squares (GLS), but he doesn't know much about methods for estimating θ, which he would need to do to form weights. He has heard that methods based on treating powers or logarithms of absolute "residuals" as "responses" in certain estimating equations can be used for this purpose. In particular, he has been told that using this approach based on the logarithm of absolute residuals is especially easy to implement under his model (1), but that this approach based on squared (absolute) residuals is the most widely used by statisticians. He has also been told that some people prefer using this approach based on the absolute residuals themselves rather than their squares or logarithms.

Your friend knows that you are taking ST 762 and asks you to explain all of this to him. What do you tell him?

This problem has to do with differences among several of the methods for estimating θ and their implications for estimation of both β and θ. Here are some of the points you may have made (not in order of importance):

- Standard large sample theory says that, insofar as the precision with which you estimate $\boldsymbol{\beta}$ goes, there is no effect of estimating θ nor of the method you use to estimate it. your sample size $n = 200$ is pretty large, so it very well might make little difference as to which of these methods you use. However, you can never be sure in any given problem that the implications of large sample theory are valid (so that $n = 200$ is really "large enough"), so it may be possible that the method you use to estimate θ will matter in terms of how well you can estimate $\boldsymbol{\beta}$. Fancier theory says that, in fact, the precision with which you can estimate $\boldsymbol{\beta}$ can be directly related to that with which you estimate θ; the worse it is for θ, the less precise the estimator for $\boldsymbol{\beta}$ can be. So it is probably worth worrying about to be safe.
- The method based on logarithms of absolute residuals regresses the log of the absolute value of $|r_j| = |Y_j - f(x_j, \hat{\boldsymbol{\beta}}^*)|$, where $\hat{\boldsymbol{\beta}}^*$ is a "preliminary" estimator for $\boldsymbol{\beta}$ (e.g., OLS or a previous GLS estimator) against the logarithm of the model for standard deviation, which in your case is $\log \sigma + \theta \log x_j$. Thus, for your model, it is really easy, as this boils down to a simple linear regression, where the estimator of θ is the estimator of slope.
- However, in the situation you have (variance model does not depend on $\boldsymbol{\beta}$, symmetric distribution, ϵ_j iid), theory says that this method can be very inefficient relative to the other methods when the data are in fact normally distributed or even when they are prone to "unusual" observations. So if you do use this method, be aware that its simplicity is offset by relative imprecision, which, as above, can translate into poor estimation of $\boldsymbol{\beta}$.

- The method based on squared residuals is like a weighted regression with $|r_j|^2$ as the "responses" and the variance $\sigma^2 x_j^{2\theta}$ as the "mean model." It is popular because it can be derived by making the assumption that $Y_j|x_j$ is normal and because, even if it is not, it will always produce a consistent estimator for θ. The method based on absolute residuals is like a weighted regression with $|r_j|$ as the "responses" and "mean model" $e^{\eta} x_j^{\theta}$ for some η. The parameter η only makes sense if you are willing to believe that the ϵ_j are iid (which you are), otherwise, no such η exists, and the result can be an inconsistent estimator for θ if you mistakenly assume that there is such an η. For this reason, the method based on squared residuals is usually preferred.
- However, in your situation, the method based on absolute residuals has some advantages. In the case just like yours (variance model does not depend on $\boldsymbol{\beta}$, symmetric distribution, ϵ_j iid), the squared residual method will yield the most precise estimator for θ if $Y_j|x_j$ is exactly normal. But this method based on absolute residuals can lead to a more precise estimator for θ if the true distribution of $Y_j|x_j$ deviates just slightly from normality, so if you are concerned that your data are prone to "unusual" observations, you might prefer this method.
- The methods based on either squared or absolute residuals are actually not hard to implement – for a model like yours, they can be carried out using standard nonlinear regression software. So they are really not appreciably more complicated to implement than the method based on logarithms. Given its inferior precision relative to these other two, I would recommend you use either the squared residual method or the absolute residual method. Which one depends on how confident you feel about the assumption that the ϵ_j are iid and how far, if at all, the distribution of the data deviates from normality.

Here is model (1) again for convenience:

$$E(Y_j|x_j) = f(x_j, \boldsymbol{\beta}); \quad \text{var}(Y_j|x_j) = \sigma^2 x_j^{2\theta}, \quad j = 1, \ldots, n$$

[5 points] (b) Your friend has decided to use the method for estimating θ described in (a) based on squared residuals in the analysis of his data. He is interested in testing the null hypothesis $H_0 : \theta = 0$ against the alternative $H_1 : \theta \neq 0$ to assess whether or not there is evidence, in the context of the variance model in (1), supporting the contention that variance really is not constant. He has been told that, if he makes the assumption that $Y_j|x_j$ is normally distributed, it should be possible to construct test statistics based on standard large sample theory that should have approximately a chi-square distribution with 1 degree of freedom, so it will be straightforward to carry out the hypothesis test.

What do you tell him?

This problem has to do with the issues associated with using the large sample theory for estimators for θ as the basis for inference on θ (hypothesis tests in this case).

This is certainly true if your data *really are* normally distributed; under the assumption of normality, you can construct either a Wald test statistic or a likelihood-based score type statistic that will indeed have a χ^2_1 sampling distribution in large samples. However, if in fact you data are not really normally distributed (but are still symmetric, as you believe, just with heavier tails), these test statistics will no longer have a χ^2_1 sampling distribution. Instead, the sampling distribution will be a scaled version of χ^2_1, where the scale factor depends on the *excess kurtosis* of the true distribution of $Y_j|x_j$ (if this distribution is exactly normal after all, excess kurtosis = 0, and scale factor is equal to 1). The upshot of all this is that, if the data really aren't normal after all, but deviate even only slightly from normality, your test statistic will not have a χ^2_1 distribution. If you go ahead and carry out your test assuming it does, the results would be erroneous, as the p-values you would obtain will not be correct.

You might think "why not estimate this excess kurtosis, then, and then correct this using the estimated scale factor?" Don't even go there – estimation of a higher moment quantity is ordinarily very imprecise unless you have *lots* of data (the sample size would probably have to be much larger than your $n = 200$ to get a reliable estimator). So, the bottom line is, unless you believe your data are exactly normal, you probably shouldn't be using these test statistics to draw inference on the true value of θ.

Here is model (1) again for convenience:

$$E(Y_j|x_j) = f(x_j, \boldsymbol{\beta}); \quad \text{var}(Y_j|x_j) = \sigma^2 x_j^{2\theta}, \quad j = 1, \ldots, n$$

[5 points] (c) Your friend wants to use his fitted model to calculate prediction intervals for values of the response that might be observed at different settings of x. He has found a SAS program online that calculates approximate $100(1-\alpha)\%$ prediction intervals for nonlinear models based on large sample theory under the usual assumption that the conditional variance of the response given x is constant for all j. Given that this program is already available, he would like to use it to obtain his prediction intervals.

What do you tell your friend?

This problem has to do with the the implications of getting the variance model incorrect for the validity of prediction (and calibration) intervals.

This is probably not a good idea. In a sample size as large as yours, the major source of variation that determines the width of prediction intervals is the variance of $Y_j|x_j$ itself - the variation due to fitting the model is dominated by this source. If the variance of $Y_j|x_j$ is no constant but instead follows your posited model, pretending that it is constant and using this program will lead to intervals that are too wide in regions of the range of x where $\sigma^2 x^{2\theta}$ is small and too narrow in regions where $\sigma^2 x^{2\theta}$ is large. This is because the width of valid prediction intervals will not be constant if the variance is not constant. It is worth your while to go the trouble to either modify this program or write your own if you want to obtain valid intervals that achieve the stated level

One more thing to be aware of – regardless of whether variance is constant or not, standard formulæ for prediction intervals are based on the assumption that the distribution of $Y_j|x_j$ is approximately normal. If the distribution deviates from normality, all of these intervals will probably be incorrect.

Here is model (1) again for convenience:

$$E(Y_j|x_j) = f(x_j, \boldsymbol{\beta}); \quad \mathrm{var}(Y_j|x_j) = \sigma^2 x_j^{2\theta}, \quad j = 1, \ldots, n$$

[5 points] (d) Your friend has data from a different, much smaller experiment ($n = 12$). He believes that model (1) and all of the assumptions about it in (a) should still hold for these data. He plans to estimate $(\boldsymbol{\beta}, \sigma, \theta)$ using GLS for $\boldsymbol{\beta}$ and one of the methods discussed in (a) to estimate θ. He is particularly interested in inference on the first component of $\boldsymbol{\beta}$, β_1, and plans to report an estimate and confidence interval for the true value of β_1 based on standard large sample theory, assuming that his model is correct.

What do you tell your friend?

This problem has to do with the performance of the GLS estimator in small samples and the implications of the second order theory.

With such a small sample size, it is very likely that the issue we discussed in part (a) will be problematic; that is, the fact that you have estimated θ will probably affect the true precision with which you estimate $\boldsymbol{\beta}$. The standard large sample theory assumes that how you estimated θ has no effect, so the standard errors you calculate based on it will probably be too small relative to the true extent of sampling variation inherent in estimation of $\boldsymbol{\beta}$, because they do not take the additional variation due to estimating θ in to account. This will lead to your confidence intervals being too narrow to achieve the state level of coverage. That is, the procedure you use based on the standard theory to obtain a 95% confidence interval for the true value of β_1 may only really be producing intervals that achieve, say, 88% coverage. So you will be misled into believing that you have more precise information on β_1 from your experiment of this size than you actually do.

One possible remedy is to instead use the bootstrap to obtain standard errors, and form your confidence interval using these. It has been shown that the bootstrap standard errors "automatically correct for the effect of estimating θ that the standard large sample theory does not capture, so this may produce a more reliable confidence interval.

One caveat: All of this assumes that your variance model is correctly specified (I am assuming that your mean model is correct). If it is not, you also have the problem that you will be using the wrong variance model in all of these calculations. Even if you didn't have the small sample issue, this would also lead to erroneous standard errors and intervals, because the standard theory (and the bootstrap) incorrectly characterize the variation in the data. It is possible to "fix up" the standard large sample theory standard errors to account for the fact that the variance model may be incorrect (the so-called "robust sandwich" standard errors); however, given your small sample size, this may not work so well, because you still have the issue of estimation of θ (which now does not even estimate a meaningful quantity, but there is still an effect of estimating whatever it is).

2. Suppose $(Y_j, \boldsymbol{x}_j)$, $j = 1, \dots, n$, are independent and such that

$$E(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \boldsymbol{\beta}); \quad \text{var}(Y_j|\boldsymbol{x}_j) = \sigma^2 g^2(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{x}_j), \tag{2}$$

Assume that $\boldsymbol{\theta}$ is *known*.

[5 points] (a) Suppose that we assume that the distribution of Y_j given $\boldsymbol{x}_j$ has coefficient of skewness ζ_j and coefficient of excess kurtosis κ_j for each j, for some constants ζ_j and κ_j, $j = 1, \dots, n$. Give the "optimal" quadratic estimating equations for $\boldsymbol{\beta}$ and σ under these conditions.

This is Problem 3(a) in Homework 3 Extra Problems, except all you needed to do is write down the equations. Using the shorthand notation in the notes, the optimal quadratic equation is

$$\sum_{j=1}^{n} \begin{pmatrix} f_{\beta j} & 2\sigma^2 g_j^2 \nu_{\beta j} \\ 0 & 2\sigma g_j^2 \end{pmatrix} \begin{pmatrix} \sigma^2 g_j^2 & \zeta_j \sigma^3 g_j^3 \\ \zeta_j \sigma^3 g_j^3 & (2+\kappa_j)\sigma^4 g_j^4 \end{pmatrix}^{-1} \begin{pmatrix} Y_j - f_j \\ (Y_j - f_j)^2 - \sigma^2 g_j^2 \end{pmatrix} = \mathbf{0}.$$

To answer this question, it would suffice to write this down.

You may have taken it further, as in the solution to Problem 3(a) in Homework 3 Extra Problems. If we take the inverse, we get

$$\sum_{j=1}^{n} \begin{pmatrix} f_{\beta j} & 2\sigma^2 g_j^2 \nu_{\beta j} \\ 0 & 2\sigma g_j^2 \end{pmatrix} \frac{1}{\sigma^2 g_j^6 (2 + \kappa_j - \zeta_j^2)} \begin{pmatrix} (2+\kappa_j)\sigma^4 g_j^4 & -\zeta_j \sigma^3 g_j^3 \\ -\zeta_j \sigma^3 g_j^3 & \sigma^2 g_j^2 \end{pmatrix} \begin{pmatrix} Y_j - f_j \\ (Y_j - f_j)^2 - \sigma^2 g_j^2 \end{pmatrix} = \mathbf{0}.$$

This can be simplified to

$$\sum_{j=1}^{n} \begin{pmatrix} f_{\beta j} & 2\sigma^2 g_j^2 \nu_{\beta j} \\ 0 & 2\sigma g_j^2 \end{pmatrix} \frac{1}{(2 + \kappa_j - \zeta_j^2)} \begin{pmatrix} \dfrac{(2+\kappa_j)}{\sigma^2 g_j^2} & -\dfrac{\zeta_j}{\sigma^3 g_j^3} \\ -\dfrac{\zeta_j}{\sigma^3 g_j^3} & \dfrac{1}{\sigma^4 g_j^4} \end{pmatrix} \begin{pmatrix} Y_j - f_j \\ (Y_j - f_j)^2 - \sigma^2 g_j^2 \end{pmatrix} = \mathbf{0}.$$

Multiplying this out yields

$$\sum_{j=1}^{n} \Big(\boldsymbol{a}_j \{Y_j - f(\boldsymbol{x}_j, \boldsymbol{\beta})\} + \boldsymbol{c}_j [\{Y_j - f(\boldsymbol{x}_j, \boldsymbol{\beta})\}^2 - \sigma^2 g^2(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{x}_j)] \Big) = \mathbf{0},$$

where

$$\boldsymbol{a}_j = \frac{1}{\sigma g_j (2 + \kappa_j - \zeta_j^2)} \begin{pmatrix} (2+\kappa_j) f_{\beta j}/(\sigma g_j) - 2\nu_{\beta j}\zeta_j \\ -2\zeta_j/\sigma \end{pmatrix},$$

$$\boldsymbol{c}_j = \frac{1}{\sigma^2 g_j^2 (2 + \kappa_j - \zeta_j^2)} \begin{pmatrix} -\zeta_j f_{\beta j}/(\sigma g_j) + 2\nu_{\beta j} \\ 2/\sigma \end{pmatrix}.$$

[10 points] (b) Suppose we are told that σ is *known* and $\sigma = 1$, $g(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{x}_j) = \sqrt{f(\boldsymbol{x}_j, \boldsymbol{\beta})}$, $\zeta_j = 1/\sqrt{f(\boldsymbol{x}_j, \boldsymbol{\beta})}$, and $\kappa_j = 1/f(\boldsymbol{x}_j, \boldsymbol{\beta})$. Find the form of the estimating equation for $\boldsymbol{\beta}$ in (a) under these conditions, and give an explanation for why your result turned out the way it did.

This is Problem 3(c) of Homework 3 Extra Problems. Here, we have $\sigma = 1$, $g_j = f_j^{1/2}$, $\nu_{\beta j} = (1/2)\partial/\partial\boldsymbol{\beta}(\log f_j) = (1/2)f_{\beta j}/f_j$, $\zeta_j = f_j^{-1/2}$, and $\kappa_j = f_j^{-1}$. Thus, $(2 + \kappa_j - \zeta_j^2) = (2 + f_j^{-1} - f_j^{-1}) = 2$. We need the first p rows of the equations in part (a) of this problem under these conditions.

We exhibit this directly. Substituting the above information, we may write the equations as

$$\sum_{j=1}^{n} \begin{pmatrix} f_{\beta j}, & 2f_j(1/2)f_{\beta j}/f_j \end{pmatrix} \begin{pmatrix} f_j & f_j^{-1/2} f_j^{3/2} \\ f_j^{-1/2} f_j^{3/2} & (2 + f_j^{-1}) f_j^2 \end{pmatrix}^{-1} \begin{pmatrix} Y_j - f_j \\ (Y_j - f_j)^2 - f_j \end{pmatrix} = \mathbf{0},$$

which simplifies to

$$\sum_{j=1}^{n} \begin{pmatrix} f_{\beta j}, & f_{\beta_j} \end{pmatrix} \frac{1}{2f_j^3 + f_j^2 - f_j^2} \begin{pmatrix} (2 + f_j^{-1}) f_j^2 & -f_j \\ -f_j & f_j \end{pmatrix} \begin{pmatrix} Y_j - f_j \\ (Y_j - f_j)^2 - f_j \end{pmatrix} = \mathbf{0}.$$

Thus, multiplying out, we get

$$\sum_{j=1}^{n} \frac{1}{2f_j^3} \begin{pmatrix} f_{\beta j}(2 + f^{-1}) f_j^2 - f_{\beta j} f_j, & -f_j f_{\beta j} + f_j f_{\beta j} \end{pmatrix} \begin{pmatrix} Y_j - f_j \\ (Y_j - f_j)^2 - f_j \end{pmatrix} = \mathbf{0},$$

which upon further simplification becomes

$$\sum_{j=1}^{n} \begin{pmatrix} f_{\beta j}/f_j + f_{\beta j}/(2f_j^2) - f_{\beta j}/(2f_j^2), & 0 \end{pmatrix} \begin{pmatrix} Y_j - f_j \\ (Y_j - f_j)^2 - f_j \end{pmatrix} = \mathbf{0}.$$

We thus obtain, multiplying out,

$$\sum_{j=1}^{n} f_j^{-1}(Y_j - f_j) f_{\beta j} = \mathbf{0}.$$

This is of course in the form of the GLS equations, so that the quadratic equations reduce to the "optimal" linear ones under the mean-variance model (2).

The reason: The variance function $g(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{x}_j) = \sqrt{f(\boldsymbol{x}_j, \boldsymbol{\beta})}$ and these values for ζ_j and κ_j correspond to those for a Poisson distribution. As in the solution to Homework 3 Extra Problem 3(e), this is a member of the scaled exponential family class, for which the maximum likelihood estimating equation for $\boldsymbol{\beta}$ is exactly of the GLS form. For these distributions, the variance is related to the mean in a specific way; thus an explanation is that, for this class, trying to gain information about $\boldsymbol{\beta}$ from the variance via quadratic equation is fruitless. The optimal quadratic equation in this case is the linear equation above, which corresponds to maximum likelihood, over which we cannot expect to achieve any improvement.

Even if you couldn't get the algebra to work out, you could have guessed at the result by recognizing that the values of σ, g_j, ζ_j, and κ_j correspond exactly to the first 4 moments of a Poisson distribution.

[5 points] (c) Suppose instead that σ is not known and must be estimated; that $f(\boldsymbol{x}_j, \boldsymbol{\beta})$ in (2) is correctly specified; and that $Y_j|\boldsymbol{x}_j$ are normally distributed, $j = 1, \ldots, n$, with conditional mean of the form $f(\boldsymbol{x}_j, \boldsymbol{\beta})$ and conditional variance of the form $\sigma^2 f^{2\theta}(\boldsymbol{x}_j, \boldsymbol{\beta})$. Suppose that we solve equations you wrote down in (a) for σ and $\boldsymbol{\beta}$ incorporating these conditions and assuming that the distribution of $Y_j|\boldsymbol{x}_j$ has constant coefficient of variation. Suppose that, in truth, $\theta = 0.5$.

What might be the consequences, and why?

This problem has to do with the implications for the properties of quadratic estimating equations when the variance model is misspecified.

The main issue here is that we will have specified the variance model incorrectly in setting up the equations, as the constant coefficient of variation model would have $\theta = 1$. This would lead to these estimating equations being *biased* for estimation of both $\boldsymbol{\beta}$ and σ, because the mean of the quadratic "response" would be misrepresented as $\sigma^2 f^2(\boldsymbol{x}_j, \boldsymbol{\beta})$ when it really is $\sigma^2 f(\boldsymbol{x}_j, \boldsymbol{\beta})$. The result would be inconsistent estimation of both parameters.

3. Suppose that Y_j, $j = 1, \ldots, n$, are independent, and suppose we have the model

$$E(Y_j) = e^{\beta}, \quad \text{var}(Y_j) = \sigma^2 e^{\beta}, \quad j = 1, \ldots, n, \tag{3}$$

where σ is *known*.

[5 points] (a) Find the GLS estimator $\widehat{\beta}_{GLS}$ for β for $C = \infty$ and, assuming (3) is correct with $\beta_0 =$ the true value of β, show that $\widehat{\beta}_{GLS}$ is a consistent estimator for β_0.

This problem is similar to the ones in Section 10.3 in the notes and in Homework 4 Extra Problems Problem 3 (which is much messier). Some of you forgot that σ is *known* in this problem, which made part (c) seem too hard.

Here, we have $f_{\beta j} = e^{\beta}$. Thus substituting into the GLS equations

$$\sum_{j=1}^{n} g_j^{-2}(Y_j - f_j) f_{\beta j} = 0,$$

the GLS estimator solves

$$\begin{aligned} 0 &= n^{-1} \sum_{j=1}^{n} \frac{(Y_j - e^{\beta}) e^{\beta}}{e^{\beta}} \\ &= n^{-1} \sum_{j=1}^{n} (Y_j - e^{\beta}) \end{aligned}$$

which leads to $\widehat{\beta}_{GLS} = \log(\bar{Y})$, where $\bar{Y} = n^{-1} \sum_{j=1}^{n} Y_j$. If (3) is correct, then by the WLLN, $\bar{Y} \xrightarrow{p} e^{\beta_0}$, so that, by continuity, $\log \bar{Y} \xrightarrow{p} \beta_0$, showing that the GLS estimator is consistent for β_0.

Here is model (3) again for convenience:

$$E(Y_j) = e^{\beta}, \quad \mathrm{var}(Y_j) = \sigma^2 e^{\beta}, \quad j = 1, \ldots, n.$$

[5 points] (b) Write down the optimal quadratic estimating equation for β when Y_j is assumed to have coefficient of skewness $\zeta = 0$ and excess kurtosis $= \kappa$.

We have $f_j = e^{\beta}$ and $g_j^2 = e^{\beta}$, so that $\nu_{\beta j} = \partial/\partial\beta(\log e^{\beta/2}) = (1/2)\partial/\partial\beta(\beta) = 1/2$. To obtain the equation, we can substitute these values, along with $\zeta_j = 0$ and $\kappa_j = \kappa$, in the top $p = 1$ rows of $\boldsymbol{a}_j$ and $\boldsymbol{c}_j$ in Problem 2(b) of this test to obtain immediately

$$\sum_{j=1}^{n} \frac{(Y_j - e^{\beta})e^{\beta}}{\sigma^2 e^{\beta}} + \sum_{j=1}^{n} \frac{(Y_j - e^{\beta})^2 - \sigma^2 e^{\beta}}{(2+\kappa)\sigma^2 e^{\beta}} = 0.$$

Here is model (3) again for convenience:

$$E(Y_j) = e^{\beta}, \quad \text{var}(Y_j) = \sigma^2 e^{\beta}, \quad j = 1, \ldots, n.$$

[10 points] (c) Suppose that, in truth, Y_j is normally distributed, and that (3) is correct. Find the estimator $\widehat{\beta}_Q$ solving the equation you found in (b) *that takes into account this information*, and show that $\widehat{\beta}_Q$ is a consistent estimator for β_0, where again β_0 is the true value of β.

We want to find the estimator solving the equation in part (b) that takes into account the knowledge that Y_j is normal, which implies $\kappa = 0$. This estimator is in fact the normal theory ML estimator. If we substitute $\kappa = 0$ in that equation, we obtain

$$\sum_{j=1}^{n} \frac{(Y_j - e^{\beta})e^{\beta}}{\sigma^2 e^{\beta}} + \sum_{j=1}^{n} \frac{(Y_j - e^{\beta})^2 - \sigma^2 e^{\beta}}{2\sigma^2 e^{\beta}} = 0.$$

Multiplying both sides by $\sigma^2 e^{\beta} n^{-1}$ and multiplying out and collecting terms yields, letting $\bar{Y} = n^{-1}\sum_{j=1}^{n} Y_j$,

$$e^{\beta}\bar{Y} - e^{2\beta} + n^{-1}\sum_{j=1}^{n} Y_j^2/2 - e^{\beta}\bar{Y} + e^{2\beta}/2 - \sigma^2 e^{\beta}/2,$$

which simplifies to (multiplying by -2)

$$e^{2\beta} + \sigma^2 e^{\beta} - Z = 0,$$

where $Z = n^{-1} Y_j^2$. As in the notes in Chapter 10, applying the quadratic formula and taking the positive root, we get that

$$e^{\beta} = \frac{-\sigma^2 + \sqrt{\sigma^4 + 4Z}}{2}.$$

Thus, $\widehat{\beta}_Q$ satisfies this expression, and

$$\widehat{\beta}_Q = \log\left(\frac{-\sigma^2 + \sqrt{\sigma^4 + 4Z}}{2}\right).$$

Noting that the Y_j are iid and that $E(Y^2) = \text{var}(Y) + \{E(Y)\}^{=}\sigma^2 e^{\beta_0} + e^{2\beta_0}$, we have immediately that

$$Z \xrightarrow{p} e^{2\beta_0} + \sigma^2 e^{\beta_0},$$

from whence it follows that

$$e^{\widehat{\beta}_Q} \xrightarrow{p} \frac{-\sigma^2 + \sqrt{\sigma^4 + 4e^{2\beta_0} + 4\sigma^2 e^{\beta_0}}}{2}.$$

The expression under the square root is easily seen to be a perfect square $= (\sigma^2 + 2e^{\beta_0})^2$. Substituting this and simplifying yields $e^{\widehat{\beta}_Q} \xrightarrow{p} e^{\beta_0}$, and hence $\widehat{\beta}_Q \xrightarrow{p} \beta_0$.

Here is model (3) again for convenience:

$$E(Y_j) = e^{\beta}, \quad \text{var}(Y_j) = \sigma^2 e^{\beta}, \;\; j = 1, \ldots, n.$$

[13 points] (d) Under the conditions in (c), find the asymptotic relative efficiency (ARE) of $\widehat{\beta}_{GLS}$ to $\widehat{\beta}_Q$.

We have shown in (b) and (c) that both estimators are consistent for β_0. We need to deduce the behavior of the quantity $n^{1/2}(\widehat{\beta} - \beta_0)$ for both; in particular, obtain the limit in distribution. We may then obtain an expression for the ARE by the ratio of the variances.

If you were lucky enough to remember or have written down the "folklore theorem" in the case where the variance model is correct, you could have plugged into it directly. The variance of the limiting distribution of $n^{1/2}(\widehat{\beta}_{GLS} - \beta_0)$ is the inverse of the limit as $n \rightarrow \infty$ of

$$n^{-1} \sum_{j=1}^{n} f_{\beta j} f_{\beta j}^T / g_j^2$$

evaluated at β_0, which, using the facts $f_{\beta j} = e^{\beta}$ and $g_j^2 = e^{\beta}$, reduces to $e^{2\beta_0}/e^{\beta_0} = e^{\beta_0}$. Thus, substituting in the folklore results gives immediately that

$$n^{1/2}(\widehat{\beta}_{GLS} - \beta_0) \xrightarrow{L} \mathcal{N}(0, \sigma^2 e^{-\beta_0}).$$

Similarly, if you also wrote down the large sample distributional result for quadratic estimators, you could have plugged into it under the conditions here (normality, so for the normal theory ML estimator when normality is correct). Under normality, the variance of the limiting distribution of $n^{1/2}(\widehat{\beta}_Q - \beta_0)$ is the inverse of the limit as $n \rightarrow \infty$ of

$$n^{-1} \sum_{j=1}^{n} f_{\beta j} f_{\beta j}^T / g_j^2 + 2\sigma^2 n^{-1} \sum_{j=1}^{n} \nu_{\beta j} \nu_{\beta j}^T.$$

(In specializing the general results on p. 253–254 of the class notes, because there is no $\boldsymbol{\theta}$ and σ is known in our problem, there is no $\boldsymbol{Q}$ and hence $\boldsymbol{P} = \boldsymbol{I}$.) We calculated the first piece above; the second, using $\nu_{\beta j} = 1/2$, is $\sigma^2/2$. Thus, substituting into the result, we get

$$n^{1/2}(\widehat{\beta}_Q - \beta_0) \xrightarrow{L} \mathcal{N}\{0, \sigma^2(e^{\beta_0} + \sigma^2/2)^{-1}\}.$$

From these results, the ARE of $\widehat{\beta}_{GLS}$ to $\widehat{\beta}_Q$ is immediately seen to be

$$\frac{\sigma^2(e^{\beta_0} + \sigma^2/2)^{-1}}{\sigma^2 e^{-\beta_0}} = \frac{e^{\beta_0}}{e^{\beta_0} + \sigma^2/2}.$$

Note that this quantity is < 1 unless $\sigma = 0$.

If you did not recall or write down these results, you could have derived them from scratch. For GLS, a linear Taylor series in the following about $\bar{Y} = \beta_0$.

$$\begin{aligned} n^{1/2}(\widehat{\beta}_{GLS} - \beta_0) &= n^{1/2}(\log \bar{Y} - \beta_0) \\ &\approx n^{1/2}(\beta_0 - \beta_0) + (1/\bar{Y})|_{e^{\beta_0}} n^{1/2}(\bar{Y} - \beta_0) \\ &\approx e^{-\beta_0} n^{-1/2} \sum_{j=1}^{n} (Y_j - \beta_0). \end{aligned}$$

Applying the Central Limit Theorem and Slutsky's theorem then yields the above result. (This is just the δ method.)

For the quadratic estimator, the Taylor series approach is a bit more involved, but could be done (see below). The easiest way to find the variance of the limiting distribution of $n^{1/2}(\widehat{\beta}_Q - \beta_0)$ is to recognize that $\widehat{\beta}_Q$ is the normal theory ML estimator under these conditions. Thus, even if you were unable to find $\widehat{\beta}_Q$ in (c), this variance could be found as the inverse of the information matrix (the Y_j are iid here) as follows. We know under these conditions that $\widehat{\beta}_Q$ maximizes the loglikelihood function

$$\log L = -\log \sigma - (1/2)\log e^{\beta} - \frac{1}{2\sigma^2}\frac{(Y - e^{\beta})^2}{e^{\beta}}$$

which simplifies to

$$\log L = -\log \sigma - (1/2)\beta - \frac{1}{2\sigma^2}(Y^2/e^{\beta} - 2Y + e^{\beta}).$$

Taking derivatives,

$$\partial/\partial\beta(\log L) = -1/2 - \frac{1}{2\sigma^2}(-Y^2/e^{\beta} + e^{\beta})$$

(which yields the quadratic equation solved in part (c), of course). Thus,

$$\partial^2/\partial\beta^2(\log L) = -\frac{1}{2\sigma^2}(Y^2/e^{\beta} + e^{\beta}).$$

The expectation of Y^2 is easily seen to be (found in (c))

$$E(Y^2) = \text{var}(Y) + \{E(Y)\}^2 = \sigma^2 e^{\beta_0} + e^{2\beta_0};$$

thus, the expectation of $\partial^2/\partial\beta^2(\log L)$ is, at the true value β_0,

$$-\frac{1}{2\sigma^2}\{(e^{2\beta_0} + \sigma^2 e^{\beta_0})/e^{\beta_0} + e^{\beta_0}\} = -\frac{1}{2\sigma^2}(2e^{\beta_0} + \sigma^2) = -(e^{\beta_0} + \sigma^2/2)/\sigma^2.$$

Thus, the inverse of the information matrix is

$$\sigma^2(e^{\beta_0} + \sigma^2/2)^{-1},$$

as found above.

By brute-force Taylor series: Taking a linear Taylor series about $Z = e^{2\beta_0} + \sigma^2 e^{\beta_0}$.

$$\begin{aligned}
n^{1/2}(\widehat{\beta}_Q - \beta_0) &= n^{1/2}\left\{\log\left(\frac{-\sigma^2 + \sqrt{\sigma^4 + 4Z}}{2}\right) - \beta_0\right\} \\
&\approx n^{1/2}(\beta_0 - \beta_0) + \left.\frac{(\sigma^4 + 4Z)^{-1/2}}{\left(\frac{-\sigma^2 + \sqrt{\sigma^4 + 4Z}}{2}\right)}\right|_{Z = e^{2\beta_0} + \sigma^2 e^{\beta_0}} n^{1/2}(Z - e^{2\beta_0} - \sigma^2 e^{\beta_0}) \\
&\approx \frac{\{\sigma^4 + 4(e^{2\beta_0} + \sigma^2 e^{\beta_0})\}^{-1/2}}{\left(\frac{-\sigma^2 + \sqrt{\sigma^4 + 4e^{2\beta_0} + 4\sigma^2 e^{\beta_0}}}{2}\right)} n^{1/2}(Z - e^{2\beta_0} - \sigma^2 e^{\beta_0}) \\
&\approx \frac{1}{e^{\beta_0}(2e^{\beta_0} + \sigma^2)} n^{1/2}(Z - e^{2\beta_0} - \sigma^2 e^{\beta_0})
\end{aligned}$$

Now, defining $\epsilon_j = (Y_j - e^{\beta_0})/(\sigma e^{\beta_0/2})$ as usual and thus $Y_j = (e^{\beta_0} + \sigma e^{\beta_0/2}\epsilon_j)$, we have $Y_j^2 = e^{2\beta_0} + 2\sigma e^{3\beta_0/2}\epsilon_j + \sigma^2 e^{\beta_0}\epsilon_j^2$, so that

$$\begin{aligned} n^{1/2}(Z - e^{2\beta_0} - \sigma^2 e^{\beta_0}) &= n^{-1/2}\sum_{j=1}^{n}(e^{2\beta_0} + 2\sigma e^{3\beta_0/2}\epsilon_j + \sigma^2 e^{\beta_0}\epsilon_j^2 - e^{2\beta_0} - \sigma^2 e^{\beta_0}) \\ &= n^{-1/2}\sum_{j=1}^{n}\{2\sigma e^{3\beta_0/2}\epsilon_j + \sigma^2 e^{\beta_0}(\epsilon_j^2 - 1)\} \end{aligned}$$

The term in braces obviously has mean 0 and its variance is easily seen to be (under normality)

$$4\sigma^2 e^{3\beta_0} + 2\sigma^2 e^{2\beta_0} = 2\sigma^2 e^{2\beta_0}(2e^{\beta_0} + \sigma^2).$$

Thus, applying the Central Limit Theorem to this quantity along with Slutsky's theorem in the expression for $n^{1/2}(\widehat{\beta}_Q - \beta_0)$ above yields that $n^{1/2}(\widehat{\beta}_Q - \beta_0)$ converges in distribution to a multivariate normal random vector with variance

$$\frac{2\sigma^2 e^{2\beta_0}(2e^{\beta_0} + \sigma^2)}{e^{2\beta_0}(2e^{\beta_0} + \sigma^2)^2} = \sigma^2(e^{\beta_0} + \sigma^2/2)^{-1},$$

yielding the result.

Here is model (3) again for convenience:

$$E(Y_j) = e^{\beta}, \quad \text{var}(Y_j) = \sigma^2 e^{\beta}, \quad j = 1, \dots, n.$$

[5 points] (e) Give a condition under which the ARE you found in (d) would be equal to 1, and explain what is going on under this condition to yield this result.

From the answer to (d), if we were to have let $\sigma \to 0$ with $n \to \infty$, the large sample distributional results in (d) would have been identical for the two estimators. And, of course, ARE $\to 1$ under this condition. This is simply exhibiting what we know from the discussion at the end of Chapter 10; namely, that, for σ "small," the GLS and normal theory ML estimators are asymptotically equivalent. In a practical sense, these estimators will give very similar results in the case of "high signal to noise" data, as you have observed in a previous homework problem.

4. A Monte Carlo simulation study has been conducted for which data were generated according to the following model:

$$Y_j = f(x_j, \boldsymbol{\beta}) + \sigma f^{\theta}(x_j, \boldsymbol{\beta})\epsilon_j, \quad j = 1, \dots, n, \quad f(x, \boldsymbol{\beta}) = \frac{\beta_1}{1 + (\beta_2/x)}, \tag{4}$$

where $n = 11$; $x_j = 7.8125, 15.625, 31.25, 62.5, 125, 250, 500, 1000, 2000, 4000, 8000$; ϵ_j are generated independently as $\epsilon_j \sim \mathcal{N}(0,1)$; $\boldsymbol{\beta} = (\beta_1, \beta_2)^T = (680, 400)^T$, $\theta = 0.8$, and $\sigma = 0.05$. For each of 1000 data sets, the model for $E(Y_j|x_j)$ and $\text{var}(Y_j|x_j)$ implied by (4) was fit four ways:

(i) By the GLS algorithm ($C = 10$) with θ taken as known and equal to 0.8, $GLS_{0.8}$

(ii) By the GLS algorithm ($C = 10$) with θ taken as known and equal to 0.4, $GLS_{0.4}$

(iii) By the GLS algorithm ($C = 10$) with θ estimated by pseudo-likelihood, GLS_{PL}

(iv) Joint estimation of (β, σ, θ) by normal theory maximum likelihood, ML

The following table presents the results:

	MC Mean	MC SD	MC Ave Est SEs	$\frac{\text{this MSE}}{\text{MSE}_{GLS_{0.8}}}$	$\frac{\text{this MSE}}{\text{MSE}_{GLS_{0.4}}}$	$\frac{\text{this MSE}}{\text{MSE}_{GLS_{PL}}}$	$\frac{\text{this MSE}}{\text{MSE}_{ML}}$
$GLS_{0.8}$	399.30	6.24	6.31	1.00	0.69	0.87	A
$GLS_{0.4}$	398.42	7.40	7.01	1.45	1.00	1.27	B
GLS_{PL}	399.14	6.66	5.90	1.14	0.79	1.00	C
ML	399.12	D	E	F	G	H	1.00

MC Mean = Mean of 1000 estimates, MC SD = standard deviation of 1000 estimates
MC Ave Est SEs = Mean of 1000 estimated standard errors
this MSE/$\text{MSE}_{GLS_{0.08}}$ = Monte Carlo mean square error for this estimator divided by that for $GLS_{0.8}$ (others similar)

[12 points] (a) Assuming that 1000 Monte Carlo data sets gives sufficient precision so that the competing entries in the above table can be distinguished from one another with high certainty, give a thorough interpretation of all the information in the table, explaining why things turned out the way they did, citing appropriate theoretical results (note that some of the MSE ratios are redundant.)

You should have made some or all of the following points:

- All the estimators appear to be consistent, whether θ is correctly specified, incorrectly specified, or estimated. This is to be expected from the theory, as all these estimators solve unbiased estimating equations.
- $\dfrac{MSE\ GLS_{0.08}}{MSE\ GLS_{0.4}} = 0.69 < 1$ – the true value of $\theta = 0.8$, so using $\theta = 0.4$ results in using the incorrect variance function. The "folklore theory" for GLS with misspecified variance model in Chapter 9 of the notes says that this should result in an inefficient estimator for $\boldsymbol{\beta}$ relative to using the correct variance model. Hence, we are seeing this borne out in this example. Interestingly , despite the small sample size, the implications of large sample theory are reflected very well here.

- $\dfrac{MSE\ \ GLS_{0.08}}{MSE\ \ GLS_{PL}} = 0.87 < 1$ – the "folklore theory" with the variance model specified correctly says that the estimator for $\boldsymbol{\beta}$ using GLS with θ estimated should be as efficient as that using GLS with θ known. However, the 2nd order results in Chapter 11 suggest that this theory is optimistic in small sample sizes like that here and that estimating θ will have an effect on efficiency. Here, we see that theory borne out – there is a degradation in performance of the GLS estimator when θ is estimated.
- $\dfrac{MSE\ \ GLS_{PL}}{MSE\ \ GLS_{0.4}} = 0.79 < 1$ – this says that estimating θ is better than setting it equal to the wrong thing! This makes sense from what we know both about the misspecified variance function "folklore" theory and the 2nd order theory. Getting θ wrong is definitely worse than getting it right, as we already noted, consistent with the "folklore" theory. Estimating θ, while entailing some loss in efficiency because of the small sample size (2nd order theory effect) at least uses the correct variance function, so is still more precise.
- Regarding estimation of standard errors, comparing the Monte Carlo SD to the average of estimated standard errors, we see that those using $\theta = 0.4$ are optimistic, as we'd expect from the folklore theory, while those using the correct $\theta = 0.8$ are pretty good (6.24 vs. 6.31 is pretty close). So it seems that the folklore theory isn't too bad here even though the sample size is only 11. When θ is estimated by PL, the estimated standard errors are again optimistic; this is probably because, as predicted by the 2nd order theory, the true sampling variation is larger than the "folklore" theory would suggest.

Here is the table again for convenience:

	MC Mean	MC SD	MC Ave Est SEs	$\frac{\text{this MSE}}{\text{MSE}_{GLS_{0.8}}}$	$\frac{\text{this MSE}}{\text{MSE}_{GLS_{0.4}}}$	$\frac{\text{this MSE}}{\text{MSE}_{GLS_{PL}}}$	$\frac{\text{this MSE}}{\text{MSE}_{ML}}$
$GLS_{0.8}$	399.30	6.24	6.31	1.00	0.69	0.87	A
$GLS_{0.4}$	398.42	7.40	7.01	1.45	1.00	1.27	B
GLS_{PL}	399.14	6.66	5.90	1.14	0.79	1.00	C
ML	399.12	D	E	F	G	H	1.00

MC Mean = Mean of 1000 estimates, MC SD = standard deviation of 1000 estimates
MC Ave Est SEs = Mean of 1000 estimated standard errors
this MSE/MSE$_{GLS_{0.08}}$ = Monte Carlo mean square error for this estimator divided by that for $GLS_{0.8}$ (others similar)

[5 points] (b) For each of the entries denoted by A–H in the table, provide your best guesses for their numerical values, and explain the reasons for your guesses.

The value of σ is pretty small here; thus, given we have to guess, a safe guess would be that the values of A-H would be similarly to the corresponding values for GLS_{PL}; that is we might conjecture that we are in a "'small σ" situation. As discussed at the end of Chapter 10 and in the solution to Problem 3(e) here , when $\sigma \to 0$, we know that ML and GLS have the same large-sample distribution under this condition. Thus, in a Monte Carlo study where we hope the implications of large sample results hold at least approximately, we would expect to see these 2 estimators to exhibit similar behavior. where, if performance differed, GLS-PL might look a little worse than ML. So, we might expect the value of A to be around 0.87 or a little smaller (and F to be around 1.14 or a little larger). we might expect B to be around 1.27 or a little larger (and G to be 0.79 or a little smaller), and C to be around 1 or a little larger (and H to be around 1 or a little smaller). For D and E, we would expect the sampling variation of GLS-PL and ML to be roughly the same, with possibly that of ML a little smaller, so we'd expect D and E to be about the same as the corresponding values for GLS-PL or a little smaller.

That's the long winded answer - the short version would be that we'd expect the values for A–H to be similar to those for GLS-PL.

Actually, the values of A–H for ML were *identical* to those for GLS_{PL} in this simulation, so the above explanation is in fact valid.