Longitudinal Homework 1

Tim Vigers

13 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
```

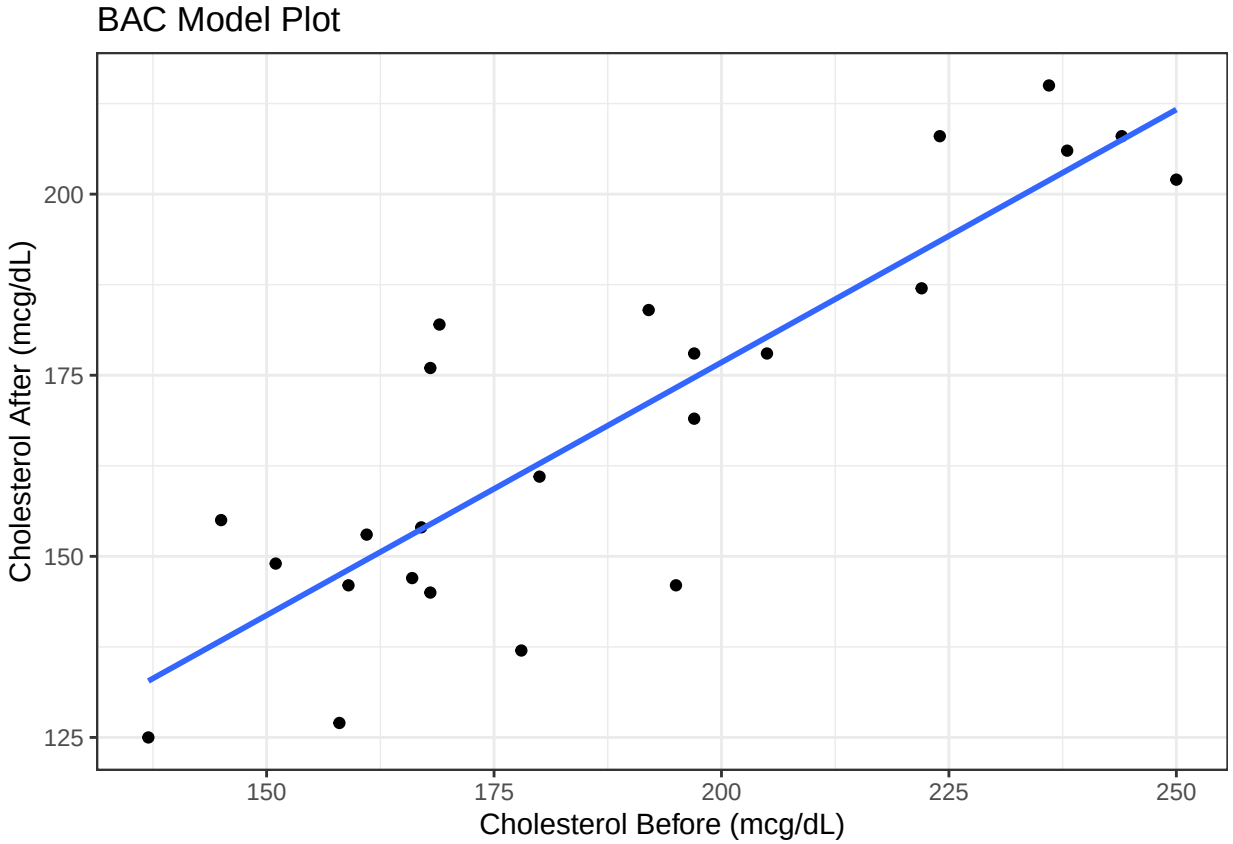
c. Baseline-as-covariate model

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

Results

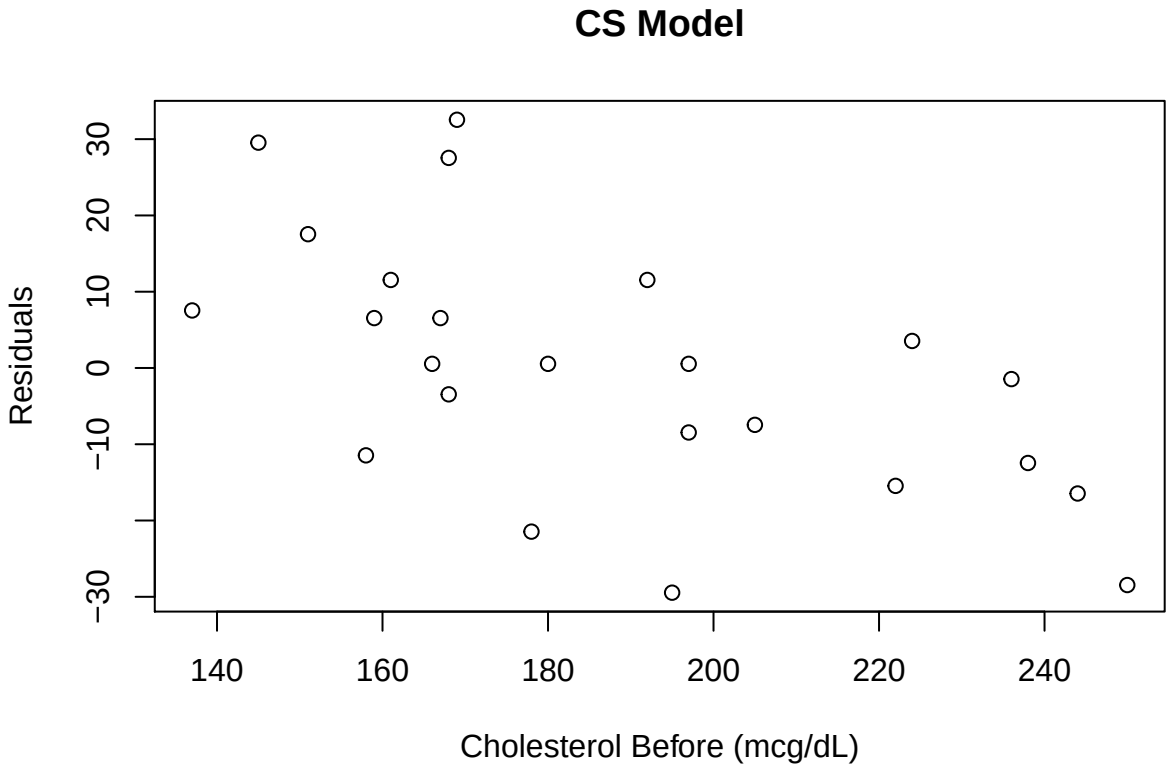
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	37.1576120	16.5393736	2.246615	0.0350309
before	0.6980735	0.0867859	8.043624	0.0000001

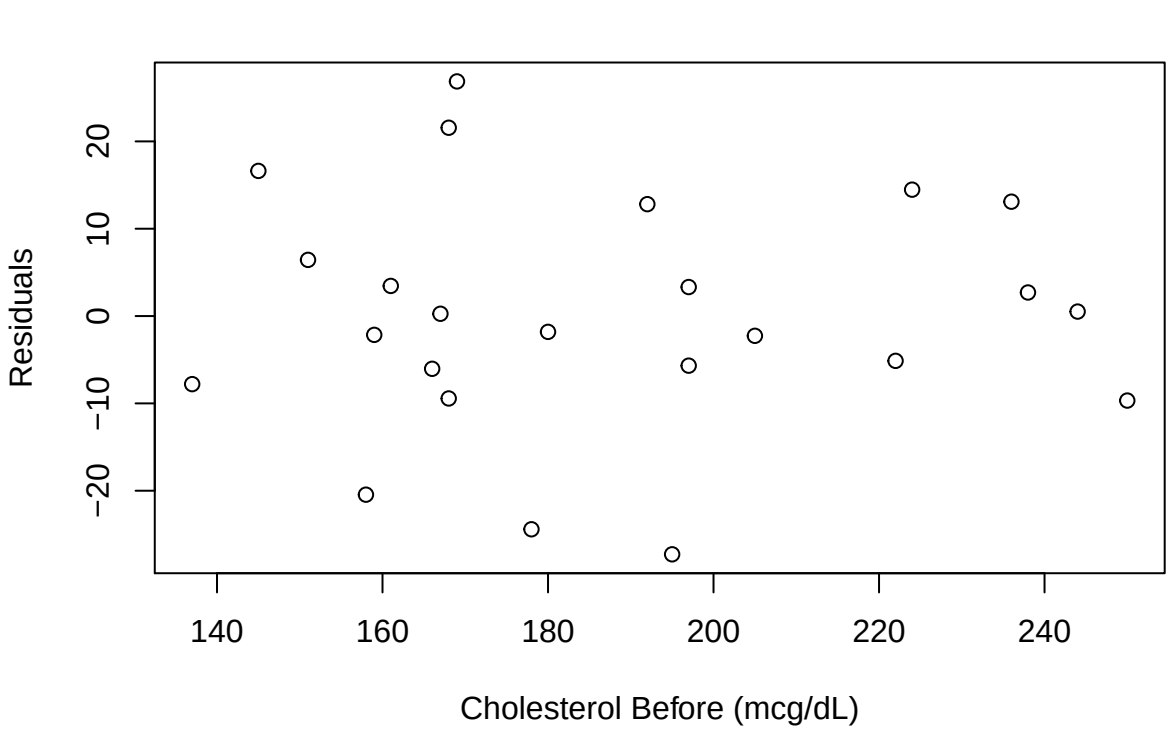
The results of this model indicate that for a theoretical starting cholesterol value of 0 mcg/dL, the average after value is 37.16 mcg/dL.For every one unit increase in the starting value, the after value increases by 0.70 (95% CI: 0.52 - 0.88, $p < 0.001$). The intercept doesn't make much sense to interpret here because a cholesterol value of 0 isn't biologically possible, but because the slope for β_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} gets its own β value.

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 will have the same β_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 ϕ is a constant and we assume the Zs to be independent (so we can apply $E(XY) = E(X)E(Y)$ for independent variables), 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

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

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 into the equation above:

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

So, using the geometric series:

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

c. Correlation

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$
$$\rho(\epsilon_t, \epsilon_{t+h}) = \frac{Cov(\epsilon_t, \epsilon_{t+h})}{\sqrt{Var(\epsilon)Var(\epsilon_{t+h})}}$$
$$Var(\epsilon_t) = E(\epsilon_t^2) - E(\epsilon_t)^2 = E(\epsilon_t^2)$$
$$E(\epsilon_t^2) = E((Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + ...)(Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + ...))$$
$$= E(Z_t^2 + \phi^2 Z_{t-1}^2 + \phi^4 Z_{t-2}^2 + ...) = E(Z_t^2) + \phi^2 E(Z_{t-1}^2) + \phi^4 E(Z_{t-2}^2) + ...$$
$$= \sum_{i=0}^{\infty} (\phi^2)^i E(Z_{t-i}^2) = \sum_{i=0}^{\infty} (\phi^2)^i \sigma^2 = \frac{\sigma^2}{1-\phi^2}$$

$Var(\epsilon_{t+h})$ 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

A stationary process $\{Y_t\}$ has a constant mean and finite second moment for all times t and the correlation between Y_t and Y_{t+h} does not depend on t for all h. This is true of $\{\epsilon_t\}$, as the correlation only depends on ϕ and h.

3. Time series data

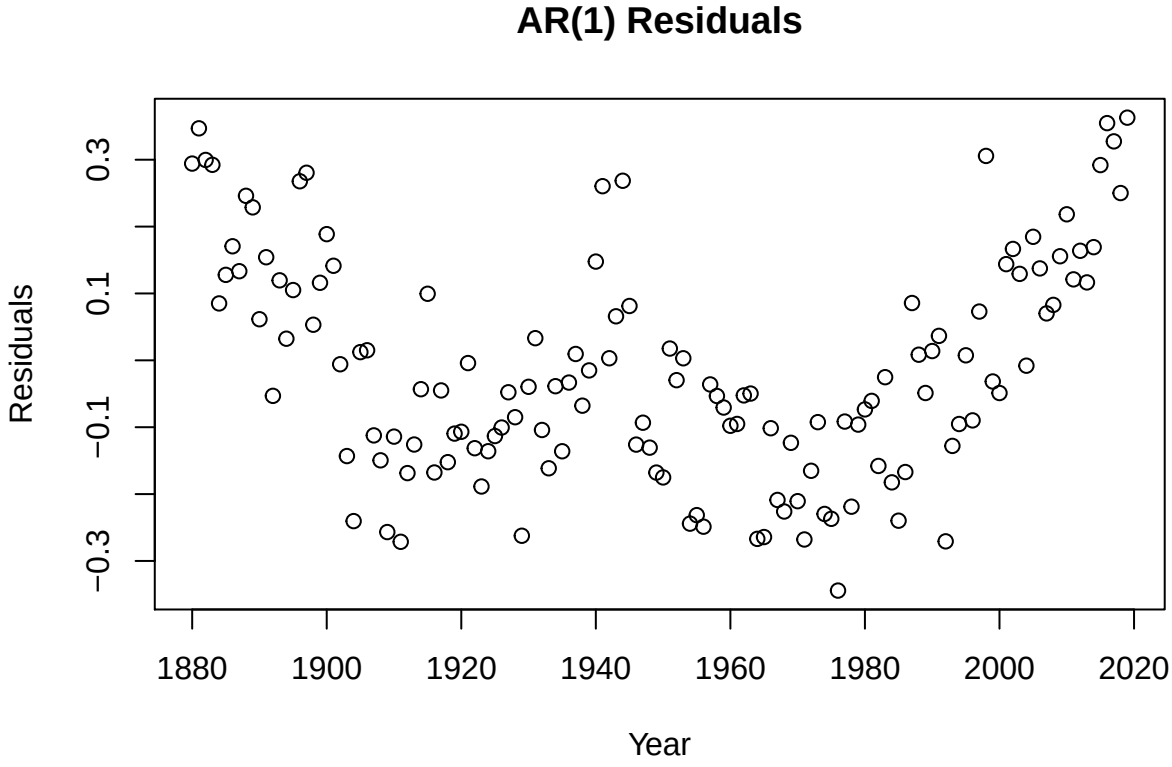
The model:

```
temps$fake <- 1
mod1 <- gls(temp ~ year,data = temps,method = "ML",
            correlation = corAR1(form = ~1|fake))
```

	Value	Std.Error	t-value	p-value
(Intercept)	-14.1026823	1.6978064	-8.306414	0
year	0.0072758	0.0008707	8.356340	0

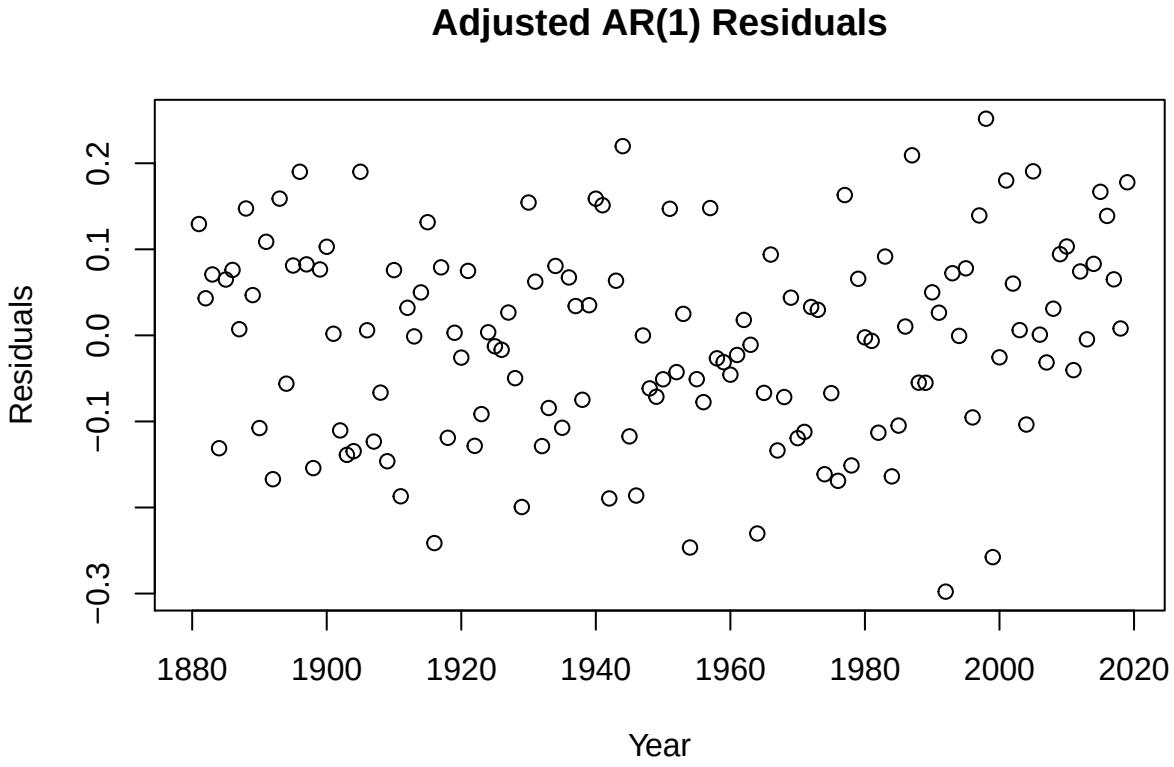
This output matches SAS, which is good!

a. Residual plot



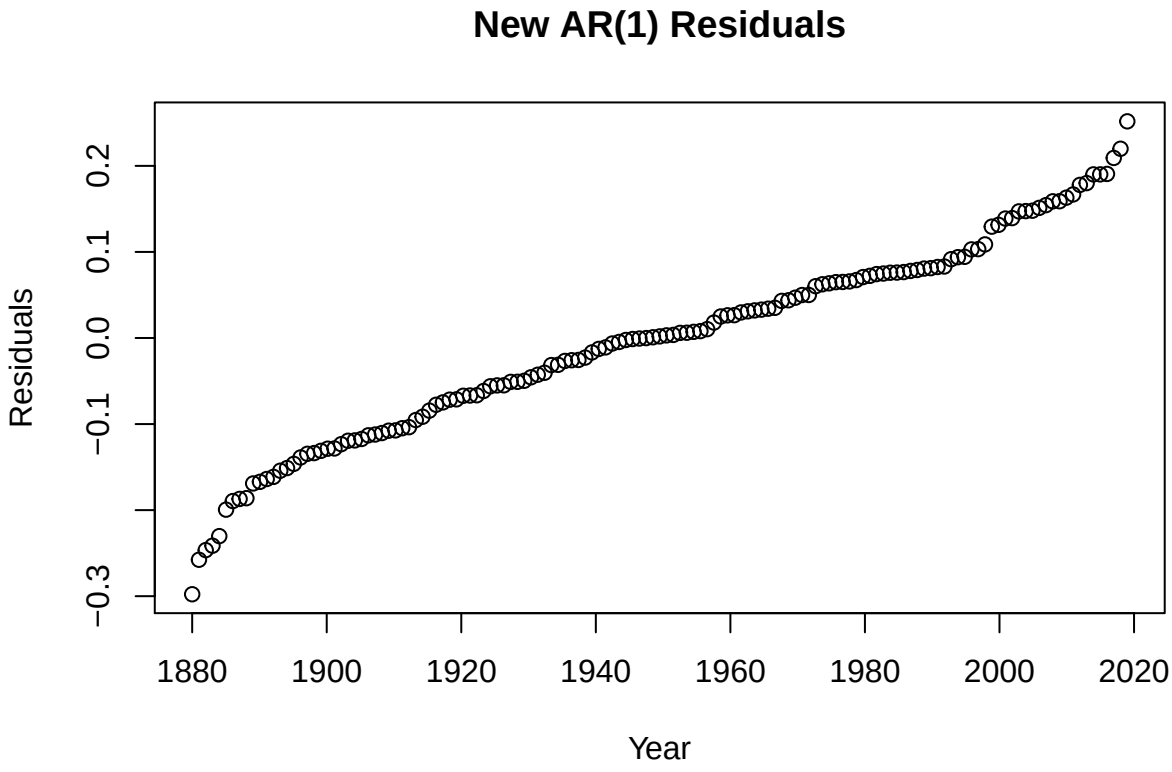
It's pretty obvious that there is a pattern in these residuals, and I think the plot suggests that there could be a quartic trend.

b. New residuals



Using the estimated correlation parameter from SAS (ϕ=0.7395) to calculate new residuals looks much better. I think this model actually fits the data realtively well given these residuals, but it isn't perfect. The "W" trend from the previous plot it still sort of visible, although I'm not sure I'd notice it if I hadn't seen the first plot.

Also, a qq plot of these residuals looks decent to me:



c. Average increase per decade

Based on this model, the average increase in temperature per decade is 0.073 degrees C (95% CI: 0.056 - 0.090).

d. Polynomial model

The first residual plot has a "W" shape, so I tried to fit a quartic model, but got a convergence warning:

```
mod2 <- gls(temp ~ year+I(year^2)+I(year^3)+I(year^4),data = temps,
            method = "ML",correlation = corAR1(form = ~1|fake))
```

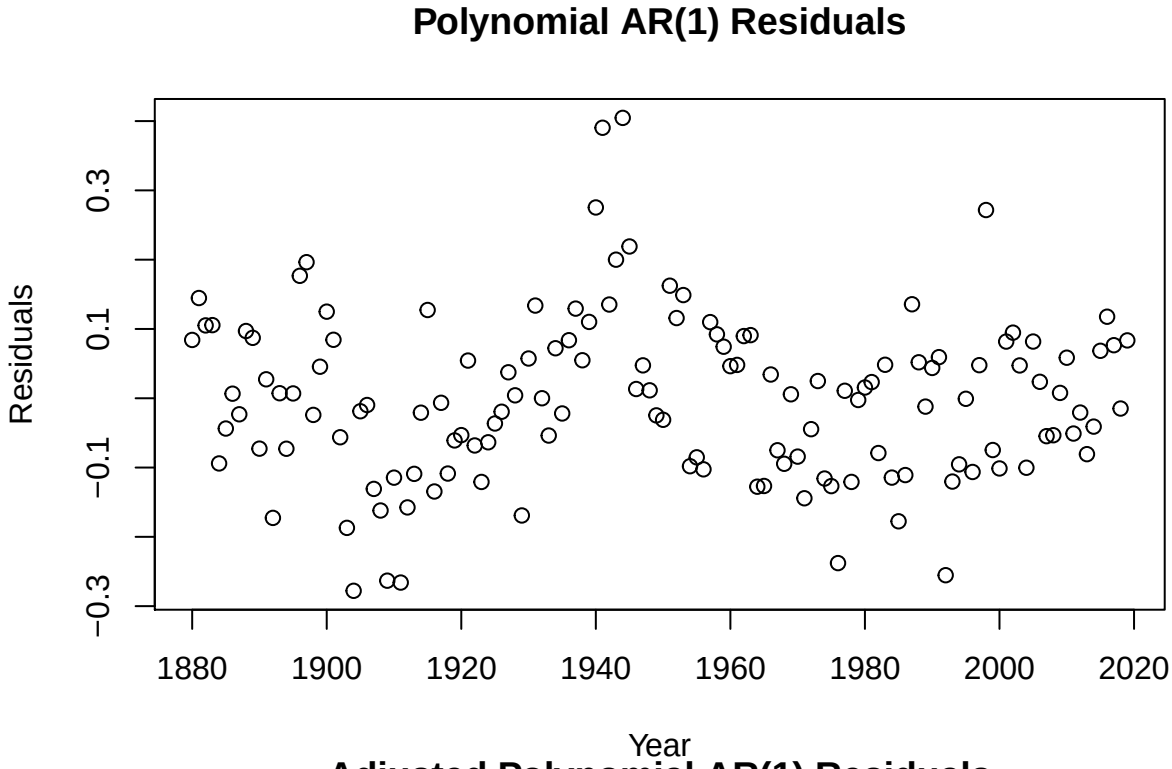
```
## Error in gls(temp ~ year + I(year^2) + I(year^3) + I(year^4), data = temps, : false convergence (8)
```

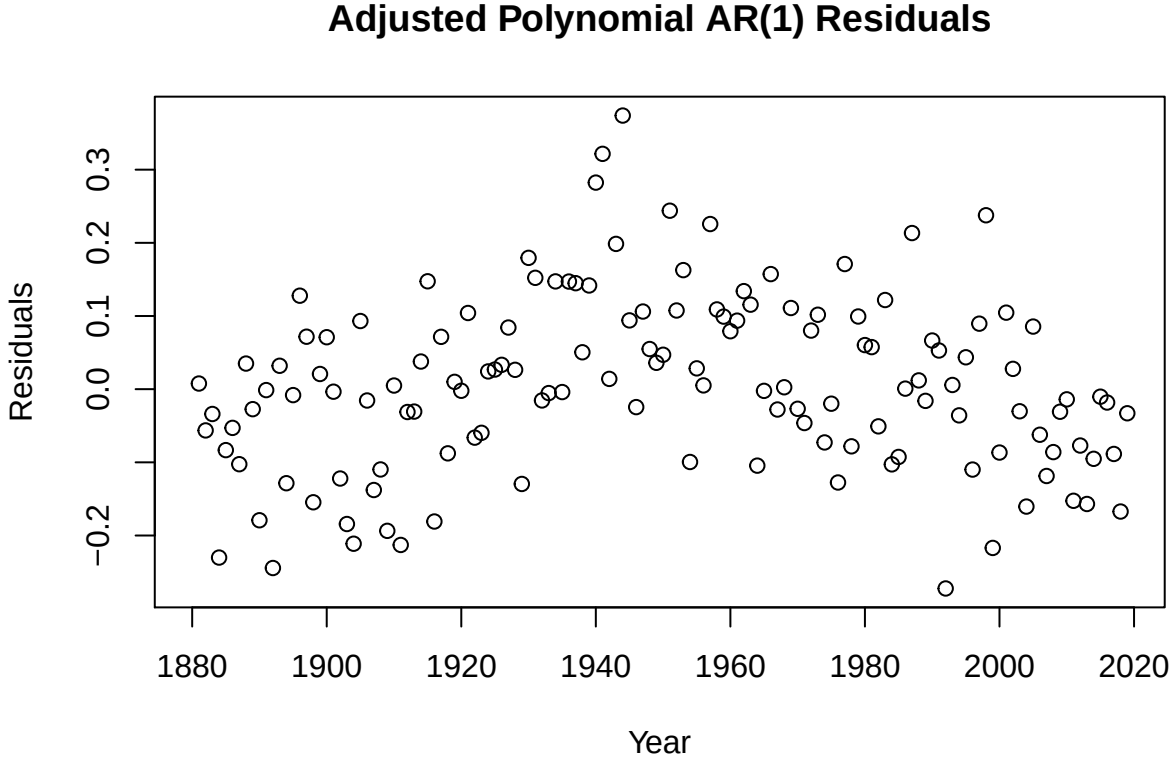
So I took it back down to a cubic model:

```
mod2 <- gls(temp ~ year+I(year^2)+I(year^3),data = temps,method = "ML",
            correlation = corAR1(form = ~1|fake))
```

	Value	Std.Error	t-value	p-value
(Intercept)	-1906.4627437	2280.8139476	-0.8358695	0.4046945
year	3.0777112	3.5113665	0.8764996	0.3823041
I(year^2)	-0.0016559	0.0018015	-0.9191344	0.3596526
I(year^3)	0.0000003	0.0000003	0.9637743	0.3368697

Polynomial model residuals





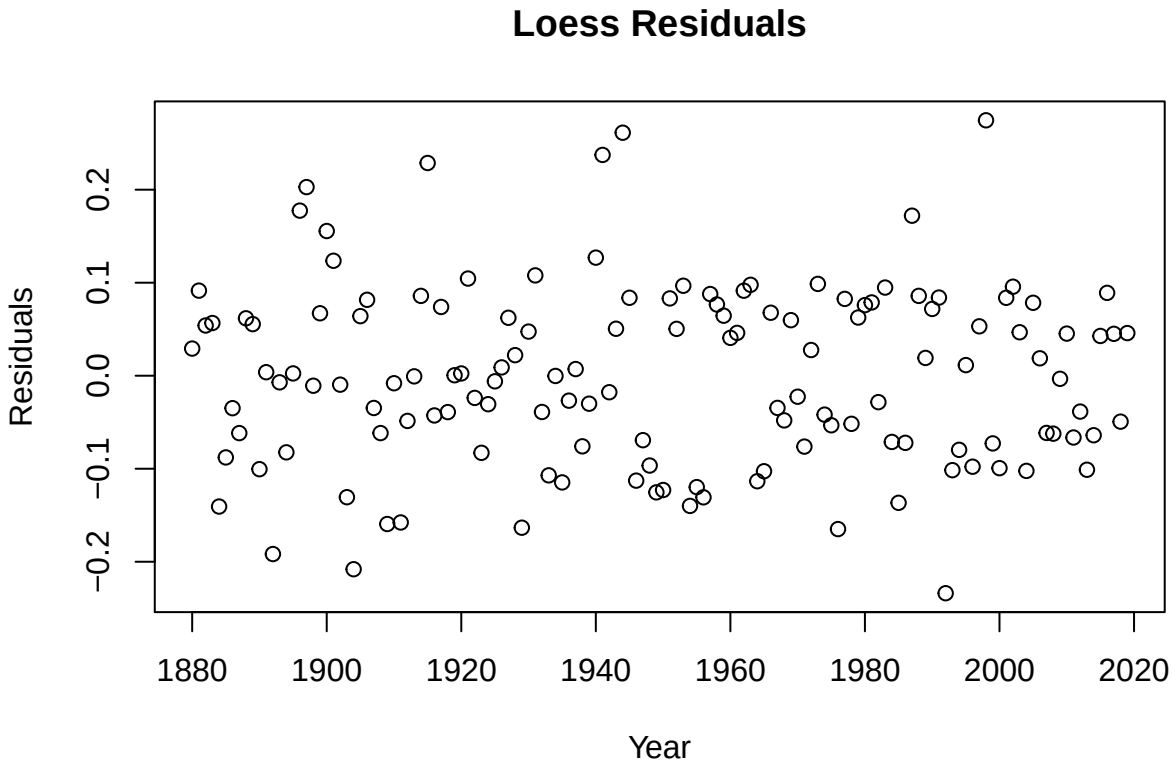
This model appears to be a better fit than the linear linear model overall, based on the regular residuals (even though there's still a pattern in the adjusted residuals, which I'm assuming is a result of the correlation parameter changing). The correlation parameter decreases from 0.7395 to 0.4655, which makes some intuitive sense. Because the fit is better, the polynomial terms are accounting for some of the correlation between

outcomes.

e. Loess model

The model

```
mod3 <- loess(temp ~ year,temps,span = 0.3)
```



The Loess residuals indicate that the fit is pretty good. They certainly look better than the polynomial residuals and the regular residual plot from the linear model. They look fairly normal for the Loess plot, whereas there were clear patterns in the linear and polynomial models. However, the "W" pattern is still vaguely visible in the most extreme points in the Loess residuals. I think the Loess plot is probably the best model overall, although the adjusted residuals that take into account both the mean and the error parts of the linear model also look good to me. Since I'm not at all comfortable interpreting the results of a Loess model, I would probably use the linear model if I was analyzing this on my own, but the Loess might be a better alternative if I was better able to interpret the output.