Statistics 435/535/711 7 February 2022
P. Shaman

t tests and partial *F* tests; interaction; Durbin-Watson statistic

1. What do the regression coefficients in a multiple regression represent? The coefficient attached to a particular x is a *partial slope* and indicates the change in y for each one unit change in that x, while at the same time holding all the other x variables in the multiple regression constant. All explanatory variables *not* included in the multiple regression are free to vary. By contrast, a *marginal slope* is obtained as the slope when we perform a simple regression, with only one x. All explanatory variables *not* included in the simple regression are free to vary. The partial slope and the marginal slope for a particular x can be very different.

This said, it should be noted that the independent variables are often themselves correlated, a condition called collinearity, and it isn't always possible to change a particular x and at the same time hold all the other x variables fixed (consider, for example, a regression with both x_1 and its square included in the model). We assess the amount of collinearity in a regression model from the correlations among the x variables, and from VIF values. For analysis of time series, however, the trigonometric dummy variables (for seasonal structure and calendar effects) and the monthly (or quarterly, or ...) dummy variables to address seasonality present very little to no collinearity.

So what is the interpretation of a partial slope? Suppose we are looking at the partial slope attached to x_1, and that $x_2, \ldots, x_k$ are other predictors in the model. Use regression to remove the influence of $x_2, \ldots, x_k$ from y, and a second regression to remove their influence from x_1. This yields residuals for y and residuals for x_1. Then the partial slope attached to x_1 in the multiple regression is the same as the slope in the simple regression of the y residuals on the x_1 residuals.

In multiple regression there is a *t* test (and a *t*-based confidence interval) for each regression coefficient. Each *t* test assesses the significance of a partial slope, which is the impact of a single predictor *beyond that of all the other predictors collectively*. That is, what does a particular x contribute to the model beyond the contributions of all the other *x*'s taken collectively? Thus, it is common to speak of a *t* test in a multiple regression as a "last-in" test. The same interpretation holds for a partial *F* test. It assesses the impact of the *r* variables being tested beyond that of all the other variables collectively.

Here is an example. A parcel delivery service would like to increase the number of packages that are sorted in each of its hub locations. Data recently sampled at 30 centers measure the number of sorts per hour. Three factors that the company can control and that influence sorting performance are the number of sorting lines, the number of sorting

workers, and the number of truck drivers. We consider additive models. The simple regression of the number of sorts per hour on the number of sorting lines yields the fit

$$\text{Sorts/Hour} = 2{,}960 + 1{,}446 \text{ Lines}.$$

The multiple regression model using all three explanatory variables gives

$$\text{Sorts/Hour} = 739 + 826 \text{ Lines} + 95 \text{ Sorters} + 118 \text{ Truckers}.$$

The marginal slope for the variable Lines is 1,446, and the partial slope for Lines is 826. That is, the addition of one line while at the same time not changing the number of sorters and truckers (thus the sorters and truckers must be redistributed among the previously existing lines) estimates 826 more sorts per hour. However, if we add one line (and do not at the same time restrict the number of sorters and truckers, thus allowing them to increase also), we can achieve on average an additional 1,446 sorts per hour.

2. Interaction is an important concept. It involves a response variable, y, and at least two independent variables. Suppose we are examining salaries of employees filling a particular job category (such as Accountant level I) in an organization. Let y be salary, x_1 timeinposition, and x_2 a dummy variable for gender (equal to 0 for males and 1 for females). Consider the relationship between salary and timeinposition for each sex. That is, perform a regression of y on x_1 for males only and a separate such regression for the females only. One can examine the two fitted lines, comparing their intercepts and their slopes. A difference in intercepts will signal a difference in entry level salaries for the two sexes, and a difference in slopes will indicate different rates of pay increase for the two sexes. If there is a difference in slopes, interaction is present. That is, the relationship between y (salary) and one of the independent variables, x_1 (timeinposition), differs according to the value of the other independent variable, x_2 (gender).

Instead of fitting the two lines separately, it is better to pool the male and female data to estimate a common error variance with more degrees of freedom than the separate regressions permit. This results in more powerful tests, and narrower confidence intervals and prediction intervals. Moreover, this approach allows one to test formally for the presence of interaction. We use the following model:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon \\ &= \beta_0 + \beta_1 \,\text{timeinposition} + \beta_2 \,\text{gender} + \beta_3 \,\text{timeinposition} \cdot \text{gender} + \varepsilon. \end{aligned}$$

For males this model is

$$y = \beta_0 + \beta_1 \,\text{timeinposition} + \varepsilon,$$

and for females it is

$$y = \beta_0 + \beta_2 + (\beta_1 + \beta_3)\,\text{timeinposition} + \varepsilon.$$

That is, for males the intercept is β_0 and the slope attached to timeinposition is β_1. For females these are $\beta_0 + \beta_2$ and $\beta_1 + \beta_3$, respectively. If the parameter β_3 is present, then we say that there is interaction between timeinposition and gender. This means that the impact of timeinposition upon salary depends on gender. Note that the interaction is built into the model by using the product of timeinposition and gender.

Notice that interaction here is a condition that involves two x's and a y. (In fact, more than two x's and a y can be involved.) Interaction should not be confused with collinearity, which involves only the x variables.

Let's use the garbage case to look at another example of interaction. Consider the model

$$\begin{aligned}\text{tonnage} = {} & \beta_0 + \beta_1\text{time} + \beta_2\text{time}^2 + \beta_3\text{time}^3 + \beta_4\text{Marlene} + \beta_5\text{George} \\ & + \beta_6\text{Marlene}\cdot\text{George} + \beta_7\text{Marlene}\cdot\text{time} + \beta_8\text{Marlene}\cdot\text{time}^2 \\ & + \beta_9\text{George}\cdot\text{time} + \beta_{10}\text{George}\cdot\text{time}^2 + \beta_{11}\text{Marlene}\cdot\text{George}\cdot\text{time} \\ & + \beta_{12}\text{Marlene}\cdot\text{George}\cdot\text{time}^2 + \text{seasonal} + \varepsilon.\end{aligned}$$

The response variable is weekly reported tonnage. Both Marlene and George are 1/0 dummy variables (e.g., Marlene is 1 for any week she is on duty, and 0 for any week she is not on duty). The model shown has two interaction components. One is the product Marlene·George. It is used to account for the weeks when both are on duty. One parameter is utilized. The second component places four different trends into the model and requires six parameters. There are trends for the weeks (i) neither Marlene nor George is on duty, (ii) only Marlene is on duty, (iii) only George is on duty, and (iv) both are on duty.

3. The Durbin-Watson statistic is used to test the null hypothesis of uncorrelated errors in a time series regression model, against the alternative that the errors follow a first-order autoregression with positive lag one correlation. The time series regression model is

(1) $$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + \varepsilon_t,$$

where the error ε_t follows a first-order autoregressive model,

$$\varepsilon_t = \rho\varepsilon_{t-1} + a_t, \quad -1 < \rho < 1,$$

and a_t is white noise. The Durbin-Watson statistic tests the null hypothesis H_0: $\rho = 0$ against the alternative that ρ is positive. It is calculated as follows: Estimate the

coefficients in (1) by ordinary least squares and calculate the residuals, e_t, $t = 1, \ldots, n$. The Durbin-Watson statistic is

$$DW = \frac{\sum_{t=2}^{n} \left(e_t - e_{t-1} \right)^2}{\sum_{t=1}^{n} e_t^2},$$

and this is approximately equal to $2(1 - r)$, where

$$r = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}$$

is the sample correlation between adjacent residuals and is an estimate of ρ. The null hypothesis is rejected if

$$DW < d_L,$$

and it is accepted if

$$DW > d_U,$$

where d_L and d_U depend on the significance level and the number of x variables. The cutoff values d_L and d_U were given by Durbin and Watson in statistical tables for selected values of n and the number of x variables, and for significance levels .05 and .01. Note that the test is inconclusive if the statistic DW falls between the two cutoff values. To test for negative autocorrelation use $4 - DW$ as the statistic and follow the above procedure.

It should be noted that the Durbin–Watson statistic addresses only lag one autocorrelation of the model errors. Given advances in software since publication of the papers by Durbin and Watson (in the early 1950s), the statistic is now mainly of historical interest.

Summary and additional remarks

1. A marginal slope is a regression coefficient in a model with only one explanatory variable. A partial slope is a regression coefficient in a model with multiple explanatory variables. A marginal slope and a partial slope for a given explanatory variable have different interpretations, and their numerical values and standard errors can differ greatly.

2. Interaction is a significant component in many statistical models. It involves a response variable and two explanatory variables. Interaction is present if the relationship between the response and one of the explanatory variables differs according to the value of the other explanatory variable. Interaction is included in models fit in the case studies which follow in these notes.

3. The Durbin–Watson ratio is calculated from the residuals of a fitted regression model. It is equal to the lag 1 residual autocorrelation. Thus, the ratio can sometimes be used to signal that a model fit to data has not provided reduction to white noise. Properties of the Durbin–Watson ratio were developed in the early 1950s. Today it is mainly of historical interest, as we will calculate residual autocorrelations at multiple lags.

Weekly Garbage Deposits

The data analyzed are weekly garbage deposits, in tons, at a Delaware Solid Waste Authority facility. There are three years of data, beginning with the week of 12/30/84 and ending with the week of 12/13/87 (each date listed is a Sunday). The data arose from a criminal trial. A refuse collector (Atlas) and two employees of the Solid Waste Authority (Marlene and George, both weighmasters) were charged with falsifying weights. When an Atlas truck appeared at the facility, Marlene and George either waved it through or recorded a lesser amount of weight. The goal of the model construction presented here is estimation of the lost tonnage, and thus of the lost revenue, attributable to the criminal activity.

The tipping fee amounts per ton were as follows:

Fiscal year	Amount
1985	$28.26
1986	$29.50
1987	$32.84
1988	$35.26

Each fiscal year begins July 1st. The first model below incorporates a third-degree polynomial trend, a seasonal component with a fundamental and three overtones (the periods of these four trigonometric components are one year, ½ year, 1/3 year, and ¼ year), and dummy variables for Marlene and George. Let y_t denote the tonnage recorded in week *t*. The model is

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 M_t + \beta_5 G_t + \beta_6 MG_t + \sum_{j=1}^{4}\left(\gamma_j \cos\frac{2\pi 7jt}{365} + \delta_j \sin\frac{2\pi 7jt}{365}\right) + \varepsilon_t .$$

The dummy variables use the 0, 1 convention. Specifically,

M_t = 1 if Marlene is on duty during week *t*
= 0 otherwise,

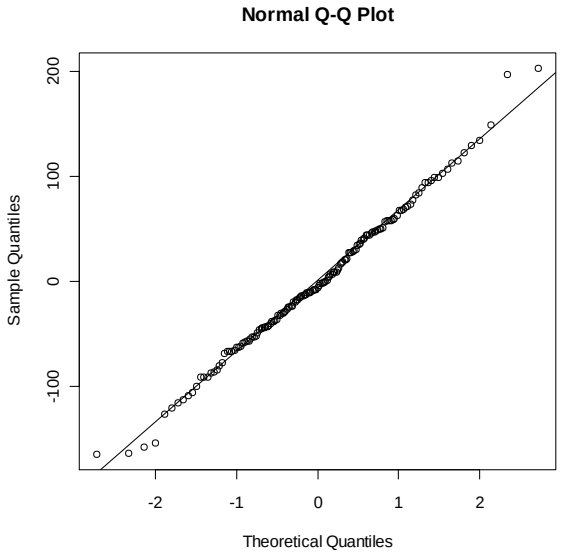
G_t = 1 if George is on duty during week *t*
= 0 otherwise,

MG_t = 1 if both Marlene and George are on duty during week *t*
= 0 otherwise.

Note that $MG_t = M_t G_t$ is an interaction variable. Call this formulation model 1.

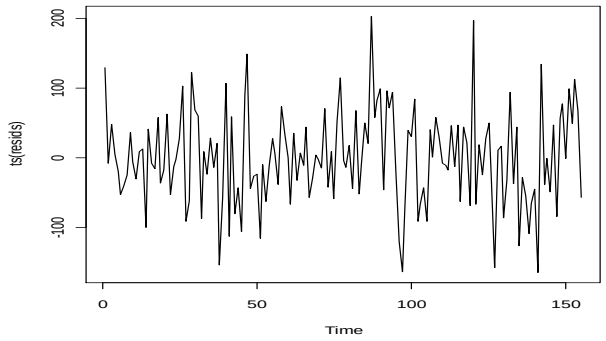
```
> garbage<-read.csv("G:/Stat71122Spring/Garbage.txt")
> attach(garbage)
> head(garbage)
    date fiscalyr   fee tonnage marlene george mg
1 123084       85 28.26  321.14       0      1  0
2  10685       85 28.26  319.75       0      0  0
3  11385       85 28.26  199.35       1      0  0
4  12085       85 28.26  151.59       1      0  0
5  12785       85 28.26  137.98       0      1  0
6  20385       85 28.26  106.76       0      1  0

> freq<-14*pi/365
> time<-as.numeric(1:length(tonnage));time2<-time*time;time3<-
time*time2
> cosm<-matrix(nrow=length(tonnage),ncol=8)
> sinm<-matrix(nrow=length(tonnage),ncol=8)
> for(j in 1:8){
+ cosm[,j]<-cos(freq*j*time)
+ sinm[,j]<-sin(freq*j*time)
+ }
> model1<-
lm(tonnage~time+time2+time3+marlene+george+mg+cosm[,1]+sinm[,1]+cosm[,2
]+sinm[,2]+cosm[,3]+sinm[,3]+cosm[,4]+sinm[,4]);summary(model1)

Call:
lm(formula = tonnage ~ time + time2 + time3 + marlene + george +
    mg + cosm[, 1] + sinm[, 1] + cosm[, 2] + sinm[, 2] + cosm[,
    3] + sinm[, 3] + cosm[, 4] + sinm[, 4])

Residuals:
    Min      1Q  Median      3Q     Max
-163.84  -43.74   -5.90   46.93  202.97

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.936e+02  2.925e+01  13.456  < 2e-16 ***
time         6.567e-01  1.463e+00   0.449 0.654308
time2       -2.851e-02  2.176e-02  -1.311 0.192157
time3        1.753e-04  9.197e-05   1.906 0.058641 .
marlene     -1.649e+02  1.595e+01 -10.337  < 2e-16 ***
george      -1.525e+02  1.584e+01  -9.629  < 2e-16 ***
mg           1.455e+02  4.147e+01   3.509 0.000606 ***
cosm[, 1]   -7.276e+00  8.179e+00  -0.890 0.375205
sinm[, 1]   -2.373e+01  9.201e+00  -2.579 0.010943 *
cosm[, 2]   -8.517e-01  8.164e+00  -0.104 0.917058
sinm[, 2]   -5.748e+01  8.422e+00  -6.824 2.46e-10 ***
cosm[, 3]    9.862e+00  8.188e+00   1.204 0.230464
sinm[, 3]   -3.512e+01  8.245e+00  -4.260 3.73e-05 ***
cosm[, 4]   -2.748e+01  8.205e+00  -3.349 0.001041 **
sinm[, 4]    4.516e+00  8.163e+00   0.553 0.581010
---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 71.33 on 140 degrees of freedom
Multiple R-squared: 0.639,      Adjusted R-squared: 0.6029
F-statistic:  17.7 on 14 and 140 DF,  p-value: < 2.2e-16
```

```
> resids<-resid(model1)
> qqnorm(resids)
> qqline(resids)
```



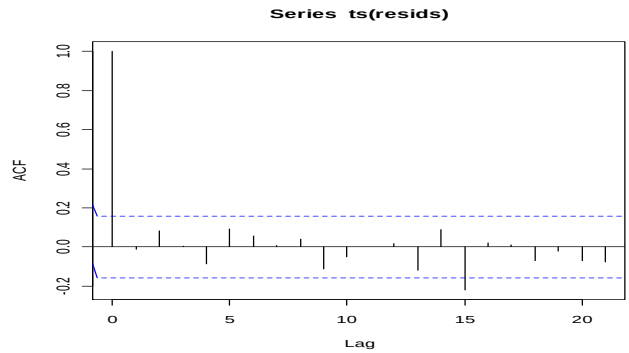
The normal quantile plot of the residuals shows decent agreement with normality. Plots of the residuals vs. time and of their autocorrelations follow.

```
> plot(ts(resids))
```



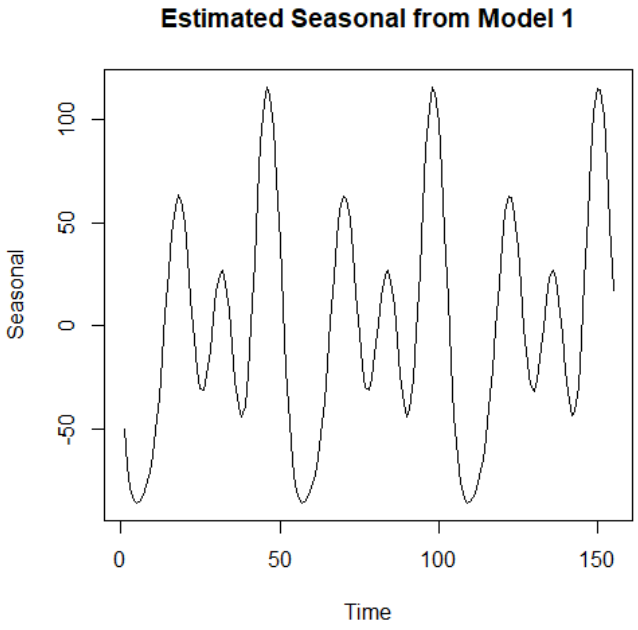
Some aspects of the trend structure have not been captured by the model, and there is evidence of very mild heteroscedasticity.

```
> acf(ts(resids))
```

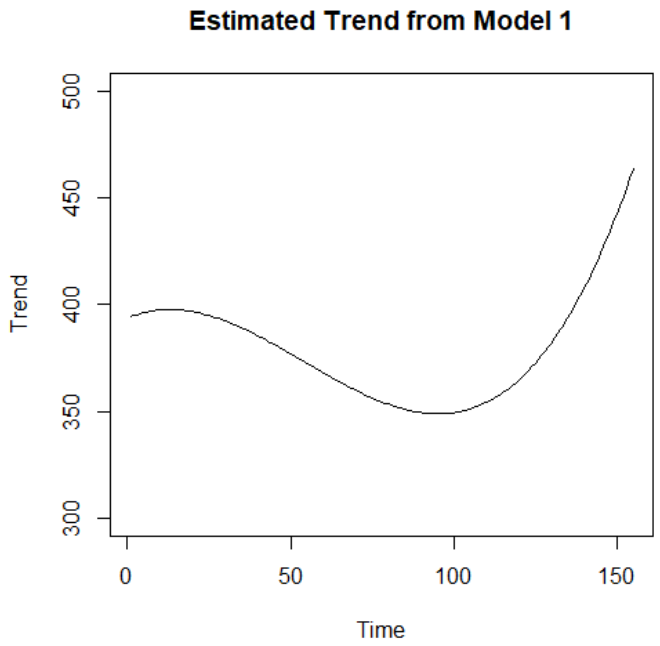


There is a modestly significant lag 15 residual autocorrelation. Otherwise the plot reveals no remaining autocorrelation. The next plots show the estimated seasonal and trend structures.

```
> xmatrix<-
matrix(c(cosm[,1],sinm[,1],cosm[,2],sinm[,2],cosm[,3],sinm[,3],cosm[,4]
,sinm[,4]),ncol=8)
> betaest<-coef(model1)[8:15]
>
plot(ts(xmatrix%*%betaest),xlab="Time",ylab="Seasonal",main="Estimated
Seasonal from Model 1")
```



```
>
plot(ts(xmatrix2%*%betaest2),ylim=c(300,500),xlab="Time",ylab="Trend",
main="Estimated Trend from Model 1")
```



The next fit, model 2, adds four more overtone pairs to the seasonal component.

```
> model2<-
lm(tonnage~time+time2+time3+marlene+george+mg+cosm[,1]+sinm[,1]+cosm[,
2]+sinm[,2]+cosm[,3]+sinm[,3]+cosm[,4]+sinm[,4]+cosm[,5]+sinm[,5]+cosm
[,6]+sinm[,6]+cosm[,7]+sinm[,7]+cosm[,8]+sinm[,8]);summary(model2)

Call:
lm(formula = tonnage ~ time + time2 + time3 + marlene + george +
    mg + cosm[, 1] + sinm[, 1] + cosm[, 2] + sinm[, 2] + cosm[,
    3] + sinm[, 3] + cosm[, 4] + sinm[, 4] + cosm[, 5] + sinm[,
    5] + cosm[, 6] + sinm[, 6] + cosm[, 7] + sinm[, 7] + cosm[,
    8] + sinm[, 8])

Residuals:
     Min       1Q   Median       3Q      Max
-159.800  -44.346   -5.717   46.003  193.304
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.904e+02  2.892e+01  13.499  < 2e-16 ***
time         8.096e-01  1.451e+00   0.558 0.577700
time2       -3.007e-02  2.157e-02  -1.394 0.165651
time3        1.798e-04  9.119e-05   1.972 0.050723 .
marlene     -1.669e+02  1.578e+01 -10.576  < 2e-16 ***
george      -1.533e+02  1.560e+01  -9.831  < 2e-16 ***
mg           1.554e+02  4.132e+01   3.761 0.000254 ***
cosm[, 1]   -8.123e+00  8.027e+00  -1.012 0.313382
sinm[, 1]   -2.306e+01  9.057e+00  -2.546 0.012059 *
cosm[, 2]   -1.841e+00  8.014e+00  -0.230 0.818646
sinm[, 2]   -5.747e+01  8.270e+00  -6.949 1.52e-10 ***
cosm[, 3]    8.664e+00  8.043e+00   1.077 0.283352
sinm[, 3]   -3.498e+01  8.091e+00  -4.323 3.01e-05 ***
cosm[, 4]   -2.870e+01  8.061e+00  -3.560 0.000516 ***
sinm[, 4]    4.633e+00  8.009e+00   0.579 0.563898
cosm[, 5]   -1.277e+01  8.023e+00  -1.591 0.113912
sinm[, 5]   -5.617e+00  8.005e+00  -0.702 0.484070
cosm[, 6]   -1.055e+01  8.003e+00  -1.318 0.189744
sinm[, 6]    6.238e+00  7.953e+00   0.784 0.434210
cosm[, 7]   -1.559e+01  8.022e+00  -1.943 0.054104 .
sinm[, 7]    7.110e+00  7.952e+00   0.894 0.372864
cosm[, 8]   -9.858e+00  8.036e+00  -1.227 0.222149
sinm[, 8]    1.298e+01  7.948e+00   1.634 0.104717
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69.96 on 132 degrees of freedom
Multiple R-squared:  0.6726,    Adjusted R-squared:  0.618
F-statistic: 12.32 on 22 and 132 DF,  p-value: < 2.2e-16
```

The partial *F* test for the four additional overtones indicates at most marginal significance (the *p* value is 0.1063). (Rather than test each cosine-sine pair individually, we have tested the four additional pairs collectively.)

```
> anova(model1,model2)
Analysis of Variance Table

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    140 712345
2    132 646145  8     66200 1.6905 0.1063
```

The *R*-square values for model 1 and model 2 are as follows:

```
> summary(model1)$r.squared
[1] 0.6390179
> summary(model2)$r.squared
[1] 0.672565
```

The partial F statistic can then also be calculated as

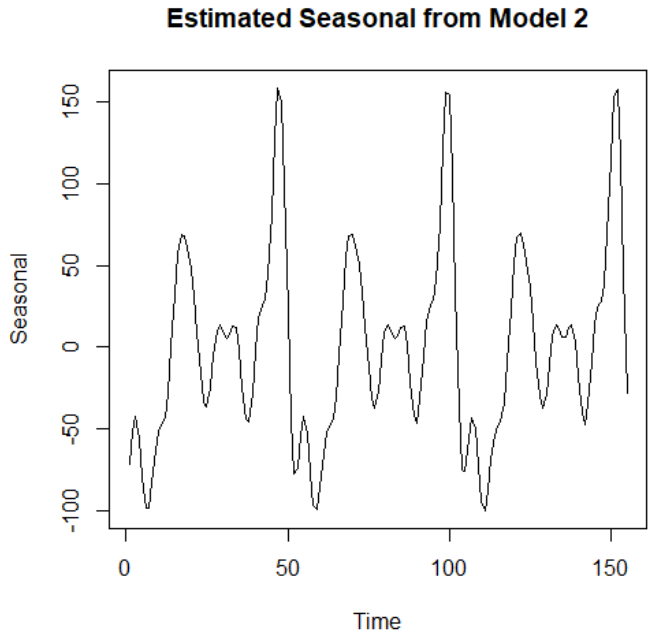
$$\frac{(0.672565 - 0.639018)/8}{(1-0.672565)/(155-22-1)} = 1.6905.$$

The p value calculation follows. It obtains the tail area beyond the value 1.6905 for an F distribution with 8 numerator degrees of freedom and 132 denominator degrees of freedom.

```
> 1-pf(1.6905,8,132)
[1] 0.1063435
```

The next plot shows the seasonal structure estimated from model 2. The additional four trigonometric pairs provide extra nuance to the interpretation of the seasonal structure, and the picture does suggest that retention of these pairs is useful.

```
> xmatrix3<-
matrix(c(cosm[,1],sinm[,1],cosm[,2],sinm[,2],cosm[,3],sinm[,3],cosm[,4]
,sinm[,4],cosm[,5],sinm[,5],cosm[,6],sinm[,6],cosm[,7],sinm[,7],cosm[,8
],sinm[,8]),ncol=16)
> betaest<-coef(model2)[8:23]
>
plot(ts(xmatrix3%*%betaest),xlab="Time",ylab="Seasonal",main="Estimated
Seasonal from Model 2")
```



The next model, model 3, tests for time trends in the behavior of Marlene and George. It adds six interaction terms: Marlene·time, Marlene·time^2, George·time, George·time^2, Marlene·George·time, and Marlene·George·time^2. The fit provides four different trends, one when Marlene alone is duty, one when George alone is on duty, one when both are on duty, and one when neither is on duty. The model does not contain terms giving the interaction between the trend and seasonal components (these terms are not significant).

```
> model3<-
lm(tonnage~time+time2+time3+marlene+george+mg+cosm[,1]+sinm[,1]+cosm[,2
]+sinm[,2]+cosm[,3]+sinm[,3]+cosm[,4]+sinm[,4]+marlene*time+marlene*tim
e2+george*time+george*time2+mg*time+mg*time2);summary(model3)

Call:
lm(formula = tonnage ~ time + time2 + time3 + marlene + george + mg +
cosm[, 1] + sinm[, 1] + cosm[, 2] + sinm[, 2] + cosm[,3] + sinm[, 3] +
cosm[, 4] + sinm[, 4] + marlene * time + marlene * time2 + george *
time + george * time2 + mg * time + mg * time2)

Residuals:   Min       1Q   Median       3Q      Max
          -165.393  -37.033    1.136   43.853  210.728

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.251e+02  4.589e+01   7.084 7.18e-11 ***
time           2.583e+00  1.805e+00   1.431  0.15481
time2         -3.712e-02  2.265e-02  -1.639  0.10355
time3          1.611e-04  9.126e-05   1.766  0.07970 .
marlene       -9.766e+01  4.949e+01  -1.973  0.05052 .
george        -5.181e+01  4.954e+01  -1.046  0.29746
mg             1.044e+03  5.181e+02   2.016  0.04581 *
cosm[, 1]     -4.710e+00  8.160e+00  -0.577  0.56481
sinm[, 1]     -2.563e+01  9.080e+00  -2.823  0.00548 **
cosm[, 2]     -3.089e+00  8.162e+00  -0.378  0.70572
sinm[, 2]     -5.918e+01  8.287e+00  -7.141 5.31e-11 ***
cosm[, 3]      1.023e+01  8.076e+00   1.267  0.20728
sinm[, 3]     -3.180e+01  8.185e+00  -3.885  0.00016 ***
cosm[, 4]     -2.648e+01  8.100e+00  -3.269  0.00137 **
sinm[, 4]      4.637e+00  8.063e+00   0.575  0.56617
time:marlene  -2.437e+00  1.408e+00  -1.731  0.08575 .
time2:marlene  1.540e-02  8.499e-03   1.812  0.07229 .
time:george   -2.622e+00  1.420e+00  -1.847  0.06695 .
time2:george   1.299e-02  8.625e-03   1.506  0.13437
time:mg       -3.524e+01  2.027e+01  -1.738  0.08443 .
time2:mg       2.543e-01  1.477e-01   1.721  0.08749 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69.85 on 134 degrees of freedom
Multiple R-squared:  0.6686,    Adjusted R-squared:  0.6192
F-statistic: 13.52 on 20 and 134 DF,  p-value: < 2.2e-16
```

```
> anova(model1,model3)
Analysis of Variance Table

  Res.Df    RSS Df Sum of Sq     F  Pr(>F)
1    140 712345
2    134 653877  6     58468 1.997 0.07036 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The additional variables (six product terms) are marginally significant ($p = 0.07$), and we choose to retain them. The four trends given by this model are as follows:

Neither Marlene nor George is on duty:

$$325.082 + 2.5825t - 0.0371196t^2 + 0.00016114t^3$$

Marlene alone is on duty:

$$227.422 + 0.14524t - 0.021722t^2 + 0.00016114t^3$$

George alone is on duty:

$$273.269 - 0.03985t - 0.024129t^2 + 0.00016114t^3$$

Both Marlene and George are on duty:

$$1219.964 - 37.714t + 0.246t^2 + 0.00016114t^3$$

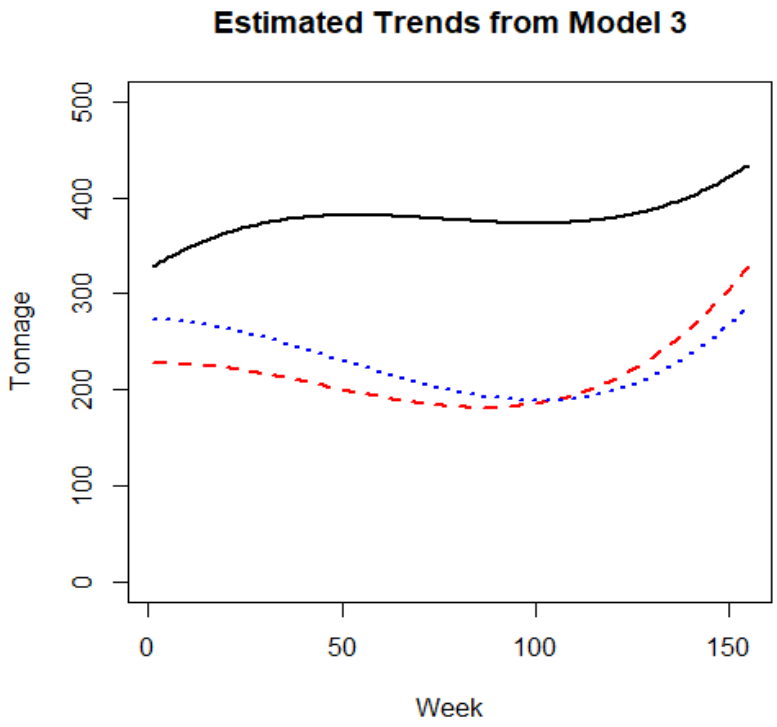
There are only four weeks during which both Marlene and George are on duty:

Week	t	Tonnage
7/14/85	29	346.84
8/18/85	34	201.89
9/22/85	39	118.11
12/21/86	104	118.64

Therefore the last trend (when both Marlene and George are on duty) should be interpreted only in the range $29 \le t \le 39$ and at $t = 104$.

Let's plot three of the trends, those when Marlene alone is on duty, George alone is on duty, and neither is on duty.

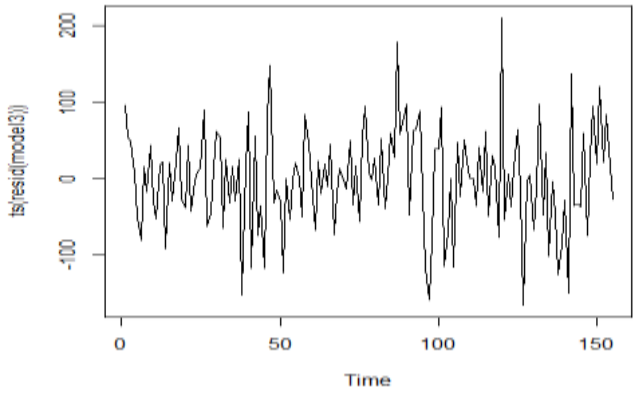
```
> nmatrix<-matrix(c(rep(1,length(tonnage)),time,time2,time3),ncol=4)
> nbeta<-c(325.082,2.5825,-0.0371196,0.00016114)
> mbeta<-c(227.422,0.14524,-0.021722,0.00016114)
> gbeta<-c(273.269,-0.03985,-0.024129,0.00016114)
>
plot(ts(nmatrix%*%nbeta),type="l",lty=1,lwd=2,col="black",ylim=c(0,500
),xlab="Time",ylab="Trend",main="Estimated Trends from Model 3")
> lines(ts(nmatrix%*%mbeta),type="l",lty=2,lwd=2,col="red")
> lines(ts(nmatrix%*%gbeta),type="l",lty=3,lwd=2,col="blue")
```



The red dashed curve is the estimated trend when Marlene alone is on duty, the blue dotted curve is the estimated trend when George alone is on duty, and the solid black curve is the estimated trend when neither is on duty.

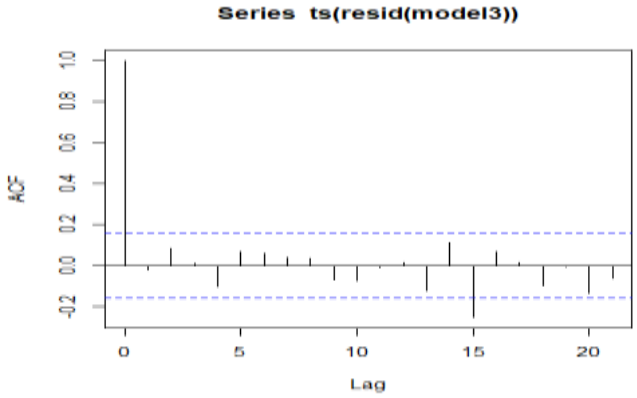
The following residual plot from model 3 displays some uncaptured seasonal and trend structure.

```
> plot(ts(resid(model3)))
```



Next is the plot of the model 3 residual correlations—it's similar to the plot on page 9 for model 1.

```
> acf(ts(resid(model3)))
```



Let's fit a model extending both the second and third models, by including eight trigonometric pairs and the interactions between the weighmasters and the time trend (giving four different trends).

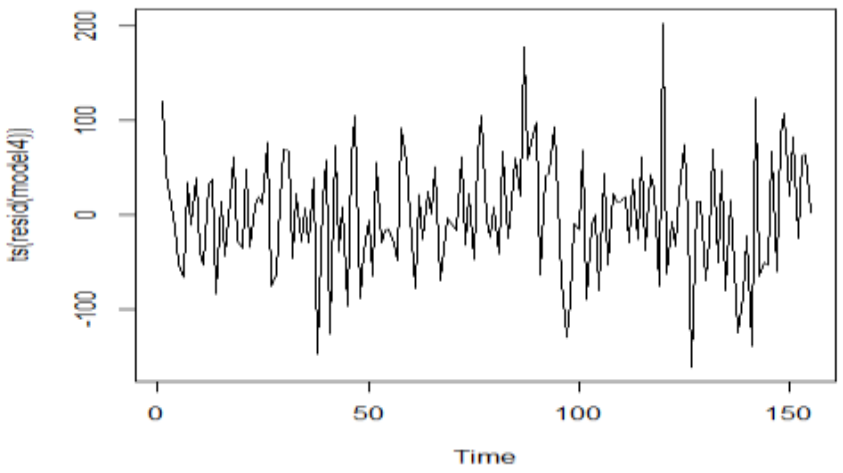
```
> model4<-
lm(tonnage~time+time2+time3+marlene+george+mg+cosm[,1]+sinm[,1]+cosm[,
2]+sinm[,2]+cosm[,3]+sinm[,3]+cosm[,4]+sinm[,4]+cosm[,5]+sinm[,5]+cosm
[,6]+sinm[,6]+cosm[,7]+sinm[,7]+cosm[,8]+sinm[,8]+marlene*time+marlene
*time2+george*time+george*time2+mg*time+mg*time2);summary(model4)

Call:
lm(formula = tonnage ~ time + time2 + time3 + marlene + george +
    mg + cosm[, 1] + sinm[, 1] + cosm[, 2] + sinm[, 2] + cosm[,
    3] + sinm[, 3] + cosm[, 4] + sinm[, 4] + cosm[, 5] + sinm[,
    5] + cosm[, 6] + sinm[, 6] + cosm[, 7] + sinm[, 7] + cosm[,
    8] + sinm[, 8] + marlene * time + marlene * time2 + george *
    time + george * time2 + mg * time + mg * time2)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-161.715  -40.721    0.693   43.050  202.251

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.214e+02  4.543e+01   7.075 9.18e-11 ***
time           2.478e+00  1.787e+00   1.386 0.168165
time2         -3.744e-02  2.242e-02  -1.670 0.097428 .
time3          1.751e-04  9.036e-05   1.938 0.054883 .
marlene       -1.018e+02  4.932e+01  -2.063 0.041145 *
george        -5.130e+01  4.885e+01  -1.050 0.295639
mg             1.122e+03  5.218e+02   2.149 0.033514 *
cosm[, 1]     -6.896e+00  8.031e+00  -0.859 0.392149
sinm[, 1]     -2.557e+01  8.932e+00  -2.863 0.004912 **
cosm[, 2]     -5.319e+00  8.029e+00  -0.663 0.508861
sinm[, 2]     -5.851e+01  8.132e+00  -7.195 4.93e-11 ***
cosm[, 3]      8.395e+00  7.943e+00   1.057 0.292563
sinm[, 3]     -3.154e+01  8.036e+00  -3.925 0.000142 ***
cosm[, 4]     -2.853e+01  7.973e+00  -3.579 0.000491 ***
sinm[, 4]      4.784e+00  7.908e+00   0.605 0.546281
cosm[, 5]     -1.395e+01  7.957e+00  -1.753 0.081997 .
sinm[, 5]     -5.315e+00  7.875e+00  -0.675 0.500986
cosm[, 6]     -1.035e+01  7.995e+00  -1.294 0.198057
sinm[, 6]      5.763e+00  7.820e+00   0.737 0.462533
cosm[, 7]     -1.616e+01  8.054e+00  -2.006 0.046967 *
sinm[, 7]      1.078e+01  7.889e+00   1.366 0.174256
cosm[, 8]     -6.667e+00  8.100e+00  -0.823 0.412025
sinm[, 8]      1.351e+01  7.892e+00   1.712 0.089353 .
time:marlene  -1.870e+00  1.409e+00  -1.327 0.186946
time2:marlene  1.044e-02  8.569e-03   1.219 0.225306
time:george   -2.216e+00  1.403e+00  -1.579 0.116765
time2:george   9.076e-03  8.564e-03   1.060 0.291293
time:mg       -3.957e+01  2.043e+01  -1.936 0.055057 .
time2:mg       2.945e-01  1.489e-01   1.977 0.050219 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 68.47 on 126 degrees of freedom
Multiple R-squared:  0.7007,    Adjusted R-squared:  0.6342
F-statistic: 10.53 on 28 and 126 DF,  p-value: < 2.2e-16
```

The variables included in this model are significant. For example, consider the following partial *F* test comparing model 4 and model 1.

```
> anova(model1,model4)
Analysis of Variance Table

  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    140 712345
2    126 590640 14    121705 1.8545 0.03762 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The residual time series plot follows.



This residual plot from model 4 resembles the residual plots from the previous models, and the same is true for the residual autocorrelation plot (not shown).

The main purpose of the analysis of this garbage data is to estimate the amount of revenue lost from the criminal activity. We provide estimates of the total lost tonnage and of the lost revenue from each of the fitted models. The calculations are illustrated for model 1. Similar calculations for models 3 and 4 are more complicated because of the presence of interactions between the weighmasters and the time trend.

Estimation of the lost tonnage and lost revenue are shown. In addition, a 95 per cent confidence interval estimate of the lost revenue is presented.

```
> xmatrixtons<-matrix(c(marlene,george,mg),ncol=3)
> betaesttons<-coef(model1)[5:7]
> -sum(xmatrixtons%*%betaesttons)
[1] 19568.49

> xmatrixrev<-matrix(c(marlene*fee,george*fee,mg*fee),ncol=3)
> -sum(xmatrixrev%*%betaesttons)
[1] 611479.7
```

Calculations to give a 95 per cent confidence interval estimate for the lost revenue are more technical statistically.

```
> r<-cbind(sum(marlene*fee),sum(george*fee),sum(mg*fee))
> sdest<-sqrt(r%*%vcov(model1)[5:7,5:7]%*%t(r))
> sdest
         [,1]
[1,] 55363.1
```

The confidence interval calculation follows.

```
> lcl<--sum(xmatrixrev%*%betaesttons)-1.96*sdest
> ucl<--sum(xmatrixrev%*%betaesttons)+1.96*sdest

> c(lcl,ucl)
[1] 502968.0 719991.4
```

Thus, model 1 estimates the total lost tonnage to be 19,568, the lost revenue to be $611,480, and the 95 per cent confidence interval estimate of the lost revenue to be ($502,968, $719,991). The table below summarizes the results for all four models.

	Estimated lost tonnage	Estimated lost revenue	95 per cent CI estimate of lost revenue
Model 1	19,568	$611,480	($502,968, $719,991)
Model 2	19,706	$615,794	($508,785, $722,804)
Model 3	19,364	$608,772	($502,279, $715,265)
Model 4	19,255	$607,888	($502,911, $712,866)

Although the four models do exhibit significant differences, the estimates of lost tonnage and lost revenue do not differ much among them.

Summary and additional remarks

1. The main purpose of the model fit to the weekly garbage deposit data series is to estimate the amount of revenue lost from the criminal activity perpetrated by Atlas and the two weighmasters.

2. The unusual work schedule of the weighmasters, consecutive weeks of working followed by weeks of not working, and only a few weeks during which both worked, allow the estimation of lost revenue to be constructed. In particular, only one of the crooked weighmasters worked during 119 of the 155 weeks (60 weeks for one and 59 for the other), both worked during 4 weeks, and neither worked during 32 weeks. The weeks during which neither worked are distributed throughout time in a manner that permits estimation of trend and seasonal patterns which would have been obtained in the absence of the criminal activity. This baseline level then facilitates estimation of the weekly revenue shortfall stemming from the criminal activity.

3. Four models are fit to the data. All are additive decompositions. All contain a 1, 0 dummy variable for each weighmaster to identify the weeks she/he worked.

4. Seasonal estimation in the models requires care because the data are weekly observations. Trigonometric pairs are used. The fundamental frequency employed is $2\pi 7/365$ ($2\pi 7/365.25$ could also be used), and in two of the models three overtone pairs are added, and in two others seven overtone pairs are appended.

4. Trend structure is included in the models with a polynomial in time. In addition, to determine whether the lost revenue patterns over time differed among the two weighmasters, the interaction between the trend and the weighmaster dummies is included in two of the models. The interaction is marginally significant in each model.

5. Estimates of the lost revenue and a 95 per cent confidence interval are obtained for each of the four models. The results obtained are consistent among the four. The estimated lost revenue ranges between $608,000 and $616,000 among the models.

U.S. Variety Store Retail Sales

The file retail.txt gives monthly retail sales, in millions of dollars, for variety stores in the U.S. for the period January 1967 through September 1979. The data have been adjusted for trading day and holiday effects.

In April 1976 W. T. Grant, a large chain, went out of business. The file contains a dummy variable which is 0 prior to April 1976 and is 1 beginning in April 1976. There are also dummy variables to permit estimation of seasonal indices and to give a frequency decomposition of the seasonal component (i.e., cosine and sine dummies).

```
> retail<-read.csv("G:/Stat71122Spring/retail.txt")
> attach(retail)
> head(retail)
  sales time month logsales obs96           c1           s1   c2           s2
1   296    1     1 5.690359     0  8.66025e-01 5.00000e-01  0.5  8.66025e-01
2   303    2     2 5.713733     0  5.00000e-01 8.66025e-01 -0.5  8.66025e-01
3   365    3     3 5.899897     0  6.12000e-17 1.00000e+00 -1.0  1.22000e-16
4   363    4     4 5.894403     0 -5.00000e-01 8.66025e-01 -0.5 -8.66030e-01
5   417    5     5 6.033086     0 -8.66030e-01 5.00000e-01  0.5 -8.66030e-01
6   421    6     6 6.042633     0 -1.00000e+00 1.22000e-16  1.0 -2.40000e-16
         c3           s3   c4           s4           c5           s5 c6 wtgrant
1  6.12e-17  1.00e+00 -0.5  8.66025e-01 -8.66030e-01  5.0000e-01 -1       0
2 -1.00e+00  1.22e-16 -0.5 -8.66030e-01  5.00000e-01 -8.6603e-01  1       0
3 -1.80e-16 -1.00e+00  1.0 -2.40000e-16  3.06000e-16  1.0000e+00 -1       0
4  1.00e+00 -2.40e-16 -0.5  8.66025e-01 -5.00000e-01 -8.6603e-01  1       0
5  3.06e-16  1.00e+00 -0.5 -8.66030e-01  8.66025e-01  5.0000e-01 -1       0
6 -1.00e+00  3.67e-16  1.0 -4.90000e-16 -1.00000e+00  6.1200e-16  1       0
```

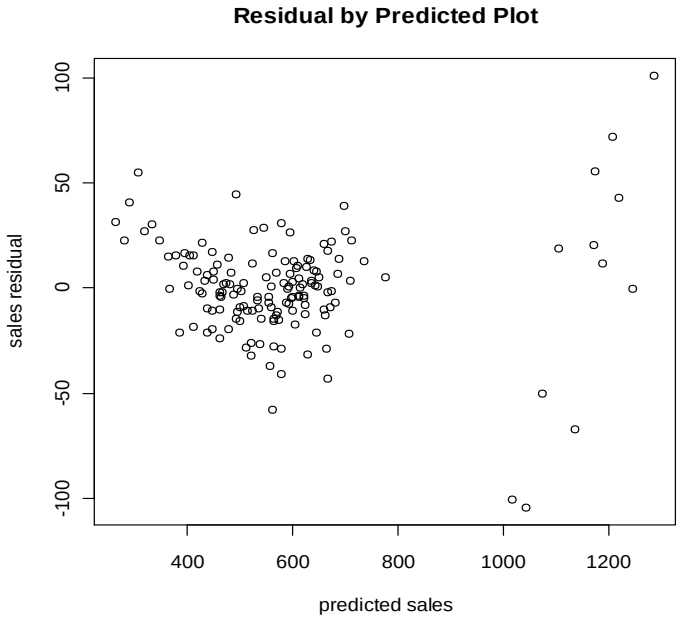
A plot of sales vs. time shows a trend component and a pronounced seasonal pattern. The impact of W. T. Grant's demise is clearly visible.

```
> sales.ts<-ts(sales,start=c(1967,1),freq=12)
> plot(sales.ts,xlab="time",ylab="sales",main="variety store monthly
sales, in millions")
```

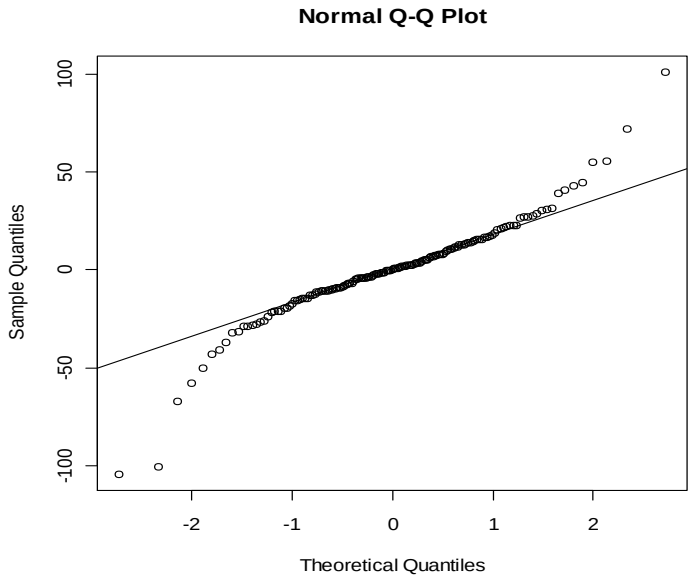
The plot also suggests it is advisable to log the sales data before proceeding with a model fit involving trend and seasonal components. However, the change in volatility does not appear to be very severe, and we will begin by trying to fit a model without logging the data. The following model has a second-degree polynomial for the trend, a seasonal component, the dummy variable for W. T. Grant, and the interaction between the W. T. Grant dummy and the trend. The dummy for W. T. Grant accounts for the sudden drop in level, and the interaction allows the shape of the trend (i.e., its slope) to change after the drop in level.

```
> fmonth<-as.factor(month)
> time<-as.numeric(time);time2<-time*time

> model<-
lm(sales~time+time2+fmonth+wtgrant+wtgrant*time+wtgrant*time2);summary(
model)

Call:
lm(formula = sales ~ time + time2 + fmonth + wtgrant + wtgrant *
    time + wtgrant * time2)

Residuals:
     Min       1Q   Median       3Q      Max
-104.037  -10.734    0.402   12.583  100.846
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   262.527114  10.120004  25.941  < 2e-16 ***
time            2.056086   0.315487   6.517 1.29e-09 ***
time2           0.006260   0.002729   2.293   0.0234 *
fmonth2        13.411342  10.310653   1.301   0.1956
fmonth3        96.632212  10.312129   9.371  < 2e-16 ***
fmonth4       113.473174  10.359085  10.954  < 2e-16 ***
fmonth5       164.995654  10.353761  15.936  < 2e-16 ***
fmonth6       147.550395  10.351017  14.255  < 2e-16 ***
fmonth7       125.291246  10.350735  12.105  < 2e-16 ***
fmonth8       167.448974  10.352909  16.174  < 2e-16 ***
fmonth9       123.792810  10.357640  11.952  < 2e-16 ***
fmonth10      154.878407  10.538115  14.697  < 2e-16 ***
fmonth11      221.696168  10.537139  21.040  < 2e-16 ***
fmonth12      727.558629  10.537037  69.048  < 2e-16 ***
wtgrant       969.282000 547.285610   1.771   0.0788 .
time:wtgrant  -15.624940   8.315833  -1.879   0.0624 .
time2:wtgrant   0.052230   0.031467   1.660   0.0993 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.29 on 136 degrees of freedom
Multiple R-squared: 0.984,      Adjusted R-squared: 0.9821
F-statistic: 521.5 on 16 and 136 DF,  p-value: < 2.2e-16

> plot(predict(model),resid(model),xlab="predicted sales",ylab="sales
residual",main="Residual by Predicted Plot")
```



```
> qqnorm(resid(model))
> qqline(resid(model))
```



The residual by predicted plot shows clear heteroscedasticity. And the normal quantile plot of the residuals is not satisfactory. Clearly, we need to log the sales data.

To be explicit, we fit the model

$$\text{logsales} = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 WTGrant + \beta_4 t \cdot WTGrant + \beta_5 t^2 \cdot WTGrant$$
$$\log S_t + \varepsilon_t .$$

```
> model2<-
lm(logsales~time+time2+fmonth+wtgrant+wtgrant*time+wtgrant*time2);summa
ry(model2)
Call:
lm(formula = logsales ~ time + time2 + fmonth + wtgrant + wtgrant *
time + wtgrant * time2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.072088 -0.015601 -0.000901  0.014866  0.076454
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.705e+00  1.106e-02 515.896  < 2e-16 ***
time           5.655e-03  3.447e-04  16.404  < 2e-16 ***
time2         -5.803e-06  2.982e-06  -1.946  0.05372 .
fmonth2        3.409e-02  1.127e-02   3.026  0.00296 **
fmonth3        2.083e-01  1.127e-02  18.485  < 2e-16 ***
fmonth4        2.378e-01  1.132e-02  21.008  < 2e-16 ***
fmonth5        3.310e-01  1.131e-02  29.257  < 2e-16 ***
fmonth6        3.025e-01  1.131e-02  26.747  < 2e-16 ***
fmonth7        2.625e-01  1.131e-02  23.209  < 2e-16 ***
fmonth8        3.350e-01  1.131e-02  29.617  < 2e-16 ***
fmonth9        2.602e-01  1.132e-02  22.993  < 2e-16 ***
fmonth10       3.146e-01  1.151e-02  27.325  < 2e-16 ***
fmonth11       4.247e-01  1.151e-02  36.888  < 2e-16 ***
fmonth12       1.003e+00  1.151e-02  87.083  < 2e-16 ***
wtgrant        1.709e+00  5.980e-01   2.858  0.00493 **
time:wtgrant  -2.860e-02  9.087e-03  -3.148  0.00202 **
time2:wtgrant  1.045e-04  3.438e-05   3.038  0.00285 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02872 on 136 degrees of freedom
Multiple R-squared: 0.9908,     Adjusted R-squared: 0.9898
F-statistic: 920.3 on 16 and 136 DF,  p-value: < 2.2e-16
```

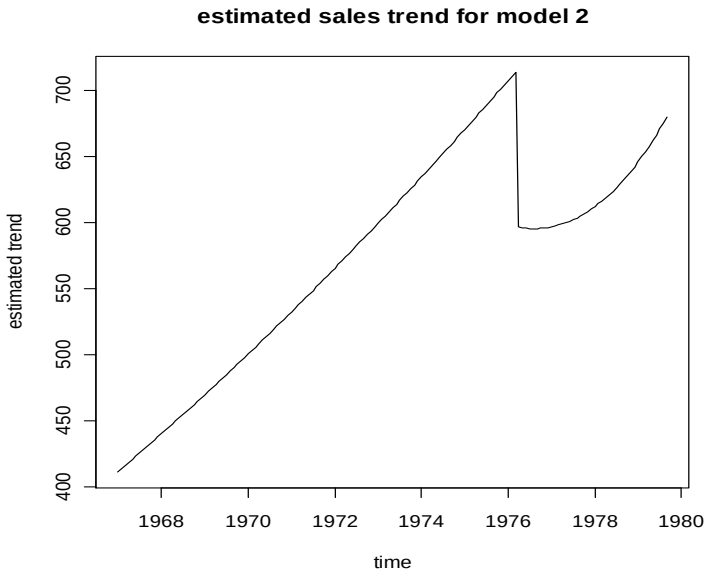
Below is a plot of the estimated trend component, shown for sales vs. time (not log sales vs. time). Prior to the demise of W. T. Grant there is steady upward growth with very slight curvature (growth is accelerating slowly). The closing of the W. T. Grant stores leads to a sudden drop in the level of monthly sales of about 117 million dollars, followed by stagnation for a year, and then a gradual recovery to approximately the same rate of growth as previously, but at a level of volume lower than that projected by the original trajectory of the trend.

Calculation of this trend is tricky because of the way R specifies the month dummy variables. The R estimate of the intercept is actually the intercept value we want plus the log of the January estimated seasonal index. We need to begin by removing the log of the January estimated seasonal index from the R intercept estimate.

```
> b1<-coef(model2)[1]
> b2<-coef(model2)[4:14]+b1
> b3<-c(b1,b2)
> newintrcpt<-b1-(b3-mean(b3))[1]
> newintrcpt
(Intercept)
   6.014189
```

This is the intercept value we want. Now we proceed to calculate and plot the trend.

```
> wtgrant2<-wtgrant*time;wtgrant3<-wtgrant*time2
> xmatrix<-
matrix(c(rep(1,length(sales)),time,time2,wtgrant,wtgrant2,wtgrant3),nco
l=6)
> sub<-c(2,3,15,16,17)
> betaest<-c(newintrcpt,coef(model2)[sub])
> plot(ts(exp(xmatrix%*%betaest)),xlab="time",ylab="estimated
trend",main="estimated sales trend for model 2")
```



The estimated trend for sales prior to April 1976 is

$$\exp(6.0142 + 0.005655\, t - 0.0000058\, t^2).$$

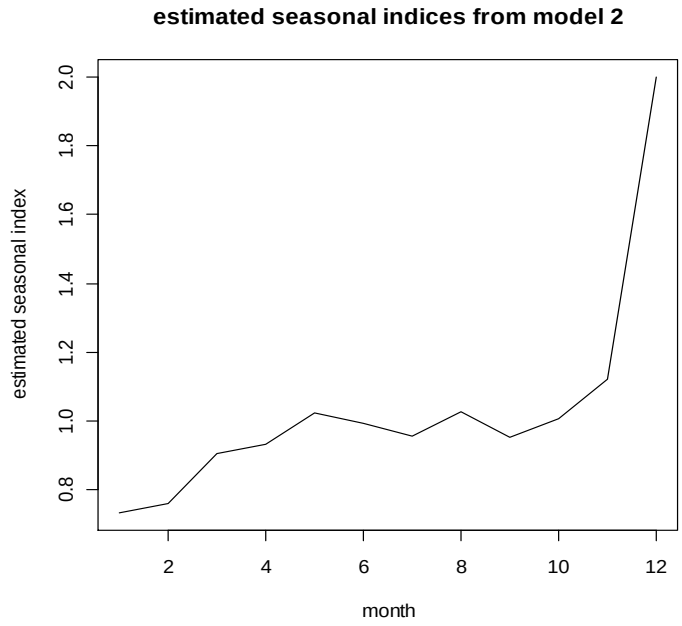
After W. T. Grant went out of business the estimated trend is

$$\exp(6.0142 + 1.7094 + (0.005655 - 0.028602)\, t + (-0.0000058 + 0.0001045\, t^2)$$

$$= \exp(7.7236 - 0.022947\, t + 0.0000987\, t^2).$$

Here are the estimated seasonal indices and their plot:

```
> seas<-exp(b3-mean(b3))
> seas
(Intercept)     fmonth2     fmonth3     fmonth4     fmonth5     fmonth6
  0.7338478   0.7592983   0.9037831   0.9308487   1.0217711   0.9930875
    fmonth7     fmonth8     fmonth9    fmonth10    fmonth11    fmonth12
  0.9541214   1.0259164   0.9519653   1.0052024   1.1221733   2.0000783
```



It is interesting to construct the seasonal component using the trigonometric basis.

```
> cosm<-matrix(nrow=length(sales),ncol=6)
> sinm<-matrix(nrow=length(sales),ncol=5)
> for(j in 1:5){
+ cosm[,j]<-cos(freq*j*time)
+ sinm[,j]<-sin(freq*j*time)
+ }
> cosm[,6]<-cos(freq*6*time)

> model3<-
lm(logsales~time+time2+wtgrant+wtgrant2+wtgrant3+cosm[,1]+sinm[,1]+cosm
[,2]+sinm[,2]+cosm[,3]+sinm[,3]+cosm[,4]+sinm[,4]+cosm[,5]+sinm[,5]+cos
m[,6]);summary(model3)

Call:
lm(formula = logsales ~ time + time2 + wtgrant + wtgrant2 + wtgrant3 +
    cosm[, 1] + sinm[, 1] + cosm[, 2] + sinm[, 2] + cosm[, 3] +
    sinm[, 3] + cosm[, 4] + sinm[, 4] + cosm[, 5] + sinm[, 5] +
    cosm[, 6])

Residuals:
      Min        1Q    Median        3Q       Max
-0.072088 -0.015601 -0.000901  0.014866  0.076454
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.014e+00  8.362e-03 719.193  < 2e-16 ***
time         5.655e-03  3.447e-04  16.404  < 2e-16 ***
time2       -5.803e-06  2.982e-06  -1.946  0.05372 .
wtgrant      1.709e+00  5.980e-01   2.858  0.00493 **
wtgrant2    -2.860e-02  9.087e-03  -3.148  0.00202 **
wtgrant3     1.045e-04  3.438e-05   3.038  0.00285 **
cosm[, 1]    7.365e-02  3.345e-03  22.022  < 2e-16 ***
sinm[, 1]   -9.287e-02  3.286e-03 -28.265  < 2e-16 ***
cosm[, 2]    1.475e-01  3.288e-03  44.851  < 2e-16 ***
sinm[, 2]   -9.765e-02  3.294e-03 -29.648  < 2e-16 ***
cosm[, 3]    1.540e-01  3.301e-03  46.662  < 2e-16 ***
sinm[, 3]   -5.071e-02  3.278e-03 -15.469  < 2e-16 ***
cosm[, 4]    1.340e-01  3.288e-03  40.744  < 2e-16 ***
sinm[, 4]   -4.473e-02  3.287e-03 -13.611  < 2e-16 ***
cosm[, 5]    1.224e-01  3.298e-03  37.109  < 2e-16 ***
sinm[, 5]    1.619e-02  3.278e-03   4.938 2.27e-06 ***
cosm[, 6]    6.167e-02  2.324e-03  26.531  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02872 on 136 degrees of freedom
Multiple R-squared: 0.9908,     Adjusted R-squared: 0.9898
F-statistic: 920.3 on 16 and 136 DF,  p-value: < 2.2e-16
```

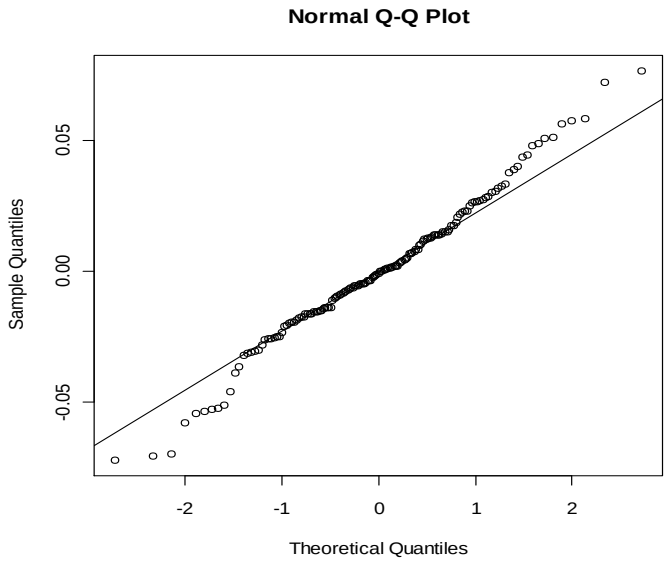
Amplitude and Phase Estimates

	Amplitude	Phase		Peak
		Degrees	Radians	t (period)
Fundamental	0.119	51.58	0.900	10.28
Second harmonic	0.177	33.51	0.585	5.44, 11.44
Third harmonic	0.162	18.22	0.318	3.80, 7.80, 11.80
Fourth harmonic	0.141	18.46	0.322	2.85, 5.85, 8.85, 11.85
Fifth harmonic	0.123	352.46	6.152	0.05, 2.45, 4.85, 7.25, 9.65
Sixth harmonic	0.062	0.00	0.000	0, 2, 4, 6, 8, 10

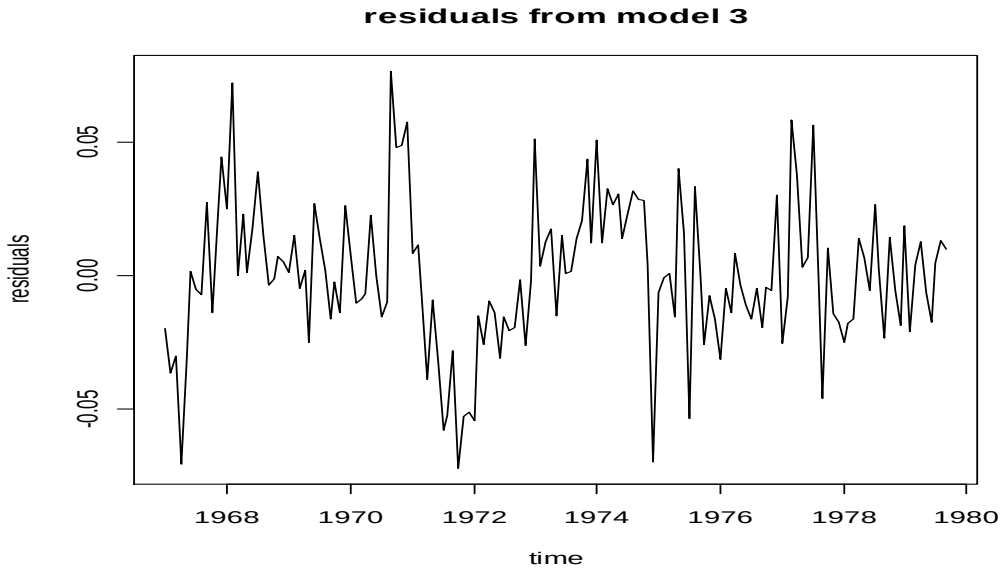
The peak calculations show the heavy concentration of sales in December and, to a lesser extent, in November. Note that the first five harmonic components are similar in intensity.

The residual distribution is not normal—both tails are long. Also, there are significant residual correlations. Further, the residual plot shows some remaining trend structure.

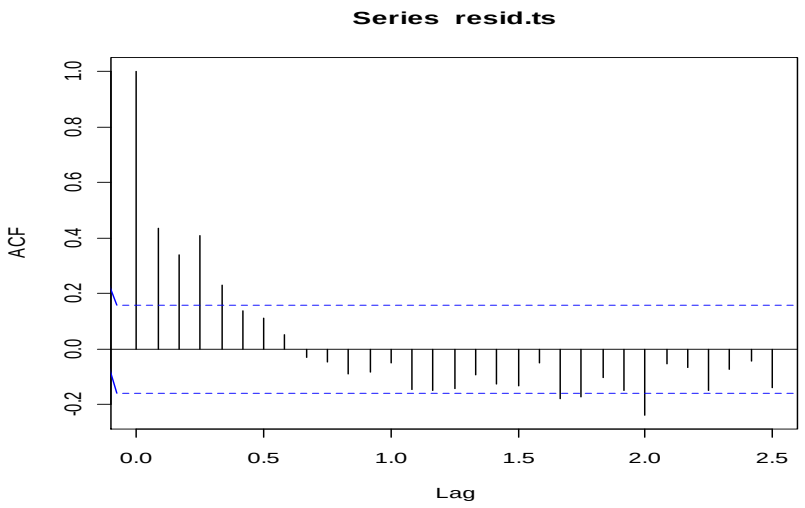
```
> qqnorm(resid(model3))
> qqline(resid(model3))
```




```
> resid.ts<-ts(resid(model3),start=c(1967,1),freq=12)
> plot(resid.ts,xlab="time",ylab="residuals",main="residuals from model
3")
```



Moreover, the residual autocorrelation plot, following, shows failure to reduce to white noise.



Some of the remaining structure can be captured by fitting a higher degree polynomial trend and adding a dummy for observation 96 (December 1974).

```
> time3<-time*time2;time4<-time*time3;time5<-time*time4
> model4<-
lm(logsales~time+time2+time3+time4+time5+wtgrant+wtgrant2+wtgrant3+obs9
6+fmonth);summary(model4)

Call:
lm(formula = logsales ~ time + time2 + time3 + time4 + time5 +
    wtgrant + wtgrant2 + wtgrant3 + obs96 + fmonth)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.647e+00  1.491e-02 378.866  < 2e-16 ***
time         1.663e-02  2.020e-03   8.233 1.55e-13 ***
time2       -4.943e-04  9.168e-05  -5.391 3.13e-07 ***
time3        8.306e-06  1.753e-06   4.739 5.49e-06 ***
time4       -5.822e-08  1.511e-08  -3.853 0.000181 ***
time5        1.372e-10  4.907e-11   2.797 0.005935 **
wtgrant      2.555e+00  2.530e+00   1.010 0.314420
wtgrant2    -4.913e-02  4.362e-02  -1.126 0.262035
wtgrant3     2.225e-04  1.892e-04   1.176 0.241648
obs96       -9.089e-02  2.673e-02  -3.400 0.000892 ***
fmonth2      3.366e-02  9.869e-03   3.411 0.000860 ***
fmonth3      2.076e-01  9.875e-03  21.018  < 2e-16 ***
fmonth4      2.348e-01  9.949e-03  23.597  < 2e-16 ***
fmonth5      3.274e-01  9.936e-03  32.952  < 2e-16 ***
fmonth6      2.985e-01  9.932e-03  30.055  < 2e-16 ***
fmonth7      2.581e-01  9.935e-03  25.979  < 2e-16 ***
fmonth8      3.303e-01  9.946e-03  33.214  < 2e-16 ***
fmonth9      2.553e-01  9.966e-03  25.615  < 2e-16 ***
fmonth10     3.102e-01  1.012e-02  30.662  < 2e-16 ***
fmonth11     4.203e-01  1.011e-02  41.557  < 2e-16 ***
fmonth12     1.006e+00  1.035e-02  97.194  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.02516 on 132 degrees of freedom
Multiple R-squared: 0.9932,     Adjusted R-squared: 0.9922
F-statistic:   962 on 20 and 132 DF,  p-value: < 2.2e-16
```

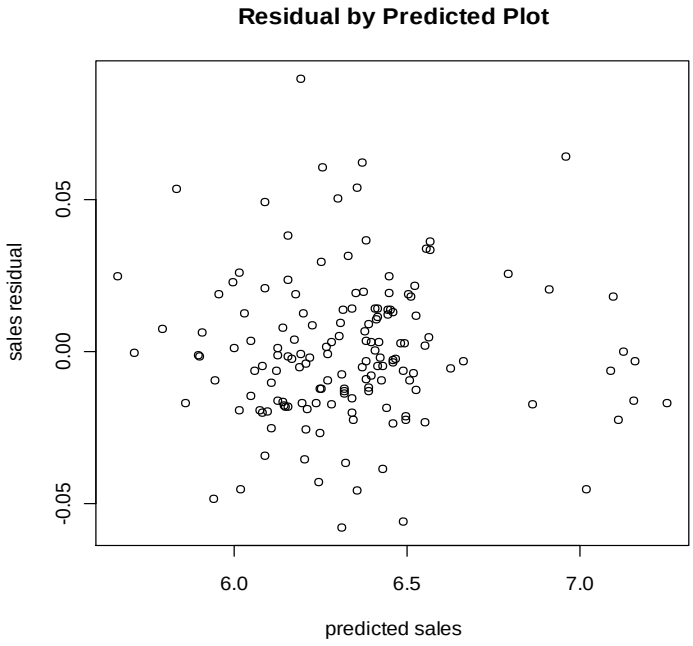
The estimated seasonal indices are essentially the same as those in the previous model. However, the interactions between *wtgrant* and trend components do not show significant *t* statistics. Let's use a partial *F* test for these two interactions.

```
> model5<-
lm(logsales~time+time2+time3+time4+time5+wtgrant+obs96+fmonth)
> anova(model5,model4)
Analysis of Variance Table

Model 1: logsales ~ time + time2 + time3 + time4 + time5 + wtgrant +
obs96 +
    fmonth
Model 2: logsales ~ time + time2 + time3 + time4 + time5 + wtgrant +
wtgrant2 +
    wtgrant3 + obs96 + fmonth
  Res.Df      RSS Df Sum of Sq     F Pr(>F)
1    134 0.085307
2    132 0.083540  2  0.001767 1.396 0.2512
```

We see that the two interactions are not needed—the *p* value for the partial *F* test is 0.2512. The added polynomial terms have accounted for the variation described in model 2 by the interactions between *wtgrant* and the trend components. The fit without the interactions, model 5, follows.

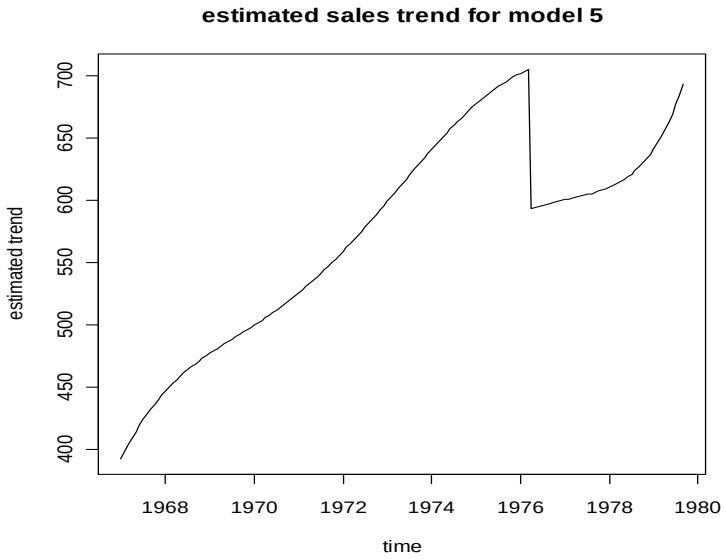
```
> summary(model5)

Call:
lm(formula = logsales ~ time + time2 + time3 + time4 + time5 +
    wtgrant + obs96 + fmonth)

Residuals:
      Min        1Q    Median        3Q       Max
-0.057748 -0.016314 -0.001951  0.012911  0.089597

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.650e+00  1.425e-02 396.405  < 2e-16 ***
time         1.606e-02  1.729e-03   9.286 3.80e-16 ***
time2       -4.775e-04  7.055e-05  -6.768 3.71e-10 ***
time3        8.326e-06  1.194e-06   6.972 1.29e-10 ***
time4       -6.232e-08  8.852e-09  -7.041 9.01e-11 ***
time5        1.647e-10  2.360e-11   6.981 1.23e-10 ***
wtgrant     -1.749e-01  1.399e-02 -12.502  < 2e-16 ***
obs96       -8.674e-02  2.669e-02  -3.250 0.001459 **
fmonth2      3.353e-02  9.898e-03   3.388 0.000925 ***
fmonth3      2.073e-01  9.902e-03  20.931  < 2e-16 ***
fmonth4      2.357e-01  9.947e-03  23.696  < 2e-16 ***
fmonth5      3.282e-01  9.949e-03  32.984  < 2e-16 ***
fmonth6      2.990e-01  9.954e-03  30.041  < 2e-16 ***
```

```
fmonth7      2.584e-01  9.962e-03  25.938  < 2e-16 ***
fmonth8      3.304e-01  9.973e-03  33.128  < 2e-16 ***
fmonth9      2.550e-01  9.986e-03  25.536  < 2e-16 ***
fmonth10     3.107e-01  1.014e-02  30.653  < 2e-16 ***
fmonth11     4.208e-01  1.014e-02  41.510  < 2e-16 ***
fmonth12     1.006e+00  1.038e-02  96.950  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02523 on 134 degrees of freedom
Multiple R-squared: 0.993,      Adjusted R-squared: 0.9921
F-statistic:  1062 on 18 and 134 DF,  p-value: < 2.2e-16
```



A plot of the trend estimate from this model follows.

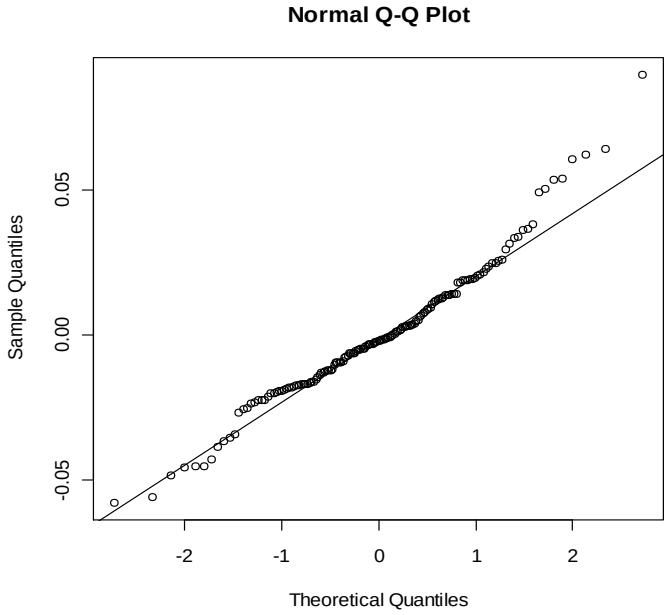
```
> b1<-coef(model5)[1]
> b2<-coef(model5)[9:19]+b1
> b3<-c(b1,b2)
> newintrcpt<-b1-(b3-mean(b3))[1]
> xmatrix<-
matrix(c(rep(1,length(sales)),time,time2,time3,time4,time5,wtgrant),nco
l=7)
> betaest<-c(newintrcpt,coef(model5)[2:7])
> plot(ts(exp(xmatrix%*%betaest)),xlab="time",ylab="estimated
trend",main="estimated sales trend for model 5")
```



The picture shows that the higher order polynomial terms have indeed accounted for the change in slope and curvature after the demise of W. T. Grant.

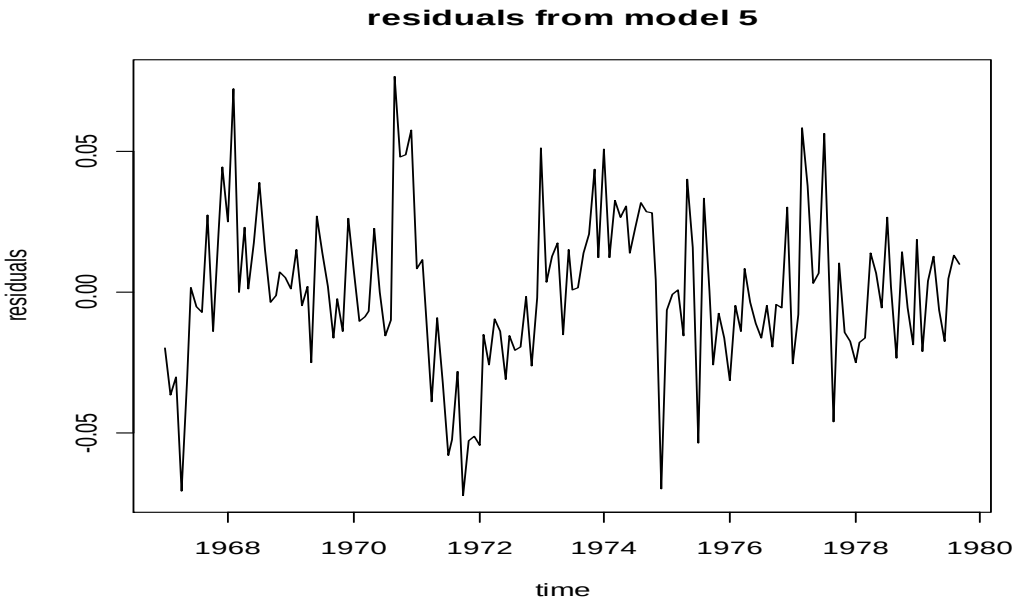
Residual analysis follows.

```
> qqnorm(resid(model5))
> qqline(resid(model5))
```



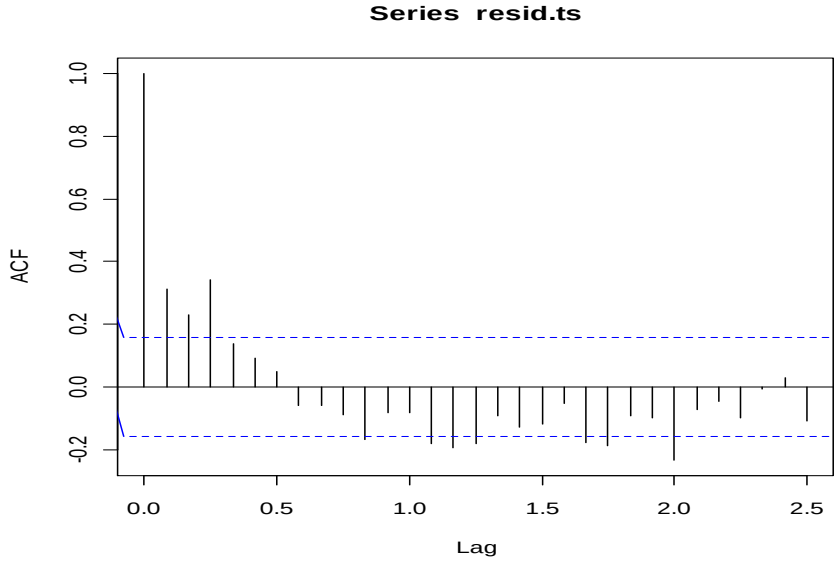
The upper tail of the residual distribution is long relative to normality.

```
> resid.ts<-ts(resid(model5),start=c(1967,1),freq=12)
> plot(resid.ts,xlab="time",ylab="residuals",main="residuals from model
5")
```



There is still uncaptured trend structure.

```
> acf(resid.ts,30)
```



There are still significant residual autocorrelations. However, this correlogram does show some improvement relative to the plot for model 3 (see page 30).

Spectral density. The spectral density function (also referred to as the spectrum) of a stationary time series is a very useful diagnostic tool. Mathematically, it is, up to scale, the Fourier transform of the autocorrelation function.

The usefulness of the spectral density lies in its ability to display key features of the structure of a time series. The horizontal axis for a plot of a spectral density is frequency, representing cycles per unit time, with values from 0 to 0.5. The measure on the vertical axis is spectral power. However, estimates of the spectral density are plotted on a logarithmic scale vertical axis. Any base logarithm can be used. It is common to use the natural logarithm, or 10 times the logarithm to the base 10, giving the decibel scale. A log transformation is used because it equalizes the standard error of the estimates across different frequencies.

A spectral density plot gives a decomposition of the variance of the time series according to frequency components forming the time series. For example, seasonal frequencies for a monthly time series with an annual cycle are 1/12, 2/12, 3/12, 4/12, 5/12, and 6/12. The spectral plot for a monthly time series with seasonal structure shows peaks at some or all of these frequencies. A peak at the frequency band near 0 is associated with trend structure of the time series. For a monthly time series for a flow variable, spectral peaks at frequencies 0.220, 0.348, and 0.432 are associated with variation attributable to calendar structure.

The goal in fitting a model to a time series is to reduce to white noise. The spectral density of a white noise time series is flat. Thus, one plots an estimate of the spectral density of the residuals from a model fit (one plots the log of the spectral estimate). If the plot is sufficiently flat, one can judge that there has been successful reduction to white noise.

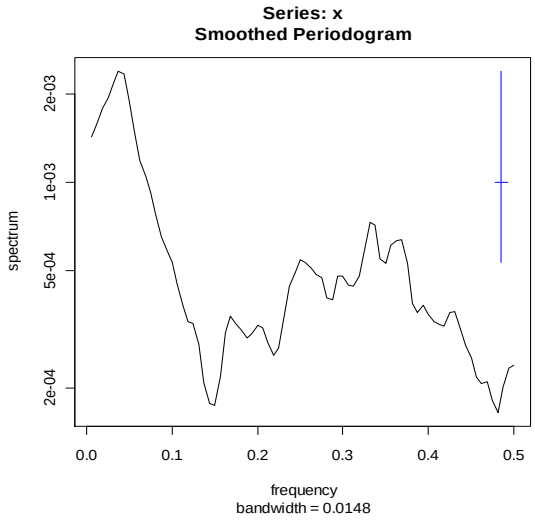
To judge for reduction to white noise via a fitted model, an alternative to use of the spectrum is examination of the residual autocorrelations. If these are not significant at all lags, one decides there is adequate reduction to white noise. The spectral plot, however, is more informative than the autocorrelation plot about pinpointing what kind of structure is remaining if there is failure to reduce to white noise.

An aside: If white light is passed through a prism, the spectrum one sees includes all colors—there are no interruptions forming gaps. One sees a constant height spectrum. White light contains all colors of the visual spectrum. This is the motivation for use of the term white noise as applied to time series.

Calculation of a spectral estimate in R requires specification of a bandwidth for the estimate. We need to choose how many raw estimates (given by the periodogram) are averaged across frequency (we need to choose the span) to produce the estimate. It is common to choose the span to be a fraction of the square root of the length of the time series. The ability to choose the span is an advantage, because it lets us see how the

appearance of the estimate changes as we change the bandwidth. In the plot below, the spectrum for the residuals of model 5 fit to the variety store sales data is shown with the span set equal to 8.

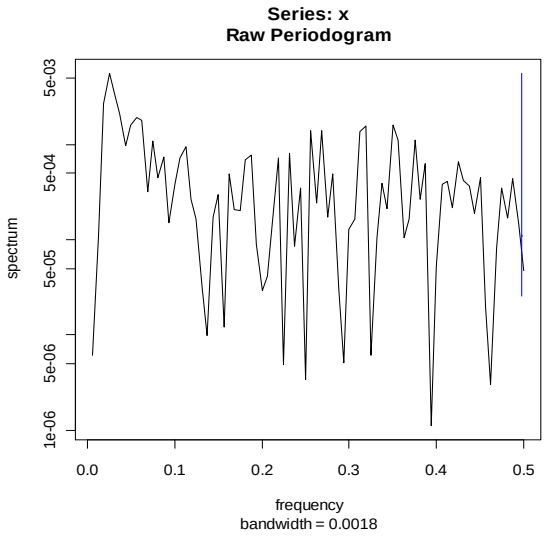
```
> spectrum(resid(model5),span=8)
```



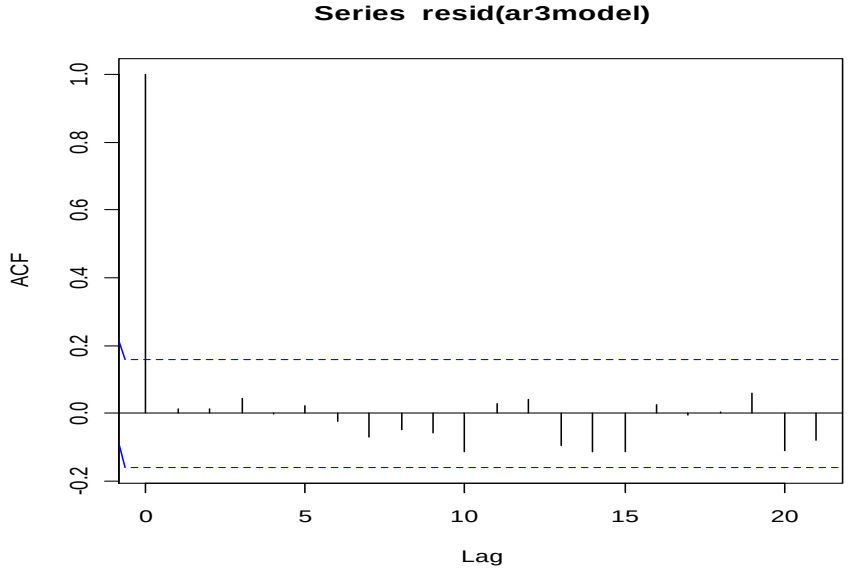
This spectral plot is not flat, indicating that the residuals do not conform to white noise structure. The spectral peak at low frequency is for slow movement of the time series, essentially remaining trend structure (as the plot on page 34 shows). There is also some spectral activity around frequencies 0.348 and 0.432, perhaps signaling modest calendar effects. The vertical blue line in the plot indicates how to draw a confidence interval band around the estimate. If, after adding the band, one can draw a horizontal line which lies entirely within the band, one can assert that the residuals conform to white noise, and thus the model has successfully reduced the time series to white noise.

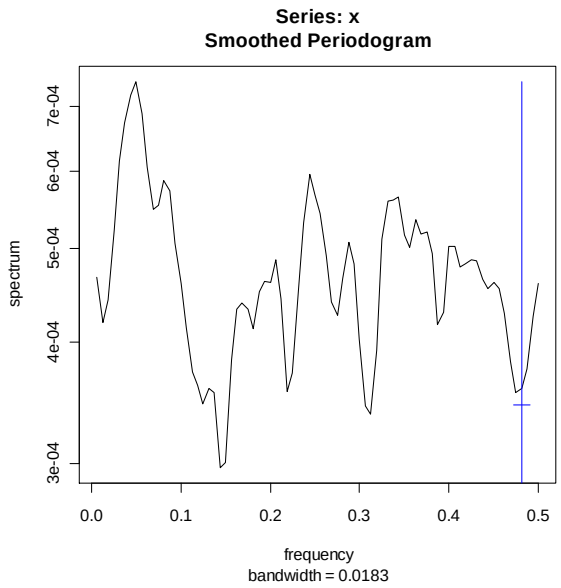
If a bandwidth is not specified in giving the R command, the unsmoothed periodogram is plotted, as the following shows.

```
> spectrum(resid(model5))
```



An autoregressive model of order 3 is a useful fit for the model 5 residual time series. Analysis of the residuals from this autoregressive fit follows.





The autocorrelation and spectral density plots for the residuals from the autoregressive fit show adequate reduction to white noise. That is, modeling the residuals as autoregressive of order 3 has achieved reduction to white noise. Later in the course we will study autoregressive models and their use.

Let's explore further the usefulness of spectral density plots. Consider the Australian beer data. The first model below fits only a trend (a sixth-degree polynomial) and a dummy for an outlier at observation 317 to the log of monthly production.

```
> ausbeer<-read.csv("G:/Stat71122Spring/beeraustralia.txt",header=T)
> attach(ausbeer)
> fmonth<-factor(month)

> head(ausbeer)
  year month beer    dlogbeer obs317 obs318
1 1956     1 93.2          NA      0      0
2 1956     2 96.0  0.02960047      0      0
3 1956     3 95.2 -0.00836825      0      0
4 1956     4 77.1 -0.21087666      0      0
5 1956     5 70.9 -0.08383285      0      0
6 1956     6 64.8 -0.08996483      0      0

> time<-as.numeric(1:length(beer))
```

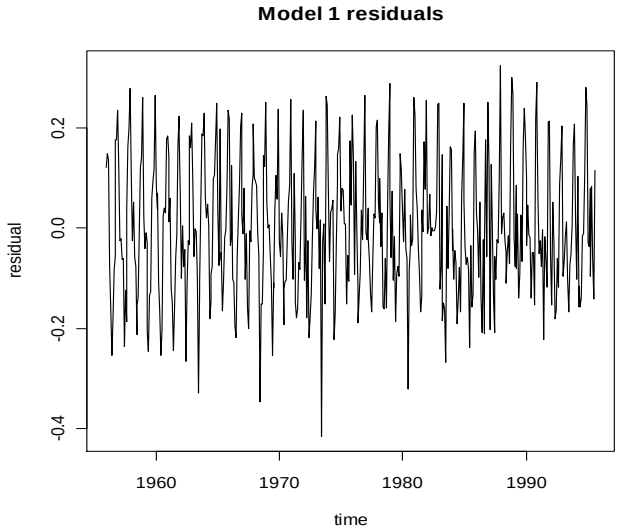
The following model includes a sixth-degree polynomial for trend estimation and a dummy variable for the May 1982 outlier, *but no seasonal estimation.*

```
> model1<-
lm(log(beer)~time+I(time^2)+I(time^3)+I(time^4)+I(time^5)+I(time^6)+obs
317);summary(model1)
```

```
Call:
lm(formula = log(beer) ~ time + I(time^2) + I(time^3) + I(time^4) +
    I(time^5) + I(time^6) + obs317)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41511 -0.09595 -0.01442  0.08669  0.32406
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.410e+00  4.414e-02  99.908  < 2e-16 ***
time         2.798e-03  2.551e-03   1.097  0.27332
I(time^2)   -4.905e-05  4.642e-05  -1.057  0.29119
I(time^3)    6.837e-07  3.648e-07   1.874  0.06156 .
I(time^4)   -3.352e-09  1.390e-09  -2.411  0.01628 *
I(time^5)    6.831e-12  2.530e-12   2.700  0.00718 **
I(time^6)   -5.012e-15  1.762e-15  -2.844  0.00465 **
obs317      -3.381e-01  1.348e-01  -2.508  0.01248 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1341 on 468 degrees of freedom
Multiple R-squared: 0.747,      Adjusted R-squared: 0.7432
F-statistic: 197.4 on 7 and 468 DF,  p-value: < 2.2e-16
```

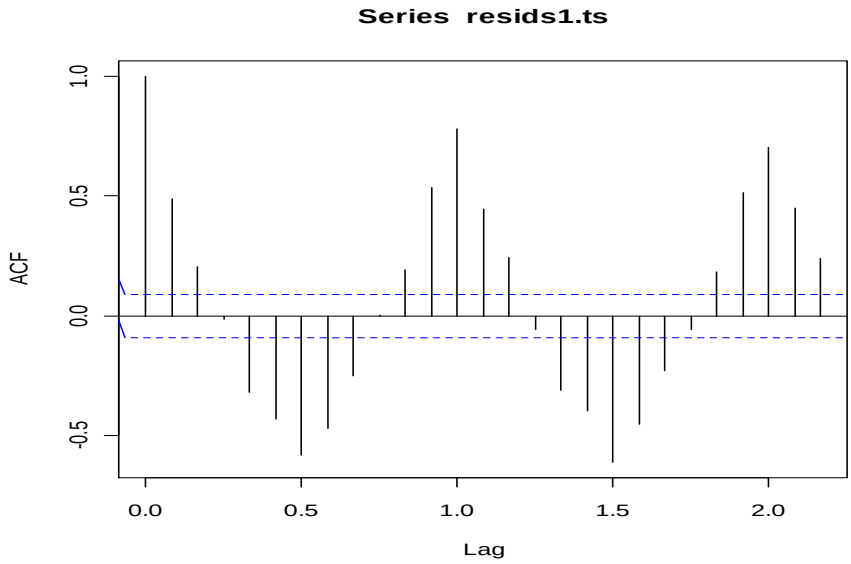
Let’s look at the residual diagnostics for this fit.

```
> resids1.ts<-ts(resid(model1),start=c(1956,1),freq=12)
> plot(resids1.ts,xlab="time",ylab="residual",main="Model 1 residuals")
```



The sixth-degree polynomial has largely captured the trend structure. However, the seasonal structure is still included in the model residuals, as the time plot above shows.

And the residual acf plot, below, also illustrates the presence of remaining seasonal structure.

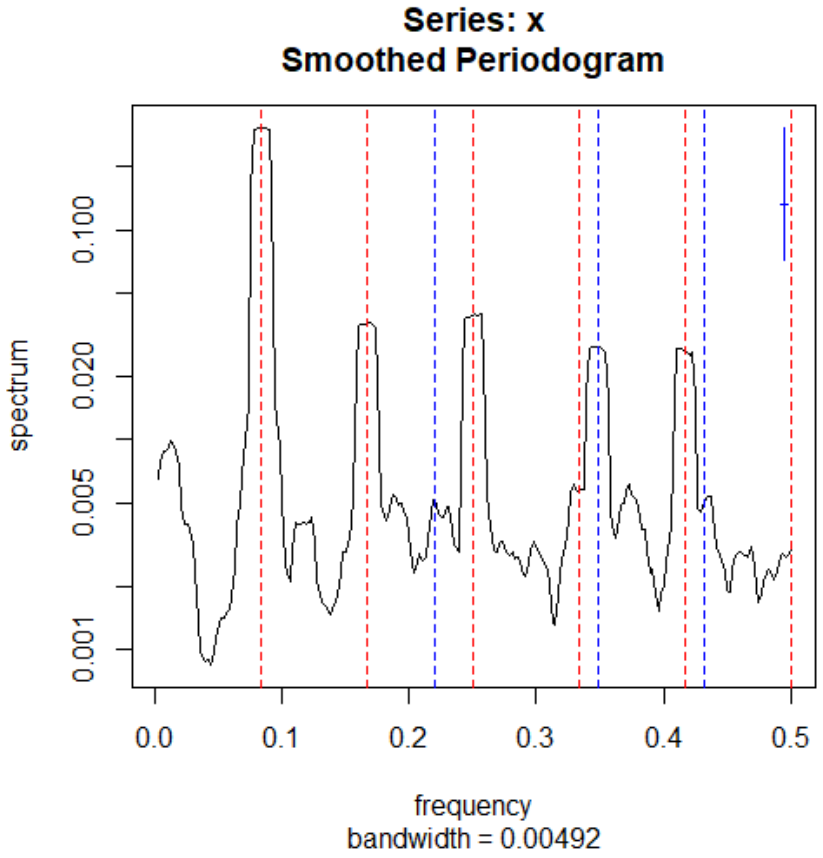


The residual spectral density is especially informative:

```
> resids1<-resid(model1)
> class(resids1)
[1] "numeric"
> spectrum(resids1,span=8)
```

I've defined the residuals as numeric class so that the spectral plot is pictured with frequency ranging from 0 to 0.5 on the horizontal axis. The red lines mark seasonal frequencies, and the blue lines mark calendar frequencies.

```
resids1<-resid(model1)
class(resids1)
[1] "numeric"

spectrum(resids1,span=8)
abline(v=c(1/12,2/12,3/12,4/12,5/12,6/12),col="red",lty=2)
abline(v=c(0.220,0.348,0.432),col="blue",lty=2)
```



The spectral plot red lines clearly show prominent spectral activity at frequencies 1/12, 2/12, 3/12, and 5/12, and perhaps at frequency 4/12, precisely indicating the presence of unmodeled seasonal structure in the model 1 residuals. The calendar structure at frequency 0.348 is also prominently present in the model 1 residuals, and perhaps calendar structure at frequency 0.432 is also included. At low frequency the plot shows there is some trend remaining in the residuals.

Let's add (static) seasonal structure and the calendar trigonometric pairs with frequencies 0.220, 0.348, and 0.432 to the model, and look at the spectral density of the new set of residuals. Let's use the trigonometric basis for the seasonal structure.

```
> cosm<-matrix(nrow=length(time),ncol=6)
> sinm<-matrix(nrow=length(time),ncol=5)
> for(i in 1:5){
+ cosm[,i]<-cos(2*pi*i*time/12)
+ sinm[,i]<-sin(2*pi*i*time/12)
+ }
> cosm[,6]<-cos(pi*time)
> c1<-cosm[,1];c2<-cosm[,2];c3<-cosm[,3];c4<-cosm[,4];c5<-cosm[,5];c6<-
cosm[,6]
> s1<-sinm[,1];s2<-sinm[,2];s3<-sinm[,3];s4<-sinm[,4];s5<-sinm[,5]
> c220<-cos(0.440*pi*time);s220<-sin(0.440*pi*time);c348<-
cos(0.696*pi*time);s348<-sin(0.696*pi*time);c432<-
cos(0.864*pi*time);s432<-sin(0.864*pi*time)
```

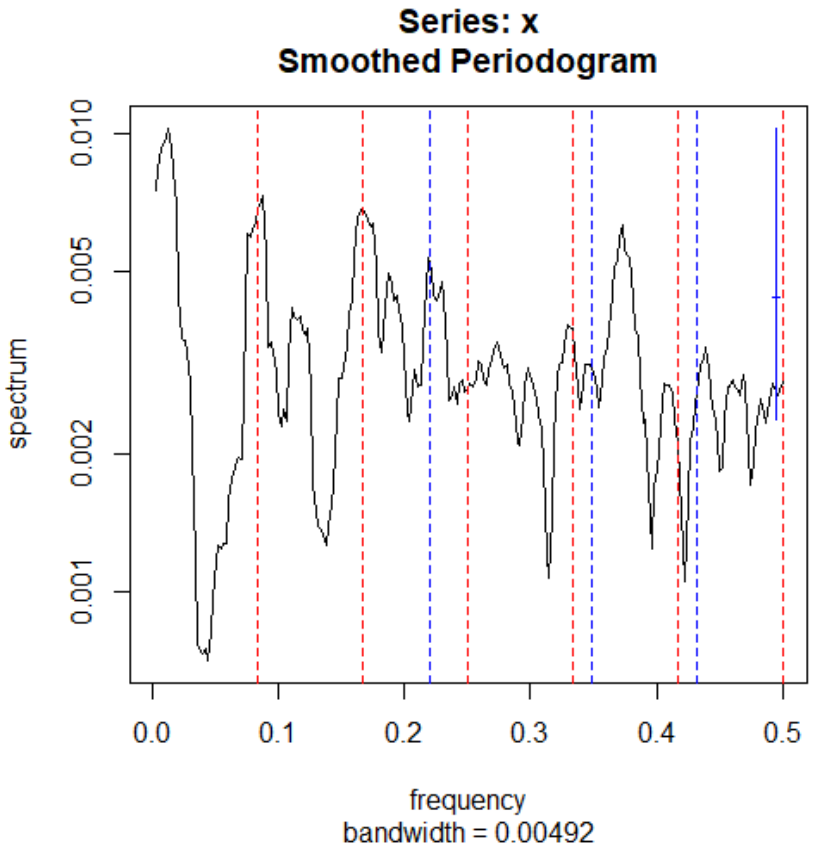
```
> model2<-
lm(log(beer)~time+I(time^2)+I(time^3)+I(time^4)+I(time^5)+I(time^6)+obs
317+c1+s1+c2+s2+c3+s3+c4+s4+c5+s5+c6+c220+s220+c348+s348+c432+s432);sum
mary(model2)

Call:
lm(formula = log(beer) ~ time + I(time^2) + I(time^3) + I(time^4) +
    I(time^5) + I(time^6) + obs317 + c1 + s1 + c2 + s2 + c3 +
    s3 + c4 + s4 + c5 + s5 + c6 + c220 + s220 + c348 + s348 +
    c432 + s432)

Residuals:
      Min        1Q    Median        3Q       Max
-0.185913 -0.036013 -0.000215  0.037196  0.168246

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.430e+00  1.959e-02 226.107  < 2e-16 ***
time         1.680e-03  1.132e-03   1.485 0.138326
I(time^2)   -2.909e-05  2.059e-05  -1.413 0.158380
I(time^3)    5.270e-07  1.618e-07   3.257 0.001213 **
I(time^4)   -2.750e-09  6.165e-10  -4.460 1.04e-05 ***
I(time^5)    5.718e-12  1.122e-12   5.096 5.11e-07 ***
I(time^6)   -4.221e-15  7.815e-16  -5.402 1.07e-07 ***
obs317      -2.244e-01  6.083e-02  -3.688 0.000254 ***
c1           1.447e-01  3.862e-03  37.465  < 2e-16 ***
s1          -9.851e-03  3.858e-03  -2.554 0.010987 *
c2           5.935e-03  3.859e-03   1.538 0.124803
s2          -4.430e-02  3.856e-03 -11.488  < 2e-16 ***
c3           3.989e-02  3.854e-03  10.352  < 2e-16 ***
s3          -2.805e-02  3.862e-03  -7.263 1.67e-12 ***
c4           1.210e-02  3.859e-03   3.135 0.001828 **
s4          -2.992e-03  3.855e-03  -0.776 0.438038
c5           3.046e-02  3.861e-03   7.887 2.35e-14 ***
s5           2.450e-02  3.858e-03   6.350 5.25e-10 ***
c6          -2.209e-03  2.728e-03  -0.810 0.418400
c220        -3.353e-03  3.857e-03  -0.869 0.385145
s220         3.743e-04  3.858e-03   0.097 0.922772
c348        -2.239e-02  3.858e-03  -5.803 1.23e-08 ***
s348        -3.342e-02  3.856e-03  -8.667  < 2e-16 ***
c432        -1.228e-02  3.869e-03  -3.174 0.001608 **
s432         5.222e-03  3.850e-03   1.357 0.175603
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05943 on 451 degrees of freedom
Multiple R-squared:  0.9521,    Adjusted R-squared:  0.9496
F-statistic: 373.8 on 24 and 451 DF,  p-value: < 2.2e-16
```

All of the seasonal trigonometric components except c6 are significant. The calendar pair at frequency 0.220 is not significant, but the 0.348 and 0.432 pairs are.

The spectral plot of the residuals from this new fit, model 2, should eliminate much of the spectral activity at the seasonal frequencies (not all, because dynamic seasonal structure

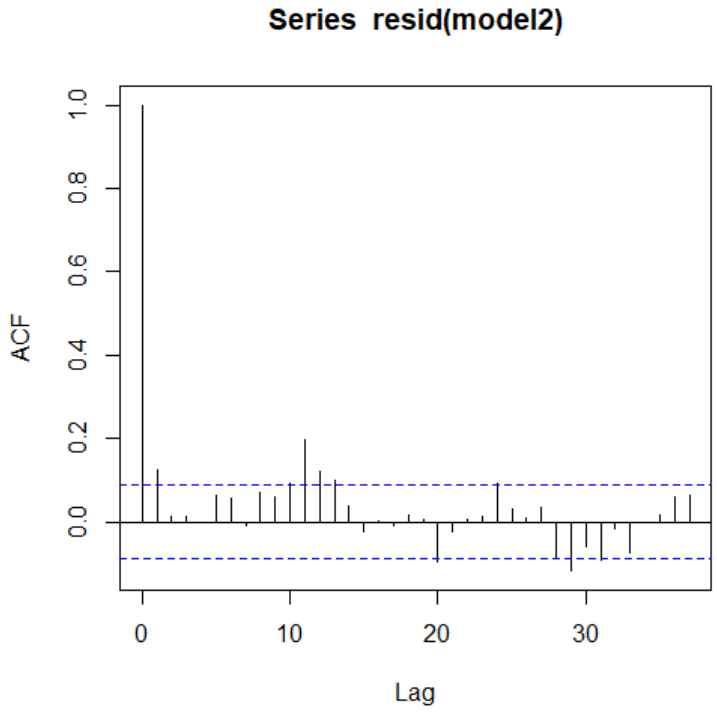
is not addressed by model 2), and all of the spectral activity at the calendar frequencies 0.348 and 0.432.

```
> spectrum(resid(model2),span=8)
> abline(v=c(1/12,2/12,3/12,4/12,5/12,6/12),col="red",lty=2)
> abline(v=c(0.220,0.348,0.432),col="blue",lty=2)
```



First note that the spectral power is greatly diminished. In the plot on page 41 it ranges from 0.001 to about 0.4. In the plot directly above the range is 0.001 to 0.01. The plot above shows the remaining trend (at low frequency). There are peaks at the seasonal frequencies 1/12, 2/12, and 4/12. These indicate remaining dynamic seasonal structure. However, the static structure has been removed by model 2. In addition, model 2 has properly accounted for calendar structure at frequencies 0.348 and 0.432. The plot shows that model 2 has not produced reduction to white noise—the remaining trend and dynamic seasonal structure have precluded reduction to white noise. However, model 2 has accounted for much of the structure in the beer production time series.

Let's also look at the residual acf plot for model 2.



Although this plot also shows failure to reduce to white noise, it indicates that not much significant structure remains. The residual spectral plot clearly provides more useful information than this residual acf plot. It identifies much more clearly what structure remains. The blue bar in the spectral plot shows the width of a 95 per cent confidence interval around the spectral curve. As mentioned above, the confidence interval width is constant across frequency when we plot the spectral estimate on a log scale. R uses the log scale (base e) as the default.

Let's end by relating the amplitude calculations from model 2 to the spectral power results in the plot on page 41. Here are the amplitude calculations for model 2.

```
> ampltd<-c(rep(0,times=6))
> b2<-coef(model2)[9:19]
> for(i in 1:5){
+ i1<-2*i-1
+ i2<-i1+1
+ ampltd[i]<-sqrt(b2[i1]^2+b2[i2]^2)
+ }
> ampltd[6]<-abs(b2[11])

> ampltd
[1] 0.145011537 0.044699068 0.048768508 0.012464686 0.039084803
0.002209124
```

We compare these squared amplitudes to the peaks of spectral power at the seasonal frequencies in the plot on page 41.

```
> ampltd^2
[1] 2.102835e-02 1.998007e-03 2.378367e-03 1.553684e-04 1.527622e-03
[6] 4.880229e-06
```

Consider the following table.

	Frequency	Amplitude
Fundamental	1/12	0.145012
Second harmonic	2/12	0.044699
Third harmonic	3/12	0.048769
Fourth harmonic	4/12	0.012465
Fifth harmonic	5/12	0.039085
Sixth harmonic	6/12	0.002209

Compare the values in the second column to the heights of the spectral peaks at the seasonal frequencies for the plot on page 41. The plot on page 41 gives the spectral activity for the residuals from model 1, and the table above shows that portion of the spectrum accounted for by the static seasonal structure.

The table above shows that the fundamental seasonal component at frequency 1/12 is dominant, and the components at frequencies 3/12 and 2/12 are next most prominent. The component at frequency 5/12 also plays a role, and those at frequencies 4/12 and 6/12 are less important. The results here in the table do not precisely match those in the plot on page 41, but there is an approximate correspondence. One reason for the failure to match precisely is that the second column in the table describes only static seasonal structure, and the plot on page 41 includes both static and dynamic seasonal structure. The spectral summary of just the dynamic seasonal structure is described in the spectral plot on page 43.

Summary and additional remarks

1. The time series of monthly sales for variety stores has a sudden drop in the trend, attributable to the bankruptcy and closure of the chain W. T. Grant. To model the series, we include a dummy which is equal to 0 when W. T. Grant is in business, and equal to 1 after it ceases to operate. Use of the interaction between this dummy and the powers of time permits estimation of two trend structures, one before and one after the cessation of operation.

2. The estimate of the spectral density of the residual time series is used as a tool to help determine whether a model has adequately reduced to white noise. If there is reduction to white noise, the residual spectrum appears flat. If the residual spectrum is judged to be not flat and has peaks, the peak locations on the frequency axis can often be used to identify components which need to added to a fitted model. For the variety store time series, residual spectral calculations reveal peaks at the calendar frequencies 0.348 and 0.432, and also peaks at several seasonal frequencies. The latter (seasonal) peaks signal less than complete estimation of the seasonal structure.

3. The spectral estimation methodology is also used to show clearly the need to include seasonal components and calendar trigonometric pairs in modeling the Australian beer data series.