# Longitudinal Homework 1

*Tim Vigers*

*12 September 2019*

## 1. The simplest longitudinal analysis

### a. Change-score model

```
# Calculate change
chol$change <- chol$after - chol$before
# Model
change_mod <- lm(change ~ 1, data = chol)
```

**Results**

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -19.54167 | 3.430458 | -5.696519 | 8.4e-06 |

Regressing on the intercept essentially just calculates the average change score and tests whether or not this value is equal to 0.

### b. Simple test

The test on the intercept is the same as a simple one-sample t test on the change scores, or a paired t test on the before and after values.

```
t.test(chol$change)
```

```
##
##  One Sample t-test
##
## data:  chol$change
## t = -5.6965, df = 23, p-value = 8.435e-06
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -26.63811 -12.44522
## sample estimates:
## mean of x
## -19.54167
```

```
t.test(chol$after,chol$before,paired = T)
```

```
##
##  Paired t-test
##
## data:  chol$after and chol$before
## t = -5.6965, df = 23, p-value = 8.435e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -26.63811 -12.44522
```

```
## sample estimates:
## mean of the differences
##               -19.54167
```

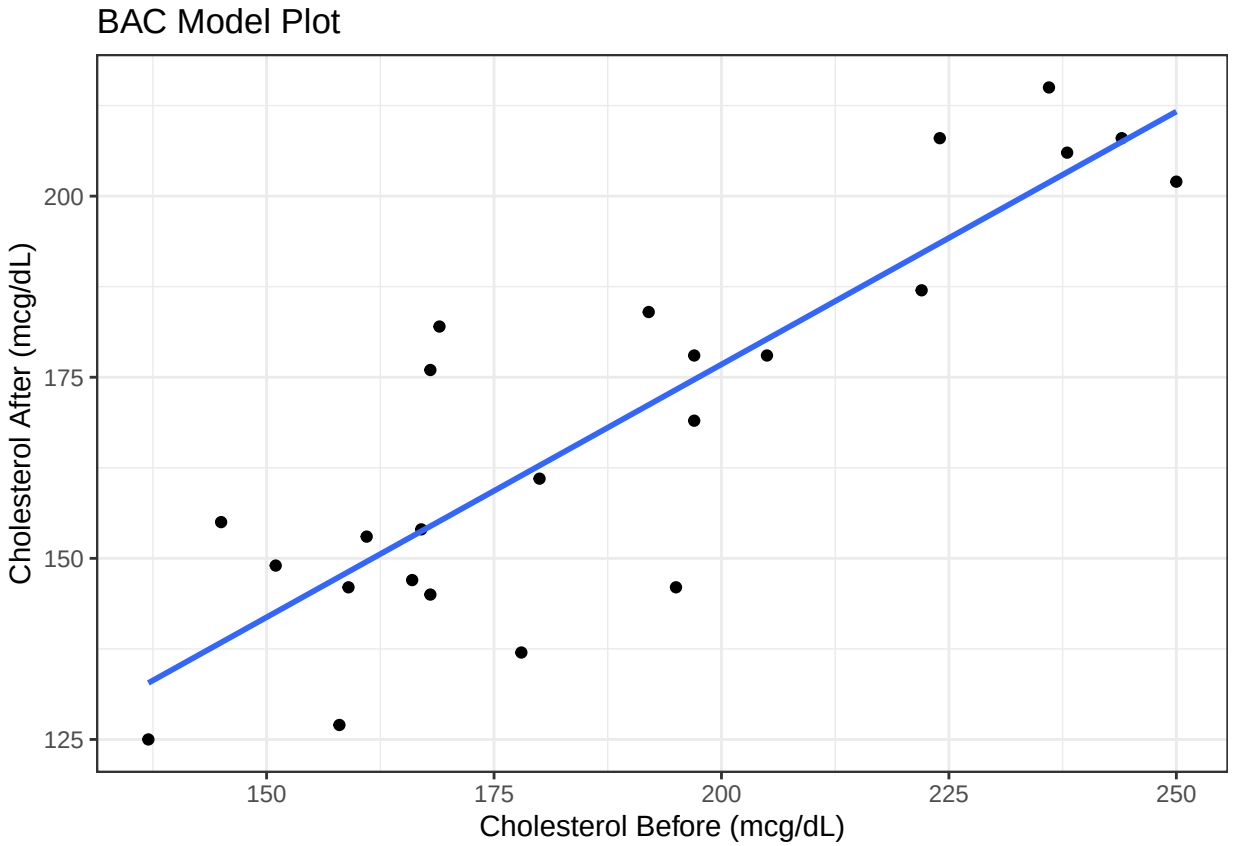
### c. Baseline-as-covariate model

```
baseline_mod <- lm(after ~ before, data = chol)
```

**Results**

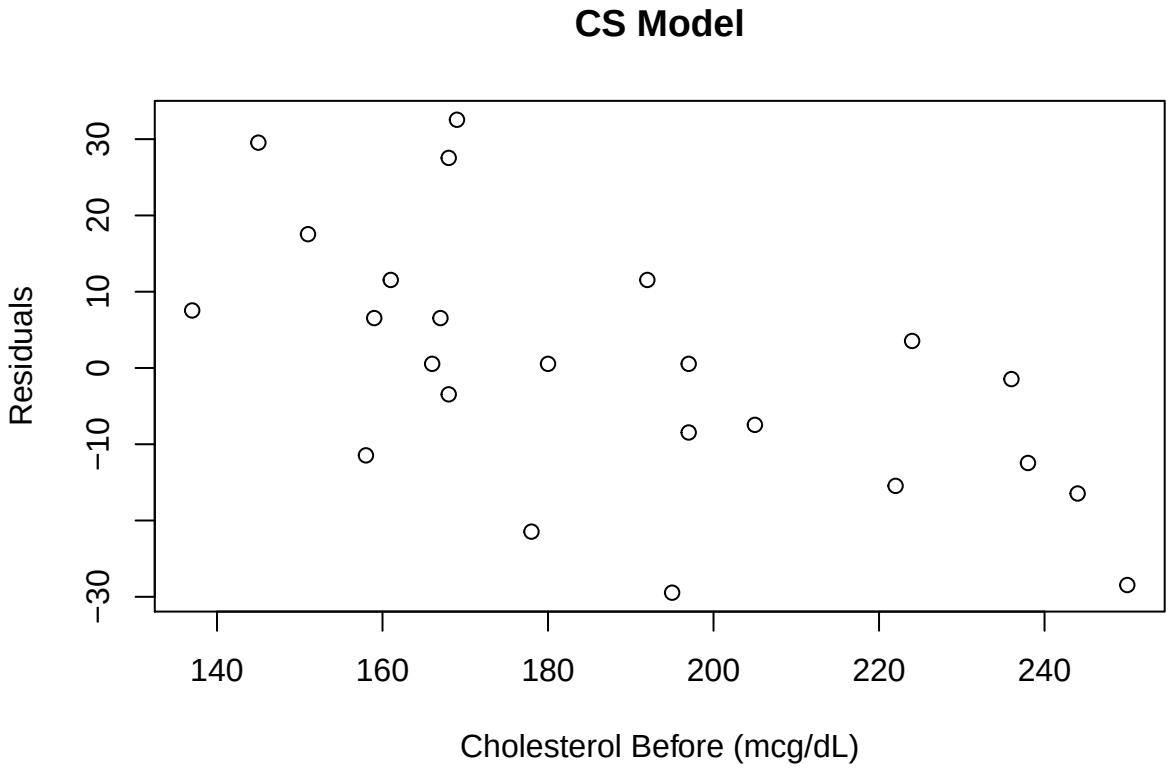
| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 37.1576120 | 16.5393736 | 2.246615 | 0.0350309 |
| before | 0.6980735 | 0.0867859 | 8.043624 | 0.0000001 |

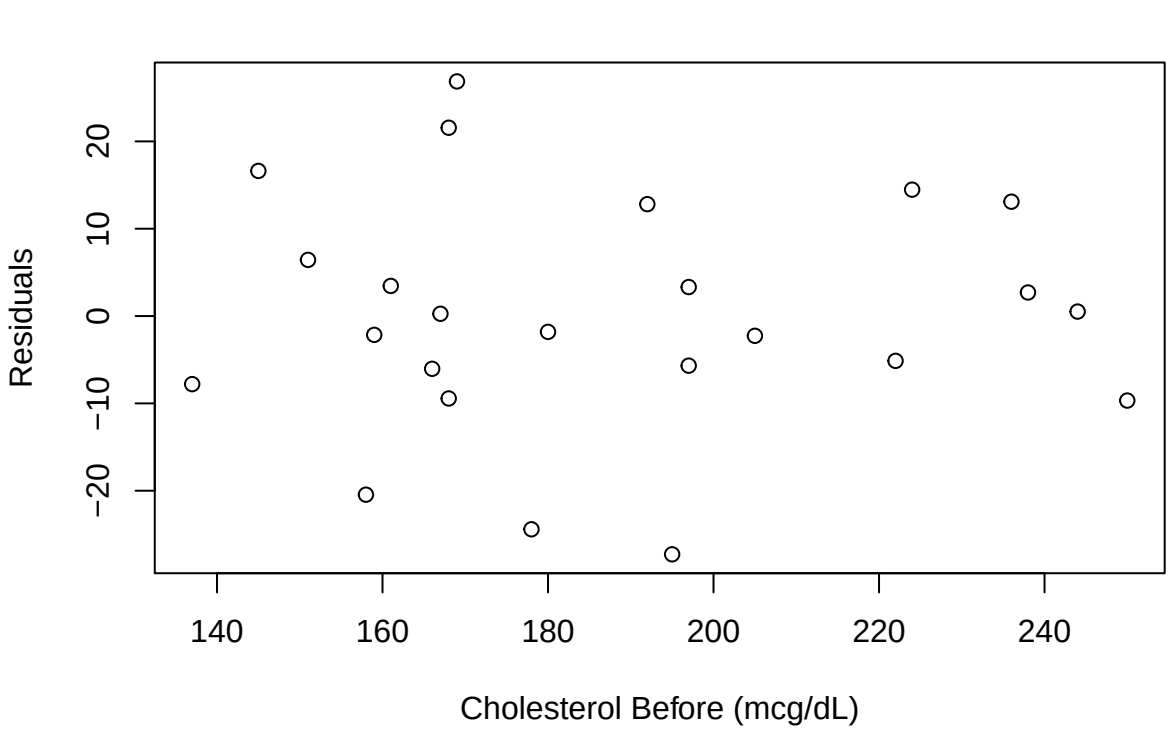
The results of this model indicate that for a theoretical starting cholesterol value of 0 mcg/dL, the average after value is 37.16 mcg/dL.For every one unit increase in the starting value, the after value increases by 0.70 (95% CI: 0.52 - 0.88, $p < 0.001$). Th intercept doesn't make much sense to interpret here, but because the slope for $\beta_1$ is less than 1 we can conclude that the vegetarian diet significantly lowered cholesterol (i.e. after was lower on average than before).



### d. Compare CS and BAC

**Plot the residuals**




The biggest advantage of the CS model is that the results are easy to interpret and explain, while the BAC model is a little bit trickier (for example if you only look at the intercept you might falsely conclude that

cholesterol was higher after the diet). However, in the residual plot for the CS model you can clearly see that there's an association between the residuals and the starting cholesterol value. The BAC model takes care of this association by adjusting for the baseline value, which makes it the preferable model overall (provided you feel comfortable interpreting it).

The CS model forces the baseline value to have a slope of 1, which is avoided in the BAC model:

Change score model:$Y_{i2} - Y_{i1} = \beta_0 + \epsilon_i Y_{i2} = Y_{i1} + \beta_0 + \epsilon_i$Baseline-as-covariate model:$Y_{i2} = \beta_0 + \beta_1 Y_{i1} + \epsilon_i$In the BAC model, $Y_{i1}$

## Hybrid model

### i. Beta coefficients and model fit

$$Y_{i2} - Y_{i1} = \beta_0 + \beta_1' Y_{i1} + \epsilon_i Y_{i2} = Y_{i1} + \beta_0 + \beta_1' Y_{i1} + \epsilon_i Y_{i2} = \beta_0 + Y_{i1}(\beta_1' + 1) + \epsilon_i \beta_1 = \beta_1' + 1 \beta_1' = \beta_1 - 1$$

Based on this, it's clear that the hybrid model with have the same $\beta_0$ , and that the slope of $Y_{i1}$ in the hybrid model is $\beta_1 - 1$ from the BAC model. You can confirm this using the model output.

Table 3: BAC model results

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 37.1576120 | 16.5393736 | 2.246615 | 0.0350309 |
| before | 0.6980735 | 0.0867859 | 8.043624 | 0.0000001 |

Table 4: Hybrid model results

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 37.1576120 | 16.5393736 | 2.246615 | 0.0350309 |
| before | -0.3019265 | 0.0867859 | -3.478979 | 0.0021287 |

### ii. Hypotheses

The null hypothesis for the test of the "before" variable is:

$$H_0\text{: } \beta_1' = (\beta_1 - 1) = 0 \text{ or } \beta_1 = 1$$

## f. Mixed model

```
mixed_mod <- lme(change ~ 1, data = chol,random = ~1|id)
```

**Results**

| | Value | Std.Error | DF | t-value | p-value |
|---|---|---|---|---|---|
| (Intercept) | -19.54167 | 3.430458 | 24 | -5.696519 | 7.2e-06 |

The mixed model with unstructured variance is the exact same as the linear model (in this case I used the change score model as a comparison). This is because the mixed model is just a special case of the GLM.

## 2. A first-order autoregressive process

### a. Expected value

First, expand the expected value:

$$E(\epsilon_t) = E(Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + ...)$$

Because $\phi$ is a constant and we assume the Zs to be independent, this becomes:

$$E(\epsilon_t) = E(Z_t) + \phi E(Z_{t-1}) + \phi^2 E(Z_{t-2}) + ...) = 0 + 0 + ... + 0$$

An infinite sum of zeroes is zero, so $E(\epsilon_t) = 0$

### b. Covariance

$Cov(\epsilon_t, \epsilon_{t+h}) = E(\epsilon_t \epsilon_{t+h}) - E(\epsilon_t)E(\epsilon_{t+h}) = E(\epsilon_t \epsilon_{t+h})E(\epsilon_t \epsilon_{t+h}) = E((Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + ...)(Z_{t+h} + \phi Z_{t+h-1} + \phi^2 Z_{t+h-2} + ...$

As long as the indices are different, the Z terms are independent, and the expected value of each Z is 0. So, using h = 1 you get:

$$E(\epsilon_t \epsilon_{t+1}) = E(Z_t Z_{t+1} + \phi Z_t Z_t + \phi^2 Z_t Z_{t-1} + \phi Z_{t-1} Z_{t+h} + ...) = \phi Z_t Z_t + \phi^3 Z_{t-1} Z_{t-1} + ... = \phi \sum_{i=0}^{\infty} (\phi^2)^i Z_{t-i}^2$$

And h=2 gives you:

$$E(\epsilon_t \epsilon_{t+2}) = \phi^2 Z_t Z_t + \phi^4 Z_{t-1} Z_{t-1} + ... = \phi^2 \sum_{i=0}^{\infty} (\phi^2)^i Z_{t-i}^2$$

And so on, giving you:

$$\phi^h \sum_{i=0}^{\infty} (\phi^2)^i Z_{t-i}^2$$

The Zs are identically distributed, so we only need to calculate $E(Z_t^2)$ and plug that value in:

$$Var(Z_t) = E(Z_t^2) - E(Z_t)^2 E(Z_t)^2 = 0 E(Z_t^2) = Var(Z_t) = \sigma^2$$

So, using the geometric series:

$$\phi^h \sum_{i=0}^{\infty} (\phi^2)^i Z_{t-i}^2 = \frac{\phi^h \sigma^2}{1 - \phi^2}$$

### c. Correlation

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \rho(\epsilon_t, \epsilon_{t+h}) = \frac{Cov(\epsilon_t, \epsilon_{t+h})}{\sqrt{Var(\epsilon)Var(\epsilon_{t+h})}} Var(\epsilon_t) = E(\epsilon_t^2) - E(\epsilon_t)^2 = E(\epsilon_t^2)E(\epsilon_t^2) = E((Z_t + \phi Z_{t-1} + \phi^2 Z_t$$

$Var(\epsilon_{t+h})$ will be the same, so

$$\sqrt{Var(\epsilon)Var(\epsilon_{t+h})} = \frac{\sigma^2}{1-\phi^2}$$

.

Therefore:

$$\rho(\epsilon_t, \epsilon_{t+h}) = \frac{\phi^h \sigma^2}{1-\phi^2} * \frac{1-\phi^2}{\sigma^2} = \phi^h$$

### d. Stationary process

$\{\epsilon_t\}$ is not a stationary process, because if you say that h is time (e.g. number of time points in a longitudinal experiment) then the correlation between $\epsilon_t$ and $\epsilon_{t+h}$ decreases over time.

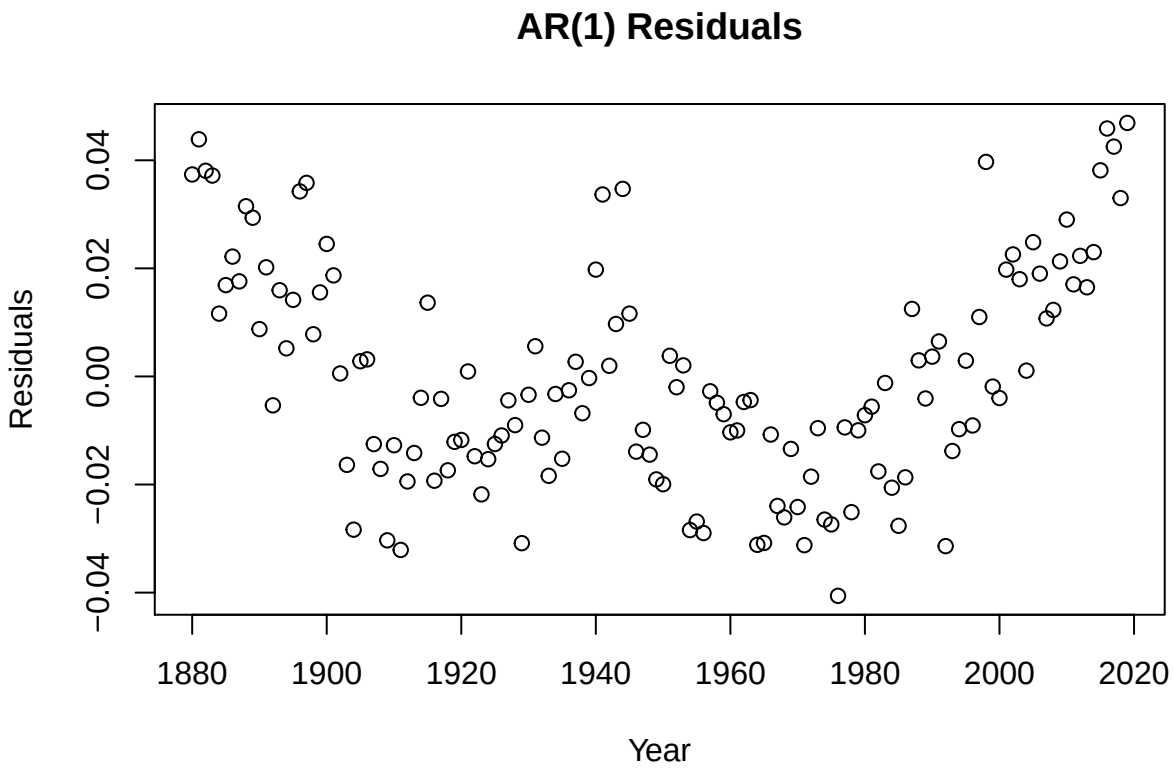
## 3. Time series data

The model:

```
mod1 <- lme(temp ~ year,data = temps,method = "ML",random = ~1|year,correlation = corAR1())
```

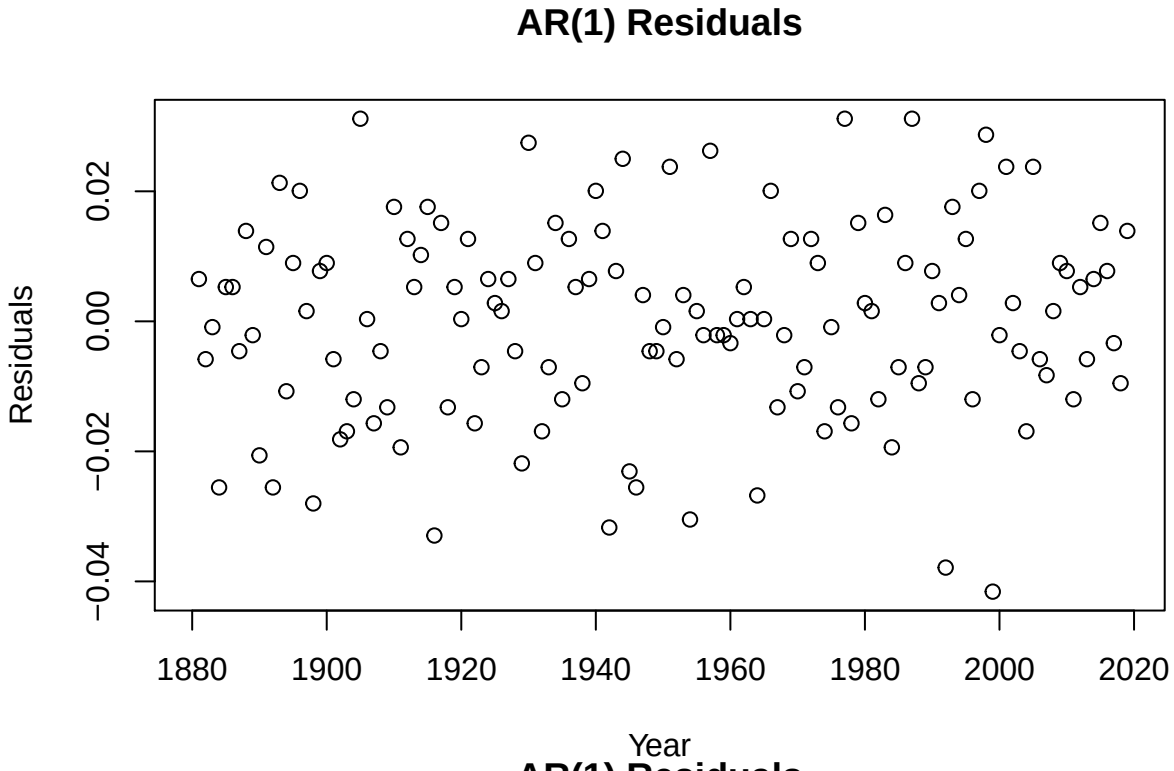
| | Value | Std.Error | DF | t-value | p-value |
|---|---|---|---|---|---|
| (Intercept) | -13.9941249 | 0.6781232 | 138 | -20.63655 | 0 |
| year | 0.0072133 | 0.0003478 | 138 | 20.74150 | 0 |

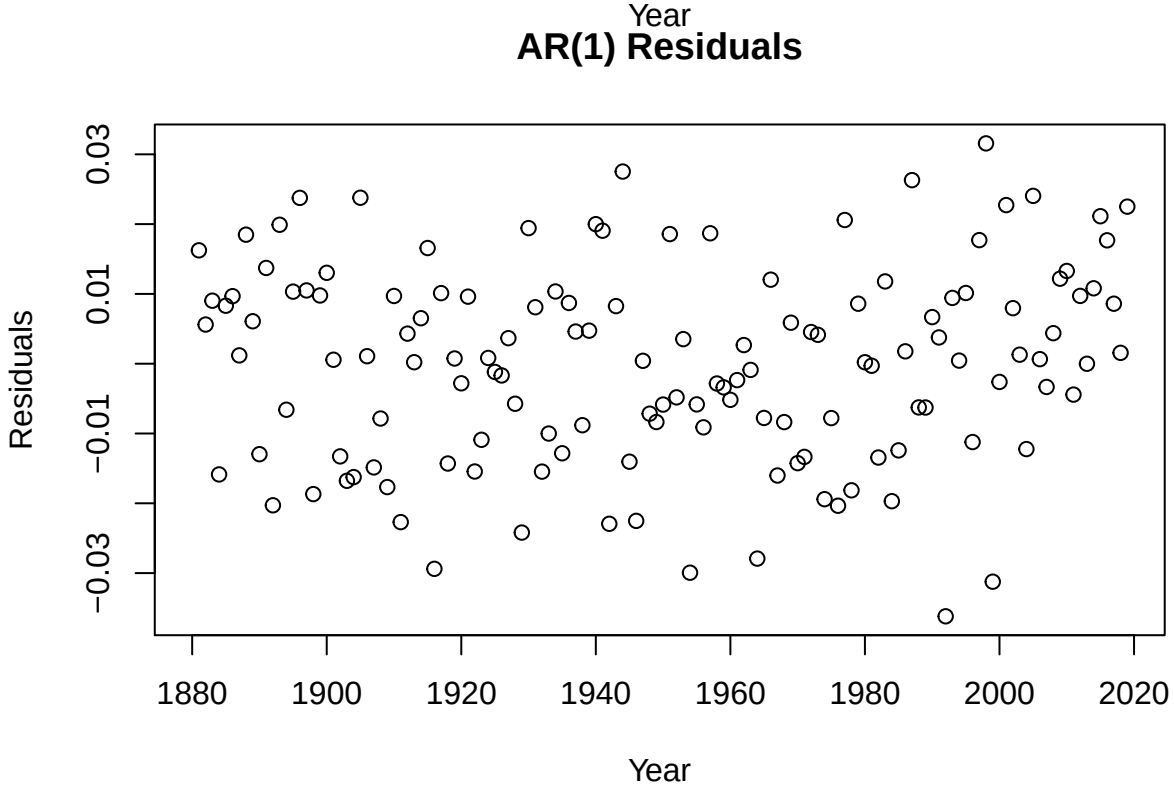
**a. Residual plot**




It's pretty obvious that there is a pattern in these residuals, which means that the model is violating the assumptions of a linear model. So, it might be worth looking into alternative models that capture global temperature trends better.

**b.**


AR(1) Residuals
Residuals
0.02
0.00
−0.02
−0.04
1880
1900
1920
1940
1960
1980
2000
2020
Year


AR(1) Residuals
Residuals
0.03
0.01
−0.01
−0.03
1880
1900
1920
1940
1960
1980
2000
2020
Year