36-467 Final Exam

Eu Jing Chua
eujingc

December 11, 2018

Question 0

```
lsial <- read.csv("http://www.stat.cmu.edu/~cshalizi/dst/18/exams/2/sial.csv")
rownames(lsial) <- lsial$X
lsial$X <- NULL
lsial[1:25, ] <- log(lsial[1:25, ])
```

Question 1

$$
\begin{aligned}
S_i(t+1) &= g + S_i(t) + \eta_i(t+1) && (1)\\
\log N_i(t+1) &= g + \log N_i(t) + \eta_i(t+1) && (2)\\
N_i(t+1) &= e^{g+\log N_i(t)+\eta_i(t+1)} && (3)\\
&= N_i(t)e^{g+\eta_i(t+1)} && (4)\\
\frac{N_i(t+1)}{N_i(t)} &= g + \eta_i(t+1) && (5)\\
N_i(t) &= N_i(0)e^{(g+\eta_i(t))t} && (6)\\
&= N_i(0)(1+r(t))^t && (7)
\end{aligned}
$$

As we can see above, the population size for territory i at $N_i(t)$ is of the form of an exponential growth, with the growth rate being $r(t) = 1 - e^{g+\eta_i(t)}$, or $r = 1 - e^g$ on average.

Question 2

Let $\log E_i(t) = R_i(t)$, where $E_i(t)$ is the number of slaves exported from territory i.

$$
\begin{aligned}
R_i(t) &= c + S_i(t) + \phi_i(t) && (8)\\
\log E_i(t) &= c + \log N_i(t) + \phi_i(t) && (9)\\
\log \frac{E_i(t)}{N_i(t)} &= c + \phi_i(t) && (10)\\
\frac{E_i(t)}{N_i(t)} &= e^{c+\phi_i(t)} && (11)
\end{aligned}
$$

We can see that on average, the trend for exported slaves follows a fixed proportion of the total population, with proportion $\frac{E_i(t)}{N_i(t)} = e^{c+\phi(t)}$, or $\frac{E_i(t)}{N_i(t)} = e^c$ on average.

Question 3

Let $M_{rel,i}$ be the reported relative error margin and $E_{obs,i}(t)$ be the observed number of slaves exported from territory i.

$$1 - M_{rel,i} \leq \frac{E_{obs,i}(t)}{E_i(t)} \leq 1 + M_{rel,i} \tag{12}$$

$$\log(1 - M_{rel,i}) \leq \log(E_{obs,i}(t)) - \log(E_i(t)) \leq \log(1 + M_{rel,i}) \tag{13}$$

$$\log(1 - M_{rel,i}) \leq X_i(t) - R_i(t) \leq \log(1 + M_{rel,i}) \tag{14}$$

We can then use the approximation that if $|a| \ll 1, \log(b(1+a)) \approx \log(b) + a$ to get

$$\log(1) - M_{rel,i} \leq X_i(t) - R_i(t) \leq \log(1) + M_{rel,i} \tag{15}$$

$$-M_{rel,i} \leq \epsilon_i(t) \leq M_{rel,i} \tag{16}$$

$$\tag{17}$$

Thus $M_{rel,i}$ approximates the standard error of $\epsilon_i(t)$.

Question 4

a)

$$\begin{aligned}
\mathrm{Cov}\left[S_i(t), X_i(t)\right] &= \mathrm{Cov}\left[S_i(t), R_i(t) + \epsilon_i(t)\right] && (18)\\
&= \mathrm{Cov}\left[S_i(t), R_i(t)\right] + \mathrm{Cov}\left[S_i(t), \epsilon_i(t)\right] && (19)\\
&= \mathrm{Cov}\left[S_i(t), c + S_i(t) + \phi_i(t)\right] + 0 && (20)\\
&= \mathrm{Cov}\left[S_i(t), c\right] + \mathrm{Cov}\left[S_i(t), S_i(t)\right] + \mathrm{Cov}\left[S_i(t), \phi_i(t)\right] && (21)\\
&= \mathrm{Var}\left[S_i(t)\right] && (22)
\end{aligned}$$

b)

$$\begin{aligned}
\mathbb{E}\left[X_i(t)\right] &= \mathbb{E}\left[R_i(t) + \epsilon_i(t)\right] && (23)\\
&= \mathbb{E}\left[R_i(t)\right] && (24)\\
&= \mathbb{E}\left[c + S_i(t) + \phi_i(t)\right] && (25)\\
&= \mathbb{E}\left[S_i(t)\right] + c && (26)
\end{aligned}$$

c)

The optimal linear predictor of $S_i(t)$ from $X_i(t)$ follows the form:

$$\begin{aligned}
\hat{S}_i(t) &= \alpha + \beta X_i(t) && (27)\\
&= \mathbb{E}\left[S_i(t)\right] + \left(\frac{\mathrm{Cov}\left[X_i(t), S_i(t)\right]}{\mathrm{Var}\left[X_i(t)\right]}\right)\left(X_i(t) - \mathbb{E}\left[X_i(t)\right]\right) && (28)
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{Var}\left[X_i(t)\right] &= \mathrm{Var}\left[R_i(t) + \epsilon_i(t)\right] && (29)\\
&= \mathrm{Var}\left[c + S_i(t) + \phi_i(t)\right] + \sigma_\epsilon^2 && (30)\\
&= \mathrm{Var}\left[S_i(t)\right] + \sigma_\phi^2 + \sigma_\epsilon^2 && (31)
\end{aligned}$$

so the optimal linear predictor is

$$\begin{aligned}
\hat{S}_i(t) &= \mathbb{E}\left[S_i(t)\right] + \left(\frac{\mathrm{Cov}\left[X_i(t), S_i(t)\right]}{\mathrm{Var}\left[X_i(t)\right]}\right)\left(X_i(t) - \mathbb{E}\left[X_i(t)\right]\right) && (32)\\
&= \mathbb{E}\left[S_i(t)\right] + \left(\frac{\mathrm{Var}\left[S_i(t)\right]}{\mathrm{Var}\left[S_i(t)\right] + \sigma_\phi^2 + \sigma_\epsilon^2}\right)\left(X_i(t) - \mathbb{E}\left[S_i(t)\right] - c\right) && (33)
\end{aligned}$$

Question 5

a)

$$\begin{aligned}
\text{Cov}\left[S_i(t), S_i(t+h)\right] &= \text{Cov}\left[S_i(t), hg + S_i(t) + \sum_{j=1}^{h} \eta_i(t+j)\right] && (34)\\
&= \text{Var}\left[S_i(t)\right] + \sum_{j=1}^{h} \text{Cov}\left[S_i(t), \eta_i(t+j)\right] && (35)\\
&= \text{Var}\left[S_i(t)\right] && (36)
\end{aligned}$$

b)

$$\begin{aligned}
\text{Cov}\left[S_i(t), X_i(t+h)\right] &= \text{Cov}\left[S_i(t), R_i(t+h) + \epsilon(t+h)\right] && (37)\\
&= \text{Cov}\left[S_i(t), R_i(t+h)\right] + 0 && (38)\\
&= \text{Cov}\left[S_i(t), c + S_i(t+h) + \phi_i(t+h)\right] && (39)\\
&= 0 + \text{Cov}\left[S_i(t), S_i(t+h)\right] + 0 && (40)\\
&= \text{Cov}\left[S_i(t), S_i(t+h)\right] && (41)\\
&= \text{Var}\left[S_i(t)\right] && (42)
\end{aligned}$$

c)

$$\begin{aligned}
\text{Cov}\left[X_i(t), X_i(t+h)\right] &= \text{Cov}\left[R_i(t) + \epsilon_i(t), X_i(t+h)\right] && (43)\\
&= \text{Cov}\left[R_i(t), X_i(t+h)\right] + 0 && (44)\\
&= \text{Cov}\left[c + S_i(t) + \phi_i(t), X_i(t+h)\right] && (45)\\
&= 0 + \text{Cov}\left[S_i(t), X_i(t+h)\right] + 0 && (46)\\
&= \text{Var}\left[S_i(t)\right] && (47)
\end{aligned}$$

d)

$$\begin{aligned}
\text{Var}\left[S_i(t)\right] &= \text{Var}\left[tg + S_i(0) + \sum_{j=1}^{t} \eta_i(j)\right] && (48)\\
&= \text{Var}\left[S_i(0)\right] + \sum_{j=1}^{t} \text{Var}\left[\eta_i(j)\right] && (49)\\
&= \text{Var}\left[S_i(0)\right] + t\sigma_\eta^2 && (50)
\end{aligned}$$

e)

$$\begin{aligned}
\text{Var}\left[X_i(t)\right] &= \text{Var}\left[R_i(t) + \epsilon_i(t)\right] && (51)\\
&= \text{Var}\left[c + S_i(t) + \phi_i(t) + \epsilon_i(t)\right] && (52)\\
&= \text{Var}\left[S_i(t)\right] + \text{Var}\left[\phi_i(t)\right] + \text{Var}\left[\epsilon_i(t)\right] && (53)\\
&= \text{Var}\left[S_i(0)\right] + t\sigma_\eta^2 + \sigma_\phi^2 + \sigma_\epsilon^2 && (54)
\end{aligned}$$

f)

$$\mathbb{E}\left[S_i(t)\right] = \mathbb{E}\left[tg + S_i(0) + \sum_{j=1}^{t} \eta_i(j)\right] \tag{55}$$

$$= gt + \mathbb{E}\left[S_i(0)\right] + \sum_{j=1}^{t} \mathbb{E}\left[\eta_i(j)\right] \tag{56}$$

$$= \mathbb{E}\left[S_i(0)\right] + gt \tag{57}$$

g)

$$\mathbb{E}\left[X_i(t)\right] = \mathbb{E}\left[c + S_i(t) + \phi_i(t) + \epsilon_i(t)\right] \tag{58}$$

$$= c + \mathbb{E}\left[S_i(t)\right] + 0 + 0 \tag{59}$$

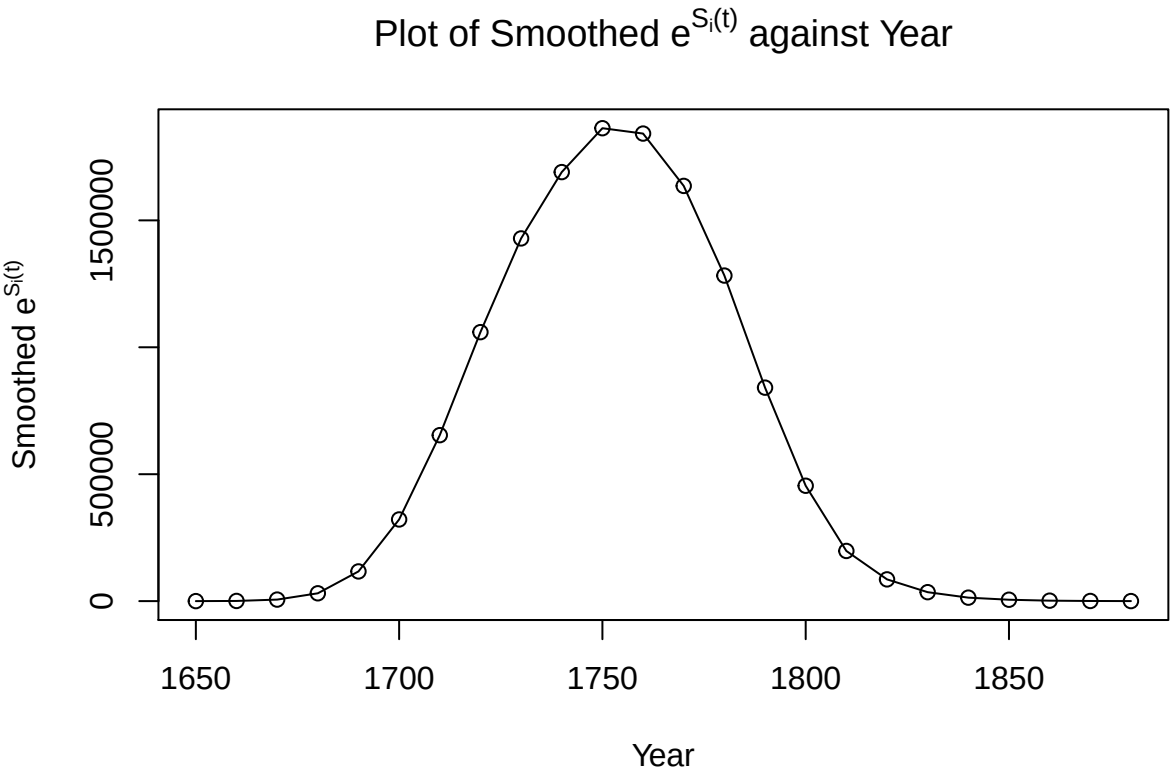
$$= \mathbb{E}\left[S_i(0)\right] + gt + c \tag{60}$$

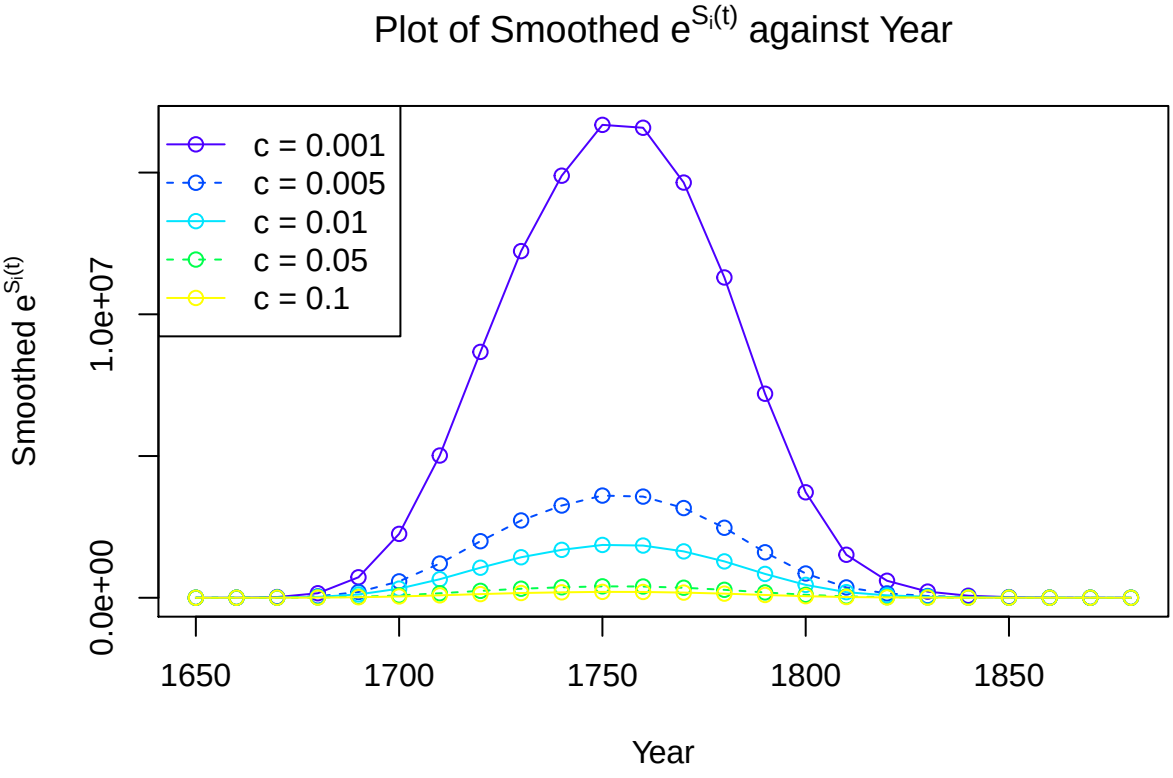
Question 6

```
# Finds the smoothed value of S_i at time t using the optimal linear prediction from
# n observations of X_i
#
# t: Time to get prediction of S_i for
# obs: Observations of X_i to predict from
# vs0: Variance of S_i(0)
# sig.eta: Std. dev. of eta
# sig.phi: Std. dev. of phi
# sig.eps: Std. dev. of eps
# g: Constant term in growth of S_i
# c: Constant term in growth of R_i
# Returns: Smoothed value of S_i(t)

smoother <- function(t, obs, vs0, sig.eta, sig.phi, sig.eps, g, c) {
    n <- length(obs)
    x <- head(obs, -1)
    sfinal <- tail(obs, 1)

    # Variance of S from t = 1 to n
    # Var[S_i(1:n)] = Var[S_i(0)] + (1:n) * sig.eta^2
    vs <- vs0 + (1:n)*sig.eta^2

    # Variance of each observed X from t = 1 to n
    # Var[X_i(1:n)] = Var[S_i(1:n)] + sig.phi^2 + sig.eps^2
    vx <- vs + sig.phi^2 + sig.eps^2

    # Variance matrix of X
    # Non-diagonal entries of Var[X] are Cov[X_i(s), X_i(s + h)]
    vobs <- outer(1:n, 1:n, function(x,y) { vs[pmin(x,y)] })
    # Diagonal entries of Var[X] are Var[X_i(s)] (h = 0)
    diag(vobs) <- vx
    # ???
    diag(vobs)[n] <- vs[n]
```

```
    # Covariance vector between each observed X and S_i(t)
    # Cov[X, S]
    C <- matrix(vs[pmin(t,1:n)], nrow=n, ncol=1)

    # Slope of the multivariate optimal linear predictor
    beta <- solve(vobs) %*% C

    # Expected value of S_i(t)
    # E[S_i(t)] = E[S_i(0)] + g * t, assumes E[S_i(0)] to be 0
    es <- t*g

    # Expected value of observed X
    # E[X_i(1:n)] = E[S_i(0)] + (1:n) * g + c, assumes E[S_i(0)] to be 0
    ex <- (1:n)*g + c
    eobs <- ex
    # ???
    eobs[n] <- n*g

    # Intercept of multivariate optimal linear predictor
    alpha <- es - eobs %*% beta

    # Prediction from multivariate optimal linear predictor
    return(alpha + obs %*% beta)
}

# Applies smoother to each observation of X_i we have to get corresponding smoothed
# and optimal linear predictions of S_i from all n observations of X_i
simultaneous.smoother <- function(obs, vs0, sig.eta, sig.phi, sig.eps, g, c) {
    n <- length(obs)
    sapply(1:n, smoother, obs=obs, vs0=vs0, sig.eta=sig.eta, sig.phi=sig.phi,
           sig.eps=sig.eps, g=g, c=c)
}
```

Question 7


Plot of Smoothed e^Si(t) against Year
Smoothed e^Si(t)
0
500000
1500000
1650
1700
1750
1800
1850
Year

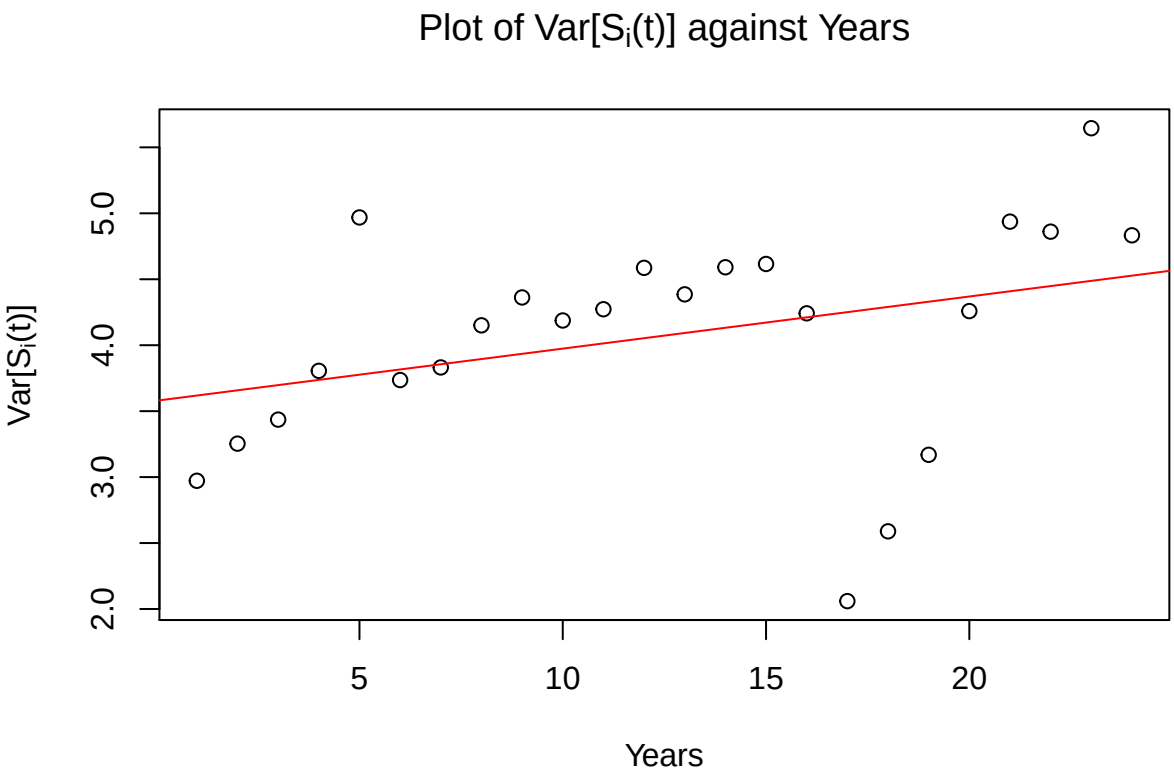

Plot of Smoothed e^Si(t) against Year
c = 0.001
c = 0.005
c = 0.01
c = 0.05
c = 0.1
Smoothed e^Si(t)
0.0e+00
1.0e+07
1650
1700
1750
1800
1850
Year

The results look sensible. It seems that as we decrease c, the population numbers stay higher across the years, while increasing c causes the population numbers to stay lower across the years. The peak in population is also much more dramatic for lower values of c as compared to higher values of c.

Question 8

It assumes $\text{Var}\left[S_i(0)\right] \neq 0$ and it is the same across all territories, $\text{Var}\left[S_i(0)\right] = \sigma_0^2 \quad \forall i$.

Question 9



From the scatter plot, we can see that there seems to be an approximate linear trend in the earlier years, from 1650 to 1800. However, this trend drops off from 1810 onwards, no longer resembling a continuation of the early linear trend. Hence we choose to refit a linear trend to just the earlier years, from 1650 to 1800:

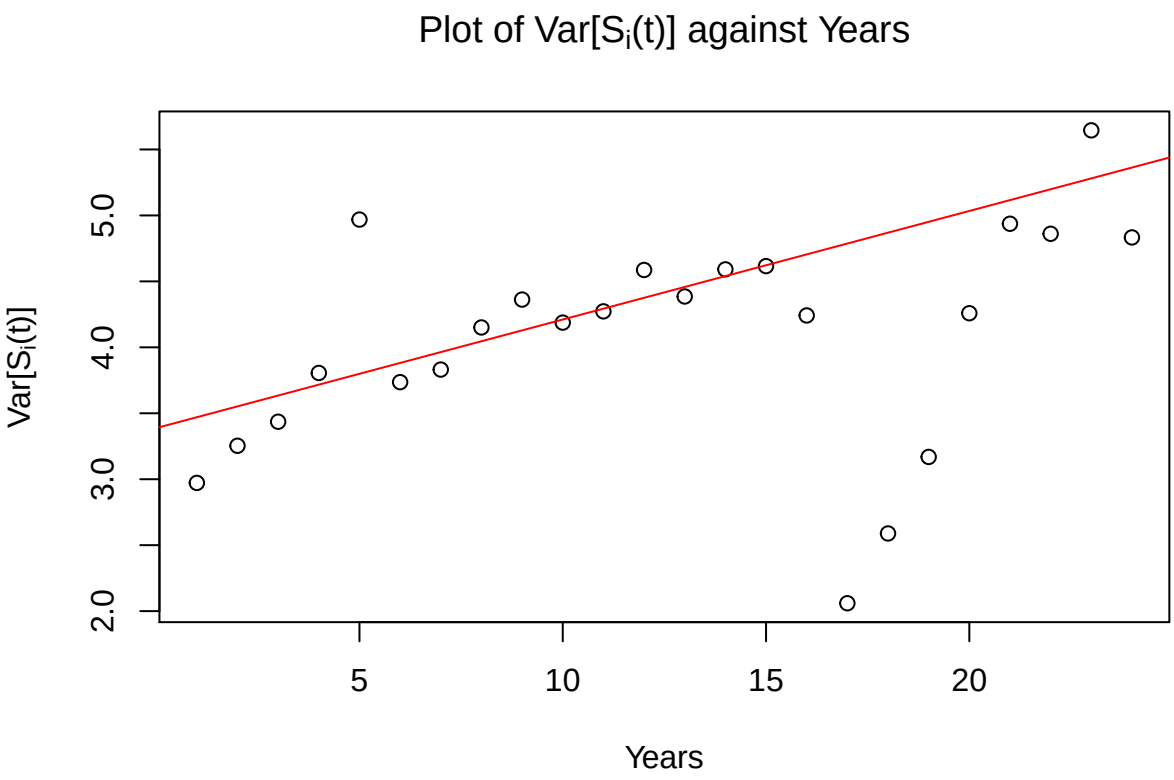
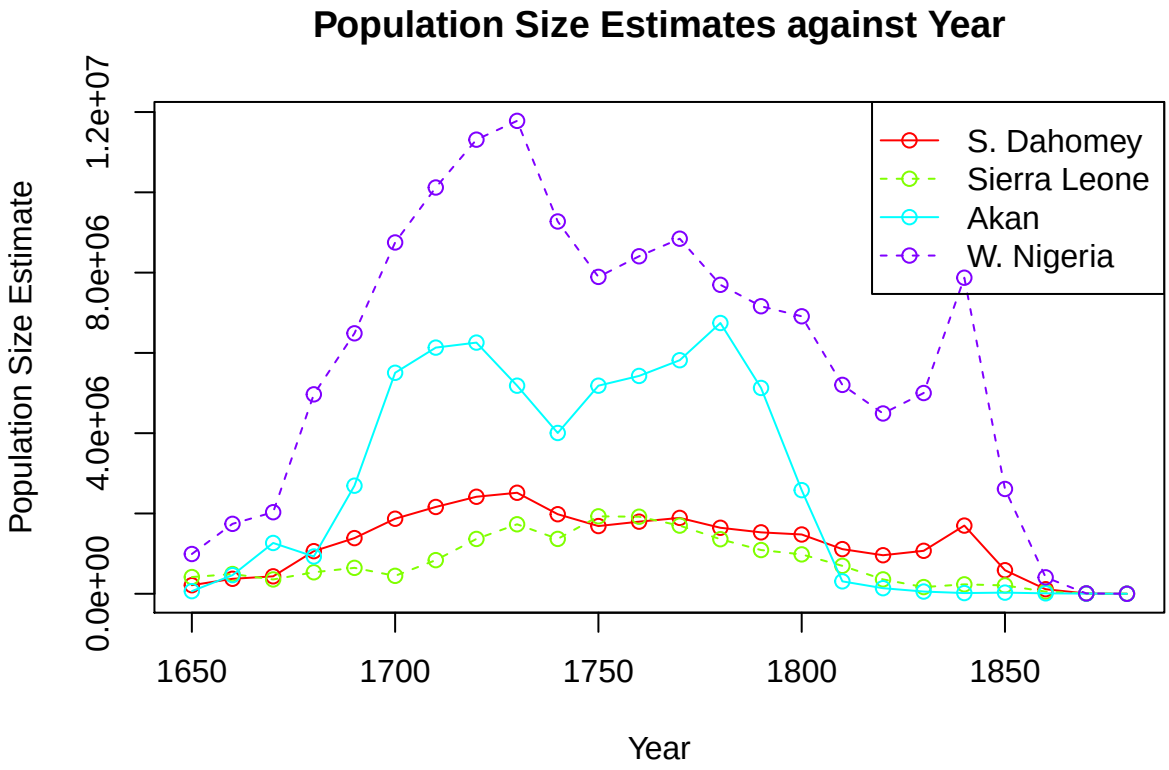


Table 1: Estimates

	x
$\text{Var}\left[S_i(0)\right]$	3.388
σ_η	0.287

Question 10



Question 11

```
internal.jitter <- function(data) {
    jitter <- function(x) { rnorm(n=length(x)-2, mean=head(x,-2),
                                  sd=x["exportErrorMargin"]) }
    data[1:(nrow(data)-2),] <- apply(data, 2, jitter)
    return(data)
}
```

The code above simulates the noise from measuring $R_i(t)$ to get an observed $X_i(t)$, hence it implements the simulation of $\epsilon_i(t)$. Assuming the measurement noise is reasonably modelled as gaussian, this would give us a good idea of the uncertainty in the estimates that will be consistent with the reported error margins.