Homework 1

Joseph Froelicher

September 20, 2021

Question 1

I acknowledge the homework expectation set forth by the instructor in conjunction with the honor code for the Colorado School of Public Health.

Question 2

Part A

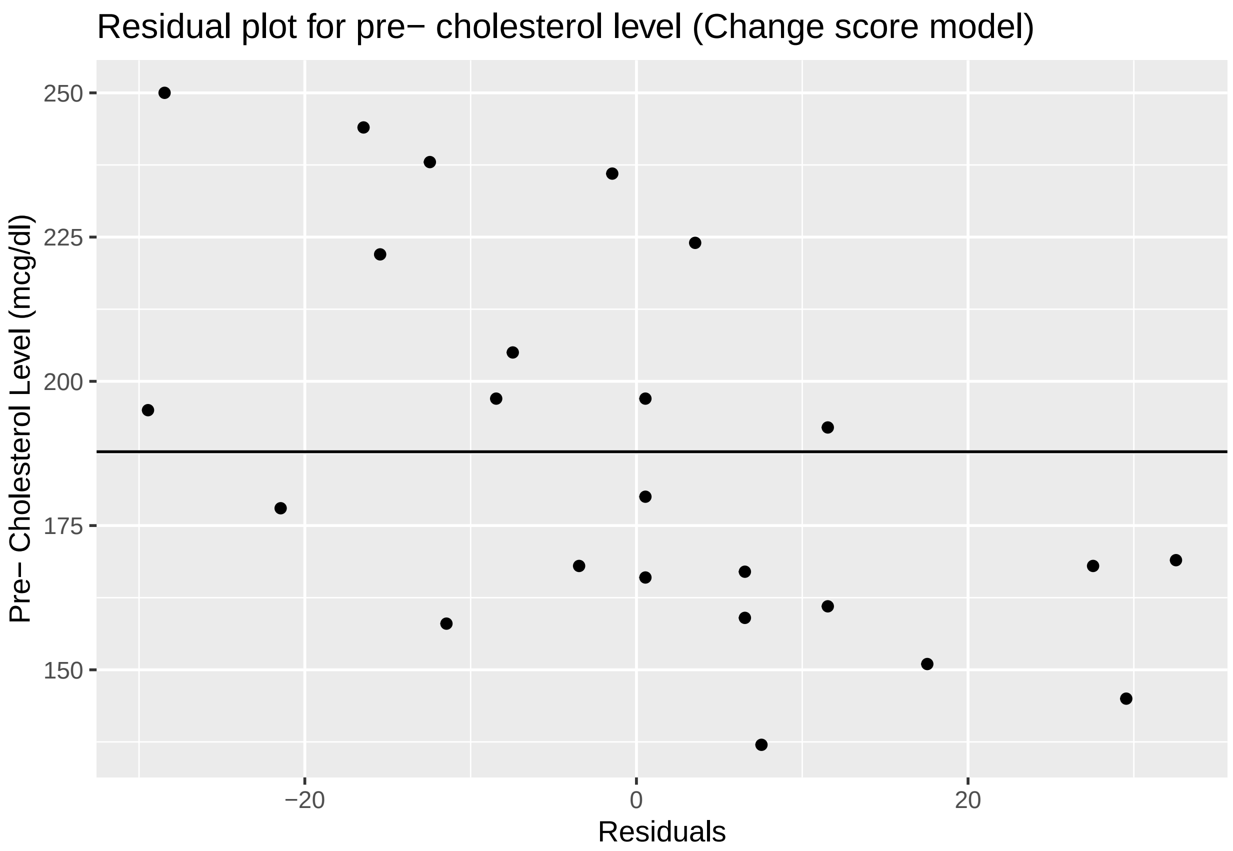
After a linear regression where the outcome was change in cholesterol level (mcg/dl), it was determined that the average difference between post- and pre-cholesterol level was significant ($\text{p} = 8 \times 10^{-6}$). The cholesterol level (mcg/dl) was 19.54 (mcg/dl) lower on average for the 24 hospital employees who switched from the standard American diet to a vegetarian diet for one month.

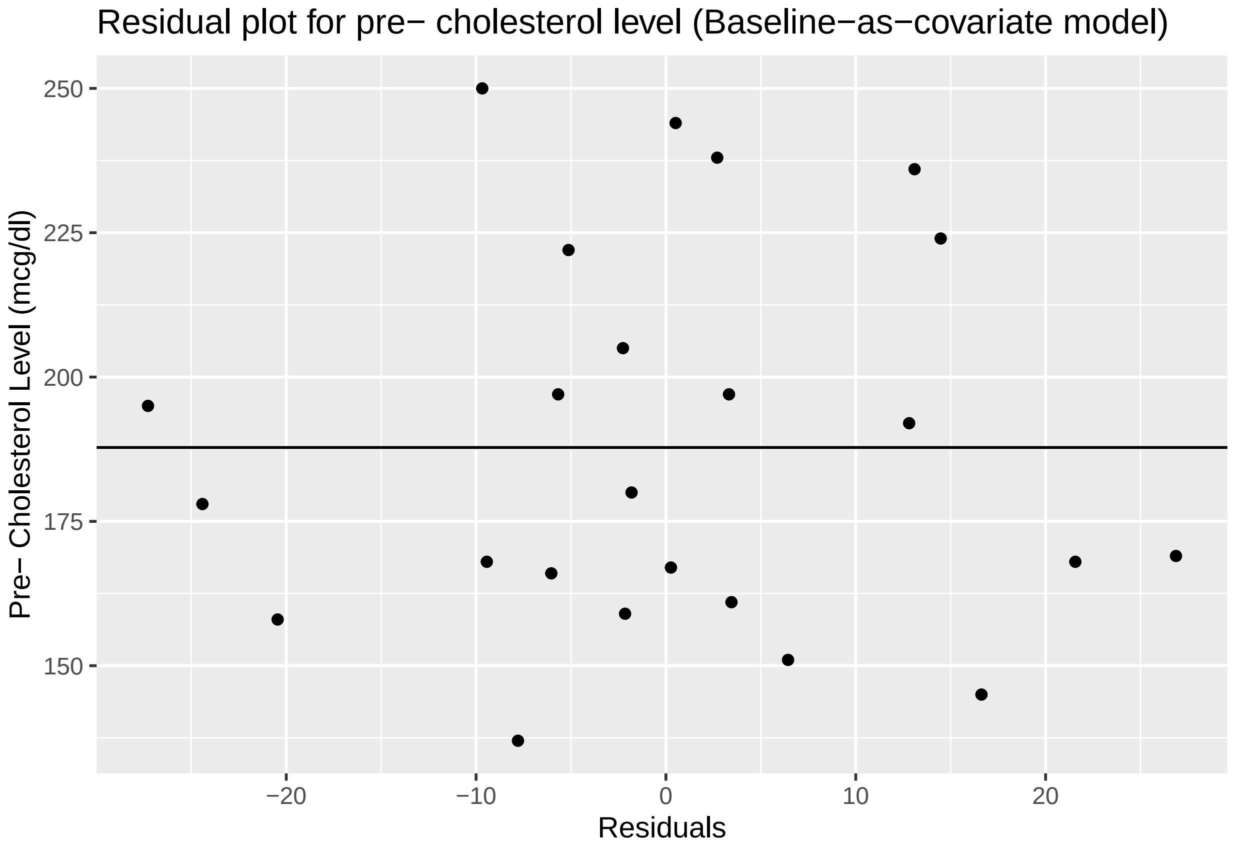
Part B

The simple test (see appendix) that could be performed to yield the same result as the previous is a paired t-test to detect differences between the two time points ($t = 5.6965187$).

Part C

After a linear regression where the outcome was pre- cholesterol level (mcg/dl), it was determined that the average post- cholesterol level was significant in the model of pre- cholesterol (mcg/dl) ($\text{p} = 5 \times 10^{-8}$). For a one mcg/dl increase in cholesterol before the start of the vegetarian diet, there was a corresponding increase of 0.698 mcg/dl in cholesterol after the conclusion of the vegetarian diet in 24 hospital employees.


Residual plot for pre− cholesterol level (Change score model)
Pre− Cholesterol Level (mcg/dl)
250
225
200
175
150
−20
0
20
Residuals



Part D

The change score model is very easy to convey interpretations (just as a simple t-test is). However, the baseline-as-covariate model does seem to have normally distributed residuals, suggesting maybe it is fitting the data a bit better. Additionally, the AIC value for the baseline-as-covariate model was significantly lower than the change-score model (198.0097704 < `r`AIC(model0)‘). The interpretation for the baseline-as-covariate model is slightly more complex. We are interpreting the baseline cholesterol level using the post cholesterol value as a predictor.

Part E

$$y_{i2} - y_{i1} = \beta_{intercept} + \beta_{after} x_{i2} + \epsilon_i$$

$$y_{i2} = \beta_{intercept} + \beta_{after} x_{i2} + \epsilon_i + y_{i1}$$

The differnece between the hybrid and the baseline-as-covariate model is the addition of the baseline cholesterol level on the right side of the equation. This should shift the intercept of the equation for each patient (the difference from a random intercept here is we are not estimating the random intercept term, rather we are providing it, the random intercept is the baseline cholesterol value). This seems to result in a more conservative model (lower estimate for the coefficient estimating the slope, and a larger p-value). It seems to me that we are not allowing for the baseline cholesterol to account for as much variation in the hybrid model than as in the baseline as covariate model.

$$H_0 : \hat{\beta}_1' = 0, H_1 : \hat{\beta}_1' \neq 0$$

Part F

Using unstructured covariance for the mixed model containing repeated measures yielded similar results to the Hybrid model in terms of statistical results. However, it did indeed yield a smaller standard error for the intercept estimate. It seems to me that in general the mixed model approach is a more robust modeling technique, however, in this specific scenario we do seem to be over complicating a simple problem. If we go all the way back to the mention of a *t*-test, we have carried the same result throughout the analysis, that cholesterol level was affected by the vegetarian diet. Nothing about the 4 models that followed provided any new clinical information. In general, unless there is some mathematical reason to use a more robust model (in the case of several repeated measures for example), we should stand by the simple, easy to interpret, and easy to convey model.

Question 3

a. Determine $E[\epsilon_t]$

$$E[\epsilon_t] = E\left[\sum_{j=0}^{\infty} \phi^j Z_{t-j}\right]$$

$$E[\epsilon_t] = \sum_{j=0}^{\infty} \phi^j E\left[Z_{t-j}\right]$$

$$E[\epsilon_t] = \sum_{j=0}^{\infty} \phi^j \ (0)$$

$$E[\epsilon_t] = 0$$

b. determine $Cov[\epsilon_t, \epsilon_{t+h}]$

$$Cov[\epsilon_t, \epsilon_{t+h}] = E[\epsilon_t \epsilon_{t+h}] - E[\epsilon_t]E[\epsilon_{t+h}]$$

$$Cov[\epsilon_t, \epsilon_{t+h}] = E[\epsilon_t \epsilon_{t+h}]$$

$$Cov[\epsilon_t, \epsilon_{t+h}] = E\left[\left(\sum_{j=0}^{\infty} \phi^j Z_{t-j}\right)\left(\sum_{k=0}^{\infty} \phi^k Z_{t+h-k}\right)\right]$$

$$Cov[\epsilon_t, \epsilon_{t+h}] = \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \phi^j \phi^k E[Z_{t-j} Z_{t+h-k}]$$

for $r \neq s$, $E[Z_r Z_s] = E[Z_r]E[Z_s] = 0$ for $r \neq s$, $E[Z_r Z_s] = Var[Z_r]E[Z_r^2] = \sigma_z^2$

$$Cov[\epsilon_t, \epsilon_{t+h}] = \sigma_z^2 \sum_{j=0}^{\infty} \phi^j \phi^{j+h}$$

$$Cov[\epsilon_t, \epsilon_{t+h}] = \sigma_z^2 \phi^h \sum_{j=0}^{\infty} (\phi^j)^2$$

$$\forall\, |\phi| \geq 1, Cov[\epsilon_t, \epsilon_{t+h}] \to \infty$$

$$\forall\, |\phi| < 1, Cov[\epsilon_t,\ \epsilon_{t+h}] = \frac{\sigma_z^2 \phi^h}{1 - \phi^2}$$

c. determine $Corr[\epsilon_t, \epsilon_{t+h}]$

$$\rho_h = Corr[\epsilon_t, \epsilon_{t+h}]$$

$$\rho_h = \frac{\gamma_h}{\gamma_0}$$

$$\rho_h = \frac{\frac{\sigma_z^2 \phi^h}{1-\phi^2}}{\frac{\sigma_z^2}{1-\phi^2}}$$

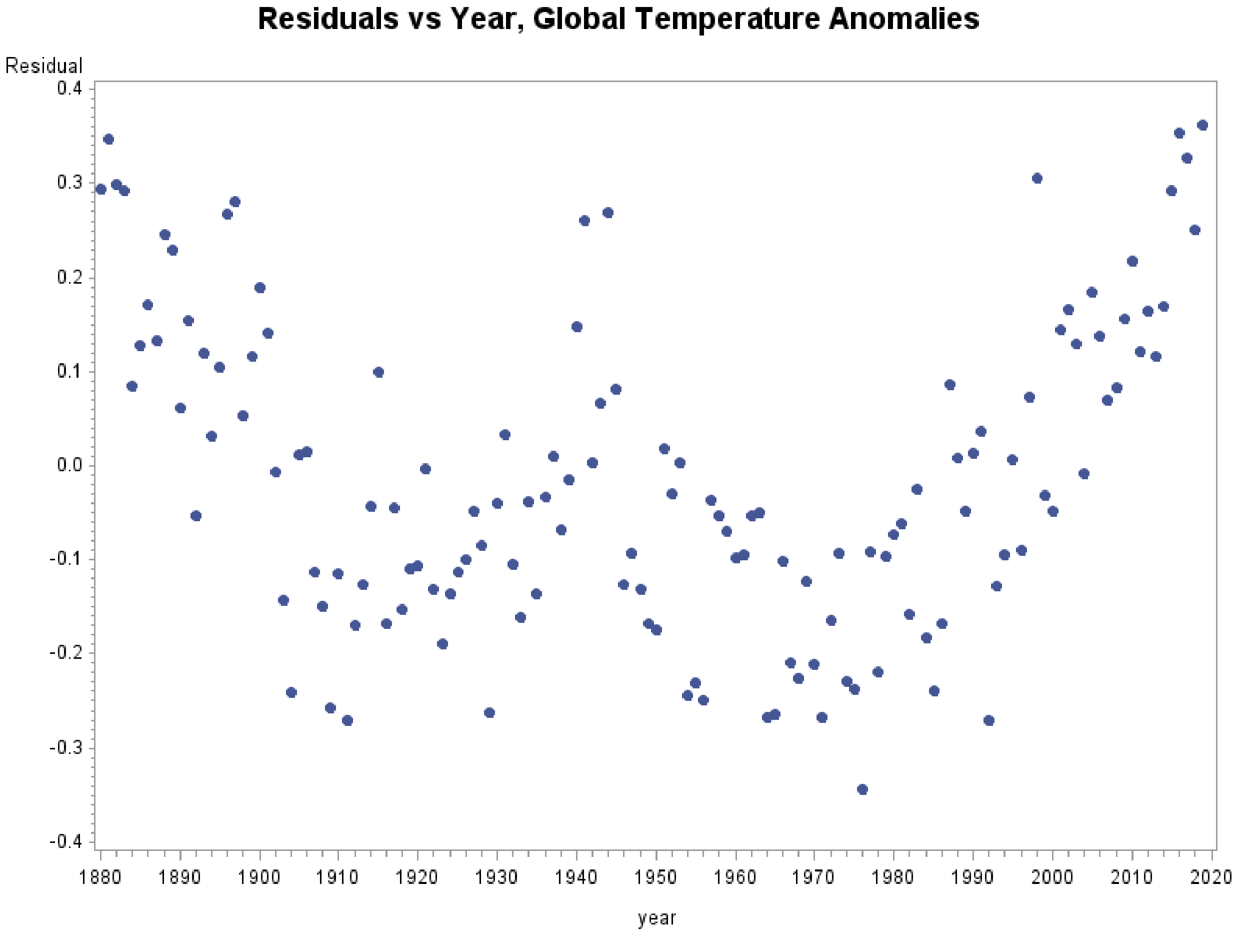
$$\rho_h = \phi^{|h|}$$

d. is $\{\epsilon_t\}$ a stationary process?

This process is mostly stationary. When $|\phi| < 1$, the mean and variance are finite and constant, and the autocorrelation only depends on h.

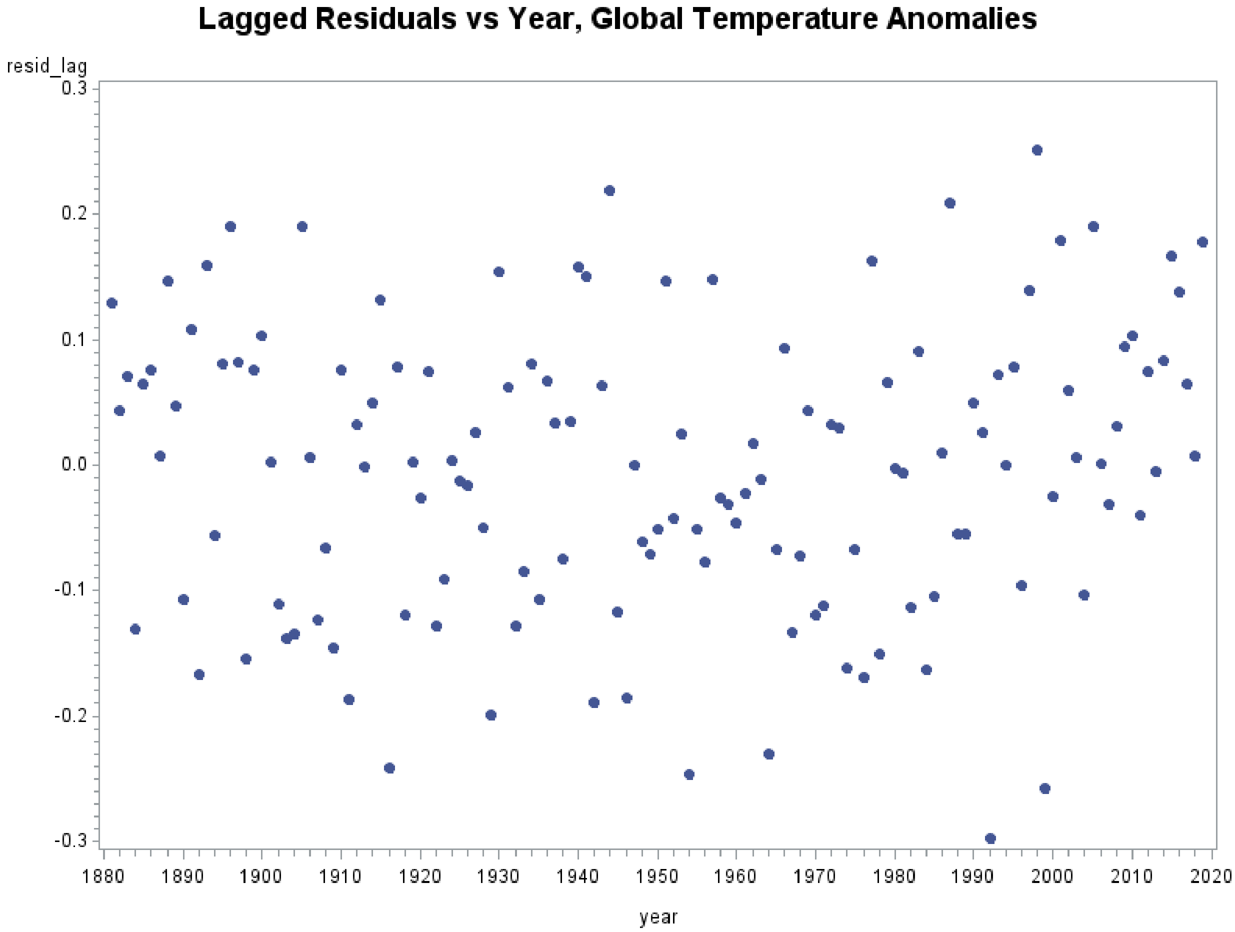
Question 4

Part A



The residuals do not have constant variance over time. It seems as though our assumption $\epsilon_i \sim N(0, \sigma)$ is violated in this model. A linear model does not seem to be fitting the data as well as it could be.

Part B



After accounting for correlation between time points, it seems as though there is a better spread about the data. This is reasonable, because we are scaling the residuals based on their previous time point. Our assumption of homoscedasticity seems to be met using this model. The errors are normally disdributed with equal variance about 0.

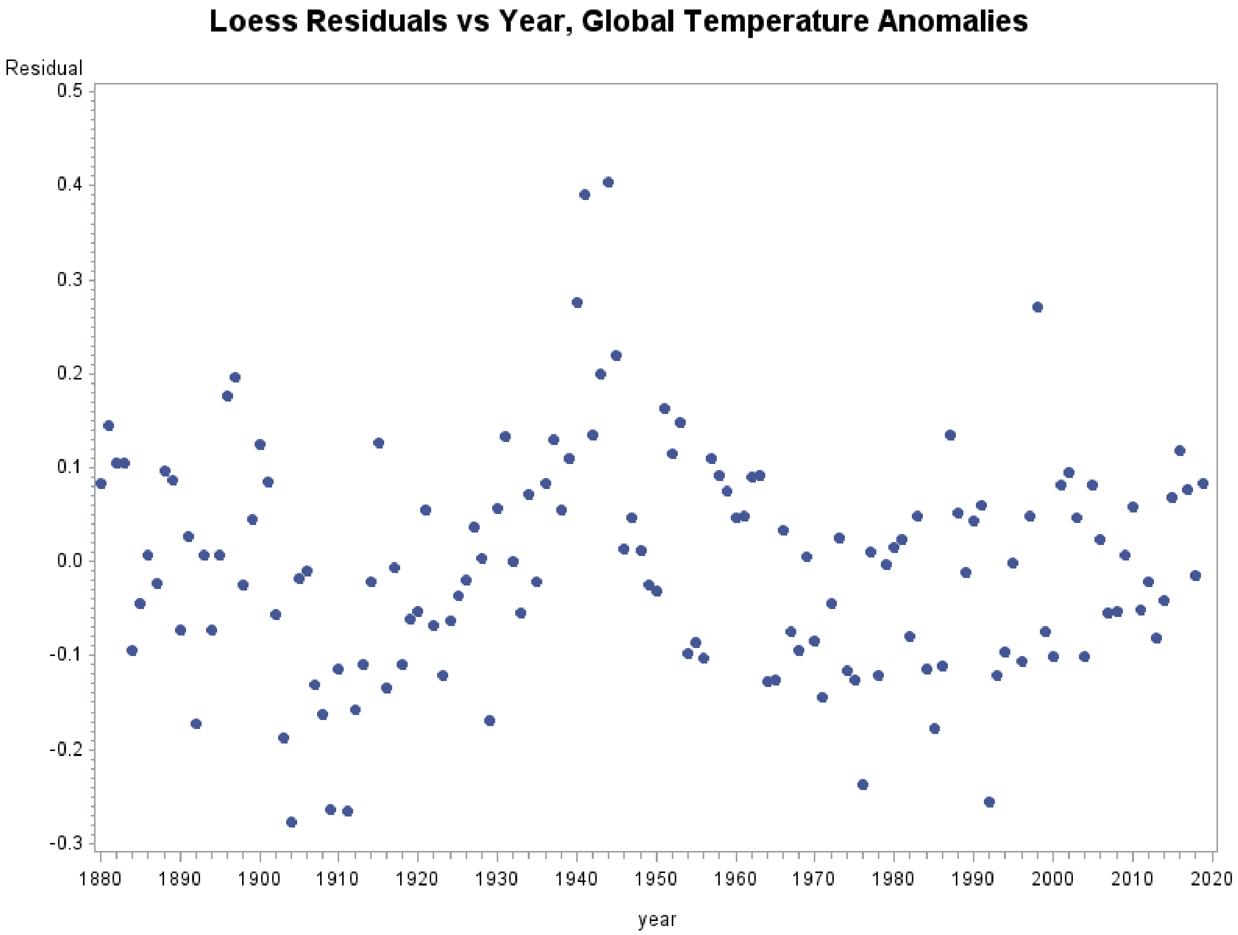
Part C

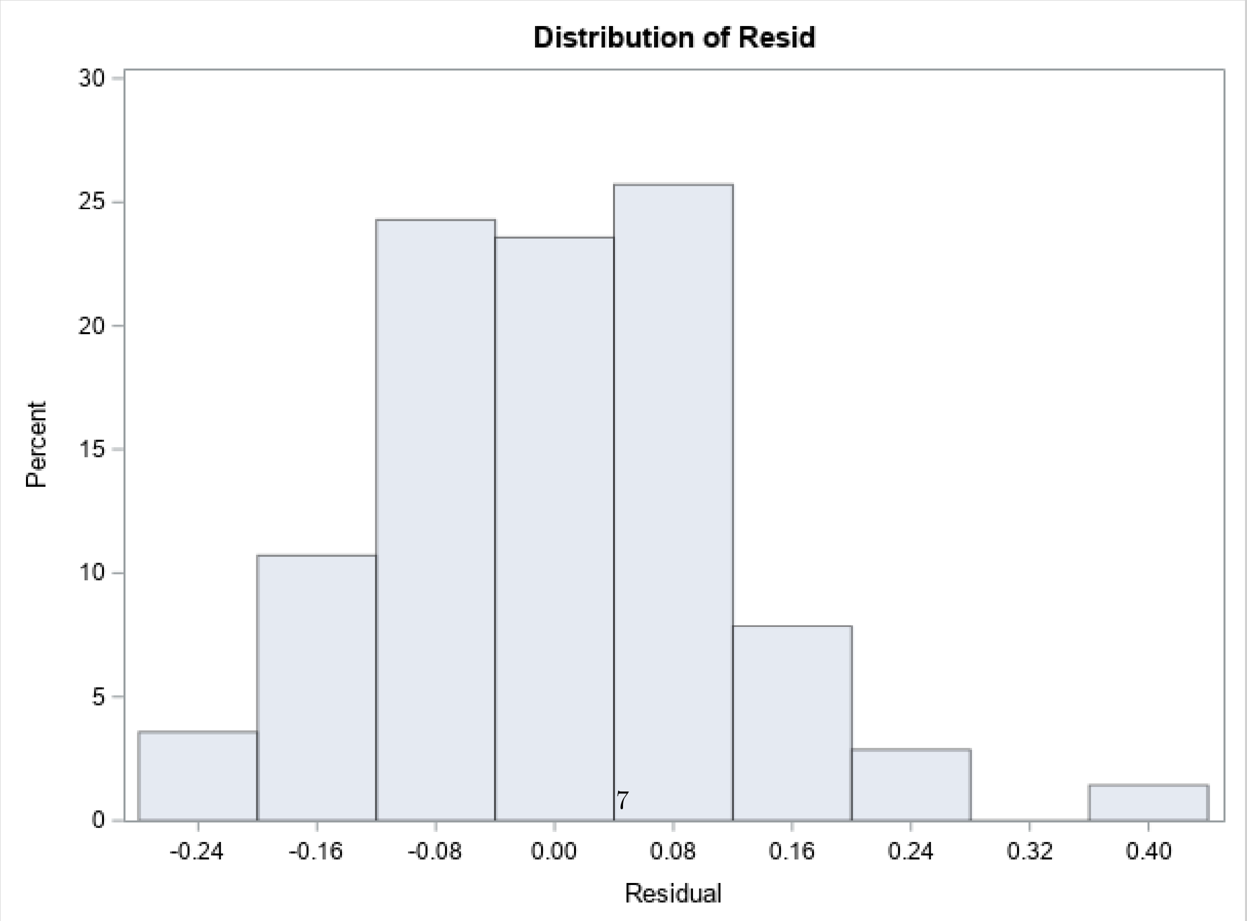
Based on the model fitted above, the average increase in temperature per decade is 0.07 degrees (0.007 per year x 10).

Part D

Based on looking at the plot, I chose a cubic function to model the time trend. The model fit in terms of information criterion is much better for the cubic function than it is for the linear time trend (>6 difference for AIC and BIC). Furthermore, the estimate for the correlation parameter has decreased from 0.72 to 0.47. This seems resonable because we are allowing for a more flesible model to explain changes between time points. This seems like a fairly good fit for this model. It terms of fit alone I would cetainly choose the cubic time trend model over the linear time trend model, it is explaining the variability in the data much better than the linear model. That being said, the cubic trend is extremely difficult to interpret, and would have to put more thought into interpreting that model.

Part E


Loess Residuals vs Year, Global Temperature Anomalies
Residual
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
year


Distribution of Resid
Percent
30
25
20
15
10
5
0
-0.24
-0.16
-0.08
0.00
0.08
0.16
0.24
0.32
0.40
Residual

In general this seems to be a pretty good fit. The errors are not much worse than the model with autoregressive correlation structure. This is also a different fit, and I don't know if we compare information criterion between parametric and non-parametric models. There is an obvious non-linear trend over time, and the non-parametric seems to be modeling this non-linear trend fairly well based on the residuals, we should just be careful about how we convey results from a non-parametric method.

Appendix

```r
library(ggplot2)

data = read.csv("../data/cholesterol.csv")

# part a
model0 = lm((after - before) ~ 1, data = data)
summary(model0)
AIC(model0)


# part b
result_t = t.test(data$before, data$after, paired = T)

# part c
model1 = lm(after ~ before, data = data)
summary(model1)
AIC(model1)

# part d
aov = anova(model1, model0)

ggplot(data, aes(x = model0$residuals, y = data$before)) +
  geom_point() +
  geom_hline(yintercept = mean(data$before)) +
  labs(
    x = "Residuals",
    y = "Pre- Cholesterol Level (mcg/dl)",
    title = "Residual plot for pre- cholesterol level (Change score model)"
  )

ggplot(data, aes(x = model1$residuals, y = data$before)) +
  geom_point() +
  geom_hline(yintercept = mean(data$before)) +
  labs(
    x = "Residuals",
    y = "Pre- Cholesterol Level (mcg/dl)",
    title = "Residual plot for pre- cholesterol level (Baseline-as-covariate model)"
  )
# part e
model2 = lm((after - before) ~ before, data = data)
summary(model2)

# # load the SASmarkdown package
# library(SASmarkdown)
# # set up the options so that knit knows where you SAS executable is
# # set the linesize to be easily readable on letter size paper, portrait
# # and set the knir options using opts_chunk$set().
# saspath <- "C:/Program Files/SASHome/SASFoundation/9.4/sas.exe"
# sasopts <- "-nosplash -linesize 75"
# knitr::opts_chunk$set(engine="sas", engine.path=saspath,
#         engine.opts=sasopts, comment=NA)
# # run these commands to convince yourself that
# # within this knitr session the engine changed.
# knitr::opts_chunk$get()$engine
# knitr::opts_chunk$get()$engine.path
# knitr::opts_chunk$get()$engine.opts

/* Question 2 */
/* Import data */
proc import datafile = 'C:\Users\jofro\dev\longi\data\cholesterol.csv'
 out = data1
 dbms = CSV;
run;

/* add subject variable to data */
data wide;
 set data1;
 subject+1;
run;

/*
    transpose data by subject:
    'let' transposes only the last observation for each 'by' group
    rename _name_ and col1
*/
proc transpose data=wide out=long (rename=(_name_=time col1=cholesterol)) let;
 by subject;
run;

/* linear mixed model with unstructured covariance */
proc mixed data = long method = ml;
 class time;
 model cholesterol = time / solution;
 repeated / type=UN;
run;

/* Question 4 */
/* import data */
proc import datafile = 'C:\Users\jofro\dev\longi\data\global_temp_anomalies.csv'
 out = data2
 dbms = CSV;
run;

/* mixed model with AR(1) structure */
proc mixed data = data2 method=ml plots=all;
 model temp=year / solution outp=tempout;
 repeated / type=AR(1) subject=intercept;
run;

/* plot the residulas vs time */
symbol1 v=dot i=none;
proc gplot data=tempout;
 plot resid*year;
 title 'Residuals vs Year, Global Temperature Anomalies';
```

```
run;

/* transform the residuals */
data lag;
 set tempout;
 y_new = pred + lag1(resid)*0.7395;
 resid_lag = temp - y_new;
run;

/* plot the transformed residuals */
symbol1 v=dot i=none;
proc gplot data=lag;
 plot resid_lag*year;
 title 'Lagged Residuals vs Year, Global Temperature Anomalies';
run;

/* transform the year variable to quadratic and cubic terms */
data poly;
 set data2;
 year_sq = year*year;
 year_cu = year*year*year;
run;

/* cubic mixed model */
proc mixed data=poly method=ml plots=all;
 model temp=year year_sq year_cu / solution outp=tempout;
 repeated / type=AR(1) subject=intercept;
run;

/* non-parametric model */
proc loess data=data2;
 ods output scoreresults=scoreout
     outputstatistics=statout;
 model temp = year / smooth= 0.3 residual clm degree=1;
 score data=tempout / clm;
run;

/* plot residuals from non-parametric model */
symbol1 v=dot i=none;
proc gplot data=scoreout;
 plot resid*year;
 title 'Loess Residuals vs Year, Global Temperature Anomalies';
run;
```