P218 Problem Set 2

Will Hotten

Code for this problem set can be found at the following link.

Question 1

1.a.

Clearly for $T = 1$:

$$f_{x_1|x_0}(x_1 \mid x_0) = f_{x_1|x_0}(x_1 \mid x_0)$$

For $T = 2$:

$$\begin{aligned} f_{x_2,x_1|x_0}(x_2, x_1 \mid x_0) &= \frac{f_{x_2,x_1,x_0}(x_2, x_1, x_0)}{f_{x_o}(x_0)} \\ &= \frac{f_{x_2|x_1,x_0}(x_2 \mid x_1, x_0) \cdot f_{x_1,x_0}(x_1, x_0)}{f_{x_0}(x_0)} \\ &= f_{x_2|x_1,x_0}(x_2 \mid x_1, x_0) \cdot f_{x_1|x_0}(x_1 \mid x_0) \end{aligned}$$

For $T = 3$:

$$\begin{aligned} f_{x_3,x_2,x_1|x_0}(x_3, x_2, x_1 \mid x_0) &= \frac{f_{x_3,x_2,x_1,x_0}(x_3, x_2, x_1, x_0)}{f_{x_o}(x_0)} \\ &= \frac{f_{x_3|x_2,x_1,x_0}(x_3 \mid x_3, x_1, x_0) \cdot f_{x_2,x_1,x_0}(x_2, x_1, x_0)}{f_{x_0}(x_0)} \\ &= f_{x_3|x_2,x_1,x_0}(x_3 \mid x_2, x_1, x_0) \cdot f_{x_2,x_1|x_0}(x_2, x_1 \mid x_0) \\ &= f_{x_3|x_2,x_1,x_0}(x_3 \mid x_2, x_1, x_0) \cdot f_{x_2|x_1,x_0}(x_2 \mid x_1, x_0) \cdot f_{x_1|x_0}(x_1 \mid x_0) \end{aligned}$$

Therefore generalising for any $T \geq 1$ via iteration:

$$f_{x_T,...,x_1|x_0}(x_T, ..., x_1 \mid x_0) = \prod_{t=1}^{T} f_{x_T|x_{t-1},...,x_1}(x_T \mid x_{t-1}, ..., x_0)$$

1.b.

Using part a) we can define the likelihood and corresponding log-likelihood:

$$\mathcal{L}(x_T, ..., x_1 \mid x_0, \delta, \sigma^2) = \prod_{t=1}^{T} f_{x_T|x_{t-1},...,x_0,\delta,\sigma^2}(x_T \mid x_{t-1}, ..., x_0, \delta, \sigma^2)$$
$$L(x_T, ..., x_1 \mid x_0, \delta, \sigma^2) = \sum_{t=1}^{T} log[f_{x_T|x_{t-1},...,x_0,\delta,\sigma^2}(x_T \mid x_{t-1}, ..., x_0, \delta, \sigma^2)]$$

Using the normal distribution density:

$$L(x_T, ..., x_1 \mid x_0, \delta, \sigma^2) = \sum_{t=1}^{T} log\left[\frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x_t - x_{t-1} - \delta)^2}{2\sigma^2}\right)\right]$$
$$= \sum_{t=1}^{T} log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(x_t - x_{t-1} - \delta)^2$$

F.O.C's:

$$\frac{\partial L}{\partial \delta} = \frac{\sum_{t=1}^{T}(x_t - x_{t-1} - \hat{\delta}_{ML}^2)}{\sigma^2} = 0$$
$$\sum_{t=1}^{T}(x_t - x_{t-1} - \hat{\delta}_{ML}) = 0$$

$$\hat{\delta}_{ML} = \frac{1}{T}\sum_{t=1}^{T}(x_t - x_{t-1}) = \frac{1}{T}\sum_{t=1}^{T}(x_T - x_0) = \frac{x_T}{T}$$

$$\frac{\partial L}{\partial \sigma^2} = -\frac{T}{2\hat{\sigma}_{ML}^2} + \frac{\sum_{t=1}^{T}(x_t - x_{t-1} - \hat{\delta}_{ML})^2}{2\sigma^4} = 0$$

$$\hat{\sigma}_{ML}^2 = \frac{1}{T}\sum_{t=1}^{T}(x_t - x_{t-1} - \hat{\delta}_{ML})^2 = \frac{1}{T}\sum_{t=1}^{T}(x_t - x_{t-1} - \frac{x_t}{T})^2$$

1.c.

$$\mathcal{I}(\delta) = Var\left(\frac{d}{d\delta}L(x_T, ..., x_1 \mid x_0, \delta, \sigma^2)\right) = -E\left[\frac{d}{d\delta^2}L(x_T, ..., x_1 \mid x_0, \delta, \sigma^2)\right]$$

Calculating this second derivative (using the fact $\sigma^2 = 1$):

$$\begin{aligned}\frac{dL}{d\delta^2} &= -\frac{T}{\sigma^2}\\ &= -T\end{aligned}$$

Using the Cramer-Rao Lower Bound formula:

$$\begin{aligned}-E[T] &= T\\ Var(\hat{\delta}_{ML}) =\geq \frac{1}{T} &= \mathcal{I}^{-1}(\delta)\end{aligned}$$

When $\sigma^2 = 1$:

$$Var(\hat{\delta}_{ML}) = Var\left(\frac{x_T}{T}\right) = \frac{1}{T}$$

Therefore, $\hat{\delta}_{ML}$ does indeed achieve the Cramer Rao Lower Bound.

1.d.

Solving for the MLE estimates using R we achieve:

$$\begin{aligned}\hat{\delta}_{ML} &= 0.0024\\ \hat{\sigma}^2_{ML} &= 0.0019\\ \sqrt{\hat{\sigma}^2_{ML}} &= 0.0423\end{aligned}$$

1.e.

- GM1 holds as the X matrix is full rank.
- GM2 also holds as we know $E[\varepsilon \mid X] = E[\varepsilon] = 0$ from the fact that $E[\varepsilon] = 0$ and ε and X are independent.
- GM3 holds as we know that there is:
  - Homoskedasticity: $Var(\varepsilon_i \mid X) = \gamma^2$ for all i.
  - No serial correlation: ε is i.i.d. so there are no off-diagonal elements in $Var(\varepsilon \mid X)$.

(y_t, x_t) are not i.i.d across $t = 1, ..., T$. To show:

$$\begin{aligned}y_t &= \alpha + \beta(\delta + x_{t-1} + \eta_t) + \varepsilon_t\\ y_{t+1} &= \alpha + \beta x_{t-1} + \varepsilon_{t+1}\end{aligned}$$

However, the fact that the Gauss-Markov assumptions still hold allows us to use standard testing procedures to test the null hypothesis $H_0 : \beta = 1$.

We can construct the test statistic $T = \frac{\hat{\beta}-1}{\sqrt{\hat{\sigma^2}(X'X)^{-1}}}$ which is distrubuted with $t(n-k) = t(2)$ under the null hypothesis H_0. We then compare this to the two sided critical value for the $t(2)$ distribution at a specified significance level to reject / fail to reject the null hypothesis H_0.

Question 2

2.a.

Regression results are summarised in the table below:

Table 1: Results

	Dependent variable:
	logw0
educ_1	-3.709^{***} (1.337)
educ_3	-2.920^{**} (1.337)
educ_4	-1.763 (1.337)
educ_6	-0.932 (0.726)
educ_7	-2.140^{***} (0.689)
educ_8	-1.174^{***} (0.404)
educ_9	-1.263^{***} (0.377)
educ_10	-1.011^{***} (0.375)
educ_11	-1.174^{***} (0.367)
educ_12	-0.976^{***} (0.348)
educ_13	-0.873^{**} (0.352)
educ_14	-0.656^{*} (0.350)
educ_15	-0.664^{*} (0.355)
educ_16	-0.172 (0.345)
educ_17	-0.154 (0.362)
educ_18	-0.196 (0.371)
educ_19	0.293 (0.394)
educ_20	
exper_0	-2.227^{*} (1.337)
exper_1	-2.454^{**} (1.096)
exper_2	-2.334^{**} (1.043)
exper_3	-2.295^{**} (1.023)
exper_4	-2.038^{**} (1.015)
exper_5	-2.165^{**} (1.013)
exper_6	-2.081^{**} (1.011)
exper_7	-1.878^{*} (1.008)
exper_8	-1.821^{*} (1.007)
exper_9	-1.906^{*} (1.007)
exper_10	-1.752^{*} (1.007)
exper_11	-1.771^{*} (1.007)
exper_12	-1.701^{*} (1.007)
exper_13	-1.724^{*} (1.007)
exper_14	-1.641 (1.004)
exper_15	-1.503 (1.010)
exper_16	-1.604 (1.035)
exper_17	-1.889^{*} (1.087)
exper_18	-1.764^{*} (1.004)
exper_19	-1.265 (1.132)
exper_20	
exper_21	-0.452 (1.145)
exper_22	
exper_25	
Constant	12.226^{***} (1.063)
Observations	1,500
R^2	0.112
Adjusted R^2	0.089
Residual Std. Error	0.810 (df = 1461)
F Statistic	4.835^{***} (df = 38; 1461)
Note:	*p<0.1; **p<0.05; ***p<0.01

2.b.

Restrictions on alpha require that:

$$E(\alpha_0 + \alpha_{educ_i} \cdot i + \alpha_{exper_j} \cdot j) = E(\beta_0 + \beta_1 educ_i + \beta_2 exper_j)$$

For the α's relating to the education dummy variables:

$$\begin{aligned}
\alpha_{educ_1} &= \beta_1 \\
\alpha_{educ_2} &= 2 * \beta_1 = 2 * \alpha_{educ_1} \\
\ldots \\
\alpha_{educ_{20}} &= 20 * \beta_1 = 20 * \alpha_{educ_1}
\end{aligned}$$

Similarly for the α's relating to the experience dummy variables:

$$\begin{aligned}
\alpha_{exper_1} &= \beta_2 \\
\alpha_{exper_2} &= 2 * \beta_2 = 2 * \alpha_{exper_1} \\
\ldots \\
\alpha_{educ_{25}} &= 25 * \beta_2 = 25 * \alpha_{educ_2}
\end{aligned}$$

The corresponding R matrix is:

$$R_{43\times 46} = \begin{pmatrix}
0 & 2 & -1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\
0 & 3 & 0 & -1 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 20 & 0 & 0 & \cdots & -1 & 0 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & 0 & \cdots & 0 & 2 & -1 & 0 & \cdots & 0 \\
0 & 0 & 0 & 0 & \cdots & 0 & 3 & 0 & -1 & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & 0 & 0 & \cdots & 0 & 25 & 0 & 0 & \cdots & -1
\end{pmatrix}$$

q is a 43×1 vector of 0's.

If we drop conditions for the coefficients on the values education and experience that are either dropped for perfect collinearity (4) or result in NA estimates (4), we arrive at a $R_{35\times 46}$ restriction matrix. Computing the F-statistic:

$$F = \frac{(RSS_r - RSS_u)/p}{RSS_u/(n-k)} \overset{H_0}{\sim} F(p, n-k)$$

These values are calculated in the R code linked at the top of this document.

$$\begin{aligned}
F &= \frac{(987.56 - 958.30)/35}{958.30/(1500-39)} = 1.274 \\
F_{crit} &= F_{0.95}(35, (1500-39)) = 1.431
\end{aligned}$$

As F $(= 1.274) < F_{crit}$ $(= 1.431)$ we fail to reject the null hypothesis of a linear specification of the conditional expectation function.

2.c.

I don't believe that this is a good way of testing the assumption of a linear conditional expectation function, as the result only applies for the population exhibiting our observed values of data. This result might not generalise out of sample, meaning we wouldn't be able to conclude on the assumption of a linear conditional expectation function out of sample.

The part of the assumption we are testing is testing the linearity of education and experience affecting log(wage) independently. We are ignoring any potential interactions of education and experience which may enter the specification in a non-linear fashion.

Question 3

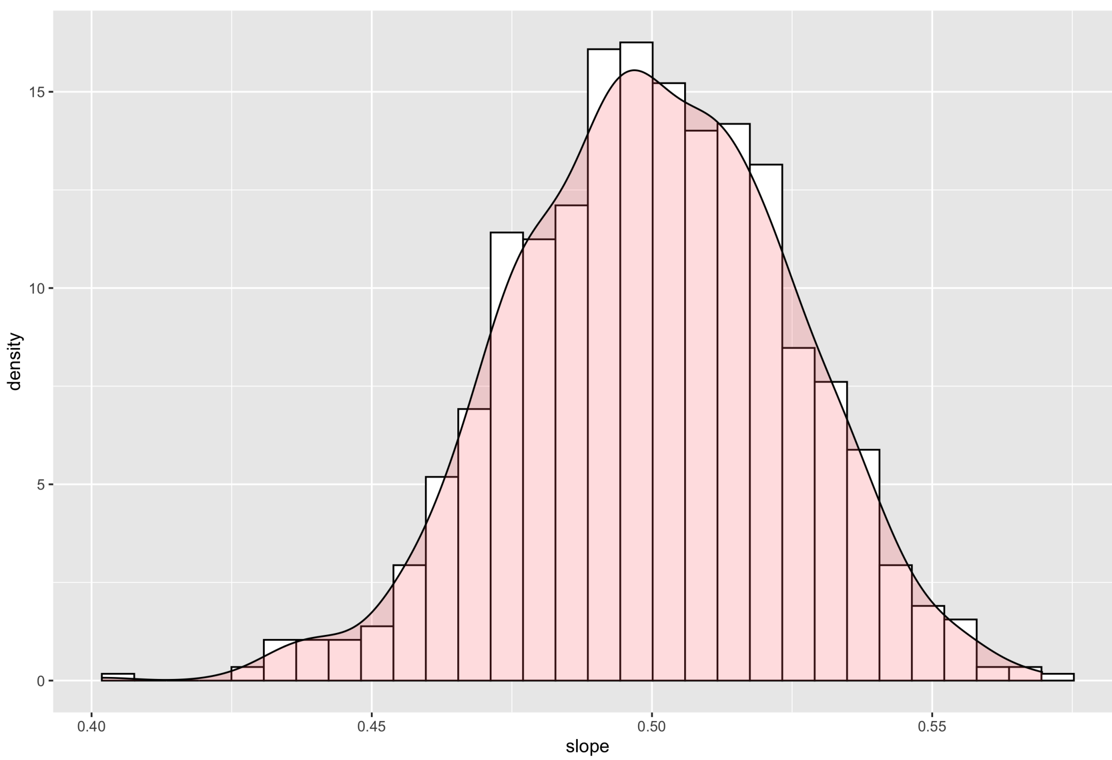
3.a.

The table below shows the summary statistics resulting from a Monte Carlo simulation of $y_i = 1 + 0.5z_i + \varepsilon_i$ for 100 simulated observations (N) and 1,000 iterations of the simulation (M).

Table 2: Monte Carlo simulation results, N = 100, M = 1,000

Statistic	N	Mean	St. Dev.	Min	Max
intercept	1,000	1.001	0.093	0.763	1.354
slope	1,000	0.500	0.025	0.402	0.570

Figure 1: Histogram for Monte Carlo simulation results, N = 100, M = 1,000



The mean values for the simulated OLS regression coefficients are very similar to the population parameters, providing evidence to suggest the unbiasedness of the OLS estimator of the regression coefficients.

3.b.

The tables below show the summary statistics resulting from a Monte Carlo simulation of y_i when increasing the number of simulated observations of N (1,000, 10,000 and 100,000). We can see that the effect of this is to reduce the standard deviation of the mean OLS parameter estimates. Therefore, the fact that the standard deviations of the distributions of our OLS parameter estimates decrease as N is increased provides evidence to suggest that the OLS estimator of the regression coefficients is consistent. The corresponding histograms for the slope coefficients visualise the increasing concentration of mean parameter estimates around the true population value (note the changing x-axis scale).

Table 3: Monte Carlo simulation results, N = 1,000, M = 1,000

Statistic	N	Mean	St. Dev.	Min	Max
intercept	1,000	1.000	0.030	0.903	1.102
slope	1,000	0.500	0.008	0.478	0.525

Figure 2: Histogram for Monte Carlo simulation results, N = 1000, M = 1,000

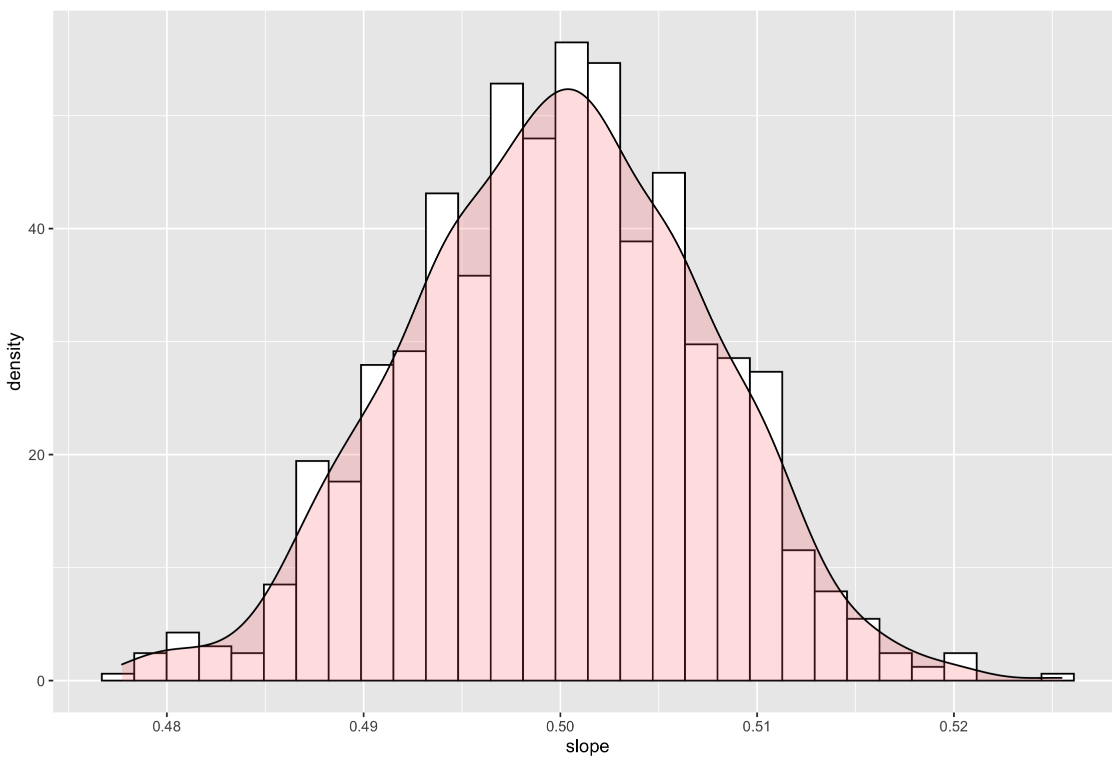


Table 4: Monte Carlo simulation results, N = 10,000, M = 1,000

Statistic	N	Mean	St. Dev.	Min	Max
intercept	1,000	1.000	0.009	0.968	1.032
slope	1,000	0.500	0.002	0.492	0.507

Figure 3: Histogram for Monte Carlo simulation results, N = 10,000, M = 1,000

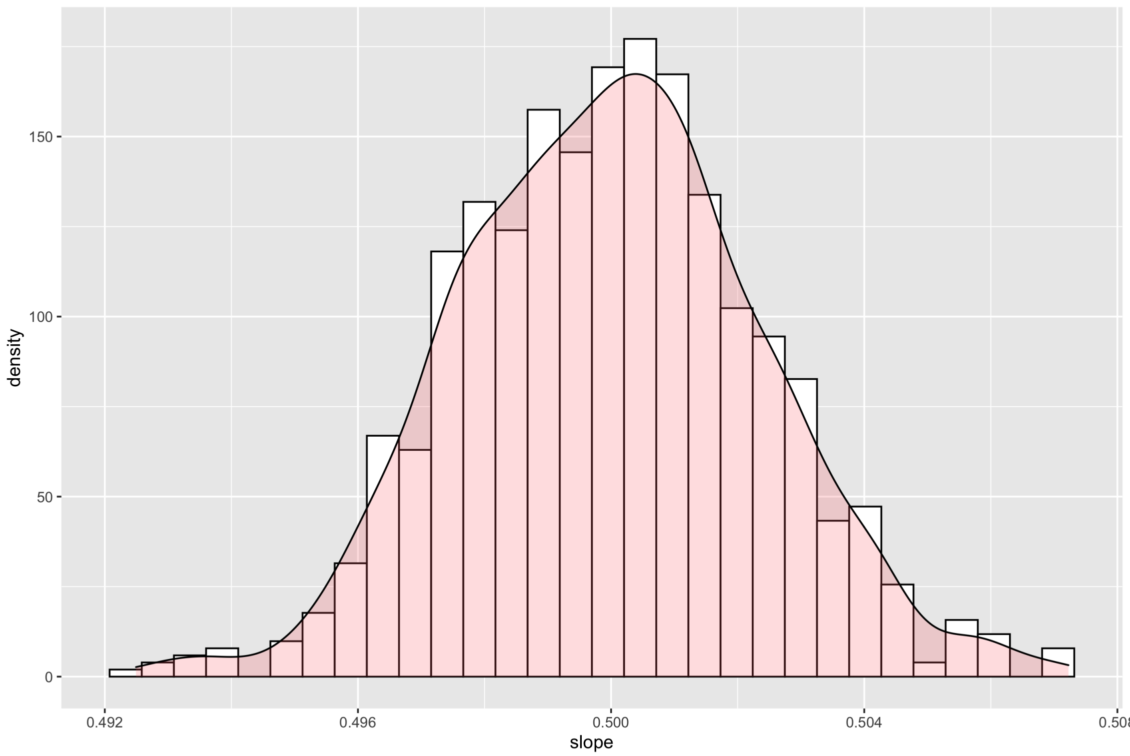
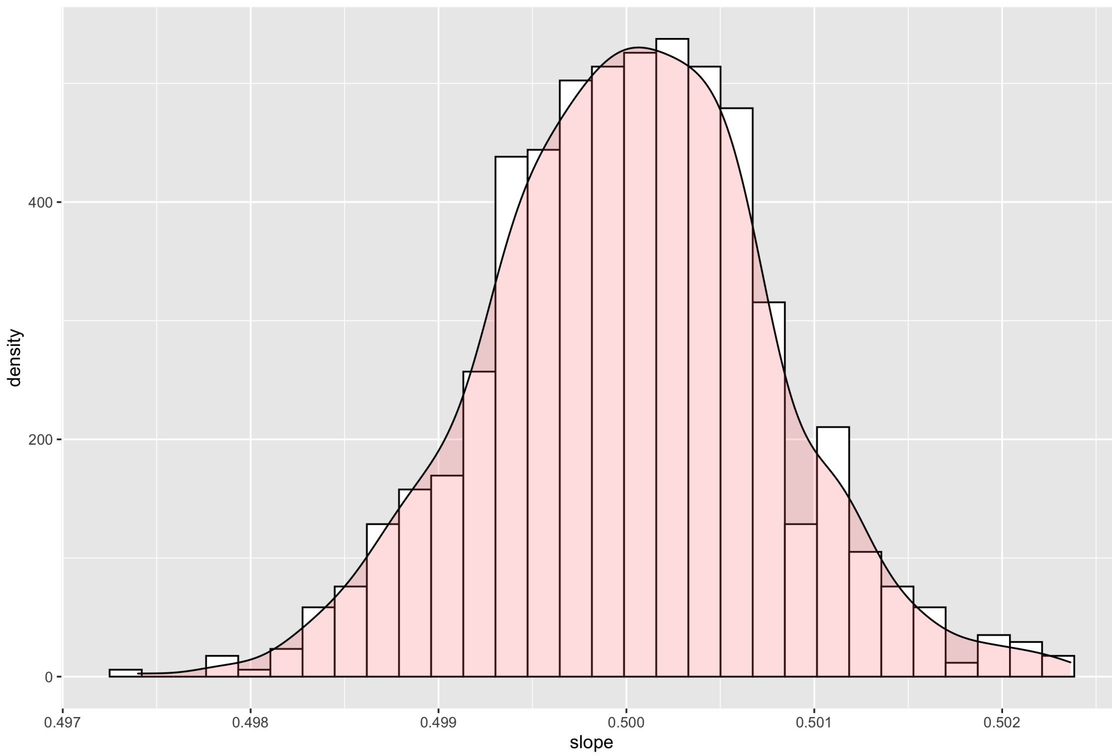


Table 5: Monte Carlo simulation results, N = 100,000, M = 1,000

Statistic	N	Mean	St. Dev.	Min	Max
intercept	1,000	1.000	0.003	0.989	1.008
slope	1,000	0.500	0.001	0.497	0.502

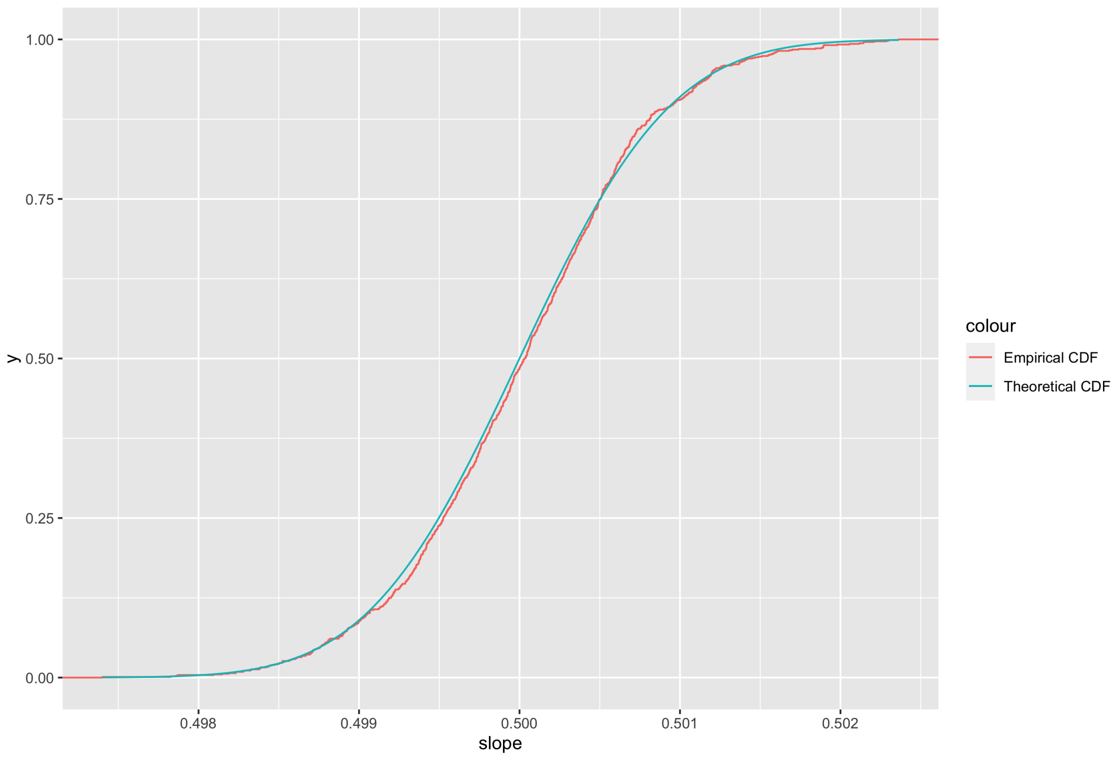
Figure 4: Histogram for Monte Carlo simulation results, N = 100,000, M = 1,000



3.c.

The asymptotic normality of the OLS estimator of the regression slope can be seen in the following figure where the empirical CDF arising from the Monte Carlo simulation is contrasted to the corresponding theoretical normal CDF for the true population slope coefficient.

Figure 5: Empirical CDF vs. Theoretical normal CDF, N = 100,000, M = 1,000



We can see that these two CDF's plotted are approximately equal, providing evidence to suggest that the OLS estimator of the regression slope is asymptotically normal with the asymptotic variance $\sigma^2/Var(z_i)$.

3.d.

True variance:

$$\sigma^2 = Var(\varepsilon_i)$$
$$= \frac{1}{3}$$

Table 6: Monte Carlo simulation results, N = 100, M = 1,000

Statistic	N	Mean	St. Dev.	Min	Max
intercept	1,000	1.001	0.093	0.763	1.354
slope	1,000	0.500	0.025	0.402	0.570
sigmasq_hat	1,000	0.334	0.031	0.224	0.425

The fact that the mean of the $\frac{RSS}{n-k}$ estimates from the Monte Carlo simulation (0.334) is very close to the true population σ^2 ($\frac{1}{3}$) provides good evidence to suggest that it is an unbiased estimator of σ^2.

Question 4

4.a.

Definition of $\hat{\beta}$ for univariate linear regression model:

$$\hat{\beta} = \beta + \frac{\sum_{i=1}^{n} x_i \varepsilon_i}{\sum_{i=1}^{n} x_i^2}$$

Simplifying the expressions on the RHS (assuming n = even for simplicity):

$$\begin{aligned}
\sum_{i=1}^{n} x_i^2 &= 0^2 + 0 + 2^2 + \ldots + n^2 \\
&= \sum_{i=1}^{\frac{n}{2}} (2i)^2 \\
&= 4 \sum_{i=1}^{\frac{n}{2}} i^2 \\
&= \frac{4(\frac{n}{2})(\frac{n}{2}+1)(n+1)}{6}
\end{aligned}$$

$$\begin{aligned}
\sum_{i=1}^{n} x_i \varepsilon_i &= 2\varepsilon_2 + 4\varepsilon_4 + \ldots + n\varepsilon_n \\
&= \sum_{i=1}^{\frac{n}{2}} 2i\varepsilon_i
\end{aligned}$$

Finding the expected value and variance of $\frac{\sum_{i=1}^{n} x_i \varepsilon_i}{\sum_{i=1}^{n} x_i^2}$:

$$\begin{aligned}
E[\sum_{i=1}^{n} x_i \varepsilon_i] &= \sum_{i=1}^{\frac{n}{2}} 2iE[\varepsilon_i] = 0 \\
\implies E\left[\frac{\sum_{i=1}^{n} x_i \varepsilon_i}{\sum_{i=1}^{n} x_i^2}\right] &= 0
\end{aligned}$$

$$
\begin{aligned}
Var(\sum_{i=1}^{n} x_i \varepsilon_i) &= Var(\sum_{i=1}^{\frac{n}{2}} 2iE[\varepsilon_i]) \\
&= \sum_{i=1}^{\frac{n}{2}} 4i^2 Var(\varepsilon_i) \\
&= 4\sum_{i=1}^{\frac{n}{2}} i^2 \\
&= \frac{4(\frac{n}{2})(\frac{n}{2}+1)(n+1)}{6}
\end{aligned}
$$

Let $a \equiv \frac{4(\frac{n}{2})(\frac{n}{2}+1)(n+1)}{6}$

Therefore:

$$
\begin{aligned}
Var\left(\frac{\sum_{i=1}^{n} x_i \varepsilon_i}{\sum_{i=1}^{n} x_i^2}\right) &= Var\left(\frac{\sum_{i=1}^{n} x_i \varepsilon_i}{a}\right) \\
&= \frac{1}{a^2} Var(\sum_{i=1}^{n} x_i \varepsilon_i) \\
&= \frac{1}{a} \\
&= \frac{6}{4(\frac{n}{2})(\frac{n}{2}+1)(n+1)}
\end{aligned}
$$

As a result:

$$
\hat{\beta} \sim \mathcal{N}\left(\beta, \, \frac{6}{4(\frac{n}{2})(\frac{n}{2}+1)(n+1)}\right)
$$

As $n \to \infty$, $Var(\hat{\beta}) \to 0$

By Chebyshev's inequality:

$$
Pr(|\hat{\beta} - \beta|) \leq \frac{Var(\hat{\beta})}{\varepsilon^2} \to 0, \text{ for any } \varepsilon > 0
$$

$$
\implies \hat{\beta} \xrightarrow{p} \beta
$$

$\hat{\beta}$ is consistent.

4.b.

As in a), definition of $\hat{\beta}$ for univariate linear regression model:

$$\hat{\beta} = \beta + \frac{\sum_{i=1}^{n} x_i \varepsilon_i}{\sum_{i=1}^{n} x_i^2}$$

Simplifying the expressions on the RHS:

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} \lambda^{2i} = \frac{\lambda^2(1-\lambda^{2n})}{1-\lambda^2}$$

$$\sum_{i=1}^{n} x_i \varepsilon_i = \sum_{i=1}^{n} \lambda^i \varepsilon_i$$

Finding the expected value and variance:

$$E[\sum_{i=1}^{n} \lambda^i \varepsilon_i] = \sum_{i=1}^{n} \lambda^i E[\varepsilon_i] = 0$$

$$Var(\sum_{i=1}^{n} \lambda^i \varepsilon_i) = \sum_{i=1}^{n} \lambda^{2i} Var(\varepsilon_i) = \sum_{i=1}^{n} \lambda^{2i} = \frac{\lambda^2(1-\lambda^{2n})}{1-\lambda^2}$$

If we define $a \equiv \frac{\lambda^2(1-\lambda^{2n})}{1-\lambda^2}$ we have exactly the same as in part a).

Therefore:

$$Var(\hat{\beta}) = \frac{1}{a} = \frac{1-\lambda^2}{\lambda^2(1-\lambda^{2n})}$$

As $n \to \infty$, $Var(\hat{\beta}) \to \frac{1-\lambda^2}{\lambda^2}$

By Chebyshev's inequality:

$$Pr(|\hat{\beta} - \beta|) \leq \frac{Var(\hat{\beta})}{\varepsilon^2} \to \frac{1-\lambda^2}{\varepsilon^2 \lambda^2}, \text{ for any } \varepsilon > 0$$

Therefore we are unable to guarantee the consistency of $\hat{\beta}$.

4.c.

- GM1 holds in both cases. In both cases there is no perfect multicollinearity as we know that $x \neq 0$.
- GM2 also holds in both cases. X is predetermined in both cases, therefore $E[\varepsilon \mid X] = E[\varepsilon] = 0$.
- GM3 holds in both cases. Again, X is predetermined in both cases. Thus, $Var(\varepsilon \mid X) = Var(\varepsilon) = I_n$.

Gauss-Markov assumptions holding does not guarantee the consistency of the OLS estimator, it only guarantees that $\hat{\beta}$ is the best linear conditionally unbiased estimator. We know that in part b) the OLS estimator must be BLUE (GM1-3 hold) but it is not consistent as the variance of the estimator does not converge to zero as n increases.

4.d.

In b) we had that:

$$\hat{\beta} = \beta + \frac{\sum_{i=1}^{n} x_i \varepsilon_i}{\sum_{i=1}^{n} x_i^2} = \beta + \frac{\sum_{i=1}^{n} \lambda^i \varepsilon_i}{\frac{\lambda^2(1-\lambda^{2n})}{1-\lambda^2}}$$

This is a constant plus a linear function of ε, meaning that it will be normally distributed.

The mean and variance of this distribution were calculated in part b). Therefore:

$$\hat{\beta} \sim \mathcal{N}\left(\beta,\, \frac{1-\lambda^2}{\lambda^2(1-\lambda^{2n})}\right)$$

We also know that:

$$\frac{1-\lambda^2}{\lambda^2(1-\lambda^{2n})} \xrightarrow{p} \frac{1-\lambda^2}{\lambda^2}$$

Thus:

$$\hat{\beta} \xrightarrow{d} \mathcal{N}\left(\beta,\, \frac{1-\lambda^2}{\lambda^2}\right)$$

Question 5

5.a.

$$
\begin{aligned}
L(\theta \mid Y) &= \sum_{i=1}^{1000} (log(\theta^{-\frac{1}{2}}) - y_i \cdot \theta^{-\frac{1}{2}}) \\
&= -500log(\theta) - \theta^{-\frac{1}{2}} \sum_{i=1}^{1000} y_i
\end{aligned}
$$

$$
\frac{\partial L(\theta \mid Y)}{\partial \theta} = -500\theta^{-1} + \frac{1}{2}\theta^{-\frac{3}{2}} \sum_{i=1}^{1000} y_i
$$

$$
\frac{\partial^2 L(\theta \mid Y)}{\partial \theta^2} = 500\theta^{-2} - \frac{3}{4}\theta^{-\frac{5}{2}} \sum_{i=1}^{1000} y_i
$$

$$
\begin{aligned}
\mathcal{I}(\theta) = -E\left[\frac{\partial^2 L(\theta \mid Y)}{\partial \theta^2}\right] &= -\left[500\theta^{-2} - \frac{3}{4}\theta^{-\frac{5}{2}} \sum_{i=1}^{1000} E[y_i]\right] \\
&= -\left[500\theta^{-2} - \frac{3}{4}\theta^{-\frac{5}{2}}(1000\theta^{\frac{1}{2}})\right] \\
&= \frac{250}{\theta^2}
\end{aligned}
$$

5.b.

$$
L(\theta \mid Y) = \sum_{i=1}^{1000} (log(\theta^{-\frac{1}{2}}) - y_i \cdot \theta^{-\frac{1}{2}})
$$

$$
\begin{aligned}
\frac{\partial L(\theta \mid Y)}{\partial \theta} &= -500\hat{\theta}^{-1} + \frac{1}{2}\hat{\theta}^{-\frac{3}{2}} \sum_{i=1}^{1000} y_i = 0 \\
500\hat{\theta}^{\frac{1}{2}} &= \frac{1}{2} \sum_{i=1}^{1000} y_i \\
\hat{\theta}^{\frac{1}{2}} &= \frac{1}{1000} \sum_{i=1}^{1000} y_i \\
\hat{\theta}_{ML} &= \bar{y}^2
\end{aligned}
$$

$$
\begin{aligned}
E[\hat{\theta}_{ML}] &= E[\bar{y}^2] \\
&= E[\bar{y}]^2 + Var(\bar{y})
\end{aligned}
$$

As y_i is a random sample:

$$
Var(\bar{y}) = \frac{\theta}{1000}
$$

Computing $E[\bar{y}]^2$:

$$
\begin{aligned}
E[\bar{y}]^2 &= E\left[\frac{1}{1000}\sum_{i=1}^{1000} y_i\right]^2 \\
&= \left(\frac{1}{1000}\sum_{i=1}^{1000} E[y_i]\right)^2 \\
&= \left(\frac{1}{1000} \cdot 1000\sqrt{\theta}\right)^2 \\
&= \theta
\end{aligned}
$$

$$
\implies E[\hat{\theta}_{ML}] = \theta + \frac{\theta}{1000}
$$

$$
\implies bias(\hat{\theta}_{ML}) = \frac{\theta}{1000}
$$

5.c.

$$
\hat{\theta}_{ML} = \bar{y}^2
$$

By LLN:

$$
\bar{y} \xrightarrow{p} E[y_i] = 1
$$

By Continuous Mapping Theorem:

$$
\bar{y}^2 \xrightarrow{p} 1^2 = 1
$$

Therefore:

$$
\hat{\theta}_{ML} \xrightarrow{p} 1
$$

5.d.

$$
\begin{aligned}
d(g, f_\theta) &= \int_0^\infty g(y) \log \frac{g(y)}{f_\theta(y)} dy \\
&= \int_0^\infty g(y) \log g(y) dy - \int_0^\infty g(y) \log f_\theta(y) dy
\end{aligned}
$$

First integral doesn't depend on θ so maximising $d(g, f_\theta)$ is equivalent to maximising $-\int_0^\infty g(y) \log f_\theta(y) dy$

$$
\begin{aligned}
\max_\theta - \int_0^\infty g(y) \log f_\theta(y) \\
\max_\theta - \int_0^\infty \frac{1}{\sqrt{2\pi y}} \exp^{-\frac{y}{2}} \left[-\frac{1}{2} \log \theta - \frac{y}{\sqrt{\theta}} \right] \\
\max_\theta \frac{1}{2} \log \theta + \frac{1}{\sqrt{\theta}}
\end{aligned}
$$

FOC:

$$
\frac{\partial}{\partial \theta} = \frac{1}{2\theta} - \frac{1}{2\theta^{\frac{3}{2}}} = 0
$$

$$
\implies \theta = 1
$$

This is the same as in part c), i.e. the value of θ that minimises the KL divergence is equal to the value that the ML estimator converges to as sample size goes to infinity.

We know that MLE will maximise $\sum_{i=1}^n \log f_\theta(y_i)$, or equivalently will maximise $\frac{1}{n} \sum_{i=1}^n \log f_\theta(y_i)$. By LLN this converges in probability to $E_g[log(f_\theta(y_i)]$, so $plim\hat{\theta}_{ML}$ minimises $E_g[log(f_\theta(y_i)]$.

$$
E_g[log(f_\theta(y_i))] = E_g[\log g(y)] - E_g \left[\log \frac{g(y)}{f_\theta(y)} \right]
$$

The RHS is maximised (with respect to θ) where $E_g[\log \frac{g(y)}{f_\theta(y)}]$ is minimised, which by definition is where the KL divergence is minimised. Therefore $plim\hat{\theta}_{ML}$ will also minimise the KL divergence.

5.e.

If correct density of y_i is $g(y)$, then:

$$
\begin{aligned}
\log f_\theta(y_i) &= \log \frac{1}{\sqrt{\theta}} - \frac{y_i}{\sqrt{\theta}} \\
&= -\frac{1}{2} \log(\theta) - \frac{y_i}{\sqrt{\theta}}
\end{aligned}
$$

$$\frac{d}{d\theta} = -\frac{1}{2\theta} + \frac{y_i}{2\theta^{\frac{3}{2}}}$$
$$\frac{d^2}{d\theta^2} = \frac{1}{2\theta^2} - \frac{3y_i}{4\theta^{\frac{5}{2}}}$$

$$\begin{aligned}
Var\left(\frac{d}{d\theta}\right) &= Var\left(-\frac{1}{2\theta} + \frac{y_i}{2\theta^{\frac{3}{2}}}\right) \\
&= Var\left(\frac{y_i}{2\theta^{\frac{3}{2}}}\right) \\
&= \frac{1}{4\theta^3} Var(y_i) \\
&= \frac{1}{2\theta^3}
\end{aligned}$$

$$\begin{aligned}
-E\left[\frac{d^2}{d\theta^2}\right] &= -E\left[\frac{1}{2\theta^2} - \frac{3y_i}{4\theta^{\frac{5}{2}}}\right] \\
&= -\frac{1}{2\theta^2} + \frac{3E[y_i]}{4\theta^{\frac{5}{2}}} \\
&= -\frac{1}{2\theta^2} + \frac{3}{4\theta^{\frac{5}{2}}}
\end{aligned}$$

Comparing:

$$\begin{aligned}
Var\left(\frac{d}{d\theta}\right) - \left(-E\left[\frac{d^2}{d\theta^2}\right]\right) &= \frac{1}{2\theta^3} - \left(-\frac{1}{2\theta^2} + \frac{3}{4\theta^{\frac{5}{2}}}\right) \\
&= 1 + \theta - \frac{3\sqrt{\theta}}{2}
\end{aligned}$$

This is greater than zero for all possible values of theta.