1. A study in the Netherlands followed men and women for up to 21 years. At three year intervals, participants answered questions about respiratory symptoms and smoking status. Pulmonary function was determined by forced expiratory volume in one second (FEV1) at each observation time.

   The data in this question consist of participants who did not change smoking status over the duration of the study. There are 32 former smokers and 101 current smokers for a total of 133 participants. Although it was intended for FEV1 to be recorded at all time points (baseline, every three years through year 15, and at year 19), not all individuals have FEV at every time point. Some possibly helpful descriptive statistics and plots are shown below.

Time	Former Smoker (N=32)	Current Smoker (N=101)
0	3.52 (23)	3.23 (85)
3	3.58 (27)	3.12 (95)
6	3.26 (28)	3.09 (89)
9	3.17 (30)	2.87 (85)
12	3.14 (29)	2.80 (81)
15	2.87 (24)	2.68 (73)
19	2.91 (28)	2.50 (74)

Table 1: Mean FEV1 (and sample size) by smoking status and time.

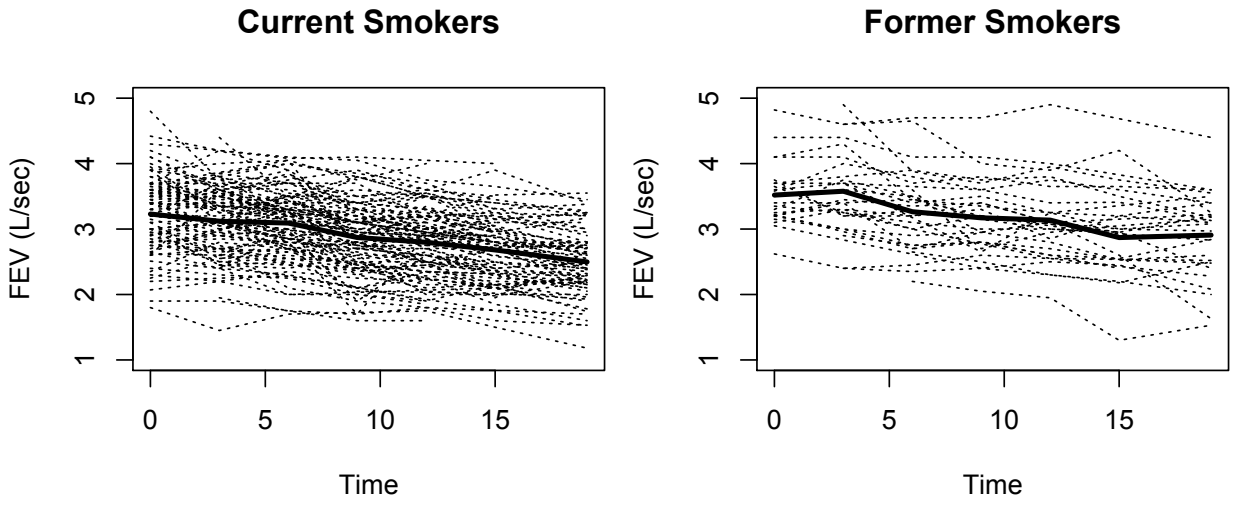

Figure 1: Mean trajectory (solid line) and individual trajectories (dashed line) by smoking status.

The results from fitting the model $E(Y_{ij}|X) = \beta_0 + \beta_1 t_{ij} + \beta_2 S_i$ using ordinary least squares are shown below. Here i indicates the subject (1 to 133), j indexes the time measurements on a specific subject, and S_i is the binary indicator of smoking status (1=current smoker).

```
. regress fev year smoke

      Source |       SS       df       MS              Number of obs =     771
-------------+------------------------------           F(  2,   768) =   87.11
       Model |  54.6251219     2  27.3125609            Prob > F      =  0.0000
    Residual |  240.805055   768  .313548249           R-squared     =  0.1849
-------------+------------------------------           Adj R-squared =  0.1828
       Total |  295.430177   770  .383675555           Root MSE      =  .55995

------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |   -.038589   .0032995   -11.70   0.000    -.0450661   -.0321119
       smoke |  -.3106707   .0469242    -6.62   0.000    -.4027857   -.2185558
       _cons |   3.562053   .0510095    69.83   0.000     3.461918    3.662188
------------------------------------------------------------------------------
```

(a) For each of the following, describe the effect on the value if the regression were run using GEE with an independent working covariance matrix. (Chose from: stay the same, increase, decrease, differ but could be in either direction, or cannot be determined). Provide a brief justification for your choice.

i. $\hat{\beta}_0$

Solution: All of these point estimates will stay the same because the equations used to solve for the estimates are the same using OLS ($\hat{\beta} = (X^T X)^{-1}(X^T Y)$) and GEE with working independence ($\hat{\beta} = (X^T W^{-1} X)^{-1}(X^T W^{-1} Y)$, $W = I$)

ii. $\hat{\beta}_1$

Solution: All point estimates stay the same.

iii. $\hat{\beta}_2$

Solution: All point estimates stay the same.

iv. Standard error of $\hat{\beta}_0$

> **Solution:** The standard errors increase for estimates of betas that must be made between subjects (effect of smoking and mean at baseline) due to "loss" of independent subjects when not assuming that all measurements are independent.

v. Standard error of $\hat{\beta}_1$

> **Solution:** The standard errors decrease for estimates of betas that can be made within subject (i.e. trend over time) due to positive correlation of measurements within individuals.

vi. Standard error of $\hat{\beta}_2$

> **Solution:** The standard errors increase for estimates of betas that must be made between subjects (effect of smoking and mean at baseline) due to "loss" of independent subjects when not assuming that all measurements are independent.

(b) A statistics student brings output from Stata and expresses concern that the results are slightly different when using different working covariance matrices in GEE. Explain to this statistics student why the results may be different using different working covariance structures. Be sure to include (a) why the point estimates may be different, (b) why the standard errors may be different, and (c) if these differences represent a substantial problem with the GEE method. You may assume that the statistics student can understand mathematical notation, but you should include written explanation as well.

> **Solution:** Estimates in GEE come from solving the equation $\hat{\beta} = (X^T W^{-1} X)^{-1}(X^T W^{-1} Y)$, where W is the working covariance matrix. Therefore, using a different form of the working covariance matrix will lead to slightly different estimates. The standard errors of $\hat{\beta}$ are given by $(X^T W^{-1} X)^{-1} X^T W^{-1}(rr^T) W^{-1} X (X^T W^{-1} X)^{-1}$, where r is the vector of residuals. Thus the standard errors will also differ based on the choice of working covariance matrix. These differences do not represent a problem with the GEE method – although choosing a working covariance structure that is close to the truth will be more efficient, all choices will result in estimates and standard errors with the correct inferential properties (e.g., unbiased estimates and nominal type I error rate).

(c) A collaborator suggests using a random effects model instead of GEE. This collaborator is satisfied with the model of a linear time effect and a fixed effect of smoking that was used above. The collaborator suggests including a random intercept and a random time effect (random slope). There are three possible models: one with random intercepts, one with independent random intercepts and slopes, and one with possibly correlated random intercepts and slopes. Which of these models is most analogous to the GEE model using working exchangeable? Explain your choice.

Solution: The model with random intercepts only is analogous to GEE with an exchangeable correlation structure; the robust standard errors do not formally assume that this model is maintained, but the exchangeable covariance structure assumes a constant correlation between measurements on the same individual, which is exactly what is assumed with the random intercept model.

(d) The STATA output from fitting the model with possibly correlated random intercepts and slopes is shown below. As alluded to previously, this model is:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 S_i + \mathrm{b}_{0i} + \mathrm{b}_{1i} t_{ij} + \epsilon_{ij}$$

As before, i indicates the subject (1 to 133), j indexes the time measurements on a specific subject, and S_i is the binary indicator of smoking status (1=current smoker).

```
.  xtmixed fev year smoke || id: year, cov(unstructured)
[some output omitted]
------------------------------------------------------------------------------
         fev |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |  -.0371543   .0015181   -24.47   0.000    -.0401298   -.0341789
       smoke |  -.3250446   .1083222    -3.00   0.003    -.5373523   -.1127369
       _cons |   3.546401    .096068    36.92   0.000     3.358111    3.734691
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Unstructured             |
                     sd(year) |    .009866   .0019306      .0067232    .0144781
                   sd(_cons) |   .5502334   .0368449      .4825569    .6274014
            corr(year,_cons) |    -.31308   .1480349     -.5687205   -.0022833
-----------------------------+------------------------------------------------
                sd(Residual) |   .2043253   .0063677      .1922183    .2171948
------------------------------------------------------------------------------
```

For each term below, (1) note the estimate provided by this model and (2) provide an interpretation suitable for a non-statistician.

	Estimate	Interpretation
$\hat{\beta}_0$	3.55	This is the estimated FEV1 at baseline for a "typical" former smoker.
$\hat{\beta}_1$	-0.037	This is the estimated change in FEV1 for each year in the study for a "typical" individual, holding smoking status constant.
$\hat{\beta}_2$	-0 .33	This is the estimated difference in mean FEV1 between current and former smokers.
$\hat{\sigma}_{\mathrm{b}_0}$	0.55	This represents a measure of the individual variability in the baseline FEV1 values.
$\hat{\sigma}_{\mathrm{b}_1}$	0.0098	This represents a measure of the individual variability in the slope.
$\hat{\rho}_{\mathrm{b}_0,\mathrm{b}_1}$	-0.31	This is the estimated correlation between the random intercepts and random slopes, indicating that individuals with higher estimated intercepts have more negative slopes (faster rate of FEV1 decline).

(e) A different collaborator suggests centering the time variable using $t^*_{ij} = t_{ij} - 9$. If the same model is run using this new "centered" time variable, describe what you know about the results based on the results of the model with the uncentered time variable and the descriptive statistics of the data. Provide justification for your response.

i. What will happen to the estimate and standard error of $\hat{\beta}_1$?

> **Solution:** The estimate of the slope and the standard error of this estimate will be unchanged due to the centering of the time values.

ii. What will happen to the estimate and standard error of $\hat{\beta}_2$?

> **Solution:** The estimate of the slope and the standard error of this estimate will be unchanged due to the centering of the time values.

iii. What will happen to the estimate and standard error of $\hat{\beta}_0$?

> **Solution:** The estimate of the intercept will be exactly $\hat{\beta}_{0old} - 9(\hat{\beta}_1)$, which is 3.212. The standard error will differ depending on variation in the data; we would expect it to be smaller if the data are truly homoscedastic.

iv. What will happen to the estimate of $\hat{\sigma}_{\mathrm{b}_0}$?

> **Solution:** The estimate of the variability of the random intercept effect may be different. If the data were truly homoscedastic, we would expect the variability of the random intercept to be slightly less after nearly centering the time values.

v. What will happen to the estimate of $\hat{\sigma}_{\mathrm{b}_1}$?

> **Solution:** The estimate of the variability of the random slopes should remain unchanged.

vi. What will happen to the estimate of $\hat{\rho}_{\mathrm{b}_0,\mathrm{b}_1}$?

> **Solution:** The correlation between the random effects will likely decrease in absolute value due to the intercept at $t = 9$. With the intercept at $t = 0$, we saw that the random intercept and slope were negatively correlated, indicating that higher that average values at $t = 0$ lead to steeper slopes (negative values of b_1). At $t = 9$, we expect this correlation to be less strong.

(f) Finally, not all participants completed the lung function assessment at each scheduled time point. A collaborator turns to you and asks, "is this a problem?" Briefly discuss the missing data mechanism assumed in the following two approaches. Use

language appropriate for an scientific collaborator.

i. A GEE approach?

Solution: The GEE approach assumes that the missing data mechanism is missing completely at random; in this scenario this might be because the machine was broken on random days. However, if the data were missing due to smoking status or rate of lung function decline (i.e., not missing completely at random), the GEE approach could be problematic. In such a scenario, estimates may be biased due to the unobserved data.

ii. A random effects model?

Solution: The random effects model assumes that the missing data mechanism is missing at random; in this scenario the missing mechanism might be related to smoking status (for instance) and the random effects model would be valid. However, if the data were missing based on unobserved factors (i.e., missing not at random), this would pose a problem similar to that described using the GEE approach.

Division of Biostatistics Q2 Exam
Day 1 Methods and Applications
Survival Question

Consider an independent censorship model in which failure time T is exponential with hazard rate θ and censoring time C follows an arbitrary nonparametric distribution free of θ. The observed data are $\{(x_i, \delta_i), i = 1, \ldots, n\}$ where $x_i = \min(t_i, c_i)$ and $\delta_i = I(t_i \leq c_i)$.

1. Write out the full likelihood function L_f. Obtain the MLE, say $\hat{\theta}_1$, of θ from the function L_f based on the observed data.

   The full likelihood function on the basis of $(x_1, \delta_1), \ldots, (x_n, \delta_n)$ is

$$\begin{aligned} \mathcal{L}_f &= \prod_{i=1}^{n} \left\{ [f(x_i;\theta) S_C(x_i)]^{\delta_i} [f_C(x_i) S(x_i;\theta)]^{1-\delta_i} \right\} \\ &\propto \prod_{i=1}^{n} f(x_i;\theta)^{\delta_i} S(x_i;\theta)^{1-\delta_i} = \prod_{i=1}^{n} (\theta e^{-\theta x_i})^{\delta_i} (e^{-\theta x_i})^{1-\delta_i} \end{aligned}$$

   Score function is $l^{'}(\theta) = \sum_i (\delta_i/\theta - x_i)$
   MLE of θ is $\hat{\theta}_1 = \frac{\sum_i \delta_i}{\sum_i x_i}$.

2. What is the asymptotic distribution of $\hat{\theta}_1$? Provide an estimator of the asymptotic variance of $\hat{\theta}_1$.

   We have $l^{''}(\theta) = -\sum_i^n \delta_i/\theta^2$, by the asymptotic normality of MLE,

$$\sqrt{n}(\hat{\theta}_1 - \theta) \rightarrow_D \mathcal{N}(0, \frac{\theta^2}{\sum_i \delta_i})$$

3. What is the role of the distribution of C_i in the maximum likelihood derivation of $\hat{\theta}_1$, and the asymptotic distribution of $\hat{\theta}_1$? Explain with details.

   Under independent censoring, both the likelihood function used to derive $\hat{\theta}_1$ and the asymptotic distribution of $\hat{\theta}_1$ does not involve C distribution. It is nuisance and does not contribute to the estimation and inference of parameter θ in T distribution.

   In the following, assume that censoring time C also follows an exponential distribution with hazard rate λ, and $\lambda = \theta$.

4. Obtain the MLE, say $\hat{\theta}_2$, of θ based on the observed data.

   The full likelihood function becomes

$$\begin{aligned}\mathcal{L}_f &= \prod_{i=1}^{n}\{\theta e^{-\theta x_i}e^{-\lambda x_i}\}^{\delta_i}(\lambda e^{-\lambda x_i}e^{-\theta x_i})^{1-\delta_i} \\ &= \prod_{i=1}^{n}\theta e^{-2\theta x_i}\end{aligned}$$

   Score function is $l_f^{'} = \sum_i(1/\theta - 2x_i)$
   MLE of θ is $\hat{\theta}_2 = \frac{n}{2\sum_i x_i}$.

5. What is the asymptotic distribution of $\hat{\theta}_2$? Be explicit with the distributional parameters.

   $l_f^{''} = -\frac{n}{\theta^2}$, so we have

$$\sqrt{n}(\hat{\theta}_2 - \theta) \rightarrow_D \mathcal{N}(0, \frac{\theta^2}{n})$$

   ($X = \min(T, C) \sim exp(2\theta)$, so $\sum_i x_i/n \rightarrow_p 1/2\theta$ and $\hat{\theta}_2 = \frac{n}{2\sum_i x_i} \rightarrow_p \theta$)

6. In this case, does the distribution of C_i provide information in estimation of θ? Compare the asymptotic variances of $\hat{\theta}_1$ and $\hat{\theta}_2$, and discuss the results (less than 50 words).

   The distribution of C contributes in the likelihood derivation and does provide information in estimation of θ.

   The ratio of the two asymptotic variance is

$$\frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)} = \frac{\sum_i \delta_i}{n}$$

   From Q4, we know $\{\delta_i, i = 1, \ldots, n\}$ are i.i.d. random variables with Bernoulli ($\frac{\theta}{\theta+\lambda} = 1/2$ as $\lambda = \theta$), thus $\frac{\sum_i \delta_i}{n} \rightarrow 1/2$ when $n \rightarrow \infty$.

   The asymptotic variance of $\hat{\theta}_2$ is half of that of $\hat{\theta}_1$ in large sample. It is reasonable because additional information on censoring time C is used in the estimation and inference of θ by $\hat{\theta}_2$.

Solution to Qual 2 Theory Question #3
2011 September Exam

We assume at least one x_i is nonzero.

(a) The log likelihood function is given by

$$\log(L) = n(\log\sqrt{2\pi}) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2,$$

and

$$\begin{aligned}
\frac{\partial \log L}{\partial \beta} &= \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - \beta x_i) \\
\frac{\partial \log L}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \beta x_i)^2 \\
\frac{\partial^2 \log L}{\partial \beta^2} &= -\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i^2 \\
\frac{\partial^2 \log L}{\partial \sigma^2} &= \frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{n}(y_i - \beta x_i)^2 \\
\frac{\partial^2 \log L}{\partial \sigma \partial \beta} &= -\frac{2}{\sigma^3}\sum_{i=1}^{n} x_i(y_i - \beta x_i)
\end{aligned}$$

The solution of the first two equations is given by:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2},$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta} x_i)^2.$$

This solution is unique and at $(\hat{\beta}, \hat{\sigma})$ the 2×2 matrix of second order derivatives is negative definite (at least one of the diagonal entries is negative and the determinant is positive) whenever at least one x_i is nonzero.

(b) Since $\hat{\beta}$ is a linear function of independent normal random variables, it is again normally distributed with mean β and variance $\sigma^2/\{\sum_{i=1}^{n} x_i^2\}$.

(c) Variance of $\hat{\beta}$ is the smallest whenever the sum of the x_i^2 is the largest. This is possible when the x_i are either $+1$ or -1. So, there are 2^{10} different possible choices for the x_i that lead to the smallest variance for $\hat{\beta}$.

(d) Likelihood ratio test statistic is given by

$$\Lambda = \frac{\sup_{\beta=1} L(\beta)}{L(\hat{\beta})}$$

and upon simplification, we get

$$\Lambda = \exp\{-\frac{1}{2}(\sum_{i=1}^{n} x_i^2)(\hat{\beta}-1)\}^2$$

when $\sigma = 1$ or

$$-2\log(\Lambda) = (\sum_{i=1}^{n} x_i^2)(\hat{\beta}-1)^2.$$

[Not asked: Clearly, when H_0 is true, this statistic has a χ^2 distribution with 1 degree of freedom since

$$Var(\hat{\beta}) = 1/(\sum_{i=1}^{n} x_i^2),$$

and $\hat{\beta}$ is normally distributed with mean β].

(e) We reject H_0 when $\hat{\beta}$ is outside the interval $(1-(1.96/10), 1+(1.96/10))$ or outside $(0.804, 1.196)$. [One can also describe the critical region in terms of $\chi^2(1)$ critical value.]

(f) No because in Model 1, the alternative hypothesis corresponds to the slope alone while the intercept is assumed to be held at 0 whereas in the test for 0 mean for the difference $Y_i - x_i$ corresponds to a test for 0 intercept where the slope parameter β is held at 1.

2011 QII - Question 4

(a) $E(X) = \int_0^1 x\,\theta(\theta+1)x^{\theta-1}(1-x)\,dx$

$= \theta(\theta+1)\int_0^1 x^{\theta}(1-x)\,dx = \theta(\theta+1)\int_0^1 x^{(\theta+1)-1}(1-x)^{2-1}\,dx$

$= \theta(\theta+1)\frac{\Gamma(\theta+1)\Gamma(2)}{\Gamma(\theta+3)}\underbrace{\int_0^1 \frac{\Gamma(\theta+3)}{\Gamma(\theta+1)\Gamma(2)}x^{(\theta+1)-1}(1-x)^{2-1}\,dx}_{=1}$

$= \frac{\theta(\theta+1)\Gamma(\theta+1)}{(\theta+2)(\theta+1)\Gamma(\theta+1)} = \frac{\theta}{\theta+2}$

MOM estimator solves: $\bar{X} = \frac{\hat{\theta}}{\hat{\theta}+2} \Rightarrow \hat{\theta} = \frac{2\bar{X}}{1-\bar{X}}$

(b) $E(X^2) = \int_0^1 x^2\,\theta(\theta+1)x^{\theta-1}(1-x)\,dx$

$= \theta(\theta+1)\int_0^1 x^{(\theta+2)-1}(1-x)^{2-1}\,dx = \theta(\theta+1)\frac{\Gamma(\theta+2)\Gamma(2)}{\Gamma(\theta+4)}$

$= \frac{\theta(\theta+1)\Gamma(\theta+2)}{(\theta+3)(\theta+2)\Gamma(\theta+2)} = \frac{\theta(\theta+1)}{(\theta+3)(\theta+2)}$

$V(X) = E(X^2) - E(X)^2 = \frac{\theta(\theta+1)}{(\theta+3)(\theta+2)} - \left(\frac{\theta}{\theta+2}\right)^2 = \frac{2\theta}{(\theta+2)^2(\theta+3)}$

by CLT, $\sqrt{n}\left(\bar{X} - \frac{\theta}{\theta+2}\right) \Rightarrow N\left(0, \frac{2\theta}{(\theta+2)^2(\theta+3)}\right)$

Let $f(x) = \frac{2x}{1-x}$ then $f'(x) = \frac{2}{1-x} + \frac{2x}{(1-x)^2} = \frac{2}{(1-x)^2}$

Thus $f'\left(\frac{\theta}{\theta+2}\right) = \frac{2}{\left(1-\frac{\theta}{\theta+2}\right)^2} = \frac{2(\theta+2)^2}{4} = \frac{(\theta+2)^2}{2}$

Delta method gives:

$\sqrt{n}\left(f(\bar{X}) - f\left(\frac{\theta}{\theta+2}\right)\right) \Rightarrow N\left(0, \underbrace{\left[f'\left(\frac{\theta}{\theta+2}\right)\right]^2 \cdot V(X)}\right)$

$= \frac{(\theta+2)^4}{4}\cdot\frac{2\theta}{(\theta+2)^2(\theta+3)} = \frac{\theta(\theta+2)^2}{2(\theta+3)}$

i.e., $\sqrt{n}(T_n - \theta) \Rightarrow N\left(0, \frac{\theta(\theta+2)^2}{2(\theta+3)}\right)$

(5) (c) $L(\theta|x_1,\dots,x_n) = \prod_{i=1}^{n} \theta(\theta+1)\, x_i^{(\theta-1)}(1-x_i)$

$$= \theta^n(\theta+1)^n\left(\prod_{i=1}^{n} x_i\right)^{\theta-1}\prod_{i=1}^{n}(1-x_i)$$

$$\log L(\theta|x_1,\dots,x_n) = n\log\theta + n\log(\theta+1) + (\theta-1)\Sigma\log x_i + \Sigma\log(1-x_i)$$

$$\frac{\partial \log L}{\partial\theta} = \frac{n}{\theta} + \frac{n}{\theta+1} + \Sigma\log x_i = 0 \qquad W_n = -\frac{1}{n}\Sigma\log x_i$$

$$\frac{n}{\theta} + \frac{n}{\theta+1} - nW_n = 0$$

$$W_n = \frac{1}{\theta} + \frac{1}{\theta+1} = \frac{2\theta+1}{\theta(\theta+1)} = \frac{2\theta+1}{\theta^2+\theta}$$

$$W_n\theta^2 + W_n\theta - 2\theta - 1 = 0$$

$$W_n\theta^2 + (W_n-2)\theta - 1 = 0$$

$$\hat{\theta} = \frac{-(W_n-2) \pm \sqrt{(W_n-2)^2 - 4(W_n)(-1)}}{2W_n}$$

$$= \frac{2-W_n + \sqrt{W_n^2+4}}{2W_n}$$

$$= \frac{1}{W_n} - \frac{1}{2} + \sqrt{\frac{W_n^2+4}{4W_n^2}} = \frac{1}{W_n} - \frac{1}{2} + \sqrt{\frac{1}{4} + \frac{1}{W_n^2}}$$

Check 2nd deriv:

$$\frac{\partial^2 \log L}{\partial\theta^2} = \frac{-n}{\theta^2} - \frac{n}{(\theta+1)^2} < 0 \Rightarrow \hat{\theta} \text{ is the MLE}$$

(d) Since $E(\log X) = \frac{-2\theta+1}{\theta(\theta+1)}$, By WLLN: $W_n \xrightarrow{P} \frac{2\theta+1}{\theta(\theta+1)}$

Thus $\hat{\theta} = \frac{1}{W_n} - \frac{1}{2} + \sqrt{\frac{1}{4} + \frac{1}{W_n^2}} \xrightarrow{P} \frac{\theta(\theta+1)}{2\theta+1} - \frac{1}{2} + \sqrt{\frac{1}{4} + \frac{\theta^2(\theta+1)^2}{(2\theta+1)^2}}$

After a bit of algebra... $\xrightarrow{P} \theta$

Thus $\hat{\theta}$ is consistent for θ.