STATS 207 (Time Series Analysis) HW1 R Code

Chih-Hsuan ‘Carolyn’ Kao (chkao831 at stanford dot edu)

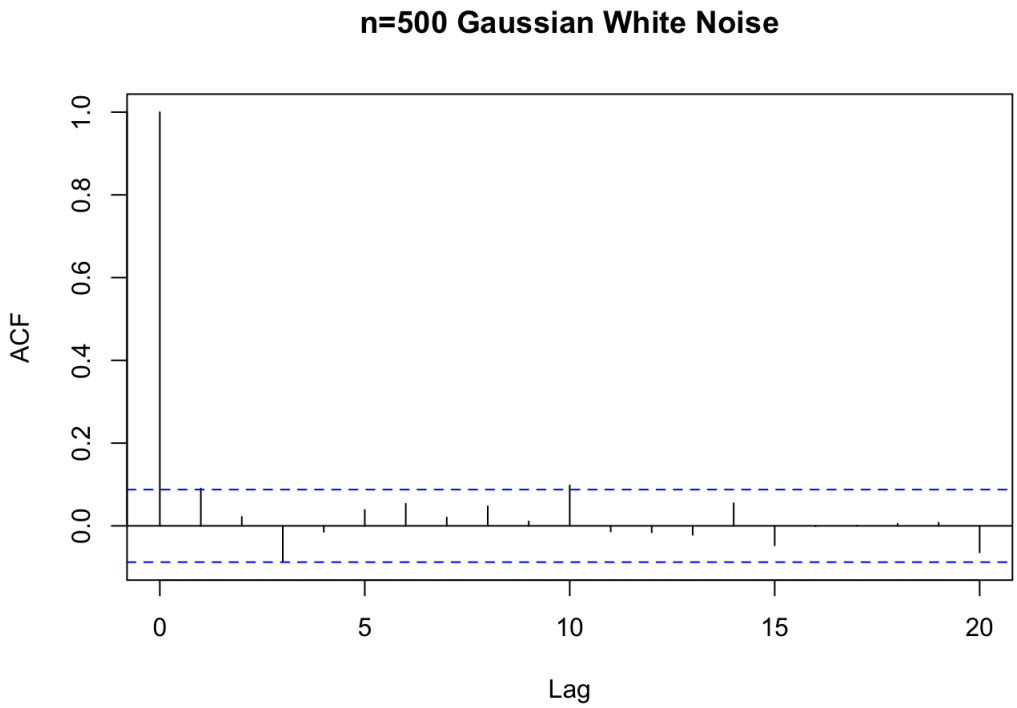
Oct 5, 2020 (Fall 2020)

All references (exercises, examples, equations, etc.) are to Shumway & Stoffer

Q3. Exercises 1.20 and 1.21

a. Simulate a series of n = 500 Gaussian white noise observations as in Example 1.8 and compute the sample ACF to lag 20. Compare the sample ACF you obtain to the actual ACF.

```
gaussian_wn = rnorm(500,0,1) #simulate Gaussian White Noise series
acf(gaussian_wn,lag=20, main="n=500 Gaussian White Noise")
```



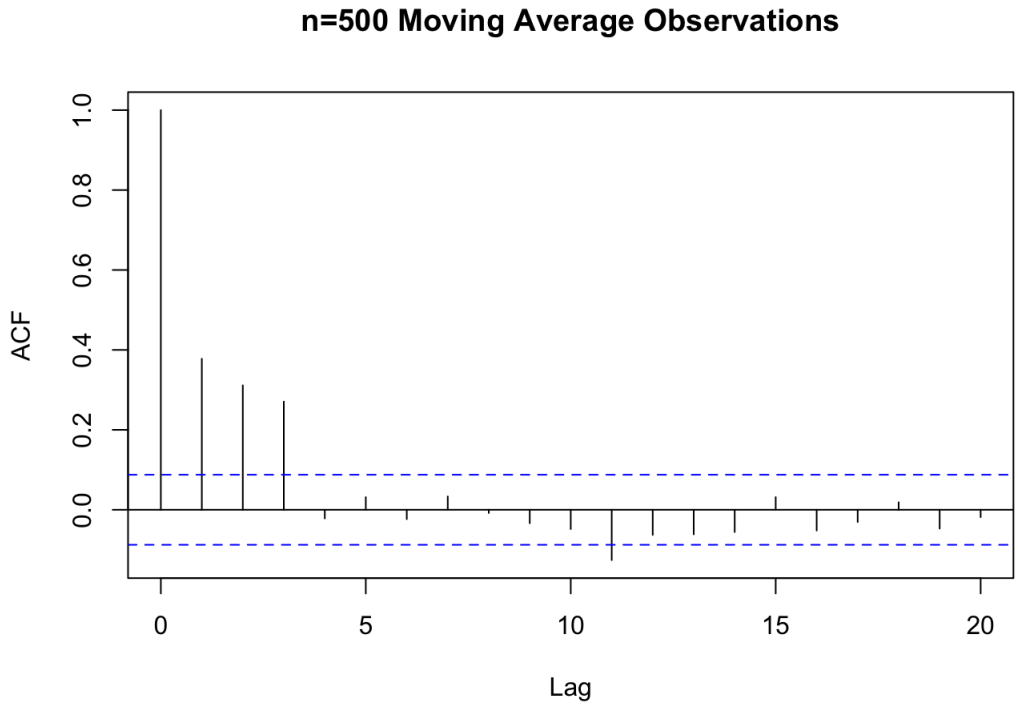
Compared to the population ACF, which is

$$\rho(h) = \begin{cases} 1, & \text{if } h = 0 \\ 0, & \text{if } h \neq 0 \end{cases}$$

The sampled Gaussian white noice observations (n=500) do not vary significantly from zero within lag 1, which roughly aligns with the theoretical results.

b. Simulate a series of n = 500 moving average observations as in Example 1.9 and compute the sample ACF to lag 20. Compare the sample ACF you obtain to the actual ACF

```
f = arima.sim(model=list(ma=rep(1,3)/3), rand.gen = rnorm, n=500)
acf(f,
    lag=20, #compute sampled rho to lag 20
    main = "n=500 Moving Average Observations",
    na.action=na.pass)
```



The system below, independent of time t, theoretically represents the population covariance function, which decreases as the lag (time separation) increases.

$$\gamma(h) = \begin{cases} \frac{3}{9}\sigma_w^2, & \text{if } h = 0 \\ \frac{2}{9}\sigma_w^2, & \text{if } h = 1 \\ \frac{1}{9}\sigma_w^2, & \text{if } h = 2 \\ 0, & \text{if } |h| > 2 \end{cases}$$

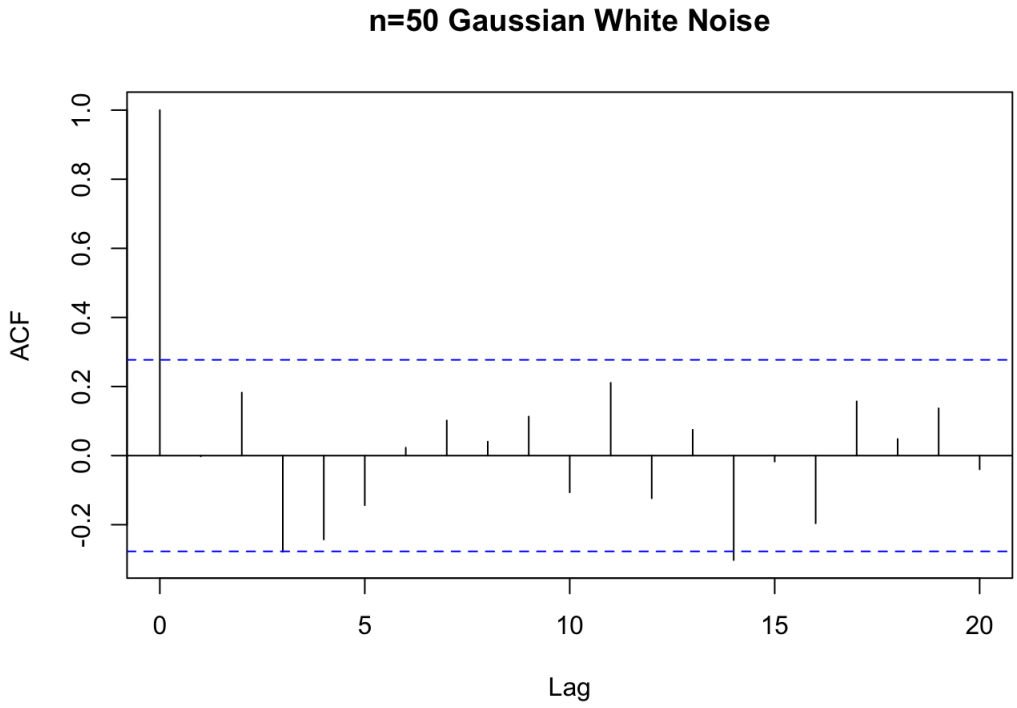
Denote $Y_t = \frac{1}{3}(w_{t-1} + w_t + w_{t+1})$, $Var(Y_t) = \frac{1}{3}$, the population autocorrelation becomes

$$\rho(h) = \begin{cases} 1, & \text{if } h = 0 \\ \frac{2}{3}, & \text{if } h = 1 \\ \frac{1}{3}, & \text{if } h = 2 \\ 0, & \text{if } |h| > 2 \end{cases}$$

From the simulation of n = 500 moving average observations as shown previously in the ACF plot, its pattern roughly aligns with the theoretical ground truth (while its values do not perfectly agree, as the sampled ρ does not vanish to within the confidence band when lag equals 3).

(c-1) Repeat parts (a) using only n = 50. How does changing n affect the results?

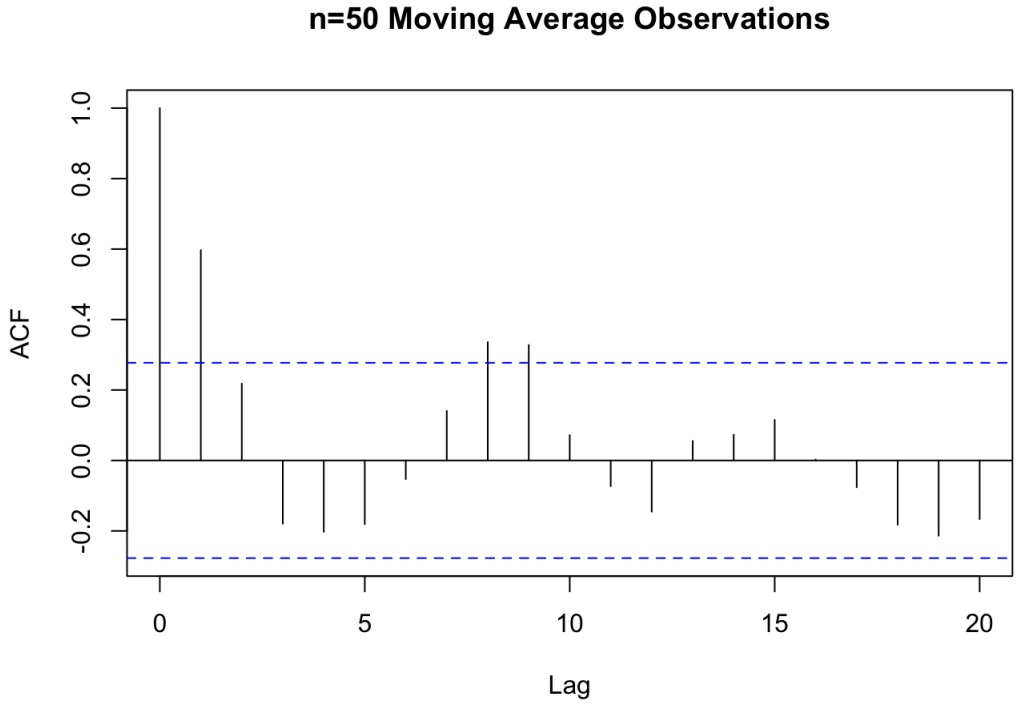
```
gaussian_wn = rnorm(50,0,1)
acf(gaussian_wn,lag=20, main="n=50 Gaussian White Noise")
```



Compared to the previous part with 500 Gaussian white noise observations, when n = 50, the ACF values do differ significantly from zero for certain lags with higher fluctuation, depending on the randomness of simulation. Hence, the decrease of sample size results in a disagreement between actual ACF and sampled ACF.

(c-2) Repeat parts (b) using only n = 50. How does changing n affect the results?

```
moving_average = rnorm(50, 0, 1)
f = filter(moving_average, sides=2, rep(1,3)/3)
acf(f,
    lag=20,
    main = "n=50 Moving Average Observations",
    na.action=na.pass)
```



Compared to the previous part with 500 moving average observations, when n = 50, the ACF values do differ significantly from zero for certain lags with higher fluctuation, depending on the randomness of simulation. Hence, the decrease of sample size similarly results in a disagreement between actual ACF and sampled ACF.

Q6. Exercises 2.2 Pollution, Temperature and Mortality

The data in this problem is extracted from a study by Shumway et al. (1988) of the possible effects of temperature and pollution on weekly mortality in Los Angeles County.

Firstly, based on formula from textbook, I practically replicate a regression model based on formula (2.21), which is

$$M_t = \beta_0 + \beta_1 t + \beta_2 (T_t - T.) + \beta_3 (T_t - T.)^2 + \beta_4 P_t + w_t$$

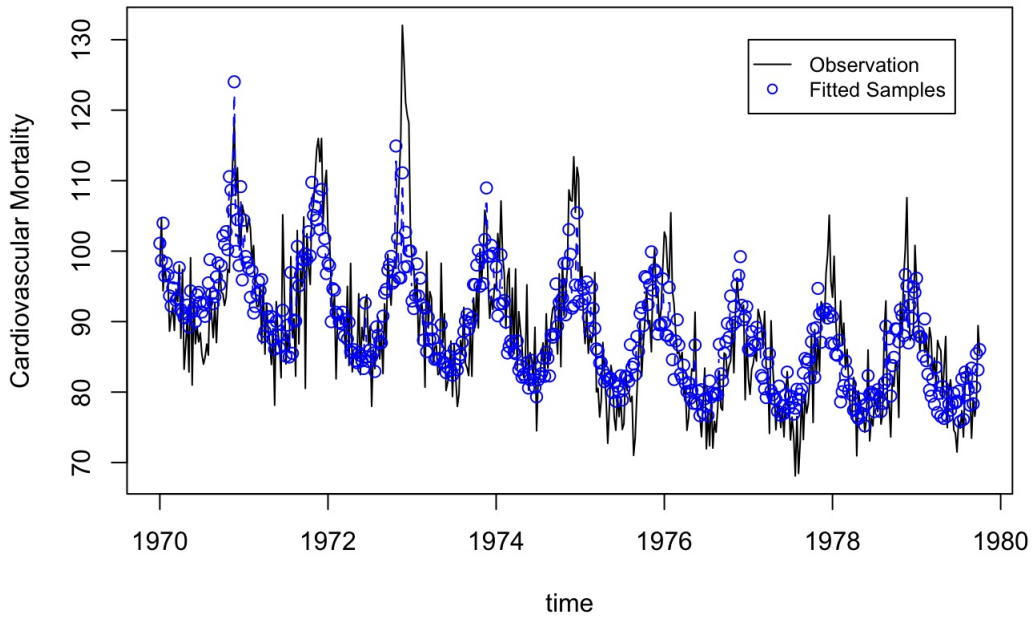
where M_t denotes cardiovascular mortality, T_t denotes temperature and P_t denotes the particulate levels. $T.$ is mean of temperature. There are 508 six-day smoothed averages obtained by filtering daily values over the 10 year period 1970-1979.

```
library(astsa)
```

```
## Warning: package 'astsa' was built under R version 3.6.2
```

```
n = length(tempr)
#here, adjust temperature for its mean to avoid collinearity problems
temp_diff = tempr - mean(tempr)
temp_diff_sqrt = temp_diff^2
t = time(cmort)

model_2_21 = lm(cmort~ t + temp_diff + temp_diff_sqrt + part, na.action=NULL)
plot(t, cmort, type="l", pch=19, col="black", xlab="time", ylab = "Cardiovascular Mortality")
points(t, model_2_21$fitted.values, pch=1, col="blue", type="b", lty=2)
legend(x=1977,y=130,c("Observation","Fitted Samples"),cex=.8,col=c("black","blue"),pch=c(NA,1), lty=1:0)
```




```
summary(model_2_21)
```

```
##
## Call:
## lm(formula = cmort ~ t + temp_diff + temp_diff_sqrt + part, na.action = NULL)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.0760  -4.2153  -0.4878   3.7435  29.2448
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.831e+03  1.996e+02   14.19  < 2e-16 ***
## t              -1.396e+00  1.010e-01  -13.82  < 2e-16 ***
## temp_diff      -4.725e-01  3.162e-02  -14.94  < 2e-16 ***
## temp_diff_sqrt  2.259e-02  2.827e-03    7.99 9.26e-15 ***
## part            2.554e-01  1.886e-02   13.54  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.385 on 503 degrees of freedom
## Multiple R-squared:  0.5954, Adjusted R-squared:  0.5922
## F-statistic:   185 on 4 and 503 DF,  p-value: < 2.2e-16
```

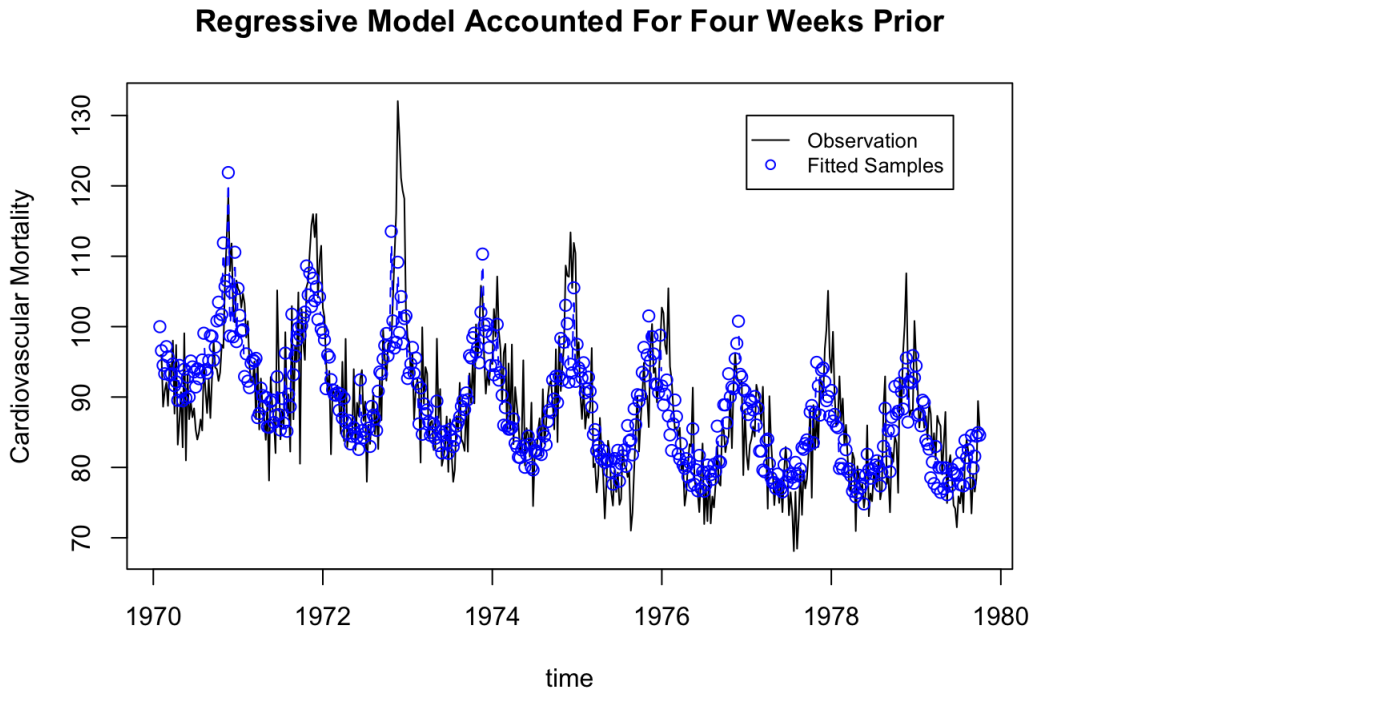
a. Add another component to the regression in equation (2.21) that accounts for the particulate count four weeks prior; that is, add P_{t-4}. Draw fitted model.

Here, I generate a regressive model on the basis of the formula

$$M_t = \beta_0 + \beta_1 t + \beta_2(T_t - T.) + \beta_3(T_t - T.)^2 + \beta_4 P_t + \beta_5 P_{t-4} + w_t$$

```
model_2_21_modified =
  lm(cmort[5:n] ~
       t[5:n]
     + temp_diff[5:n]
     + temp_diff_sqrt[5:n]
     + part[5:n]
     + part[1:(n-4)],
     na.action=NULL)

plot(t[5:n], cmort[5:n], type="l", pch=19, col="black", xlab="time", ylab = "Cardiovascular Mortality", main
="Regressive Model Accounted For Four Weeks Prior")
points(t[5:n], model_2_21_modified$fitted.values, pch=1, col="blue", type="b", lty=2)
legend(x=1977,y=130,c("Observation","Fitted Samples"),cex=.8,col=c("black","blue"),pch=c(NA,1), lty=1:0)
```



```
summary(model_2_21_modified)
```

```
##
## Call:
## lm(formula = cmort[5:n] ~ t[5:n] + temp_diff[5:n] + temp_diff_sqrt[5:n] +
##     part[5:n] + part[1:(n - 4)], na.action = NULL)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.228  -4.314  -0.614   3.713  27.800
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.808e+03  1.989e+02  14.123  < 2e-16 ***
## t[5:n]              -1.385e+00  1.006e-01 -13.765  < 2e-16 ***
## temp_diff[5:n]      -4.058e-01  3.528e-02 -11.503  < 2e-16 ***
## temp_diff_sqrt[5:n]  2.155e-02  2.803e-03   7.688 8.02e-14 ***
## part[5:n]            2.029e-01  2.266e-02   8.954  < 2e-16 ***
## part[1:(n - 4)]      1.030e-01  2.485e-02   4.147 3.96e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.287 on 498 degrees of freedom
## Multiple R-squared:  0.608,  Adjusted R-squared:  0.6041
## F-statistic: 154.5 on 5 and 498 DF,  p-value: < 2.2e-16
```

b. Report on the model order, error sum of squares, model degrees of freedom, mean-squared error, R^2, AIC, and BIC.

```
aic_2_21 = AIC(model_2_21)/n - log(2*pi)
aic_2_21_modified = AIC(model_2_21_modified)/(n-4) - log(2*pi)
bic_2_21 = BIC(model_2_21)/n - log(2*pi)
bic_2_21_modified = BIC(model_2_21_modified)/(n-4) - log(2*pi)
df = data.frame(
  model  = c("model_2_21", "model_2_21_modified"),
  order_k = c(5, 6),
  SSE = c(tail(anova(model_2_21)[, 2], 1), tail(anova(model_2_21_modified)[, 2], 1)),
  df = c(model_2_21$df, model_2_21_modified$df),
  MSE = c(anova(model_2_21)['Residuals', 'Mean Sq'],anova(model_2_21_modified)['Residuals','Mean Sq']),
  R_squared = c(summary(model_2_21)$r.squared, summary(model_2_21_modified)$r.squared),
  AIC = c(aic_2_21, aic_2_21_modified),
  BIC = c(bic_2_21, bic_2_21_modified)
)

df
```

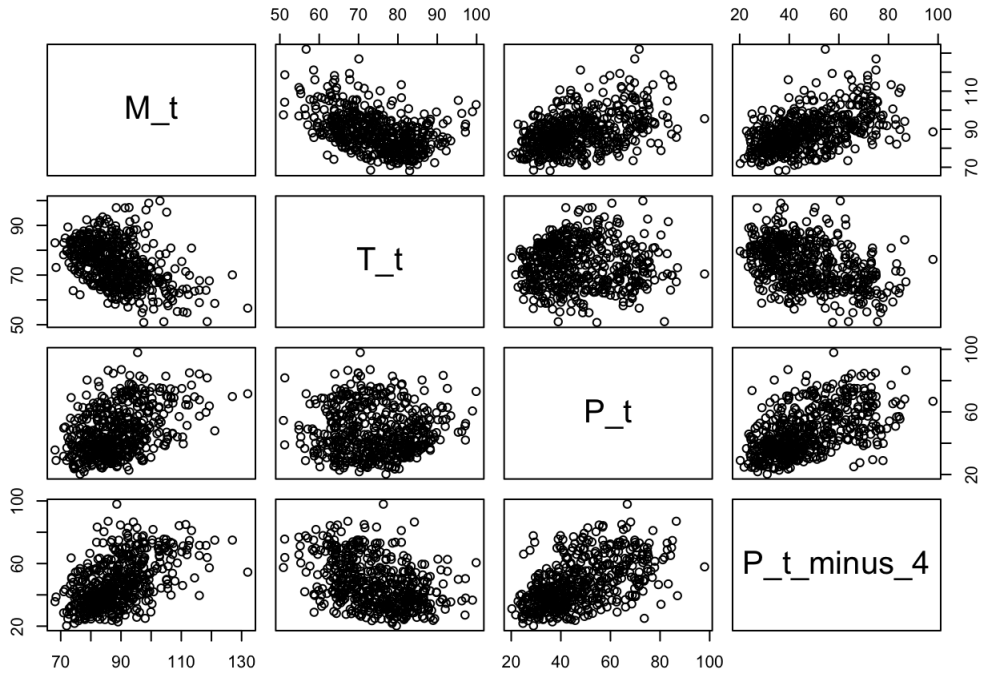
model <fctr>
model_2_21
model_2_21_modified

2 rows | 1-1 of 8 columns

c. Draw a scatterplot matrix of M_t, T_t, P_t, and P_{t-4} and then calculate the pairwise correlations between the series. Compare the relationship between M_t and P_t versus M_t and P_{t-4}.

```
M_t <- cmort[5:n]
T_t <- tempr[5:n]
P_t <- part[5:n]
P_t_minus_4 <- part[1:(n-4)]

pairs(data.frame(M_t, T_t, P_t, P_t_minus_4))
```




```
cor(data.frame(M_t, T_t, P_t, P_t_minus_4))
```

```
##                    M_t        T_t        P_t P_t_minus_4
## M_t          1.0000000 -0.4369648  0.4422896   0.5209993
## T_t         -0.4369648  1.0000000 -0.0148241  -0.3990848
## P_t          0.4422896 -0.0148241  1.0000000   0.5340505
## P_t_minus_4  0.5209993 -0.3990848  0.5340505   1.0000000
```

The correlation between cardiovascular mortality and the particulate count four weeks prior (0.52) is stronger than that of cardiovascular mortality and the pure particulate (0.44). Hence, we could conclude that adding the particulate count four weeks prior helps strengthen the