1. Consider a first-order autoregressive process:
a. Determine $E(\varepsilon_t)$
b. Determine $Cov(\varepsilon_t, \varepsilon_{t+h})$
c. Determine $Corr(\varepsilon_t, \varepsilon_{t+h})$
d. Is ε_t a stationary process?

a. Since $Z_t \sim N(0, \sigma^2)$:

$$\begin{aligned}
E(\varepsilon_t) &= E(\sum_{j=0}^{\infty} \phi^j Z_{t-j}) && (1)\\
&= \sum_{j=0}^{\infty}[E(\phi^j Z_{t-j})] && (2)\\
&= \sum_{j=0}^{\infty}[\phi^j \times E(Z_{t-j})] && (3)\\
&= E(Z_{t-j}) \times \sum_{j=0}^{\infty} \phi^j && (4)\\
&= 0 \times \sum_{j=0}^{\infty} \phi^j && (5)\\
&= 0 && (6)
\end{aligned}$$

b. Since $E(Z_t) = 0, Var(Z_t) = \sigma^2$, 5/5
we have $E(Z_t^2) = Var(Z_t)^2 + E(Z_t)^2 = \sigma^2$,
we also know that $|\phi| < 1$,

and Z_t are $i.i.d$:

$$
\begin{aligned}
Cov(\varepsilon_t, \varepsilon_{t+h}) &= E(\varepsilon_t \varepsilon_{t+h}) - E(\varepsilon_t)E(\varepsilon_{t+h}) && (7)\\
&= E(\varepsilon_t \varepsilon_{t+h}) && (8)\\
&= E[(\sum_{j=0}^{\infty} \phi^j Z_{t-j})(\sum_{i=0}^{\infty} \phi^i Z_{t+h-i})] && (9)\\
&= E[\overbrace{\sum_{j=0}^{\infty}\sum_{i=0}^{\infty}(\phi^j Z_{t-j}\phi^i Z_{t+h-i})}^{t-j\neq t+h-i} + \overbrace{\sum_{i=0}^{\infty}(\phi^{i-h+i} Z^2_{t+h-i})}^{t-j=t+h-i}] && (10)\\
&= \overbrace{E(\sum_{j=0}^{\infty} \phi^j Z_{t-j})E(\sum_{i=0}^{\infty} \phi^i Z_{t+h-i})}^{t-j\neq t+h-i} + \overbrace{E(\sum_{i=0}^{\infty} \phi^{i-h+i} Z^2_{t+h-i})]}^{t-j=t+h-i} && (11)\\
&= E(Z_{t-j})\sum_{j=0}^{\infty} \phi^j E(Z_{t+h-i})\sum_{i=0}^{\infty} \phi^i + E(Z^2_{t+h-i}))\sum_{i=0}^{\infty} \phi^{i-h+i} && (12)\\
&= 0 \quad + \quad \sigma^2(\frac{\phi^{-h}}{1-\phi^2}) && (13)\\
&= \phi^{-h}(1-\phi^2)^{-1}\sigma^2 && (14)
\end{aligned}
$$

10/10 O.k. but you can make exponent "h" instead of "-h"; see solutions.
But I think approach is o.k.

c.

$$
\begin{aligned}
Corr(\varepsilon_t, \varepsilon_{t+h}) &= \frac{Cov(\varepsilon_t, \varepsilon_{t+h})}{\sqrt{Var(\varepsilon_t)Var(\varepsilon_{t+h})}} && (15)\\
Var(\varepsilon_t) &= Cov(\varepsilon_t, \varepsilon_t) && (16)\\
&= (1-\phi^2)^{-1}\sigma^2 && (17)\\
Var(\varepsilon_{t+h}) &= Cov(\varepsilon_{t+h}, \varepsilon_{t+h}) && (18)\\
&= (1-\phi^2)^{-1}\sigma^2 && (19)\\
\text{Thus, } Corr(\varepsilon_t, \varepsilon_{t+h}) &= \frac{\phi^{-h}(1-\phi^2)^{-1}\sigma^2}{(1-\phi^2)^{-1}\sigma^2} && (20)\\
&= \phi^{-h} && (21)
\end{aligned}
$$

5/5

d. AR(1) is a weakly stationary process, since the mean and variance is the same for all t and the covariance between $\varepsilon_t, \varepsilon_{t+h}$ is the same for all t.

5/5

(2) Comparison of 4 models:

(i) change-score model, (ii) baseline-as-covariate model, (iii) hybrid model, (iv) a longitudinal model.

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -19.54       3.43  -5.697 8.43e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 282.433)
##
##     Null deviance: 6496  on 23  degrees of freedom
## Residual deviance: 6496  on 23  degrees of freedom
## AIC: 206.53
```

(i) change-score model just accounts for the intercept (mean of the change). Thus, this model not fits the data well and have the highest AIC in all 3 linear models. This model is too simple.

4/5 To be able to compare AIC's, the outcome should be the same as well as the records being fit.

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.15761   16.53937   2.247    0.035 *
## before       0.69807    0.08679   8.044 5.39e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 190.4789)
##
##     Null deviance: 16514.5  on 23  degrees of freedom
## Residual deviance:  4190.5  on 22  degrees of freedom
## AIC: 198.01
```

(ii) baseline-as-covariate model is actually the same with the (iii) hybrid model. The difference of the coefficients of “before” is 1, which is just the move of “1 before” from the left side to the right side of the equation. The residual deviance from these two models are the same, and both less than the change-score model.

5/5

```
## (Intercept) 37.15761   16.53937   2.247  0.03503 *
## before      -0.30193    0.08679  -3.479  0.00213 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 190.4789)
##
##     Null deviance: 6496.0  on 23  degrees of freedom
## Residual deviance: 4190.5  on 22  degrees of freedom
## AIC: 198.01
```

This means these 2 models include more variance. Thus they have the same AIC, lower than the first model. Both of models show the significant association between two time points. These models are reasonable and easy to interpret. But they answer slightly different hypotheses.

5/5

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: cholesterol ~ time + (1 | subject_id)
##    Data: cholesterol_l
##
## REML criterion at convergence: 421.5
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.44161 -0.42520 -0.01153  0.41158  1.53751
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  subject_id (Intercept) 767.6    27.71
##  Residual               141.2    11.88
## Number of obs: 48, groups:  subject_id, 24
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  168.250      6.154  27.342
## timebefore    19.542      3.430   5.697
##
```

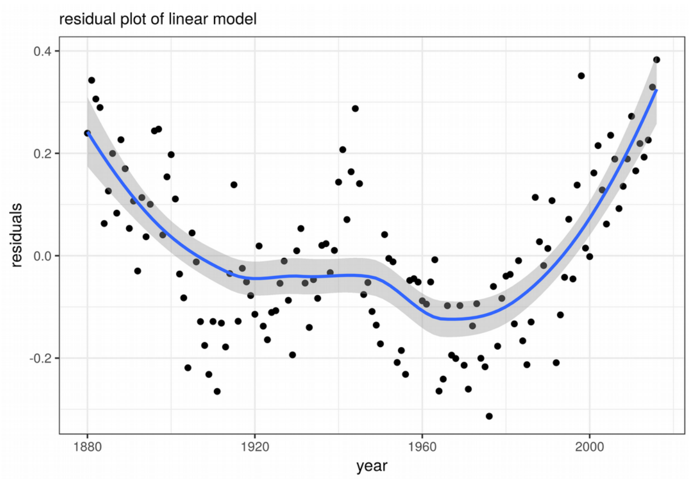
(iv) the linear mixed model with random intercept. The pre-post dataset has 2 repeat measurements, only enough for a compound symmetry covariance structure, which is the random intercept model. This model includes the “time” (before, after) as the binary fixed effect. This model also shows the significant association between cholesterol levels and time points. This model has a better fit to the data points and using the long form of dataset. The hypotheses is to test the association between cholesterol levels and time,

5/5

different from the change & baseline. However, since there are only 2 time points, this model is not necessary and too complex.

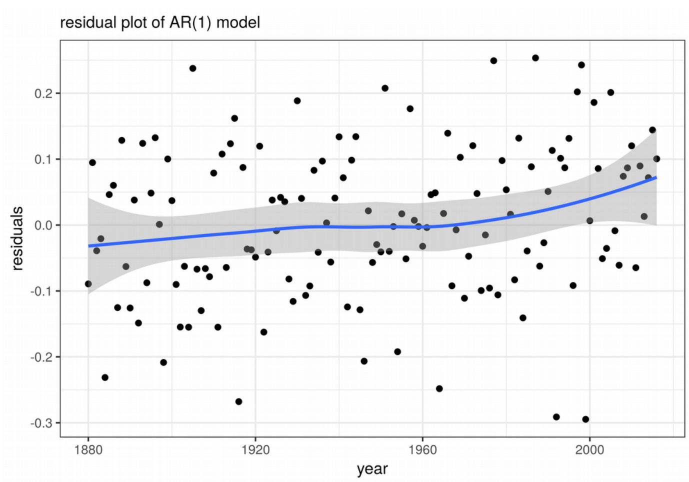
```r
paste("The AIC of lmm:", AIC(lmm_prepost))

## [1] "The AIC of lmm: 429.516076427762"
```

(3) Time series data.



a. the residual plot shows the simple linear regression not works in this case. The residuals are not normally distributed, with a “W” shape pattern. We need try a different model to get an even and balanced residual cloud.

5/5



b. and c. The estimated phi in AR(1) process is 0.9432, there is also the correlation between two neighbors. This model fits the data better compared with the simple linear model. The residual is roughly equally distributed. The complexity of the mean by AR(1) model reduces the error of the model.

13/15. What is the correlation you computed? For the fit on the data, it should be lower. See me if you have questions.

```r
coef(ar_temp)

##       ar1 intercept
## 0.9431780 0.1087432
global_temp$res_ar <- residuals(ar_temp)
```

d. the average increase in temperature per decade is 0.06 ºC.

5/5

```r
##   (Intercept)          year
## -11.650245508    0.006004107
# average change per decade
paste("average increase in temp. per decade:", round(10 * coef(lm_pred)[2],
    4))

## [1] "average increase in temp. per decade: 0.06"
```

6643 HW1

Guannan Shen

September 9, 2018

Contents

1 Question 2

The data cholesterol.txt contains cholesterol levels (adapted from Rosner, 2006). The data are a sample of cholesterol levels taken from 24 hospital employees who were on a standard American diet and who agreed to adopt a vegetarian diet for one month. Serum cholesterol measurements (mcg/dl) were made before adopting the vegetarian diet and one month after.

```
library(readr)
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------- tidyverse 1.2.1 --

## √ ggplot2 3.0.0     √ purrr   0.2.5
## √ tibble  1.4.2     √ dplyr   0.7.6
## √ tidyr   0.8.1     √ stringr 1.3.1
## √ ggplot2 3.0.0     √ forcats 0.3.0

## -- Conflicts ---------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
cholesterol <- read.csv("HW1_cholesterol.csv")
# add the subject id
dim(cholesterol)
```

```
## [1] 24  2
```

```
cholesterol <- cholesterol %>% mutate(subject_id = seq(1:24))
# add change
cholesterol_w <- cholesterol %>% mutate(change = after - before)
# now model the first 3 models change-score model
m1_change_score <- glm(change ~ 1, data = cholesterol_w, family = gaussian)
summary(m1_change_score)
```

```
##
## Call:
## glm(formula = change ~ 1, family = gaussian, data = cholesterol_w)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -29.458  -11.708    0.542    8.542   32.542
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -19.54       3.43  -5.697 8.43e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 282.433)
##
##     Null deviance: 6496  on 23  degrees of freedom
## Residual deviance: 6496  on 23  degrees of freedom
## AIC: 206.53
##
## Number of Fisher Scoring iterations: 2
```

```r
# baseline as covariate model
m2_baseline <- glm(after ~ before, data = cholesterol_w, family = gaussian)
summary(m2_baseline)
```

```
##
## Call:
## glm(formula = after ~ before, family = gaussian, data = cholesterol_w)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -27.2819   -6.4768   -0.7734    8.0280   26.8680
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.15761   16.53937   2.247    0.035 *
## before       0.69807    0.08679   8.044 5.39e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 190.4789)
##
##     Null deviance: 16514.5  on 23  degrees of freedom
## Residual deviance:  4190.5  on 22  degrees of freedom
## AIC: 198.01
##
## Number of Fisher Scoring iterations: 2
```

```r
# hybrid model
m3_hybrid <- glm(change ~ before, data = cholesterol_w, family = gaussian)
summary(m3_hybrid)
```

```
##
## Call:
## glm(formula = change ~ before, family = gaussian, data = cholesterol_w)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -27.2819   -6.4768   -0.7734    8.0280   26.8680
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 37.15761   16.53937   2.247  0.03503 *
## before      -0.30193    0.08679  -3.479  0.00213 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 190.4789)
##
##     Null deviance: 6496.0  on 23  degrees of freedom
## Residual deviance: 4190.5  on 22  degrees of freedom
## AIC: 198.01
##
## Number of Fisher Scoring iterations: 2
```

```r
# now make the long form data
cholesterol_l <- cholesterol %>% gather(key = time, value = cholesterol,
    before:after)

# longitudinal model
library(lme4)
```

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following object is masked from 'package:tidyr':
##
##     expand
```

```r
class(cholesterol_l$time)
```

```
## [1] "character"
```

```r
lmm_prepost <- lmer(cholesterol ~ time + (1 | subject_id), data = cholesterol_l)
summary(lmm_prepost)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: cholesterol ~ time + (1 | subject_id)
##    Data: cholesterol_l
##
## REML criterion at convergence: 421.5
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.44161 -0.42520 -0.01153  0.41158  1.53751
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  subject_id (Intercept) 767.6    27.71
##  Residual               141.2    11.88
## Number of obs: 48, groups:  subject_id, 24
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  168.250      6.154  27.342
## timebefore    19.542      3.430   5.697
##
```

```
## Correlation of Fixed Effects:
##            (Intr)
## timebefore -0.279
```

```
paste("The AIC of lmm:", AIC(lmm_prepost))
```

```
## [1] "The AIC of lmm: 429.516076427762"
```

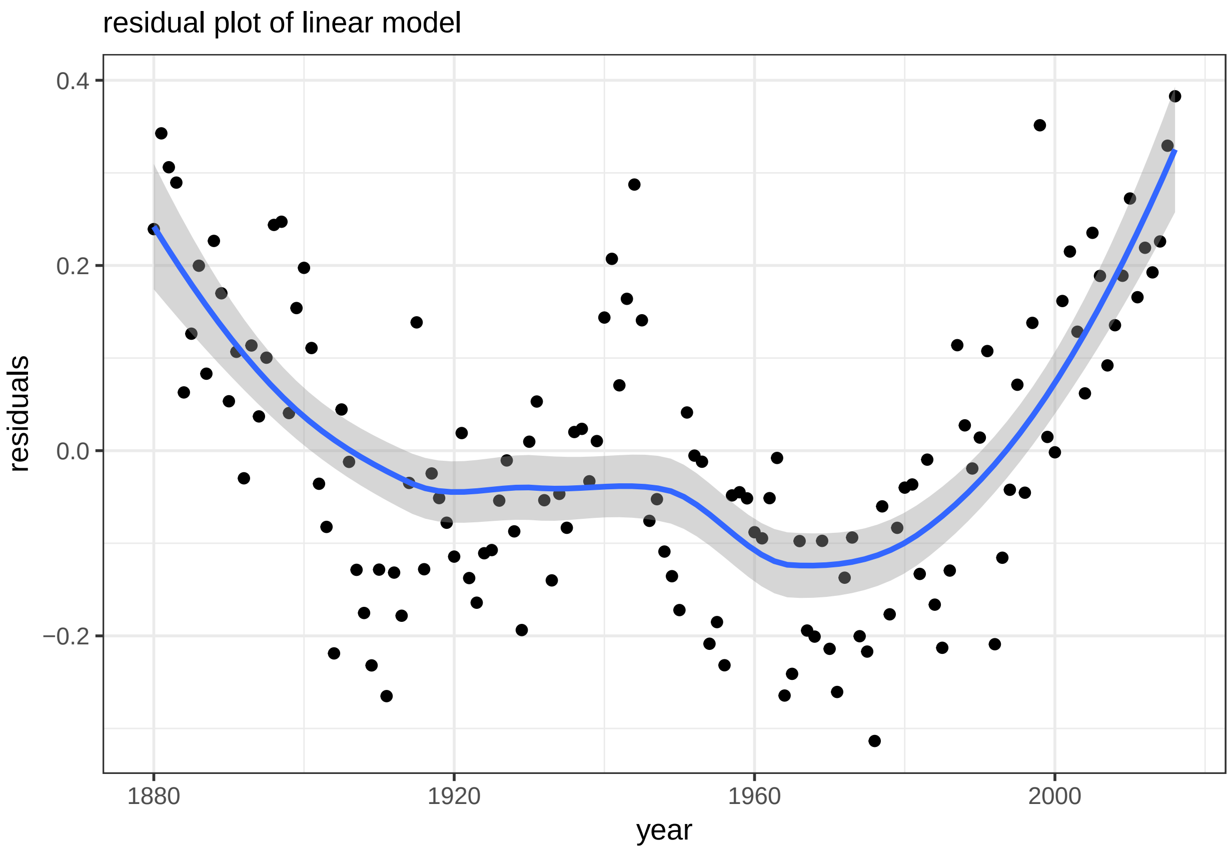
2 Question 3

```
library(astsa)
global_temp <- read.csv("HW1_global_temp_anomalies.csv", header = FALSE)
colnames(global_temp) <- c("year", "temp")
# simple linear model
lm_temp <- glm(temp ~ year, data = global_temp, family = gaussian)
summary(lm_temp)
```

```
##
## Call:
## glm(formula = temp ~ year, family = gaussian, data = global_temp)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.31355  -0.11562  -0.02463   0.11393   0.38276
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.279e+01  6.629e-01  -19.30   <2e-16 ***
## year         6.592e-03  3.402e-04   19.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.02480508)
##
##     Null deviance: 12.6597  on 136  degrees of freedom
## Residual deviance:  3.3487  on 135  degrees of freedom
## AIC: -113.67
##
## Number of Fisher Scoring iterations: 2
```

```
global_temp$res_lm <- resid(lm_temp)
ggplot(global_temp, aes(year, res_lm)) + geom_point() + geom_smooth() +
    theme_bw() + labs(y = "residuals", subtitle = "residual plot of linear model")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
ggsave("Residual_plot_of_linear_model.png", dpi = 600)
```

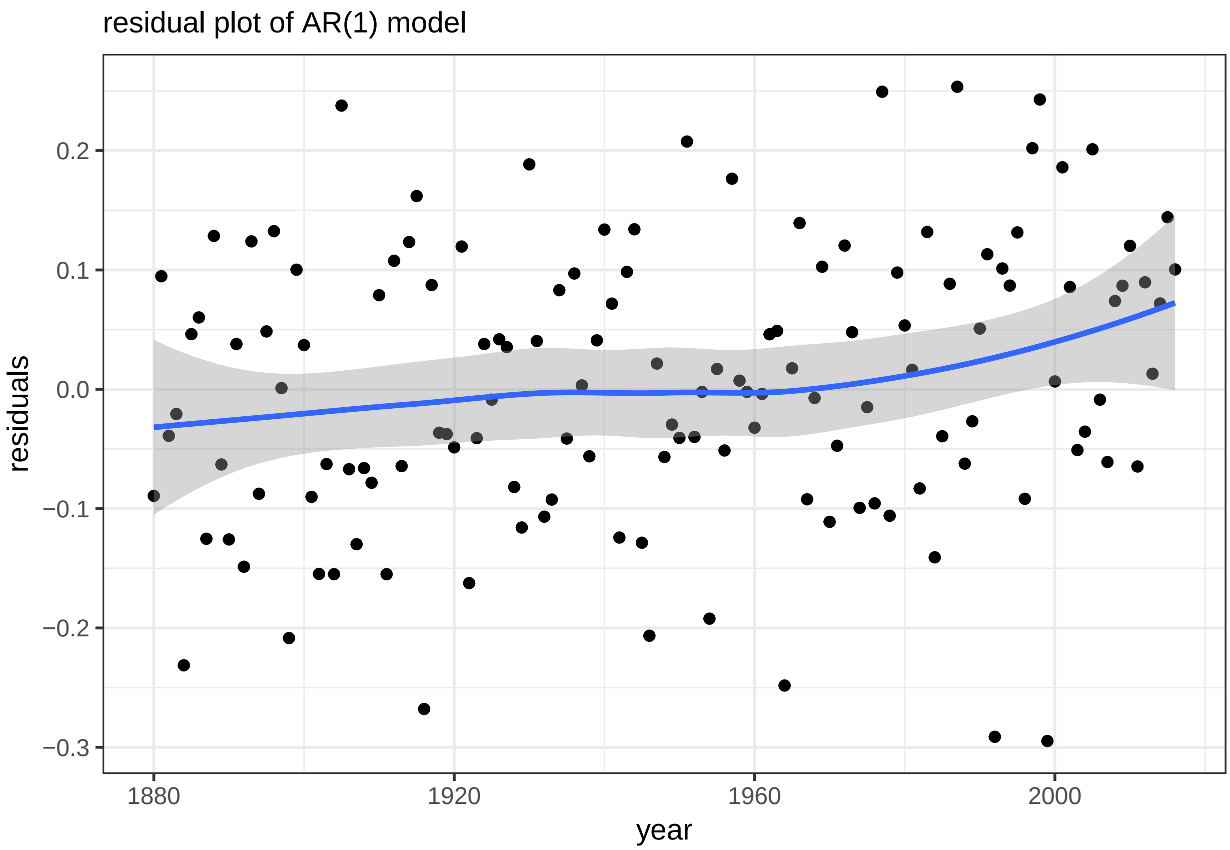
```
## Saving 6.5 x 4.5 in image
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## fit ar(1) model for time series data using ts() to make a
## univariate time series object
temp_ts <- ts(global_temp$temp)
ar_temp <- arima(temp_ts, order = c(1, 0, 0))
coef(ar_temp)
```

```
##       ar1 intercept
## 0.9431780 0.1087432
```

```
global_temp$res_ar <- residuals(ar_temp)


## residual plot
ggplot(global_temp, aes(year, res_ar)) + geom_point() + geom_smooth() +
    theme_bw() + labs(y = "residuals", subtitle = "residual plot of AR(1) model")
```

```
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```r
ggsave("Residual_plot_of_AR1_model.png", dpi = 600)
```

```
## Saving 6.5 x 4.5 in image
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```r
## predict value
global_temp$y_hat <- global_temp$temp - global_temp$res_ar
lm_pred <- glm(y_hat ~ year, data = global_temp, family = gaussian)
coef(lm_pred)
```

```
##   (Intercept)          year
## -11.650245508   0.006004107
```

```r
# average change per decade
paste("average increase in temp. per decade:", round(10 * coef(lm_pred)[2],
    4))
```

```
## [1] "average increase in temp. per decade: 0.06"
```