

6643 HW1

Guannan Shen

September 9, 2018

Contents

1	Question 2	1
2	Question 3	4

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

# 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

# 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

# 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

class(cholesterol_l$time)

## [1] "character"

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"
```

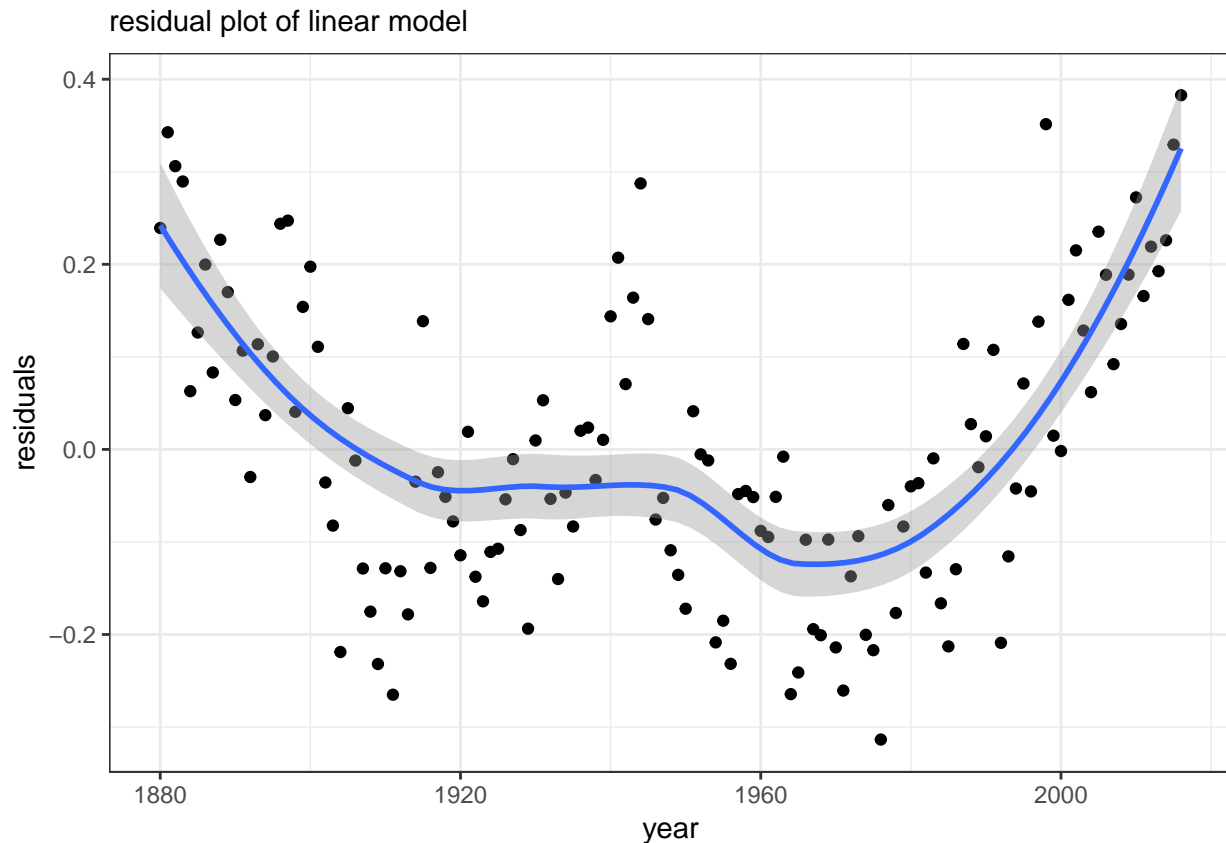
2 Question 3

```
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)
```

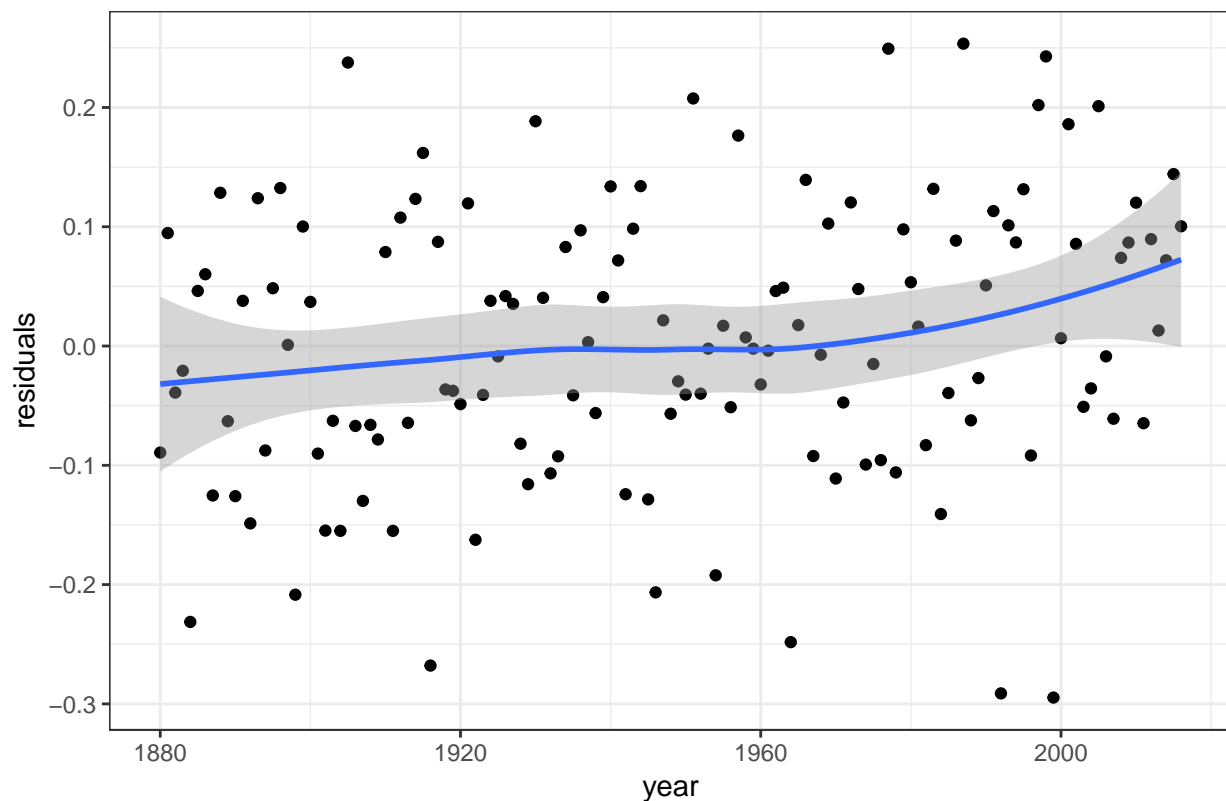
```
##      ar1 intercept
## 0.9431780 0.1087432
```

```
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'
```

residual plot of AR(1) model



```
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'
```

```
## 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
```

```
# 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"
```