1. Theory

Question 1

a) Given by the question:

- $ln\, u_i = \beta_0 + \beta_1\, ln\, b_i + v_i$
- u_i is individual i`s unemployment duration
- b_i denotes the level of unemployment benefits for which individual i was eligible upon loosing her job
- v_i is the error term

i. The use of log-estimation could have many reasons. One example to use it could be to transform a non-linear regression into a linear one. Also, if one thinks the distribution of the residual seems to be uncertain, the log transformation can lead to more normal-like distribution. In our example the interpretation of the results could be better described with logs as the interpretation of duration of unemployment based on benefits is expressed in percentage changes. In economics to estimate elasticities it is common to use the log-log regression model since it shows the relation between a 1% change in the level of unemployment benefits (X-Variable) and a 1% change in the unemployment duration (Y-Variable).

ii. On general one can say that the log transformation leads the regression to a reduction in the data and so it increases the standard errors. Another problem could be the interpretation since the coefficients no longer reflects the unit changes (only percentage changes). Also, it gives less weight to outliners, thus, could affect the estimation of the slope of β_1.

b) Some economists generally question models based on cross-section data since there is often unobserved heterogeneity across sections. There is only an observation of a one-time horizon and with IV (uncorrelated with the input with explanatory effect on the output) one could find a way around endogeneity in such cross-section data baseline. But not everything can be solved with IV. There could also be some unobserved heterogeneity between individuals or their behavior which are biasing the estimation it is correlated to employment benefits as well as unemployment duration. All this might lead our regression model from question to endogeneity and resulting in a biased estimate of the coefficients.

Tarik Benli 15-719-818, Ramon Gmür 16-705-220

c) Given by the question:

a. $ln\, u_{i,s} = \beta_0 + \beta_1\, ln\, \bar{b}_s + \beta_2 \mathrm{X}_{i,s} + v_{i,s}$

b. s indexes the state in which individual i lives

c. $\bar{b}_s$ is the average level of unemployment benefit in state s

d. $\mathrm{X}_{i,s}$ are individual controls (such as age, past employment history or education)

By taking the average benefit within a state instead of the individual benefits, one could solve some endogeneity problems. The model with the average benefits controls for differences for individual characteristics which are the same within individuals of a state but differ across states. Without labeling the average state differences in the regression they would be contained in the error term and could lead the regression into violation of assumption of uncorrelated unobservables.

By including as much variables as possible to the regression one could get rid of some OVB which are caused by endogeneity problems arising from individuals. Controlling for individual characteristics in the regression helps solving part of the endogeneity problem. But not every problem (for example measurement errors) can be solved by controlling for individual characteristics or with the average benefits method (controlling for state-specific effects).

d) There will be for sure still some endogeneity problematics in our model even by controlling for individual characteristics or state-specific effects as it is done in the regression model from c). One can still have measurement errors or reversed causality. Some economists will always have reasons to question this kind of regression models. One cannot observe everything and if it is observable it could be difficult to put those in the regression in a reasonable way and the interpretation of the results could suffer.

e) If one can assume that the benefits are uncorrelated with other unobserved variables and by using the controlling elements from c), one can say that the coefficient β_1 (and the elasticity too) is consistent. The partial elasticity can be read in the coefficient of β_1 since it is still a log-log regression model. It can be interpreted as a one percentage increase of the average benefits which effects individual duration of unemployment by β_1 percentage points. To get more consistent results, one could make a time series by record more data in

Tarik Benli 15-719-818, Ramon Gmür 16-705-220

other time periods and put in the regression time fixed effects. This would allow to get rid of the individual heterogeneity, because it does not vary across time.

Question 2

f) Given by the question:

a. Government decision: $ln\,\bar{b}_s = \gamma_0 + \gamma_1 ln\hat{u}_s + v_s$ **(3)**

i. $\hat{u}_s$ is the predicted average unemployment duration

ii. γ_0 and γ_1 are unknown parameters

iii. v_s is an approximation error satisfying $E(v_s) = 0$

b. $\hat{u}_s = \bar{u}_s \eta_s$ **(4)**

i. The prediction error in the average unemployment duration (η_s) is multiplicative

ii. $\bar{u}_s$ is the true average unemployment duration (which is only observed after the government sets $\bar{b}_s$

iii. $\eta_s > 0, E(ln\eta_s) = 0$

c. $ln\,\bar{u}_s = \beta_0 + \beta_1 ln\bar{b}_s + X_s\Gamma + \varepsilon_s$ **(5)**

i. $ln\,\bar{u}_s$ is the log average unemployment duration

ii. $E(\varepsilon_s|X_s) = 0$

iii. X_s contains other economic variables that determine the unemployment rate duration

iv. v_s and η_s are independent from each other and from X_s and ε_s

First step: (4) → (3)

$$ln\,\bar{b}_s = \gamma_0 + \gamma_1 ln\,(\bar{u}_s\eta_s) + v_s \quad \textbf{(6)}$$

In the next step solve (5) by taking $ln\bar{b}_s$ on the left-hand side:

$$ln\,\bar{u}_s = \beta_0 + \beta_1 ln\bar{b}_s + X_s\Gamma + \varepsilon_s$$

$$\beta_1 ln\bar{b}_s = ln\,\bar{u}_s - \beta_0 - X_s\Gamma - \varepsilon_s$$

$$ln\bar{b}_s = \frac{ln\,\bar{u}_s - \beta_0 - X_s\Gamma - \varepsilon_s}{\beta_1} \quad \textbf{(7)}$$

Now, set (6) and (7) equal:

$$\gamma_0 + \gamma_1\,ln(\bar{u}_s\eta_s) + v_s = \frac{ln\,\bar{u}_s - \beta_0 - X_s\Gamma - \varepsilon_s}{\beta_1}$$

$$\gamma_0\beta_1 + \gamma_1\beta_1\, ln\bar{u}_s + \gamma_1\beta_1 ln\eta_s + \beta_1 v_s = ln\,\bar{u}_s - \beta_0 - X_s\Gamma - \varepsilon_s$$

Solve now for $ln\bar{u}_s$:

$$ln\bar{u}_s - \gamma_1\beta_1 ln\bar{u}_s = \gamma_0\beta_1 + \beta_0 + X_s\Gamma + \varepsilon_s + \gamma_1\beta_1 ln\eta_s + \beta_1 v_s$$

$$ln\bar{u}_s = \frac{\gamma_0\beta_1 + \beta_0 + X_s\Gamma + \varepsilon_s + \gamma_1\beta_1 ln\eta_s + \beta_1 v_s}{1 - \gamma_1\beta_1}$$

$$ln\bar{u}_s = \frac{\beta_0 + \gamma_0\beta_1}{1 - \gamma_1\beta_1} + \frac{X_s\Gamma}{1 - \gamma_1\beta_1} + \frac{\gamma_1\beta_1 ln\eta_s}{1 - \gamma_1\beta_1} + \frac{\beta_1 v_s + \varepsilon_s}{1 - \gamma_1\beta_1}$$

g)

First step, solve (6) for $ln\bar{u}_s$:

$$ln\bar{u}_s = \frac{ln\,\bar{b}_s - \gamma_0 - \gamma_1 l\,n\,\eta_s - v_s}{\gamma_1} \textbf{ (8)}$$

In the next step set (5) and (8) equal:

$$\frac{ln\,\bar{b}_s - \gamma_0 - \gamma_1 l\,n\,\eta_s - v_s}{\gamma_1} = \beta_0 + \beta_1 ln\bar{b}_s + X_s\Gamma + \varepsilon_s$$

$$\frac{1}{\gamma_1} ln\,\bar{b}_s - \beta_1 ln\bar{b}_s = \beta_0 + X_s\Gamma + \varepsilon_s + \frac{\gamma_0}{\gamma_1} + l\,n\,\eta_s + \frac{v_s}{\gamma_1}$$

$$(1 - \gamma_1\beta_1)\, ln\,\bar{b}_s = \gamma_1\beta_0 + \gamma_1 X_s\Gamma + \gamma_1\varepsilon_s + \gamma_0 + \gamma_1 l\,n\,\eta_s + v_s$$

$$ln\,\bar{b}_s = \frac{\gamma_1\beta_0 + \gamma_0}{1 - \gamma_1\beta_1} + \frac{\gamma_1 l\,n\,\eta_s}{1 - \gamma_1\beta_1} + \frac{\gamma_1 X_s\Gamma}{1 - \gamma_1\beta_1} + \frac{v_s + \gamma_1\varepsilon_s}{1 - \gamma_1\beta_1}$$

And now signing the asymptotic bias:

$$E(ln\,\bar{b}_s\varepsilon_s) = E\left(\left(\frac{\gamma_1\beta_0+\gamma_0}{1-\gamma_1\beta_1} + \frac{\gamma_1 l\,n\,\eta_s}{1-\gamma_1\beta_1} + \frac{\gamma_1 X_s\Gamma}{1-\gamma_1\beta_1} + \frac{v_s+\gamma_1\varepsilon_s}{1-\gamma_1\beta_1}\right)\varepsilon_s\right) = 0 + E\left(\frac{v_s+\gamma_1\varepsilon_s}{1-\gamma_1\beta_1}\varepsilon_s\right) = 0 +$$

$$E\left(\left(\frac{v_s}{1-\gamma_1\beta_1} + \frac{\gamma_1\varepsilon_s}{1-\gamma_1\beta_1}\right)\varepsilon_s\right) = 0 + \frac{\gamma_1}{1-\gamma_1\beta_1}E(\varepsilon_s\varepsilon_s) = \frac{\gamma_1}{1-\gamma_1\beta_1}\sigma^2_{\varepsilon_s} > 0 \rightarrow \text{positive bias!!!}$$

because $\gamma_1 < 0, \beta_1 > 0, \sigma^2_{\varepsilon_s} > 0$

h) One can consider another strategy by taking the predicted average unemployment duration as a determinant for the level of benefit. With the reduced form equation of $ln\,\bar{u}_s$, the unemployment duration elasticity could be easily measured.
The limitations of this idea of measuring the elasticity is that we do not have a timeseries. We have only a cross-section data base. So one cannot control for unobserved

Tarik Benli 15-719-818, Ramon Gmür 16-705-220

heterogeneity between states. For that we would need more state-specific characteristics included to the regression to clear out all state differences.

i) The government has the incentive to reduce the expenditures. To get a minimum expenditure the government has to reduce the unemployment duration on one hand and in the other one it strives for a minimization in the benefits for unemployed people. It could get there by setting $\bar{b}_s$ as a function of $\hat{u}_s$. So, it can measure the elasticity of $\hat{u}_s$. By knowing the elasticity, the government can obtain the minimal $\bar{u}_s$ with a minimum level of $\bar{b}_s$.

2. Empirical Question

a) See results below

VARIABLES	(1) pooled
bin_1	0.00376***
	(0.000657)
bin_2	0.0119***
	(0.000976)
bin_3	0.00317***
	(0.000657)
bin_4	0.00636***
	(0.000457)
bin_5	0.00352***
	(0.000296)
bin_6	0.00387***
	(0.000441)
bin_8	0.00315***
	(0.000360)
bin_9	0.00241***
	(0.000273)
bin_10	-0.0104***
	(0.000911)
Constant	4.198***
	(0.00667)
Observations	26,460
R-squared	0.089

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

By looking at monthly averages, one might omit extreme days within the month (especially hot or cold temperatures) because other days with more average temperatures could make the statistical outliers disappear.
The coefficient of the hottest temperature bin is -0.01, meaning that spending one additional day in this hottest bin compared to the 60°F-69°F-bin leads to a decrease of the death rate of 1%. This coefficient does not have the sign we expected. We expected one additional day in this hottest bin to lead to an increase in the death rate.

b)

Additional to the usual OLS assumptions, for this pooled OLS coefficient to be consistent we need two additional assumptions:

- Contemporaneous Exogeneity: $Cov(x_{it}, \epsilon_{it}) = 0 \; for \; all \; t = 1, \ldots, T$

- Uncorrelated Effects: $Cov(x_{it}, \alpha_i) = 0 \; for \; all \; t = 1, \ldots, T$

In this context, Uncorrelated Effects is the crucial assumption. This means that the average days spent in a specific temperature bin should not be correlated with a state's specific (but constant over time) characteristics.

However, this assumption is likely to be violated. The reason is that average days in a specific temperature bin is definitively correlated with a state's specific characteristics. For example: a large country like the States has different climates in different regions. These different climates are definitively correlated with the average days spent in a specific temperature bin.

Random effects would not solve this problem. RE relies on the same assumptions as pooled OLS, however, contemporaneous exogeneity is not enough anymore. RE needs Strict Exogeneity to hold. But since we already argued that Uncorrelated Effects is violated, RE would not work.

c) See results below

VARIABLES	(1) s&m-FE
bin_1	0.00136***
	(0.000388)
bin_2	0.00506***
	(0.000522)
bin_3	0.00194***
	(0.000369)
bin_4	0.00212***
	(0.000271)
bin_5	0.00226***
	(0.000190)
bin_6	0.00125***
	(0.000239)
bin_8	-0.000847***
	(0.000195)
bin_9	-0.00121***
	(0.000188)
bin_10	-0.00118**
	(0.000525)
Constant	4.344***
	(0.00566)
Observations	26,460
Number of stfips	49
R-squared	0.374
State FE	YES
Month FE	YES

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The coefficient in this model with state FE and month FE for the hottest temperature bin is -0.001 and statistically significant on the 5%-level. This suggests that one additional day spent in the hottest temperature bin compared to 60°F-69°F-bin leads to a decrease in the mortality rate of 0.1%. In this model, the average state temperature and the monthly average temperature (represented by average days spent in a bin) are de-meaned.

This coefficient does not recover the causal impact of temperature on log mortality. The reason for that is the following: By separating the Fixed Effects for states and months, we are de-meaning the monthly average temperature for the USA as a whole, not for the specific states, and we are de-meaning the average state temperature for the whole time period, not for the month individually.

d) See results below

VARIABLES	(1) s per m-FE
bin_1	0.00432***
	(0.000501)
bin_2	0.00666***
	(0.000586)
bin_3	0.00422***
	(0.000433)
bin_4	0.00421***
	(0.000346)
bin_5	0.00285***
	(0.000281)
bin_6	0.00112***
	(0.000313)
bin_8	-3.08e-06
	(0.000271)
bin_9	0.000995***
	(0.000317)
bin_10	0.00250**
	(0.00103)
Constant	4.417***
	(0.0129)
Observations	26,460
Number of stfips	49
State/Month FE	YES

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The statistically significant (on the 5%-level) coefficient of the hottest temperature bin is 0.0025, meaning spending an additional average day in the hottest temperature compared to the 60°F-69°F-bin leads to an increase in the mortality rate of 0.25%. The FE added in this model de-mean the state's average temperature for each month.

e)

The estimated equation looks as follows:

$$\ddot{y}_{it} = \ddot{x}'_{it}\beta + \ddot{\epsilon}_{it}$$

where $\ddot{y}_{it}$ is the de-meaned log mortality rate in state *i* at time *t*, $\ddot{x}'_{it}$ is a vector containing the de-meaned average days spent in a specific temperature bin (1-10) in state *i* at time *t* and $\ddot{\epsilon}_{it}$ is the de-meaned error term.

In order for our estimator to be consistent, we need the following assumptions:

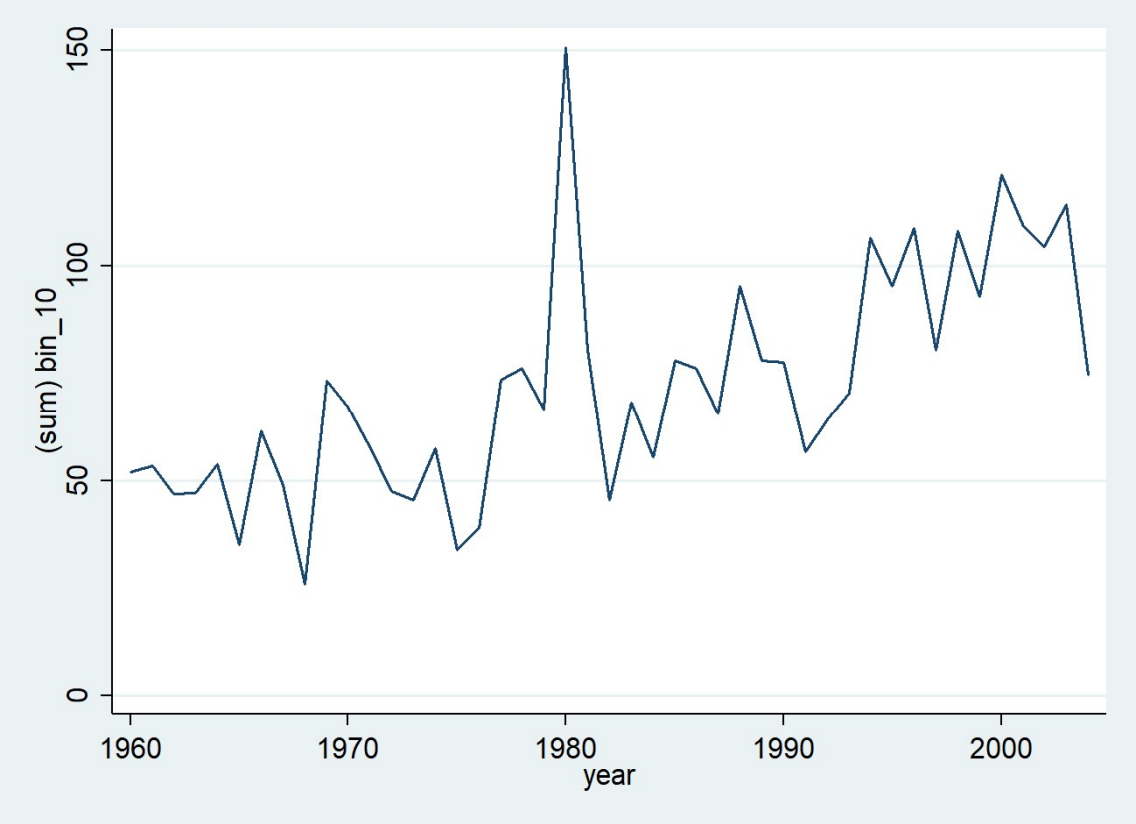
- Strict Exogeneity: $Cov(\ddot{x}_{is}, \ddot{e}_{it}) = 0 \; for \; all \; s, t = 1, \ldots, T$

- Arbitrary Effects (not really an assumption): No restrictions in the relationship between x_{it} and α_i.

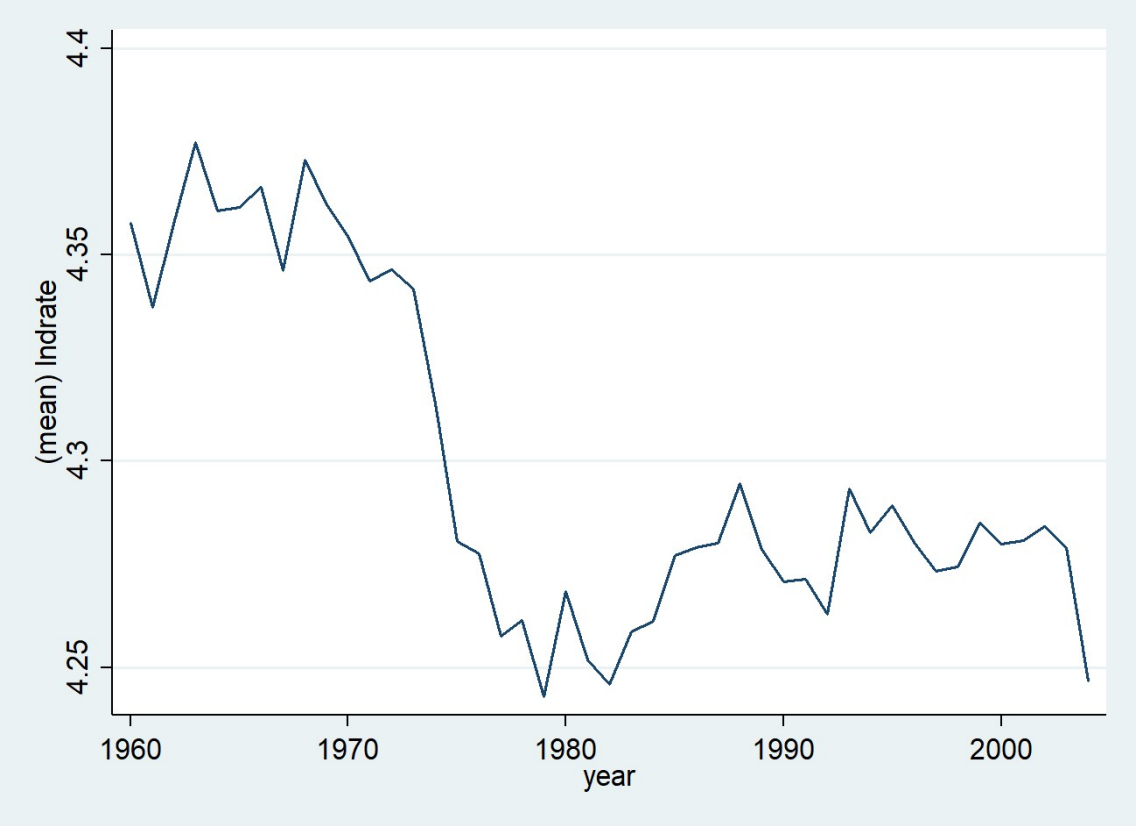
We expect the Strict Exogeneity to be violated (as itself it is already a really strong assumption). There is for example definitively a correlation between the average days spent in a specific temperature bin and unusual precipitations.

f)

i)



ii)



g) See results below

VARIABLES	(1) s per m-FE
bin_1	0.00253***
	(0.000443)
bin_2	0.00522***
	(0.000517)
bin_3	0.00241***
	(0.000383)
bin_4	0.00326***
	(0.000306)
bin_5	0.00258***
	(0.000248)
bin_6	0.00113***
	(0.000276)
bin_8	1.85e-05
	(0.000239)
bin_9	0.00256***
	(0.000283)
bin_10	0.00653***
	(0.000911)
devp25	0.00352***
	(0.000998)
devp75	-0.00425***
	(0.000967)
year	-0.458***
	(0.0105)
year2	0.000115***
	(2.64e-06)
Constant	460.9***
	(10.36)
Observations	26,460
Number of stfips	49
State/Month FE	YES

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Compared to d), the coefficient on the hottest temperature bin increases to 0.0065 and is statistically significant on the 1%-level, meaning according to this model, the mortality rate increases by 0.65% if one additional day is spent in the hottest bin compared to the 60°F-69°F-bin. Looking at ii), the inclusion of year and year^2 makes sense, as it controls for a decrease in the log mortality rate. If one looks at i), one can also explain why bin_10 yielded negative results until d). These facts suggest OVB in d).

h) See results below

VARIABLES	(1) MVA	(2) CVD
bin_1	-0.00637***	0.0190***
	(0.00221)	(0.00118)
bin_2	0.00569**	0.0166***
	(0.00259)	(0.00138)
bin_3	0.00464**	0.0132***
	(0.00191)	(0.00102)
bin_4	0.00568***	0.0105***
	(0.00153)	(0.000817)
bin_5	0.00346***	0.00562***
	(0.00124)	(0.000663)
bin_6	0.00235*	0.00316***
	(0.00138)	(0.000738)
bin_8	-0.00206*	-0.00135**
	(0.00119)	(0.000639)
bin_9	-0.00560***	-0.00486***
	(0.00140)	(0.000747)
bin_10	-0.0343***	-0.0108***
	(0.00454)	(0.00243)
Constant	-6.282***	-3.398***
	(0.0571)	(0.0305)
Observations	26,453	26,460
Number of stfips	49	49
State/Month FE	YES	YES

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

For motor-vehicle accident, the coefficient for the hottest temperature bin is -0.034, suggesting a decrease of the mortality rate due to motor-vehicle accidents of 3.4% if a day is spent in bin_10 compared to bin_7. For cardiovascular diseases, the coefficient for the hottest temperature bin is -0.011, suggesting a decrease of the mortality rate due to cardiovascular diseases of 1.1%. Both coefficients are statistically significant on the 1%-level. While this makes sense for the motor-vehicle accidents (one rather stays at home than going for a joyride if it is too hot), it does not make sense for cardiovascular diseases, as people suffering from this are at risk at higher temperatures. These results do not add credibility to a causal interpretation in d), it does the opposite.

i)

The magnitude of the effect in g (3.4% and 1.1%) is rather large. Only one additional day in these temperature bins compared to the standard temperature bin decreases the mortality rate rather large. However, the results in g) are definitively biased, meaning even though the magnitude seems large, we should not interpret these effects as causal (actually, we should not interpret them at all, especially the CVD).

3. Log-file

See attachment