# Homework #2

Professor: Pat Kline

Students: Christina Brown, Sam Leone, Peter McCrory, Preston Mui

## 2. Logit MLE

a) Derive the score of the logit-likelihood: First, noting that

$$\frac{\partial \Lambda(X_i'\beta)}{\partial \beta} = \frac{\exp(X_i'\beta)X_i'(1+X_i'\beta) - \exp(X_i'\beta)\exp(X_i'\beta)X_i'}{(1+\exp(X_i'\beta))^2} = \frac{\exp(X_i'\beta)X_i'}{(1+\exp(X_i'\beta))^2}$$

The score of the logit log-likelihood is therefore

$$\begin{aligned} s(\beta) &= \sum_i \frac{Y_i \exp(X_i'\beta)X_i'}{\Lambda(X_i'\beta)(1+\exp(X_i'\beta))^2} - \frac{(1-Y_i)\exp(X_i'\beta)X_i'}{(1-\Lambda(X_i'\beta))(1+\exp(X_i'\beta))^2} \\ &= \sum_i \Big(Y_i - (1-Y_i)\exp(X_i'\beta)\Big)\frac{X_i'}{1+\exp(X_i'\beta)} \\ &= \sum_i \Big(Y_i(1+\exp(X_i'\beta)) - \exp(X_i'\beta)\Big)\frac{X_i'}{1+\exp(X_i'\beta)} \\ &= \sum_i \Big(Y_i - \frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}\Big)X_i' \end{aligned}$$

b) The moment condition identifying $\beta_{ML}$ is

$$E\bigg[\Big(Y_i - \frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}\Big)X_i'\bigg] = 0$$
$$E\bigg[E[Y_i - \frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}|X_i]X_i'\bigg] = 0$$

c) The moment conditions identifying $\beta_{NLLS}$ are

$$\begin{aligned} 0 &= E\bigg[-2(Y_i - \Lambda(X_i'\beta))\frac{\exp(X_i'\beta)}{(1+\exp(X_i'\beta))^2}X_i'\bigg] \\ &= E\bigg[E[(Y - \Lambda(X_i'\beta))|X_i]\frac{\exp(X_i'\beta)}{(1+\exp(X_i'\beta))^2}X_i'\bigg] \end{aligned}$$

d) Under which conditions does $\beta_{NLLS}$ coincide with $\beta_{ML}$? The moment conditions will coincide when $E[Y_i - \Lambda(X_i'\beta)|X_i'] = 0$ for all $X_i$; that is, they coincide when the logit model is correctly specified.

## 3. Matlab Probit DGP

a) (Matlab program attached)

b) The ML point estimates of $\hat{\beta}^{con} = (\hat{\alpha}, \hat{b})$ are $(-0.0175, 0.9159)$. The standard errors are $(0.3030, 0.3607)$, respectively.

c) Score test: Following Wooldridge (2010), page 570, the $LM$ statistic is the ESS from the following regression

$$\frac{\hat{u}_i}{\sqrt{\hat{G}_i(1-\hat{G}_i)}} = \alpha \frac{\hat{g}_i}{\sqrt{\hat{G}_i(1-\hat{G}_i)}} x_i + \gamma \frac{\hat{g}_i}{\sqrt{\hat{G}_i(1-\hat{G}_i)}} z_i$$

where $x_i$ is the regressor matrix in the unconstrained regression (constant and $X$) and $z_i$ is the vector of $X_i^2$. The ESS from this regression was 0.2963, which has a p-value of 0.58621. So, one does not reject the null that $b_2 = 0$.

d) The unrestricted model yields point estimates of $\hat{\beta}^{unc} = (\hat{\alpha}, \hat{b}, \widehat{b_2}) = (-0.0435, 0.9175, 0.0428)$. The Wald test will test the null $g(\beta) = 0$ where

$$g(\beta) \equiv (0,0,1) \cdot \beta = b_2$$
$$G(\beta) \equiv \frac{\partial g(\beta)}{\partial \beta} = (0,0,1)$$

and the Wald statistic is given by

$$N \cdot \widehat{b_2} \cdot \left( G \cdot \frac{1}{\sqrt{N}} H^{-1} \cdot G' \right)^{-1} \cdot \widehat{b_2}$$

where $H$ is the average Hessian at the ML estimate. Because we know that the model is correctly specified, I use the inverse Hessian instead of the sandwich estimator. The Wald evaluates to 7.4592, which has a p-value of 0.0063 under the $\chi^2$ distribution with d.f. 1, so one rejects the null that $b_2 = 0$. This is starkly different from the score test result, which did not reject the null. This makes sense, as the Wald tends to reject more than the LM test. If one bumps the number of observations up (say, to 5000), both tests reject the null.

## 4. Clustered DGP in Stata

a) Regressing $Y_{ic}$ on $D_{ic}$, we get the results in table 1.

The basic SE are 0.03237. The standard errors are slightly larger (0.03244) when we correct for heteroskedasticity in column 2 and significantly larger (0.09508) when we account for the intracluster correlation of observations from the same cluster.

b) Changing $\sigma_\eta^2$ to 0. Regressing $Y_{ic}$ on $D_{ic}$, we get the results in table :

The standard errors are fairly similar for columns 1 and 2 to what we had in A. They are 0.02732 for the regular SE, 0.027731 for robust SE. The clustered SE (0.02420) are similar to the magnitude of the non-clustered version.

Table 1: Clustered DGP

| | (1)<br>Y_ic, Regular SE | (2)<br>Y_ic, Robust SE | (3)<br>Y_ic, Clustered SE |
|---|---|---|---|
| d_ic | 1.010*** | 1.010*** | 1.010*** |
| | (0.03237) | (0.03244) | (0.09508) |
| Constant | 0.0394* | 0.0394* | 0.0394 |
| | (0.02278) | (0.02037) | (0.10761) |
| $R^2$ | 0.089 | 0.089 | 0.089 |
| Observations | 10000 | 10000 | 10000 |

Standard errors in parentheses

* p<0.10, ** p<0.05, *** p<0.01

Table 2: Clustered DGP

| | (1)<br>Y_ic, Regular SE | (2)<br>Y_ic, Robust SE | (3)<br>Y_ic, Clustered SE |
|---|---|---|---|
| d_ic | 0.992*** | 0.992*** | 0.992*** |
| | (0.02732) | (0.02731) | (0.02420) |
| Constant | -0.00105 | -0.00105 | -0.00105 |
| | (0.01911) | (0.01923) | (0.09650) |
| $R^2$ | 0.117 | 0.117 | 0.117 |
| Observations | 10000 | 10000 | 10000 |

Standard errors in parentheses

* p<0.10, ** p<0.05, *** p<0.01

c) Comment on your differences in the answers to a) and b).
The SE are fairly similar in columns 1 and 2 for parts a) and b) though they are slightly smaller in part b as a result of the lower variance of $y_{ic}$ in part b. In part b since there is no intra-cluster correlation in the coefficient on $d_{ic}$ the clustered standard errors are of a similar magnitude to the non-clustered version (0.02420).

d) In table 3, we see that the simulation rejects the null that the coefficient on $d_{ic}$=1 57% of the time when we don't cluster our standard errors and about 5% (which we would expect) when we cluster. So the non-clustered version is rejecting too often and the clustered version is rejecting at about the rate of p which we set.

Table 3: Monte Carlo Simulations

| | (1) | | | | |
|---|---|---|---|---|---|
| | count | mean | sd | min | max |
| reject | 1000 | .567 | .4957386 | 0 | 1 |
| reject_cluster | 1000 | .055 | .2280943 | 0 | 1 |

e) Here we have that the regression with non-clustered standard errors rejects the test that the coefficient equals the average of the cluster level $\eta$'s 3% of the time and the clustered standard errors are never rejecting, so in this case clustering is causing us not reject often enough.

Table 4: Monte Carlo Simulations

| | (1) | | | | |
|---|---|---|---|---|---|
| | count | mean | sd | min | max |
| reject | 1000 | .03 | .1706726 | 0 | 1 |
| reject_cluster | 1000 | 0 | 0 | 0 | 0 |

f) When treatment is at the cluster level like in the case of the set up for part d then clustering standard errors will produce the accurate standard errors. However, in e, when the treatment is at the individual level but those individuals are within a cluster, for example a school or village then clustering the errors will actually cause our standard errors to be larger than we would want, leading to us to not reject enough. For the purpose of the internal validity of the study, using non-clustered standard errors when treatment is at the individual level is appropriate.

## 5. CPS and WLS

a) See results in column 1 of table 5.

Table 5: Wage Regression (Weighted)

| | (1)<br>Log wage | (2)<br>Log wage (weighted by cell size) | (3)<br>Log wage (weighted by 1/cell var) |
|---|---|---|---|
| agecat==2 | 0.314*** | 0.314*** | 0.314*** |
| | (0.00746) | (0.00073) | (0.00047) |
| agecat==3 | 0.458*** | 0.458*** | 0.458*** |
| | (0.00731) | (0.00071) | (0.00046) |
| agecat==4 | 0.494*** | 0.494*** | 0.494*** |
| | (0.00771) | (0.00075) | (0.00048) |
| agecat==5 | 0.472*** | 0.472*** | 0.472*** |
| | (0.00966) | (0.00094) | (0.00061) |
| agecat==6 | 0.295*** | 0.295*** | 0.295*** |
| | (0.01651) | (0.00161) | (0.00104) |
| educcat==2 | 0.265*** | 0.265*** | 0.265*** |
| | (0.00754) | (0.00073) | (0.00047) |
| educcat==3 | 0.415*** | 0.415*** | 0.415*** |
| | (0.00764) | (0.00074) | (0.00048) |
| educcat==4 | 0.771*** | 0.771*** | 0.771*** |
| | (0.00783) | (0.00076) | (0.00049) |
| Sex | -0.261*** | -0.261*** | -0.261*** |
| | (0.00458) | (0.00045) | (0.00029) |
| Constant | 2.153*** | 2.153*** | 2.153*** |
| | (0.00995) | (0.00097) | (0.00063) |
| $R^2$ | 0.297 | 0.978 | 0.978 |
| Observations | 56182 | 56182 | 135056 |

Standard errors in parentheses

* p<0.10, ** p<0.05, *** p<0.01

b) Collapsed results used in analysis below.

c) See column 2 of table 5. The coefficients are identical (as we would expect) but the standard errors are much smaller when we weight by bin size.

d) If wages are *iid* within a cell, then the variance of each cell is $\sigma^2/N_c$ where $\sigma^2$ is the population variance and $N_c$ is the number of observations in the cell.

e) Formula above is used to calculate the variance of the cell means.

f) See results in column 3 of table 5.

g) With an RSS of 0.133 and 38 degrees of freedom we are not able to reject this model at the 5% level. In other words we cannot reject our model explains the data up to a sampling error.

h) See results in table 7.

i) With an RSS of 0.009 and 15 degrees of freedom we cannot reject this goodness of fit of this model at the 5% level either.

j) See fig. 1. The model does a very nice job of fitting the data. The place it has a bit of trouble fitting is for older more educated individuals. The model overestimates the wages of 65+ males with some college and underestimates the wages for males with a BA or more.

k) A simple model which passes the chi-squared test is just lnwage on sex. We cannot reject that this model explains all of the variation in lnwage up to a sampling error.

Table 6: Wage Regression (Weighted)

| | (1) |
|---|---|
| | Log wage |
| agecat==2 | 0.301*** |
| | (0.00103) |
| agecat==3 | 0.414*** |
| | (0.00103) |
| agecat==4 | 0.424*** |
| | (0.00103) |
| agecat==5 | 0.407*** |
| | (0.00106) |
| agecat==6 | 0.355*** |
| | (0.00068) |
| educcat==2 | 0.109*** |
| | (0.00076) |
| educcat==3 | 0.238*** |
| | (0.00083) |
| educcat==4 | 0.757*** |
| | (0.00081) |
| Sex | -0.192*** |
| | (0.00054) |
| <=25 × Dropout | 0 |
| | (.) |
| <=25 × HS | 0.0576*** |
| | (0.00078) |
| <=25 × Some College | 0.00573*** |
| | (0.00084) |
| <=25 × BA+ | -0.135*** |
| | (0.00087) |
| 25-35 × Dropout | 0.00900*** |
| | (0.00084) |
| 25-35 × HS | 0.135*** |
| | (0.00071) |
| 25-35 × Some College | 0.152*** |
| | (0.00078) |
| 25-35 × BA+ | 0 |
| | (.) |
| 35-45 × Dropout | -0.0677*** |
| | (0.00084) |
| 35-45 × HS | 0.0784*** |
| | (0.00070) |
| 35-45 × Some College | 0.154*** |
| | (0.00078) |
| 35-45 × BA+ | 0 |
| | (.) |
| 45-55 × Dropout | -0.0676*** |
| | (0.00086) |
| 45-55 × HS | 0.112*** |
| | (0.00071) |
| 45-55 × Some College | 0.158*** |
| | (0.00079) |
| 45-55 × BA+ | 0 |
| | (.) |
| 55-65 × Dropout | -0.00166* |
| | (0.00090) |
| 55-65 × HS | 0.0884*** |
| | (0.00076) |
| $R^2$ | 0.998 |
| Observations | 135056 |

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01

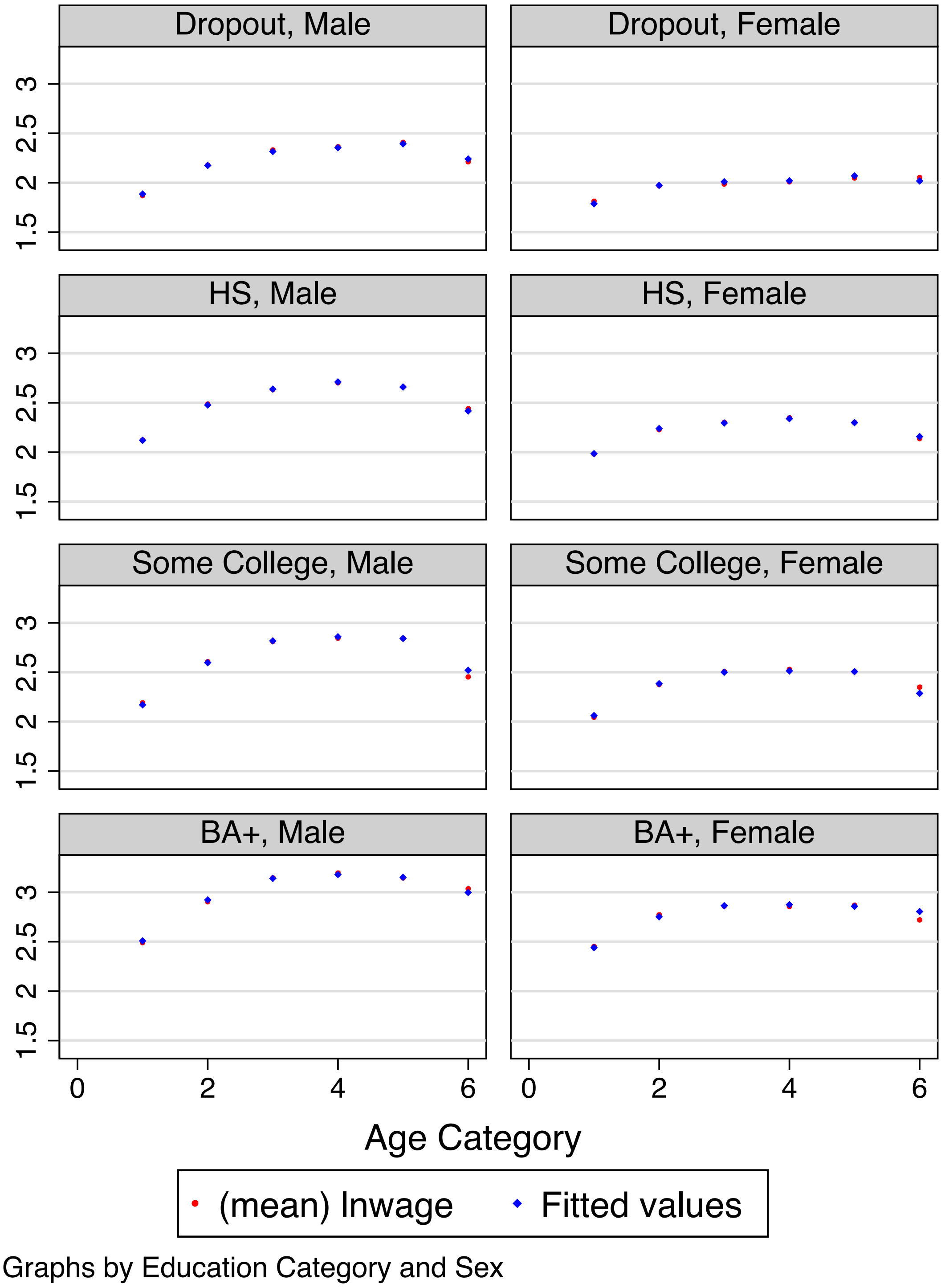
Table 7: Wage Regression (Weighted) cont

| | (1) |
|---|---|
| | Log wage |
| 55-65 × Some College | 0.167*** |
| | (0.00084) |
| 55-65 × BA+ | 0 |
| | (.) |
| >65 × Dropout | 0 |
| | (.) |
| >65 × HS | 0 |
| | (.) |
| >65 × Some College | 0 |
| | (.) |
| >65 × BA+ | 0 |
| | (.) |
| <=25 × Male | 0 |
| | (.) |
| <=25 × Female | 0.124*** |
| | (0.00055) |
| 25-35 × Male | -0.0212*** |
| | (0.00054) |
| 25-35 × Female | 0 |
| | (.) |
| 35-45 × Male | 0.0836*** |
| | (0.00054) |
| 35-45 × Female | 0 |
| | (.) |
| 45-55 × Male | 0.112*** |
| | (0.00054) |
| 45-55 × Female | 0 |
| | (.) |
| 55-65 × Male | 0.101*** |
| | (0.00058) |
| 55-65 × Female | 0 |
| | (.) |
| >65 × Male | 0 |
| | (.) |
| >65 × Female | 0 |
| | (.) |
| Dropout × Male | 0 |
| | (.) |
| Dropout × Female | -0.0302*** |
| | (0.00026) |
| HS × Male | 0.0672*** |
| | (0.00020) |
| HS × Female | 0 |
| | (.) |
| Some College × Male | 0.0410*** |
| | (0.00020) |
| Some College × Female | 0 |
| | (.) |
| BA+ × Male | 0 |
| | (.) |
| BA+ × Female | 0 |
| | (.) |
| Constant | 2.078*** |
| | (0.00057) |
| $R^2$ | 0.998 |
| Observations | 135056 |

Standard errors in parentheses

* p<0.10, ** p<0.05, *** p<0.01

Figure 1: Predicted and actual log wages by subgroup



## 6. Random Coefficient Binary Choice Model:

a) As was derived in the lecture notes, the simulated log likelihood can be derived as follows. Denote the Logistic cdf with $\Lambda$ and its pdf with $\lambda$.

First observe that

$$Pr(Y_i = 1|X_i, b_i) = Pr(\epsilon_i > -b_i x_i) = 1 - Pr(\epsilon_i < -b_i x_i) = 1 - \Lambda(-b_i x_i) = \Lambda(b_i x_i)$$

Thus, assuming that $b_i$ and $\epsilon_i$ are independent:

$$Pr(Y_i = 1|X_i) = \int \Lambda(bx_i)\frac{1}{\sigma}\phi(\frac{b_i - \mu}{\sigma})db$$

The likelihood of this model for a single observation is

$$L(Y_i, X_i; \mu, \sigma) = \left(\int \Lambda(bx_i)\frac{1}{\sigma}\phi(\frac{b_i - \mu}{\sigma})db\right)^{Y_i} \left(1 - \int \Lambda(bx_i)\frac{1}{\sigma}\phi(\frac{b_i - \mu}{\sigma})db\right)^{1-Y_i}$$

Thus, for **fixed** random draws $\{u_i\}_{m=1}^{M}$ from $\mathcal{N}(0,1)$, the simulated likelihood is

$$\hat{L}_M(Y_i, X_i, \mu, \sigma) = \left(\frac{1}{M}\sum_{m=1}^{M} \Lambda(\underbrace{(\mu + \sigma u_{im})}_{\equiv b} X_i)\right)^{Y_i} \left(1 - \frac{1}{M}\sum_{m=1}^{M} \Lambda(\underbrace{(\mu + \sigma u_{im})}_{\equiv b} X_i)\right)^{1-Y_i}$$

Clearly, taking logs (of the product of all the likelihoods) allows us to recover the simulated log-likelihood of the full sample.

Let's differentiate with respect to $\mu$ and $\sigma$!

For $\mu$:

$$\frac{\partial}{\partial \mu}\frac{1}{N}\sum_{i}^{N} Y_i log\left(\frac{1}{M}\sum_{m=1}^{M}\Lambda((\mu + \sigma u_{im})X_i)\right) + (1 - Y_i)log\left(1 - \frac{1}{M}\sum_{m=1}^{M}\Lambda((\mu + \sigma u_{im})X_i)\right)$$
$$= \frac{1}{N}\sum_{i}^{N}\left[\frac{Y_i X_i \frac{1}{M}\sum_{m=1}^{M}\lambda((\mu + \sigma u_{im})X_i)}{\frac{1}{M}\sum_{m=1}^{M}\Lambda((\mu + \sigma u_{im})X_i)} - \frac{(1 - Y_i)X_i \frac{1}{M}\sum_{m=1}^{M}\lambda((\mu + \sigma u_{im})X_i)}{1 - \frac{1}{M}\sum_{m=1}^{M}\Lambda((\mu + \sigma u_{im})X_i)}\right]$$

For $\sigma$:

$$\frac{\partial}{\partial \sigma}\frac{1}{N}\sum_{i}^{N} Y_i log\left(\frac{1}{M}\sum_{m=1}^{M}\Lambda((\mu + \sigma u_{im})X_i)\right) + (1 - Y_i)log\left(1 - \frac{1}{M}\sum_{m=1}^{M}\Lambda((\mu + \sigma u_{im})X_i)\right)$$
$$= \frac{1}{N}\sum_{i}^{N}\left[\frac{Y_i X_i \frac{1}{M}\sum_{m=1}^{M}\lambda((\mu + \sigma u_{im})X_i)u_{im}}{\frac{1}{M}\sum_{m=1}^{M}\Lambda((\mu + \sigma u_{im})X_i)} - \frac{(1 - Y_i)X_i \frac{1}{M}\sum_{m=1}^{M}\lambda((\mu + \sigma u_{im})X_i)u_{im}}{1 - \frac{1}{M}\sum_{m=1}^{M}\Lambda((\mu + \sigma u_{im})X_i)}\right]$$

b) See *max_ sml.m* file.

c) See *max_ sml.m* file.

d) These results are invariant to initial conditions:

$$\begin{bmatrix} \mu \\ ln(\sigma) \end{bmatrix} = \begin{bmatrix} 1.6011 \\ 0.5139 \end{bmatrix}$$

e) Let $\hat{\theta}_{MSL}$ denote the estimated parameters from the method of simulated likelihood and the true parameters $\theta = [\mu, \sigma]'$. Under proper specification, we know that

$$\sqrt{N}(\hat{\theta}_{MSL} - \theta) \rightarrow \mathcal{N}(\mathbf{0}, \mathbf{H}(\theta)^{-1})$$

By the delta-method, with $g(\theta) \equiv [\mu ln(\sigma)]'$, we have that

$$\sqrt{N}(g(\hat{\theta}_{MSL}) - g(\theta)) \rightarrow \mathcal{N}(\mathbf{0}, \mathbf{G}\mathbf{H}(\theta)^{-1}\mathbf{G})$$

where

$$\mathbf{G} \equiv \frac{\partial g(\theta)}{\partial \theta} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{\sigma} \end{bmatrix}$$

When we calculate the square root of the diagonal elements of $\frac{1}{N}\mathbf{H}(\hat{\theta})^{-1}$, we derive the standard errors of our estimates of $\mu$ and $\ln(\sigma)$: $(0.2633, 0.3884)$.

f) To calculate the standard errors for the incorrect model, we replace $\mathbf{H}(\theta)^{-1}$ with $\mathbf{H}(\theta)^{-1}\mathbf{V_s}\mathbf{H}(\theta)^{-1}$, where $\mathbf{V_s}$ is the population variance-covariance matrix of the scores—defined by the the criterion function we are maximizing.

Again, we use the analogy principle to replace the asymptotic variance terms with the simulated likelihood, sample averages. This yields standard errors of $\mu$ and $ln(\sigma)$: $(0.4839, 0.6417)$.