STAT 224 Autumn 2022 HW7

Matthew Zhao

Question 1

http://www.stat.uchicago.edu/~yibi/s224/data/P229-30.txt

```
stock = read.table("http://www.stat.uchicago.edu/~yibi/s224/data/P229-30.txt", header=T)
```

Q1a — 4 points

```
stock1 = lm(DJIA ~ Day, data=stock)
```

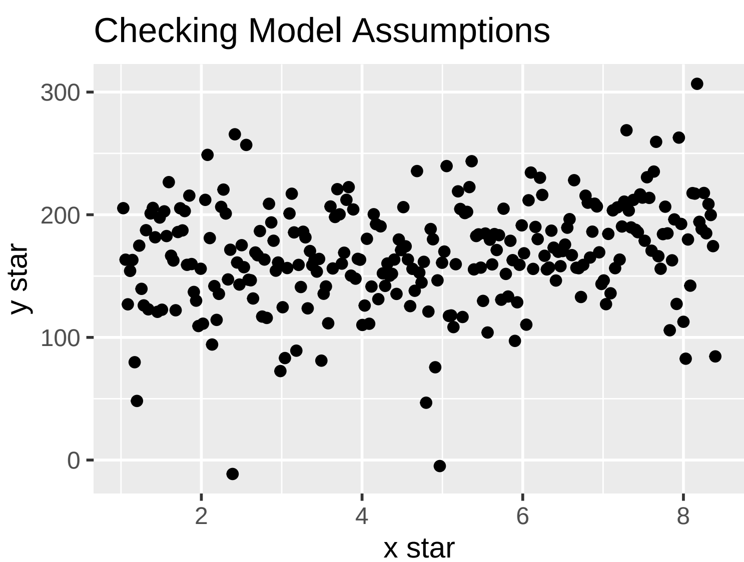
```
x = stock$Day
y = stock$DJIA
n = length(y)
n.iter = 15
rho.iter = vector("numeric", n.iter)
b0.iter = vector("numeric", n.iter)
b1.iter = vector("numeric", n.iter)
fit1 = lm(y ~ x)
res = fit1$res
rho.iter[1] = sum(res[1:(n-1)]*res[2:n]) / sum(res^2)
for(i in 2:n.iter){
  rho.iter[i] = sum(res[1:(n-1)]*res[2:n]) / sum(res^2)
  ystar = y[2:n] - rho.iter[i]*y[1:(n-1)]
  xstar = x[2:n] - rho.iter[i]*x[1:(n-1)]
  fit2 = lm(ystar ~ xstar)$coef
  b0.iter[i] = fit2[1]/(1-rho.iter[i])
  b1.iter[i] = fit2[2]
  res = y - b0.iter[i] - b1.iter[i]*x
}
data.frame(rho.iter,b0.iter,b1.iter)
##    rho.iter b0.iter b1.iter
## 1   0.96949     0.0  0.0000
## 2   0.96949  5210.4  4.2470
## 3   0.97148  5209.6  4.2637
## 4   0.97164  5209.5  4.2652
## 5   0.97166  5209.5  4.2653
## 6   0.97166  5209.5  4.2653
```

```
## 7   0.97166  5209.5  4.2653
## 8   0.97166  5209.5  4.2653
## 9   0.97166  5209.5  4.2653
## 10  0.97166  5209.5  4.2653
## 11  0.97166  5209.5  4.2653
## 12  0.97166  5209.5  4.2653
## 13  0.97166  5209.5  4.2653
## 14  0.97166  5209.5  4.2653
## 15  0.97166  5209.5  4.2653
```

$\beta_0 = 5209.5$, $\beta_1 = 4.2653$, $\rho = 0.97166$

Q1b — 2 points

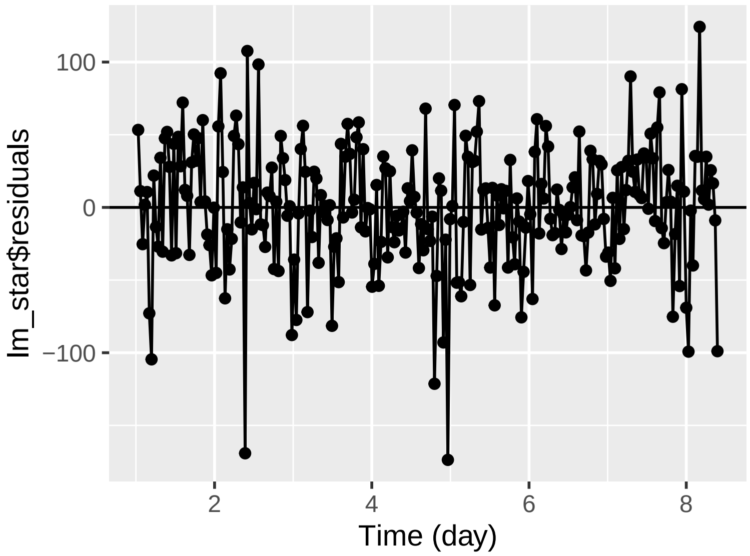
```
x = stock$Day
y = stock$DJIA
n = length(y)
rho_hat = 0.97166
x_star = x[2:n] - rho_hat*x[1:(n-1)]
y_star = y[2:n] - rho_hat*y[1:(n-1)]

ggplot(mapping = aes(x=x_star,y=y_star)) +
  geom_point() +
  labs(x='x star',y='y star', title = 'Checking Model Assumptions')
```



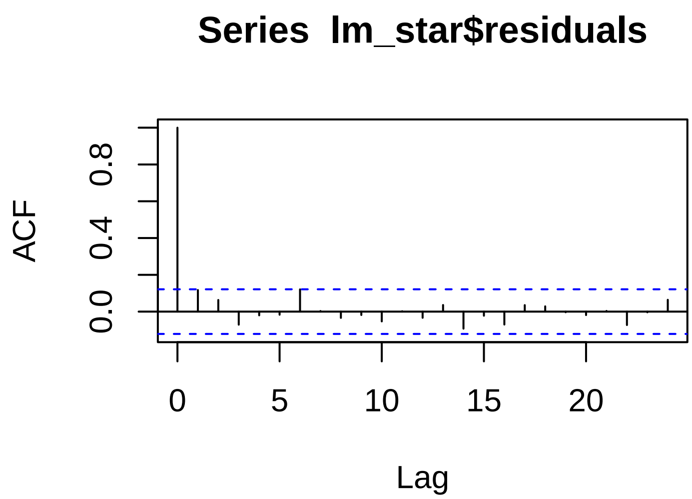
We do not see any violation of assumptions.

Q1c — 6 points

```r
lm_star = lm(y_star ~ x_star)
ggplot(mapping=aes(x=x_star,y=lm_star$residuals)) +
  geom_line() + geom_point()+
  labs(x='Time (day)','Raw Residuals') +
  geom_hline(yintercept = 0)
```




```r
library(tseries)
library(car)
runs.test(factor(lm_star$residuals > 0))
##
##  Runs Test
##
## data:  factor(lm_star$residuals > 0)
## Standard Normal = -1.17, p-value = 0.24
## alternative hypothesis: two.sided
durbinWatsonTest(lm_star)
##  lag Autocorrelation D-W Statistic p-value
##    1         0.11726         1.738   0.038
##  Alternative hypothesis: rho != 0
acf(lm_star$residuals)
```



We conclude that the residuals do not exhibit any autocorrelation.

Q1d — 3 points

We should use lm(ystar ~ xstar) since this model does not violate any SLR assumptions compared to the lm(DJIA ~ Day) model as we saw in HW 6.

```
confint(lm_star)
##                2.5 %   97.5 %
## (Intercept) 135.1902 160.0838
## x_star        1.8595   6.6711
```

95% CI for β_1: (1.8595,6.6711)

Question 2

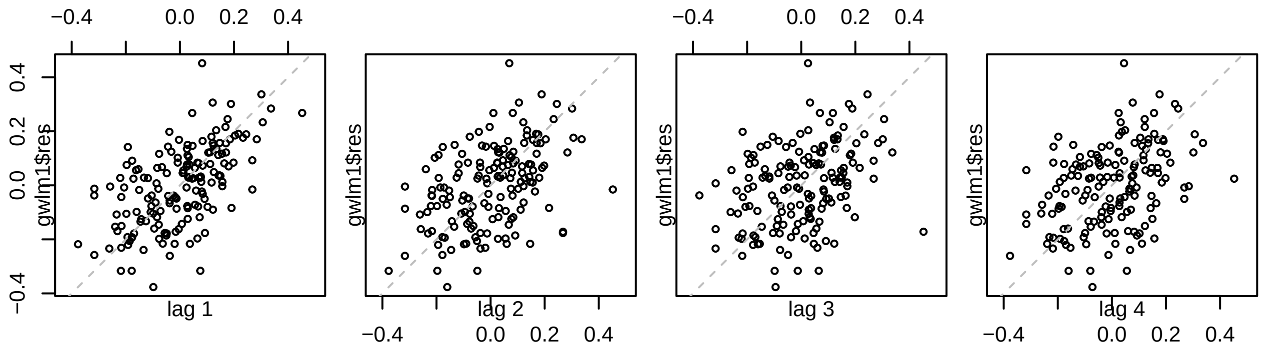
http://www.stat.uchicago.edu/~yibi/s224/data/globalwarm.txt

```
gw = read.table("http://www.stat.uchicago.edu/~yibi/s224/data/globalwarm.txt", header=T)
```

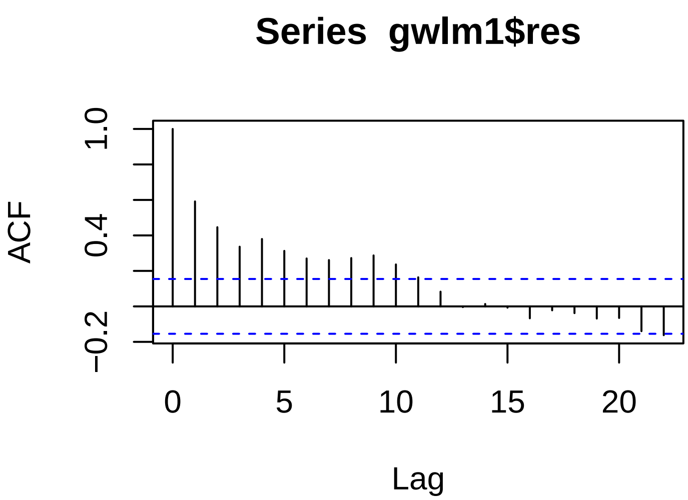
```
gwlm1 = lm(Temperature ~ Year + I(Year^2), data=gw)
```

Q2a — 3 points

```
lag.plot(gwlm1$res, lag=4, layout=c(1,4))
```




```
acf(gwlm1$res)
```



We conclude that there is autocorrelation between the residuals as seen by the lag plots for lag 1 to 4 and in the acf plot from 1-10.

Q2b — 4 points

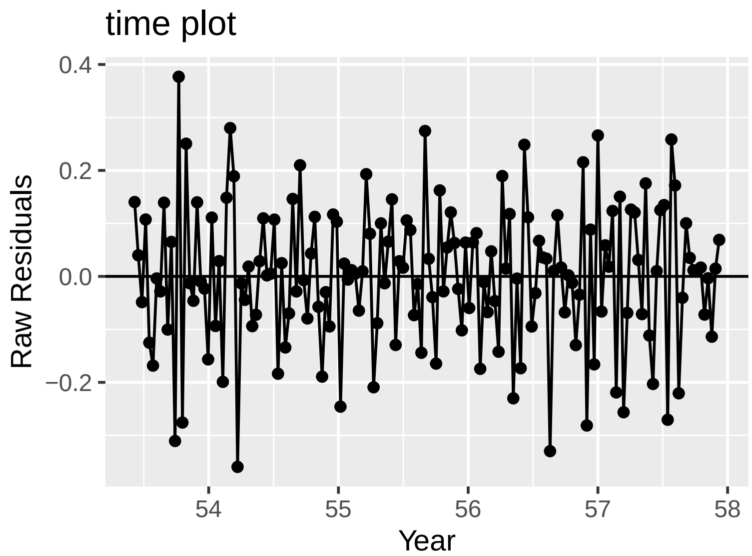
```
x = gw$Year
v = (gw$Year)^2
y = gw$Temperature
n = length(y)
n.iter = 15
rho.iter = vector("numeric", n.iter)
b0.iter = vector("numeric", n.iter)
b1.iter = vector("numeric", n.iter)
b2.iter = vector("numeric", n.iter)
fit1 = lm(y ~ x + v)
res = fit1$res
rho.iter[1] = sum(res[1:(n-1)]*res[2:n]) / sum(res^2)
for(i in 2:n.iter){
  rho.iter[i] = sum(res[1:(n-1)]*res[2:n]) / sum(res^2)
```

```
  ystar = y[2:n] - rho.iter[i]*y[1:(n-1)]
  xstar = x[2:n] - rho.iter[i]*x[1:(n-1)]
  vstar = v[2:n] - rho.iter[i]*v[1:(n-1)]
  fit2 = lm(ystar ~ xstar + vstar)$coef
  b0.iter[i] = fit2[1]/(1-rho.iter[i])
  b1.iter[i] = fit2[2]
  b2.iter[i] = fit2[3]
  res = y - b0.iter[i] - b1.iter[i]*x - b2.iter[i]*v
}
data.frame(rho.iter,b0.iter,b1.iter,b2.iter)
##    rho.iter b0.iter  b1.iter     b2.iter
## 1   0.59166    0.00  0.00000 0.000000000
## 2   0.59166  208.93 -0.22126 0.000058489
## 3   0.59249  208.96 -0.22129 0.000058498
## 4   0.59249  208.96 -0.22129 0.000058498
## 5   0.59249  208.96 -0.22129 0.000058498
## 6   0.59249  208.96 -0.22129 0.000058498
## 7   0.59249  208.96 -0.22129 0.000058498
## 8   0.59249  208.96 -0.22129 0.000058498
## 9   0.59249  208.96 -0.22129 0.000058498
## 10  0.59249  208.96 -0.22129 0.000058498
## 11  0.59249  208.96 -0.22129 0.000058498
## 12  0.59249  208.96 -0.22129 0.000058498
## 13  0.59249  208.96 -0.22129 0.000058498
## 14  0.59249  208.96 -0.22129 0.000058498
## 15  0.59249  208.96 -0.22129 0.000058498
```

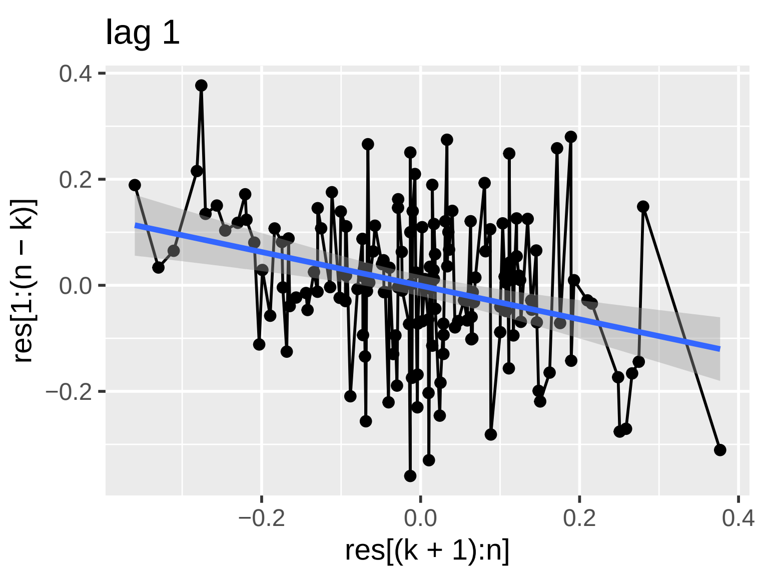
$\beta_0 = 208.96$, $\beta_1 = -0.22129$, $\beta_2 = 0.000058498$, $\rho = 0.59249$

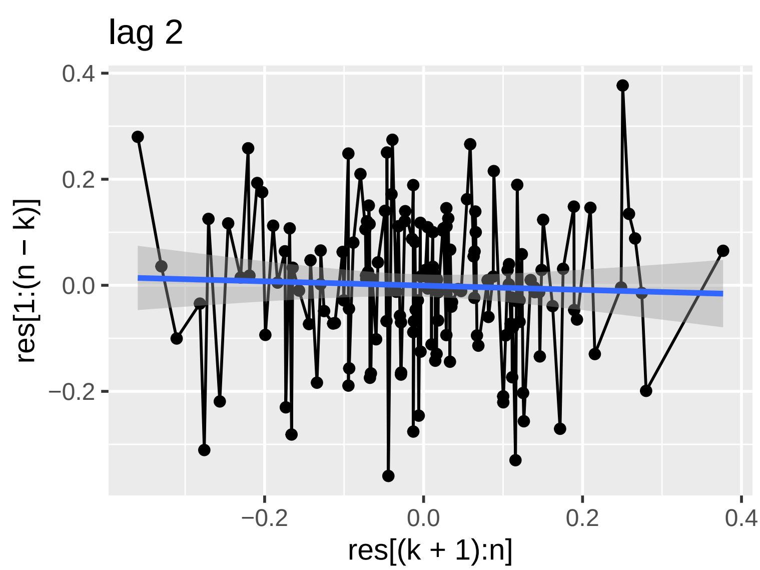
Q2c — 6 points

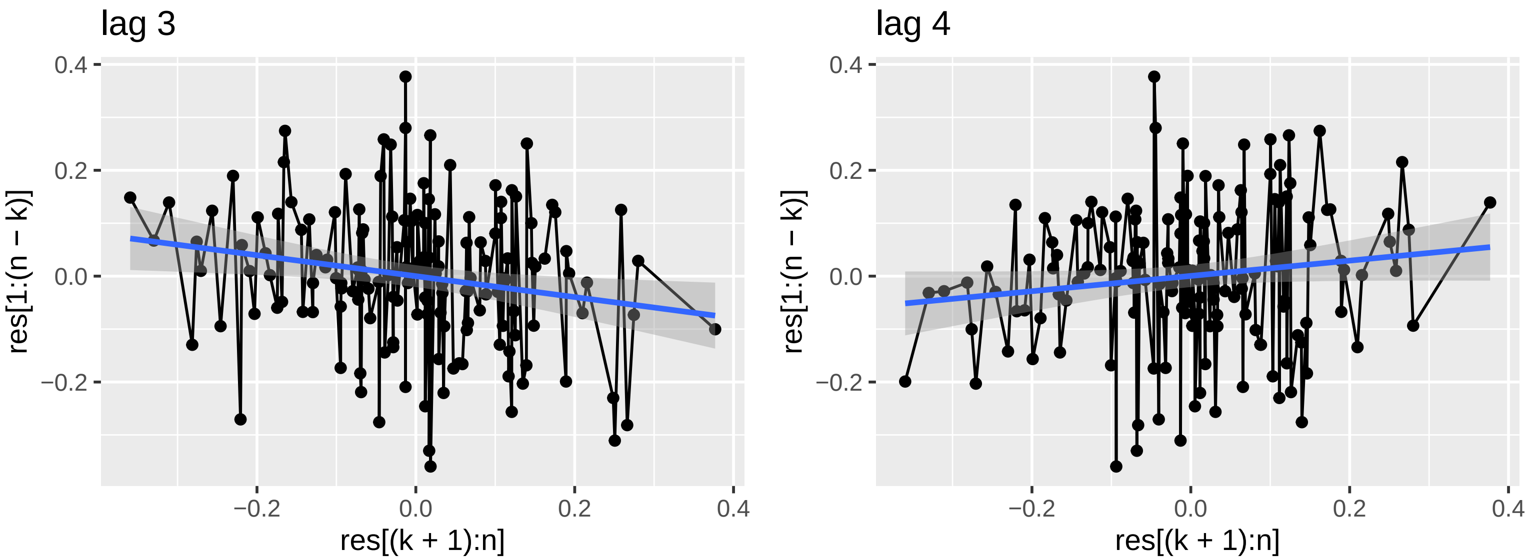
```
x_star = x[2:n] - rho_hat*x[1:(n-1)]
v_star = (x[2:n])^2 - rho_hat*(x[1:(n-1)])^2
y_star = y[2:n] - rho_hat*y[1:(n-1)]

lm_star = lm(y_star ~ x_star + v_star)
res = lm_star$residuals
ggplot(mapping=aes(x=x_star,y=res)) +
  geom_line() + geom_point()+
  labs(x='Year',y='Raw Residuals',title = 'time plot') +
  geom_hline(yintercept = 0)
```



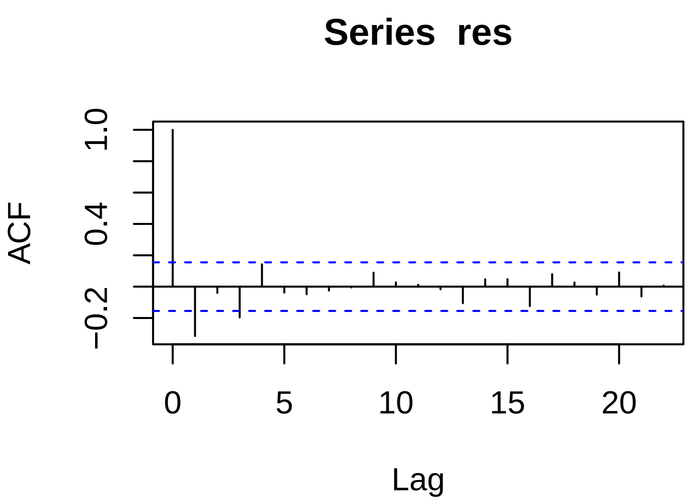

```
runs.test(factor(res > 0))
##
##  Runs Test
##
## data:  factor(res > 0)
## Standard Normal = 1.3, p-value = 0.19
## alternative hypothesis: two.sided
n = length(y_star)
for (k in c(1:4)) {
  print(ggplot(mapping=aes(x=res[(k+1):n],y=res[1:(n-k)])) +
    geom_line() + geom_point() +
    labs(title = paste('lag',k)) +
    geom_smooth(method='lm'))
}
```








```
acf(res)
```



We conclude from these tests that there is no evidence of autocorrelation.

Question 3

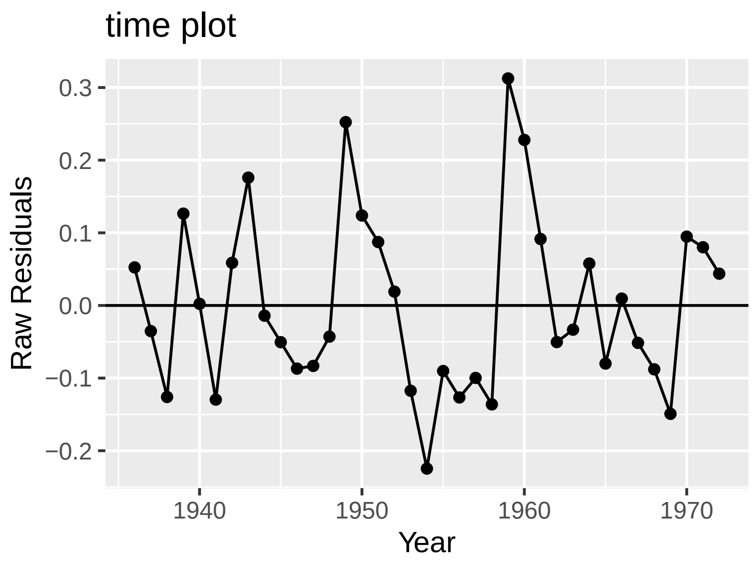
http://www.stat.uchicago.edu/~yibi/s224/data/skincancer.txt

```
skincancer = read.table("http://www.stat.uchicago.edu/~yibi/s224/data/skincancer.txt", header=
```

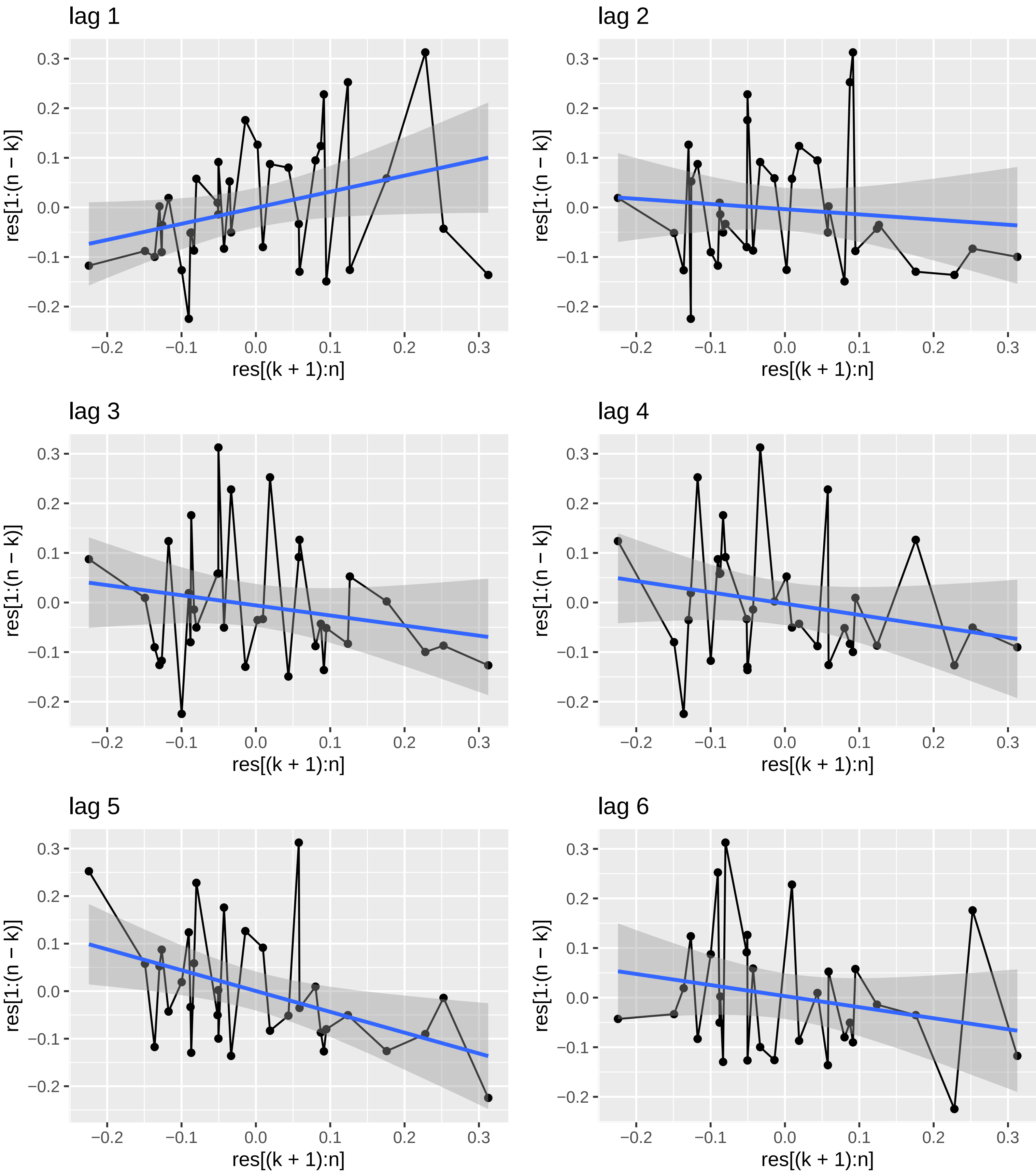
Q3a — 6 points

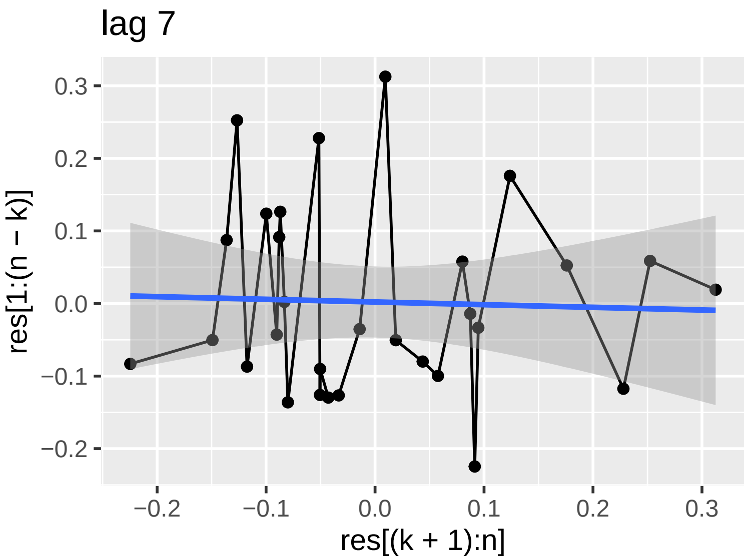
```
model2 = lm(sqrt(Melanoma) ~ Year, data=skincancer)
```

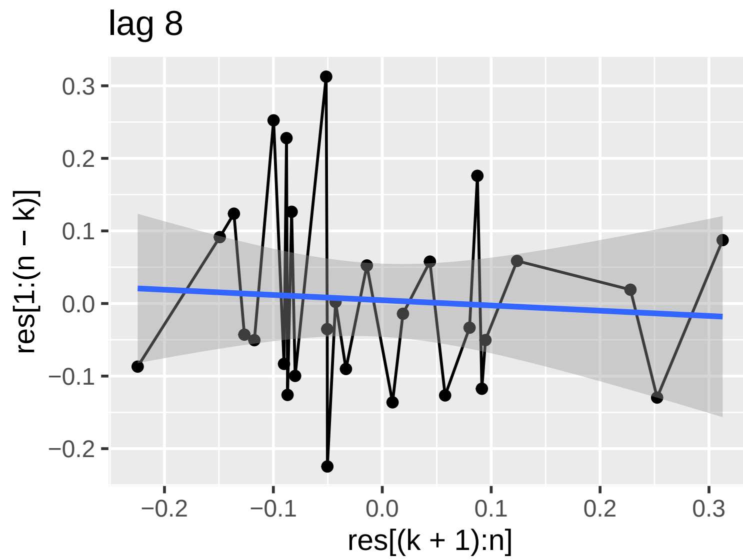
```
x=skincancer$Year
y=sqrt(skincancer$Melanoma)
res = model2$residuals
ggplot(mapping=aes(x=skincancer$Year,y=res)) +
  geom_line() + geom_point()+
  labs(x='Year',y='Raw Residuals',title = 'time plot') +
  geom_hline(yintercept = 0)
```



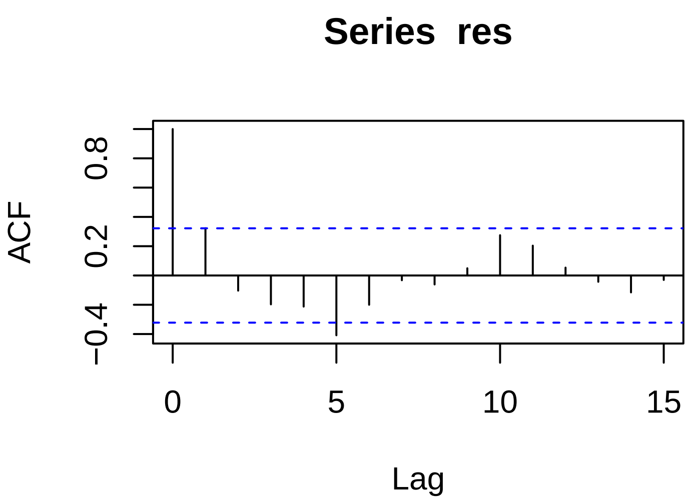

```
runs.test(factor(res > 0))
##
##  Runs Test
##
## data:  factor(res > 0)
## Standard Normal = -1.47, p-value = 0.14
## alternative hypothesis: two.sided
n = length(y)

for (k in c(1:8)) {
  print(ggplot(mapping=aes(x=res[(k+1):n],y=res[1:(n-k)])) +
    geom_line() + geom_point() +
    labs(title = paste('lag',k)) +
    geom_smooth(method='lm'))
}
```


lag 1
lag 2
lag 3
lag 4
lag 5
lag 6
res[1:(n − k)]
res[(k + 1):n]
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.2
−0.1
0.0
0.1
0.2
0.3






```
acf(res)
```



The time plot shows some evidence of runs, while the runs test fails to reject at the 5% significance level with p=0.14, indicating some potential autocorrelation. The lag-k plots appear to be showing some evidence of autocorrelation while the acf plot shows weak evidence for autocorrelation.

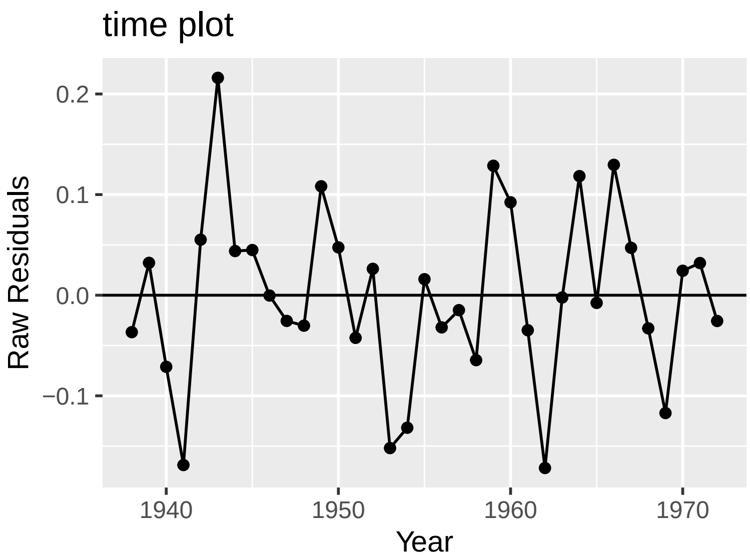
Q3b — 6 points

$$\sqrt{\text{Melanoma}_t} = \beta_0 + \beta_1 \text{Year}_t + \beta_2 \sqrt{\text{Sunspot}_t} + \beta_3 \sqrt{\text{Sunspot}_{t-1}} + \beta_4 \sqrt{\text{Sunspot}_{t-2}} + \varepsilon_t$$

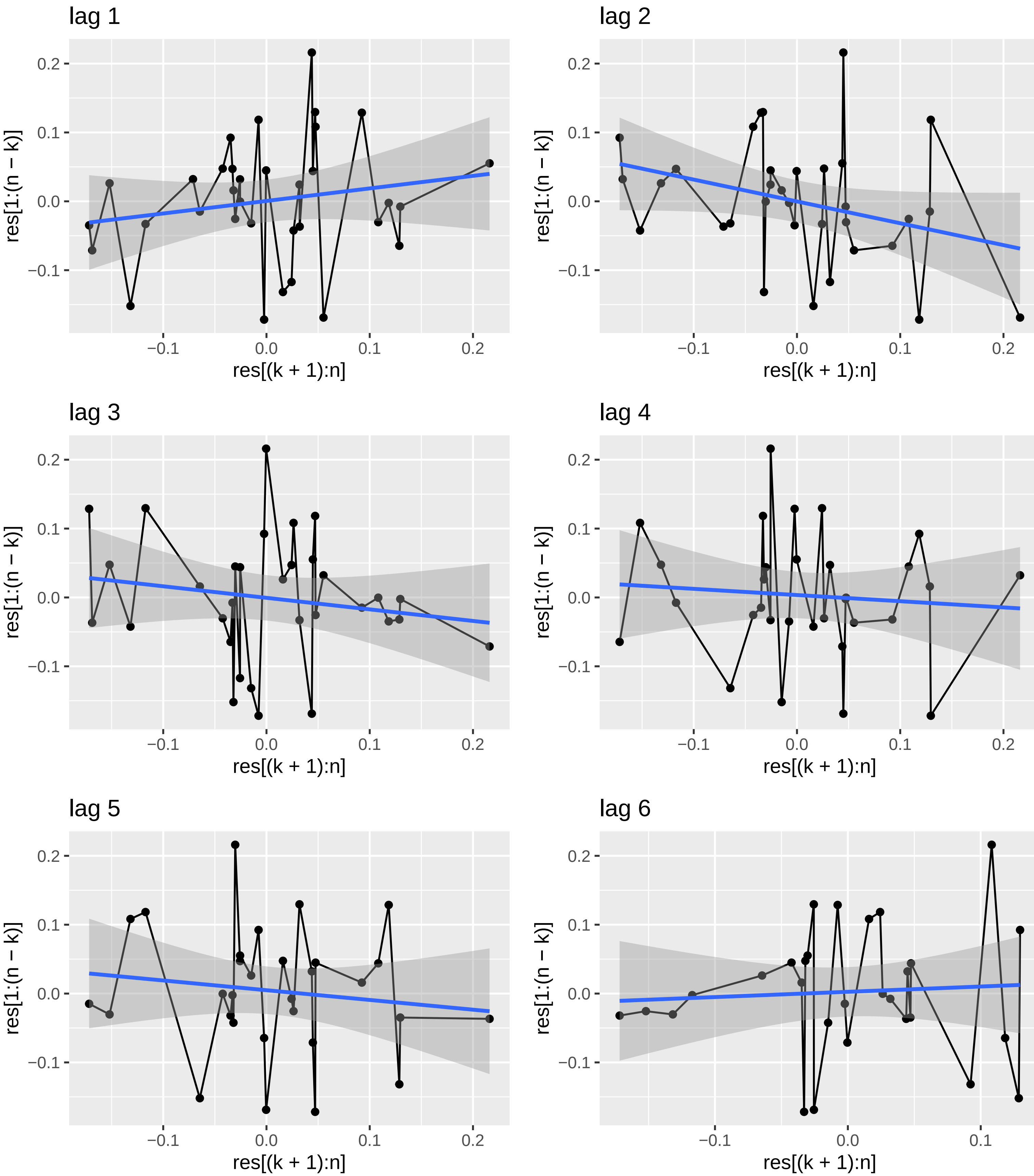
```
model3 = lm(sqrt(Melanoma[3:37]) ~ Year[3:37] + sqrt(Sunspot[3:37]) +
              sqrt(Sunspot[2:36]) + sqrt(Sunspot[1:35]), data=skincancer)
```

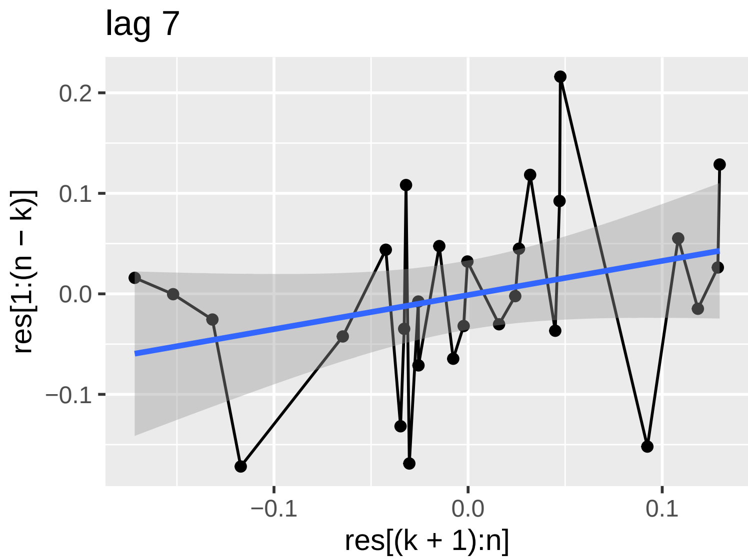
```
x=skincancer$Year[3:37]
res = model3$residuals
ggplot(mapping=aes(x=x,y=res)) +
```

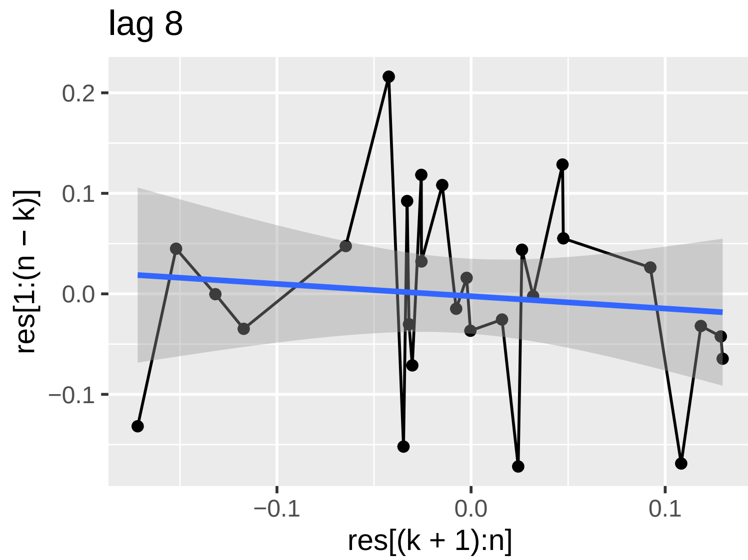
```
  geom_line() + geom_point()+
  labs(x='Year',y='Raw Residuals',title = 'time plot') +
  geom_hline(yintercept = 0)
```



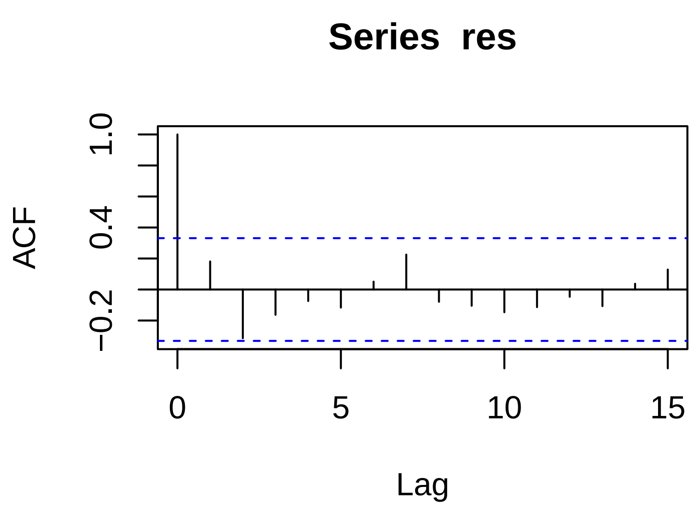

```
runs.test(factor(res > 0))
##
##  Runs Test
##
## data:  factor(res > 0)
## Standard Normal = 0.217, p-value = 0.83
## alternative hypothesis: two.sided
n = length(x)

for (k in c(1:8)) {
  print(ggplot(mapping=aes(x=res[(k+1):n],y=res[1:(n-k)])) +
    geom_line() + geom_point() +
    labs(title = paste('lag',k)) +
    geom_smooth(method='lm'))
}
```


lag 1
lag 2
lag 3
lag 4
lag 5
lag 6
res[1:(n − k)]
res[(k + 1):n]
0.2
0.1
0.0
−0.1






```
acf(res)
```



Now the tests reveal virtually no evidence of autocorrelation.

Question 4

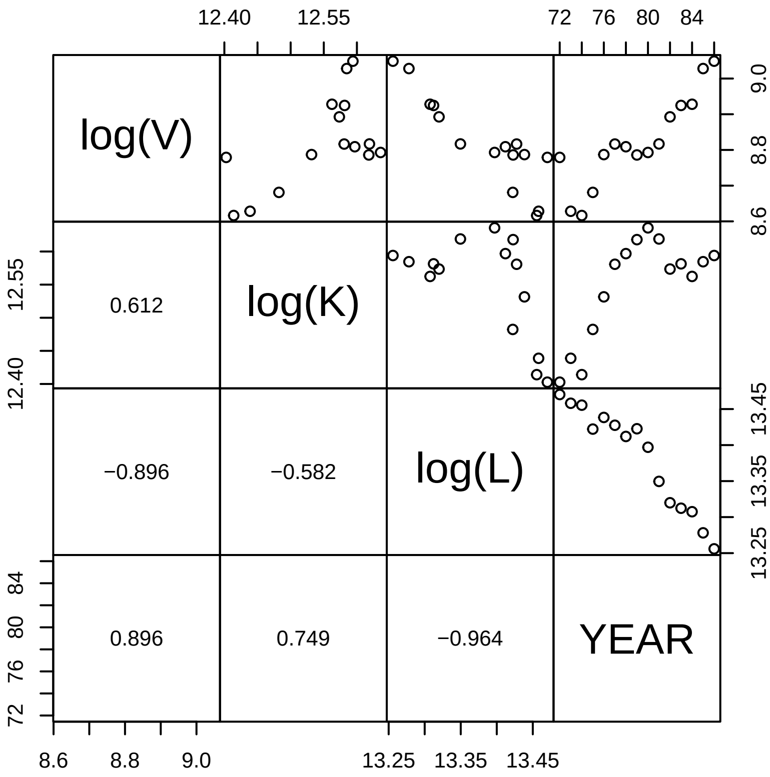
http://www.stat.uchicago.edu/~yibi/s224/data/food.txt

```
food = read.table("http://www.stat.uchicago.edu/~yibi/s224/data/food.txt", h = T)
```

Q4a — 3 points

```
lmfood = lm(log(V)~log(K)+log(L)+YEAR, data=food)
```

```
panel.cor <- function(x, y, digits = 3, prefix = "", cex.cor, ...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    text(0.5, 0.5, round(cor(x,y),digits))
}
pairs(log(V)~log(K)+log(L)+YEAR, data=food,
      gap=0,oma=c(2,2,2,2), lower.panel = panel.cor)
```




```
vif(lmfood)
##  log(K)  log(L)    YEAR
##  6.2157 38.3055 57.8073
```

From the pairwise scatterplots and the VIFs we see that labor and year are collinear (VIF > 10 and nearly -1 correlation).

Q4b — 3 points

```
summary(lm(log(V)~log(K)+log(L)+YEAR, data=food))$coef
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 19.554327  16.364689  1.194910  0.25725
## log(K)       0.044360   0.533328  0.083176  0.93521
## log(L)      -0.908236   1.427325 -0.636320  0.53758
## YEAR         0.010952   0.027843  0.393347  0.70158
summary(lm(log(V)~log(K)+log(L), data=food))$coef
##             Estimate Std. Error  t value  Pr(>|t|)
```

```
## (Intercept) 25.49288     6.0877  4.18762 0.0012593
## log(K)       0.22685     0.2536  0.89453 0.3886307
## log(L)      -1.45848     0.2734 -5.33464 0.0001780
summary(lm(log(V)~log(K)+YEAR, data=food))$coef
##              Estimate Std. Error  t value   Pr(>|t|)
## (Intercept)  9.416473  3.6443760  2.58384 0.02392806
## log(K)      -0.225628  0.3150147 -0.71625 0.48754532
## YEAR         0.028316  0.0053927  5.25079 0.00020413
```

YEAR becomes significant at the 5% level when L is removed from the model while L becomes significant when YEAR is removed from the model. I do not believe there was technological development in the food sector since its coefficient is not significant in any model. We can conclude that increasing labor input would lead to decrease in the output based on the second model.

Question 5

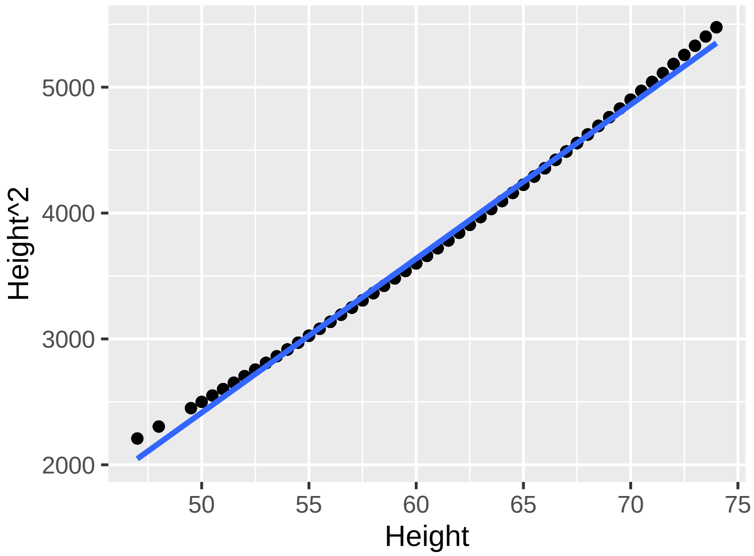
http://www.stat.uchicago.edu/~yibi/s224/data/fevdata.txt

```
fevdata = read.table("http://www.stat.uchicago.edu/~yibi/s224/data/fevdata.txt", header = TRUE
fevdata$sex = factor(fevdata$sex, labels=c("Female","Male"))
fevdata$smoke = factor(fevdata$smoke, labels=c("Nonsmoker","Smoker"))
```

Q5a — 2 points

```
m.nonsmokers = subset(fevdata, sex=="Male" & smoke=="Nonsmoker")
mod1 = lm(log(fev) ~ ht, data=m.nonsmokers)
mod2 = lm(log(fev) ~ ht + I(ht^2), data=m.nonsmokers)
summary(mod1)$coef
##              Estimate Std. Error t value    Pr(>|t|)
## (Intercept) -2.219203   0.082244 -26.983  3.9757e-83
## ht           0.051417   0.001330  38.659 3.3903e-120
summary(mod2)$coef
##                Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -1.65335452 0.76173831 -2.17050 0.030735
## ht           0.03271899 0.02505889  1.30568 0.192638
## I(ht^2)      0.00015284 0.00020455  0.74721 0.455507
```

```
vif(mod2)
##      ht I(ht^2)
##  354.47  354.47
ggplot(data = m.nonsmokers, aes(x=ht,y=ht^2)) +
  geom_point() +
  labs(x='Height',y='Height^2') +
  geom_smooth(method='lm')
```

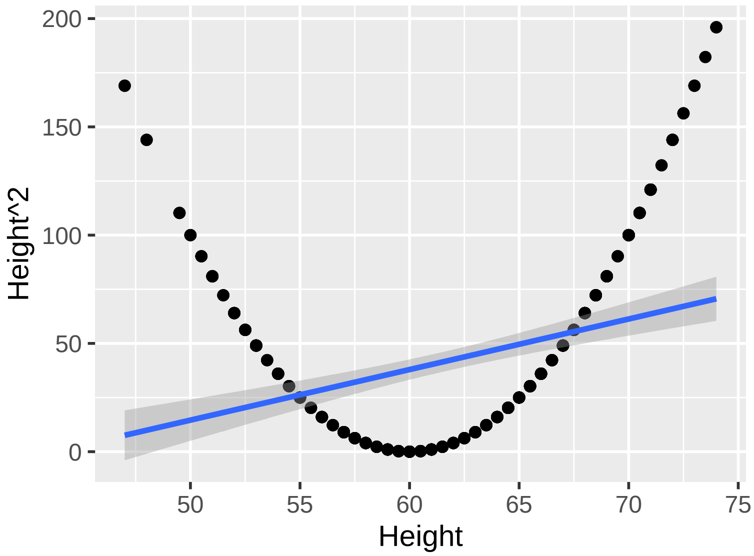


This happens because ht and ht^2 are collinear (VIF »> 10 and plot shows they are nearly exactly linearly related).

Q5b — 2 points

```
mod3 = lm(log(fev) ~ ht + I((ht-60)^2), data=m.nonsmokers)
summary(mod3)$coef
##                    Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)     -2.20358358 0.08491653 -25.94999  2.1584e-79
## ht               0.05105996 0.00141410  36.10784 1.5829e-112
## I((ht - 60)^2)   0.00015284 0.00020455   0.74721  4.5551e-01
```

```
vif(mod3)
##             ht I((ht - 60)^2)
##         1.1288         1.1288
ggplot(data = m.nonsmokers, aes(x=ht,y=(ht-60)^2)) +
  geom_point() +
  labs(x='Height',y='Height^2') +
  geom_smooth(method='lm')
```



Now there is no collinearity between the two as evidenced by a small VIF and the graph shown. It is different from Q5b because the covariates are no longer collinear/linear combinations of each other.