

# HW 04

Name: Paul Holaway, NetID: paulch2

Due: 9/22/2022 11:59pm

- Unless stated otherwise,  $W_t$  is a white noise process with variance  $\sigma_w^2$ .
  - ( $W_t$  are independent with zero means and variance  $\sigma_w^2$ .)
- Show your full work to receive full credit.

## Question 1

For the mortality data examined in class (see Example 2.2 of [TSA4] for reference):

- (a) Add another component to the regression  $M_t = \beta_0 + \beta_1 t + \beta_2(T_t - T_{\cdot}) + \beta_3(T_t - T_{\cdot})^2 + \beta_4 P_t + W_t$  that accounts for the particulate count four weeks prior; that is, add  $P_{t-4}$  to the regression. State your conclusion.

```
#Setup
cmort = cmort
tempr = tempr
temp = tempr - mean(tempr)
temp2 = temp^2
trend = time(cmort)
part = part
mort = ts.intersect(cmort, trend, temp, temp2, part, partL4 = lag(part, -4))
#Model Creation
model = lm(cmort ~ trend + temp + temp2 + part + partL4, data = mort)
summary(model)
```

```
##
## Call:
## lm(formula = cmort ~ trend + temp + temp2 + part + partL4, data = mort)
##
## 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 ***
## trend       -1.385e+00  1.006e-01 -13.765  < 2e-16 ***
## temp        -4.058e-01  3.528e-02 -11.503  < 2e-16 ***
## temp2        2.155e-02  2.803e-03   7.688  8.02e-14 ***
## part         2.029e-01  2.266e-02   8.954  < 2e-16 ***
```

```
## partL4          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
```

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: cmort
##           Df Sum Sq Mean Sq F value    Pr(>F)
## trend      1 10280.4  10280.4 260.052 < 2.2e-16 ***
## temp       1  8610.4   8610.4 217.806 < 2.2e-16 ***
## temp2      1  3475.8   3475.8  87.924 < 2.2e-16 ***
## part       1  7492.7   7492.7 189.535 < 2.2e-16 ***
## partL4     1    679.9    679.9  17.198 3.96e-05 ***
## Residuals 498 19687.0     39.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#AIC Calculation
```

```
num = length(cmort)
AIC(model)/num - log(2*pi)
```

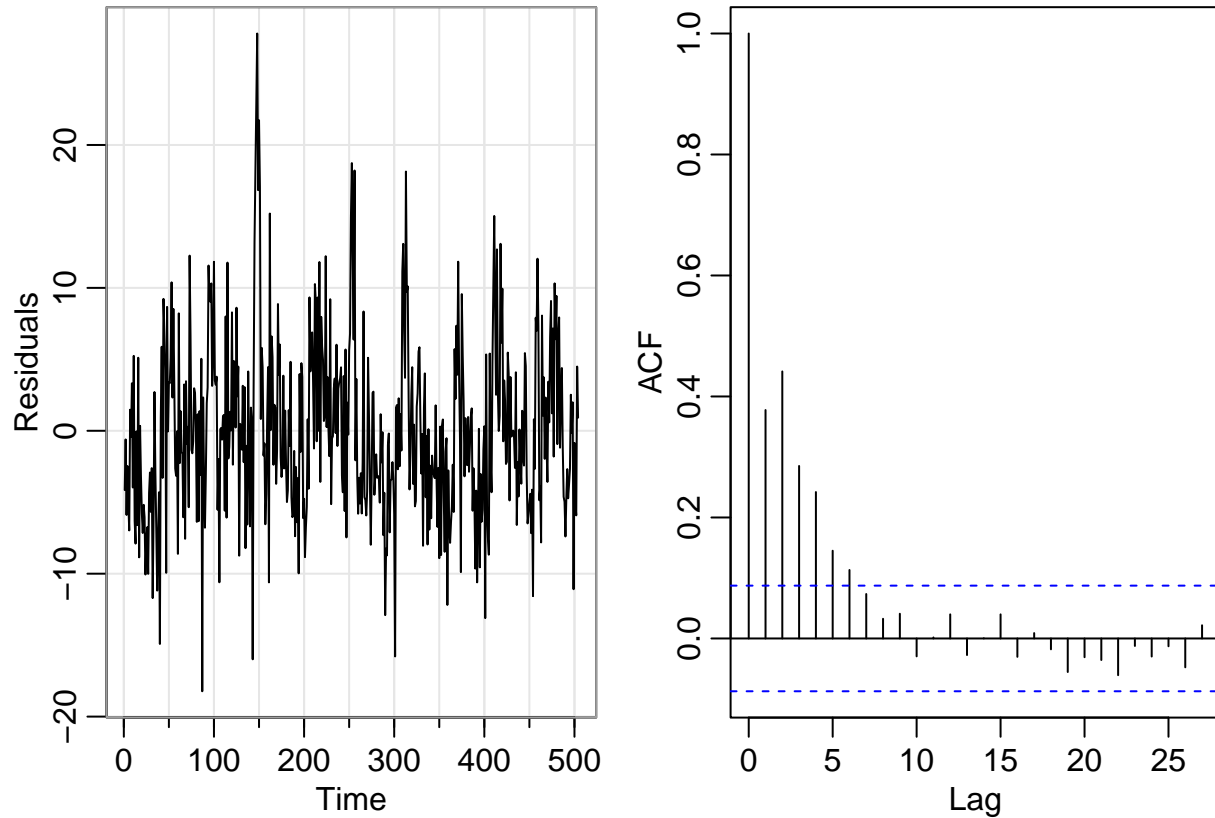
```
## [1] 4.641492
```

```
#BIC Calculation
```

```
BIC(model)/num - log(2*pi)
```

```
## [1] 4.699677
```

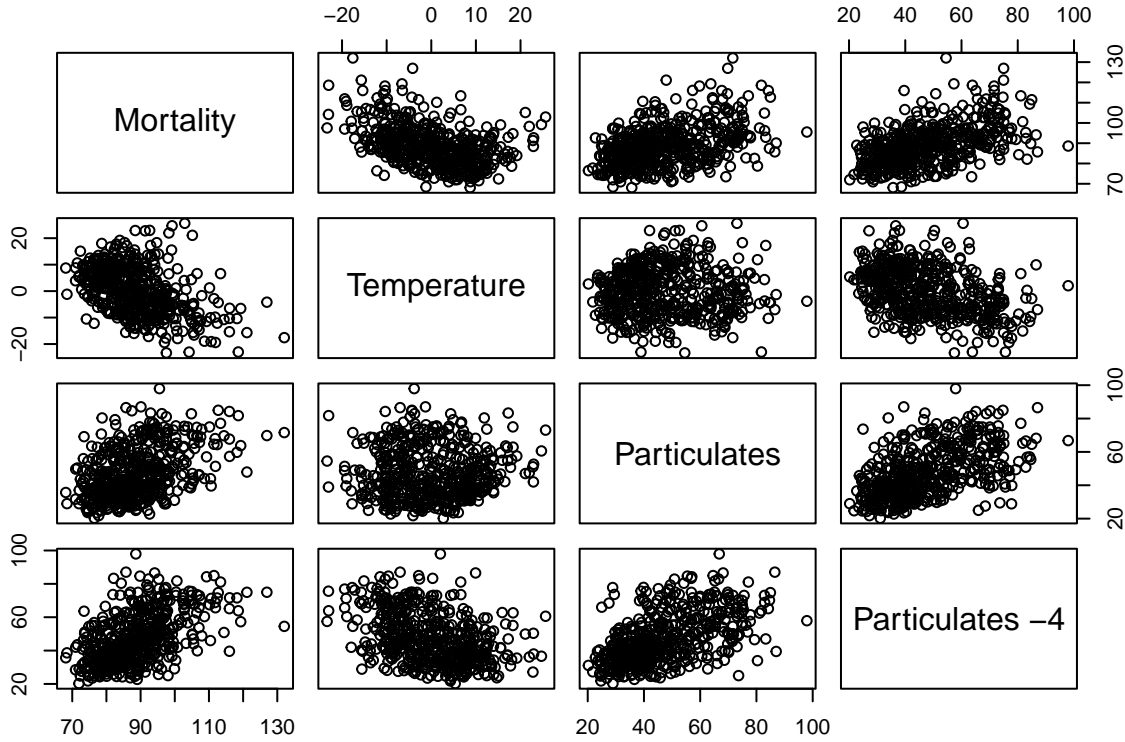
```
#Diagnostic Plots
par(mfrow = c(1,2))
tsplot(model$residuals, ylab = "Residuals")
acf(model$residuals)
```



I would leave the particulate level from four weeks ago ( $P_{t-4}$ ) in the regression. Its coefficient is statistically different than zero both individually (from the coefficient test given in the standard output;  $t$ ) and combined with the other coefficients (from the ANOVA table;  $F$ ). We then look at the AIC and BIC and see that they both are lower with it in the model than not (see table 2.2 in TSA4). The  $R^2$  is also a bit higher too. The diagnostic plots show that while we still have a little bit of a cyclical trend in the residuals over time and our ACF does not look perfectly like white noise. However, they are a bit better than without including  $P_{t-4}$  in the model. Therefore it would make sense to leave  $P_{t-4}$  in the regression model.

- (b) Draw a scatter plot 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  $P_{t-4}$ .

```
pairs(cbind(Mortality = mort[,1] , Temperature = mort[,3] , Particulates = mort[,5] ,
            `Particulates -4` = mort[,6]))
```



```
round(cor(cbind(Mortality = mort[,1] , Temperature = mort[,3] , Particulates = mort[,5] ,
                `Particulates -4` = mort[,6])), 2)
```

```
##           Mortality Temperature Particulates Particulates -4
## Mortality           1.00      -0.44          0.44          0.52
## Temperature        -0.44           1.00        -0.01        -0.40
## Particulates         0.44        -0.01           1.00          0.53
## Particulates -4      0.52        -0.40          0.53           1.00
```

- $Corr(M_t, P_t) \approx 0.44$
- $Corr(M_t, P_{t-4}) \approx 0.52$

Based on the results from calculating the pairwise correlations, we can see that there is a stronger correlation between the current mortality and the particulate levels from four weeks ago than the current level. Maybe this could mean that the time it takes to die from cardiovascular issues is a delayed response to being exposed to the particulate from the past?

## Question 2

Consider the time series

$$X_t = \beta_1 + \beta_2 t + G_t,$$

where  $\beta_1$  and  $\beta_2$  are known constants and  $G_t$  is a stationary process with mean function  $\mu_g$  and autocovariance function  $\gamma_g(h)$ . Find mean function, autocovariance function, and autocorrelation function of  $Y_t = X_t - X_{t-1}$ . (Use  $\mu_g$  and  $\gamma_g$  to describe mean function and autocovariance/autocorrelation functions.) Is  $Y_t$  stationary? Why or why not?

$$\begin{aligned}\mathbb{E}(Y_t) &= \mathbb{E}(X_t - X_{t-1}) = \mathbb{E}(X_t) - \mathbb{E}(X_{t-1}) \\ &= \mathbb{E}(\beta_1 + \beta_2 t + G_t) - \mathbb{E}(\beta_1 + \beta_2(t-1) + G_{t-1}) \\ &= \beta_1 + \beta_2 t + \mu_g - (\beta_1 + \beta_2 t - \beta_2 + \mu_g) = \beta_2 \\ &\implies \boxed{\mathbb{E}(Y_t) = \beta_2}\end{aligned}$$

$$\begin{aligned}\gamma_Y(s, t) &= \text{Cov}(Y_s, Y_t) = \text{Cov}(X_s - X_{s-1}, X_t - X_{t-1}) \\ &= \text{Cov}(\beta_1 + \beta_2 s + G_s - (\beta_1 + \beta_2 s - \beta_2 + G_{s-1}), \beta_1 + \beta_2 t + G_t - (\beta_1 + \beta_2 t - \beta_2 + G_{t-1})) \\ &= \text{Cov}(\beta_2 + G_s - G_{s-1}, \beta_2 + G_t - G_{t-1}) = \text{Cov}(G_s - G_{s-1}, G_t - G_{t-1}) \\ &= \text{Cov}(G_s, G_t) - \text{Cov}(G_{s-1}, G_t) - \text{Cov}(G_s, G_{t-1}) + \text{Cov}(G_{s-1}, G_{t-1})\end{aligned}$$

1.  $s = t \implies \gamma_Y(s, t) = \gamma_g - 0 - 0 + \gamma_g = 2\gamma_g$
2.  $s = t + 1 \implies \gamma_Y(s, t) = 0 - \gamma_g - 0 + 0 = -\gamma_g$
3.  $s = t - 1 \implies \gamma_Y(s, t) = 0 - 0 - \gamma_g + 0 = -\gamma_g$
4.  $s = t + 2 \implies \gamma_Y(s, t) = 0 - 0 - 0 + 0 = 0$
5.  $s = t - 2 \implies \gamma_Y(s, t) = 0 - 0 - 0 + 0 = 0$

$$\implies \boxed{\gamma_Y(s, t) = \begin{cases} 2\gamma_g & ; s = t \\ -\gamma_g & ; |s - t| = 1 \\ 0 & ; |s - t| \geq 2 \end{cases}}$$

$$\rho_Y(s, t) = \frac{\gamma_Y(s, t)}{\sqrt{\gamma_Y(s, s)\gamma_Y(t, t)}} = \frac{\gamma_Y(s, t)}{\sqrt{2\gamma_g * 2\gamma_g}} = \frac{\gamma_Y(s, t)}{2\gamma_g}$$

$$\implies \boxed{\rho_Y(s, t) = \begin{cases} 1 & ; s = t \\ -\frac{1}{2} & ; |s - t| = 1 \\ 0 & ; |s - t| \geq 2 \end{cases}}$$

We can see that the autocovariance function for  $Y_t$  is composed of a constant-scaled version of the autocovariance function for  $X_t$ . Since we know that  $X_t$  is stationary,  $\gamma_X$  only depends on the lag and scaling it by a constant will still have it only depend on the lag. Therefore,  $Y_t$  is stationary as the mean is constant (does not depend on  $t$ ) and the autocovariance function only depends on the lag (time difference between  $s$  and  $t$ ).

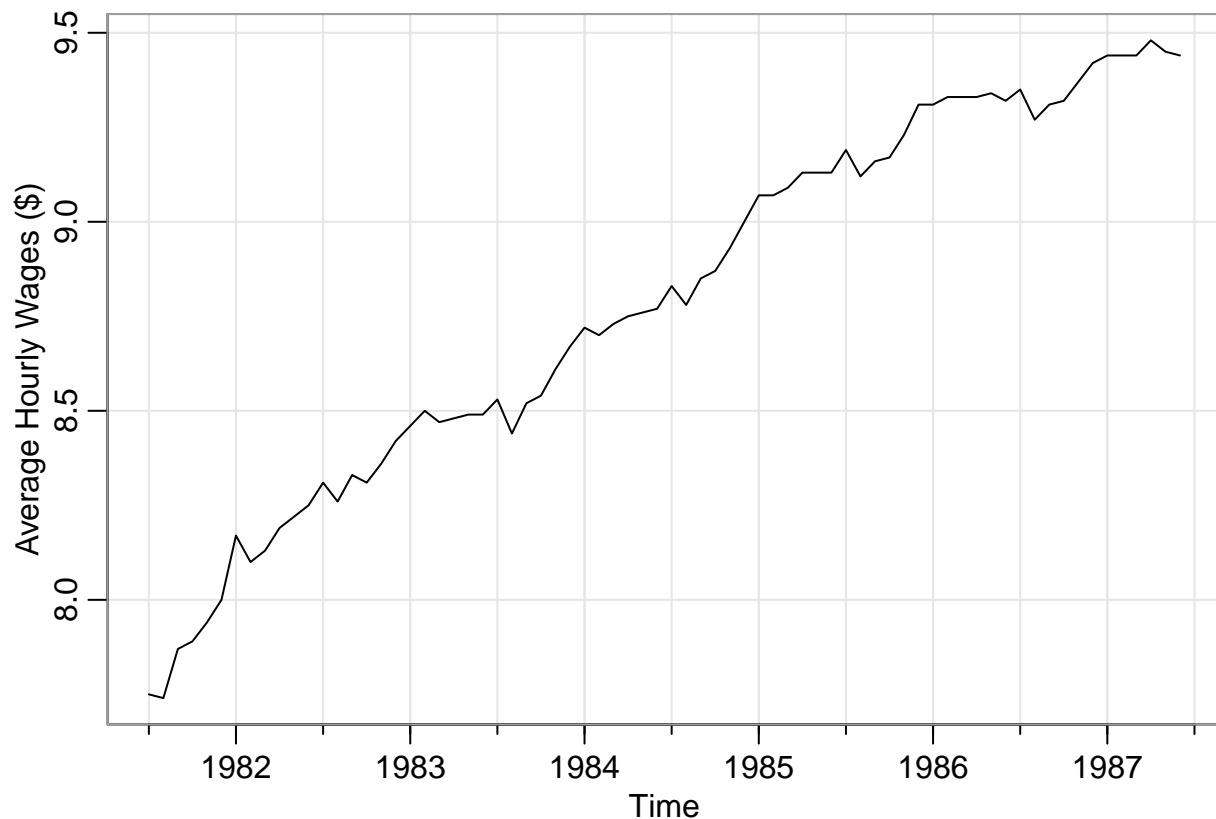
### Question 3

```
##  
## Attaching package: 'TSA'  
  
## The following objects are masked from 'package:stats':  
##  
##   acf, arima  
  
## The following object is masked from 'package:utils':  
##  
##   tar
```

The data file `wages` contains monthly values of the average hourly wages (in dollars) for workers in the U.S. apparel and textile products industry for July 1981 through June 1987.

(a) Display and interpret the time series plot for these data.

```
tsplot(wages, ylab = "Average Hourly Wages ($)")
```



From the plot we can see that the average hourly wage (in dollars) in the U.S. apparel and textile products industry has increased steadily over the six year time period. The trend is also appears to be approximately linear.

- (b) Use least squares to fit a linear time trend to this time series. Interpret the regression output. Save the standardized residuals from the fit for further analysis.

```
t = time(wages)
model2 = lm(wages ~ t)
summary(model2)

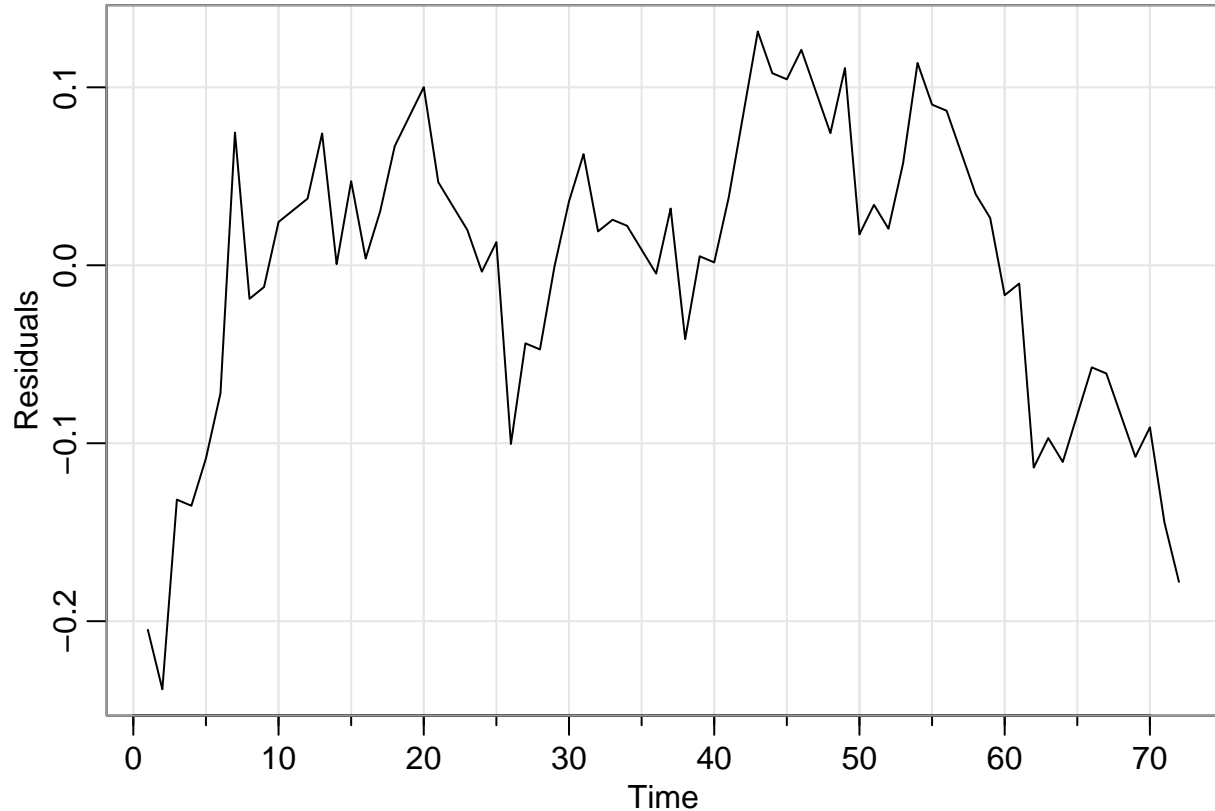
##
## Call:
## lm(formula = wages ~ t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23828 -0.04981  0.01942  0.05845  0.13136
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.490e+02  1.115e+01  -49.24  <2e-16 ***
## t            2.811e-01  5.618e-03   50.03  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08257 on 70 degrees of freedom
## Multiple R-squared:  0.9728, Adjusted R-squared:  0.9724
## F-statistic: 2503 on 1 and 70 DF, p-value: < 2.2e-16
```

$$\hat{Wage} = -549 + 0.28Year$$

From the R output we can say that the estimated average wage for workers in the U.S. apparel and textile industry will increase about 28 cents each year beyond June 1981.

(c) Construct and interpret the time series plot of the standardized residuals from part (b).

```
tsplot(model2$residuals, ylab = "Residuals")
```



From the residuals plot, we can see that there is a kind of cyclical increasing/decreasing trend, which clearly does not look like what we would like to see (White Noise). The results from this residual plot clearly shows that we do not have a good linear model for the data when we use time as the predictor.

(d) Use least squares to fit a quadratic time trend to the wages time series. Interpret the regression output. Save the standardized residuals from the fit for further analysis.

```
time = as.numeric(t)
model3 = lm(wages ~ poly(time, degree = 2, raw = TRUE))
summary(model3)
```

```
##
## Call:
## lm(formula = wages ~ poly(time, degree = 2, raw = TRUE))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.148318 -0.041440  0.001563  0.050089  0.139839
##
## Coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
##
```



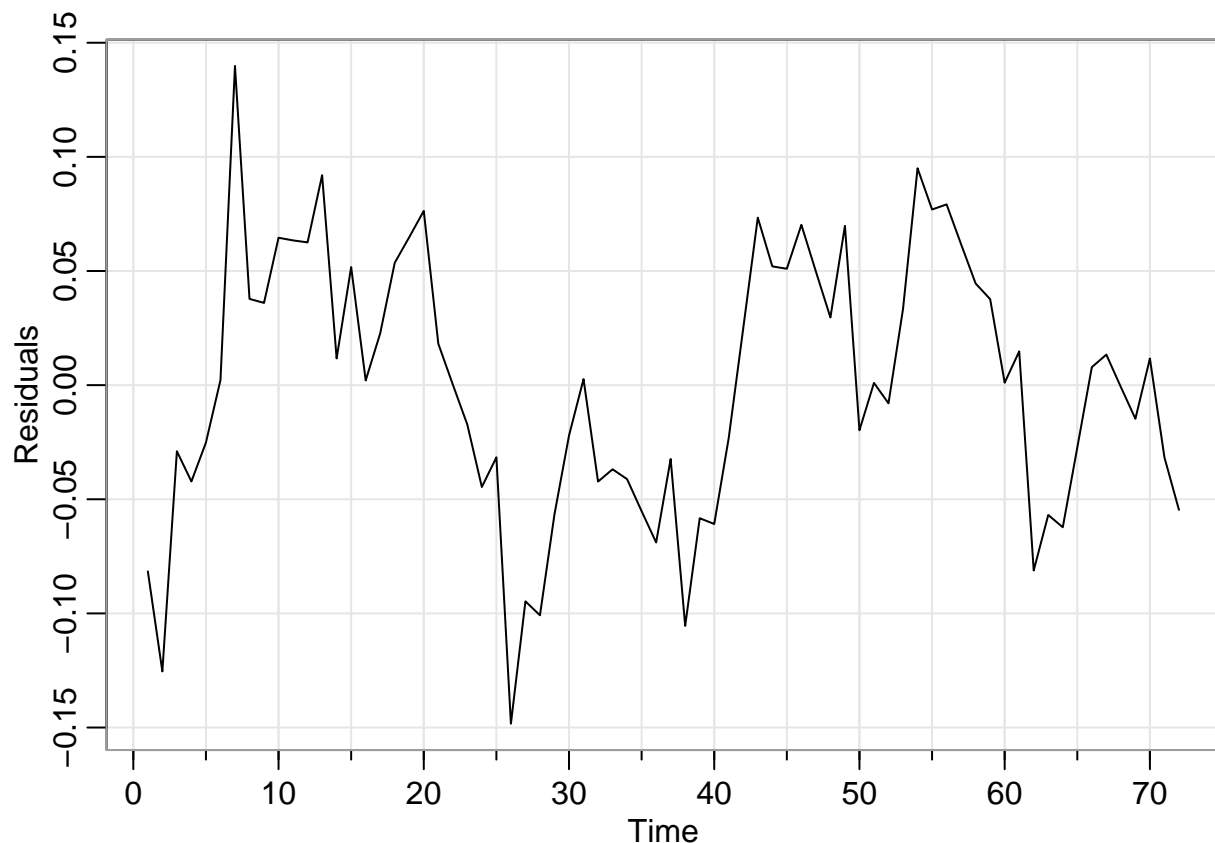
```
## (Intercept) -8.495e+04 1.019e+04 -8.336 4.87e-12 ***
## poly(time, degree = 2, raw = TRUE)1 8.534e+01 1.027e+01 8.309 5.44e-12 ***
## poly(time, degree = 2, raw = TRUE)2 -2.143e-02 2.588e-03 -8.282 6.10e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05889 on 69 degrees of freedom
## Multiple R-squared:  0.9864, Adjusted R-squared:  0.986
## F-statistic: 2494 on 2 and 69 DF, p-value: < 2.2e-16
```

$$\hat{Wage} = -84949.73 + 85.34Year - 0.02143Year^2$$

From the R output we can say that the estimated average wage for workers in the U.S. apparel and textile industry will increase about  $85.34 - 0.04286(Year)$  cents for each year beyond June 1981.

(e) Construct and interpret the time series plot of the standardized residuals from part (d).

```
tsplot(model3$residuals, ylab = "Residuals")
```



From the residuals plot, we can see that there is a kind of cyclical, which clearly does not look like what we would like to see (White Noise). However, it is much better looking when the quadratic term than just the linear term. It at least cycles around zero. The results from this residual plot shows that while this is probably not the best model possible, it is still better than the first one.