STAT 608, Spring 2022 - Assignment 7
SOLUTIONS

1. For an AR(1) process:

$$e_t = \rho e_{t-1} + \nu_t$$

for $t = 2, 3, \ldots, n$ and where $\nu_1, \nu_2, \ldots, \nu_n$ are iid $N(0, \sigma_\nu^2)$, show that

$$\text{Corr}(e_t, e_{t-2}) = \rho^2$$

NOTE THAT WE CAN WRITE

$$e_t = \rho e_{t-1} + \nu_t = \rho\left(\rho e_{t-2} + \nu_{t-1}\right) + \nu_t$$

THEN

$$\text{CORR}(e_t, e_{t-2}) = \frac{\text{COV}\left(\rho\left(\rho e_{t-2} + \nu_{t-1}\right) + \nu_t, e_{t-2}\right)}{\sigma_e^2}$$

WHERE $\sigma_e^2 = \text{VAR}(e_t)$ FOR ALL t. THIS THEN EQUALS

$$\frac{\text{COV}\left(\rho\left(\rho e_{t-2} + \nu_{t-1}\right) + \nu_t, e_{t-2}\right)}{\sigma_e^2} = \frac{\text{COV}\left(\rho^2 e_{t-2}, e_{t-2}\right) + \text{COV}\left(\rho\nu_{t-1}, e_{t-2}\right) + \text{COV}\left(\nu_t, e_{t-2}\right)}{\sigma_e^2}$$

$$= \frac{\rho^2\sigma_e^2 + 0 + 0}{\sigma_e^2} = \rho^2$$

SINCE ALL THE e AND ν TERMS ARE INDEPENDENT OF ONE ANOTHER.

2. Consider the regression model:

$$y_t = \beta_0 + \beta_1 x_t + e_t$$

where the e_t follow an AR(1) process: $e_t = \rho e_{t-1} + \nu_t$, where the ν_t are iid $N(0, \sigma_\nu^2)$. Conduct a simulation to examine the coverage probabilities of nominal 95% confidence intervals for mean response when $\rho = 0.1, 0.2, \ldots, 0.9$ and when usual least squares is used to fit the model (i.e., assuming no serial correlation). Provide either a table or plot showing the actual coverage probabilities as a function of ρ. In your simulation, use $x_1, x_2, \ldots, x_n \overset{iid}{\sim} U(0, 1)$, $\beta_0 = 0.5$, $\beta_1 = 1.5$, $n = 50$, and $\sigma_\nu^2 = 0.5$. In each simulation, compute a CI for mean response when $x = 0.5$. Use $M = 1000$ simulations for each value of ρ.

MY R CODE IS AT THE END OF THIS DOCUMENT. I OBTAINED THE RESULTS IN FIGURE 1. AS EXPECTED, OUR INFERENCE BECOMES LESS ACCURATE WITH STRONGER CORRELATIONS. WITH ρ NEARING 1, THE CONFIDENCE INTERVAL COVERAGE DROPS PRECIPITOUSLY.

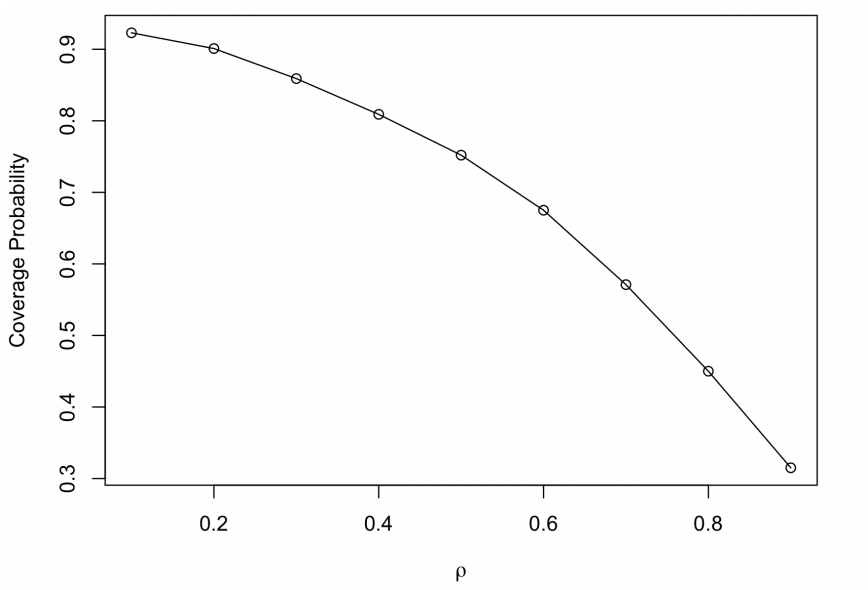

Figure 1: Actual coverage probabilities for nominal 95% confidence intervals for mean response.

3. Exercise 9.2 from the textbook.

   (a) First, it is interesting to see the seasonal pattern of sales. Figure 2 shows side-by-side boxplots of sales by month. In particular, you can see that sales peak around Christmas time. Computing the standardized residuals, we see that observations 11, 32, 33, 41, and 59 all have standardized residuals greater than 2 in absolute value and hence are potential outliers. Observations 11, 32, and 59 are all from late in the year and have relatively high sales values. Observation 41 is from the middle of the year and has a high sales value compared to other observations from that time period. Observation 33 is from late in the year and has a low sales value compared to other observations from that time period. Based on the available data, we don't have any reason to suspect the data points are errors of any kind or fundamentally different from the rest, so they will be kept in the analysis.

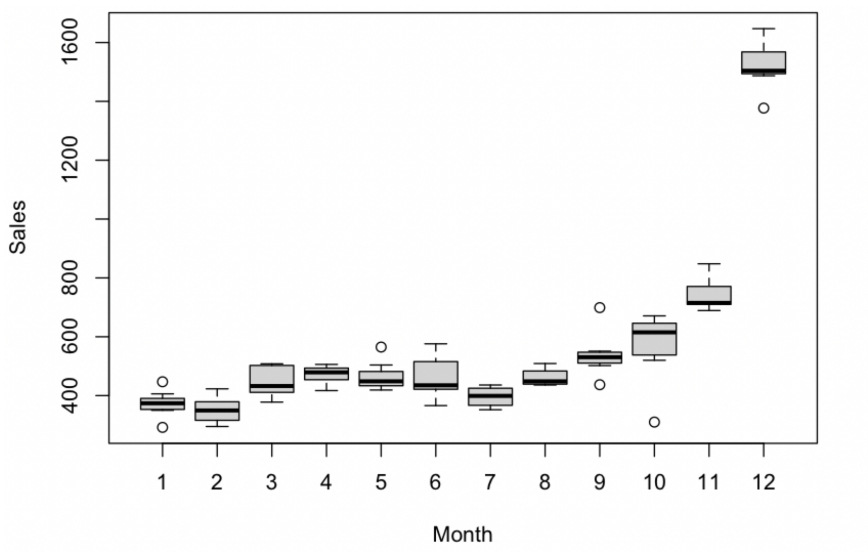

Figure 2: Side-by-side boxplots of sales by month.

HERE IS THE MODEL SUMMARY OUTPUT:

```
Call:
lm(formula = Sales ~ Time + Month_2 + Month_3 + Month_4 + Month_5 +
    Month_6 + Month_7 + Month_8 + Month_9 + Month_10 + Month_11 +
    Month_12, data = dta)

Residuals:
     Min       1Q   Median       3Q      Max
-254.638  -33.188   -7.513   31.888  167.612

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  350.3667    25.9841  13.484  < 2e-16 ***
Time           0.4298     0.2381   1.805  0.07485 .
Month_2      -18.4044    31.8128  -0.579  0.56454
Month_3       77.5408    31.8048   2.438  0.01698 *
Month_4      101.8609    31.7985   3.203  0.00195 **
Month_5       93.4311    31.7941   2.939  0.00431 **
Month_6       89.7513    31.7914   2.823  0.00600 **
Month_7       24.8214    31.7905   0.781  0.43724
Month_8       89.0166    31.7914   2.800  0.00640 **
Month_9      166.8368    31.7941   5.247 1.24e-06 ***
Month_10     199.6569    31.7985   6.279 1.66e-08 ***
Month_11     374.2882    32.8366  11.399  < 2e-16 ***
Month_12    1150.7155    32.8340  35.047  < 2e-16 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 61.43 on 80 degrees of freedom
Multiple R-squared:  0.9634,Adjusted R-squared:  0.9579
F-statistic: 175.3 on 12 and 80 DF,  p-value: < 2.2e-16
```

Time is marginally statistically significant, months 2 and 7 are insignificant, but all other terms are clearly significant. Figure 3 shows the default diagnostic plots for the fitted model object. The QQ plot reveals non-Normality. The scale-location plot suggests some minor non-constant variance, but it so minor that we will ignore it. The Cook's distance plot does not suggest and "bad" leverage points. Figure 4 shows a marginal model plot for variable Time. It looks good, since the two curves closely match each other. Figure 5 shows the autocorrelation function for the model residuals. There is apparently a weak correlation at lag 1. We might choose to ignore it and stick with this simpler model rather than fitting a GLS model with correlated errors.

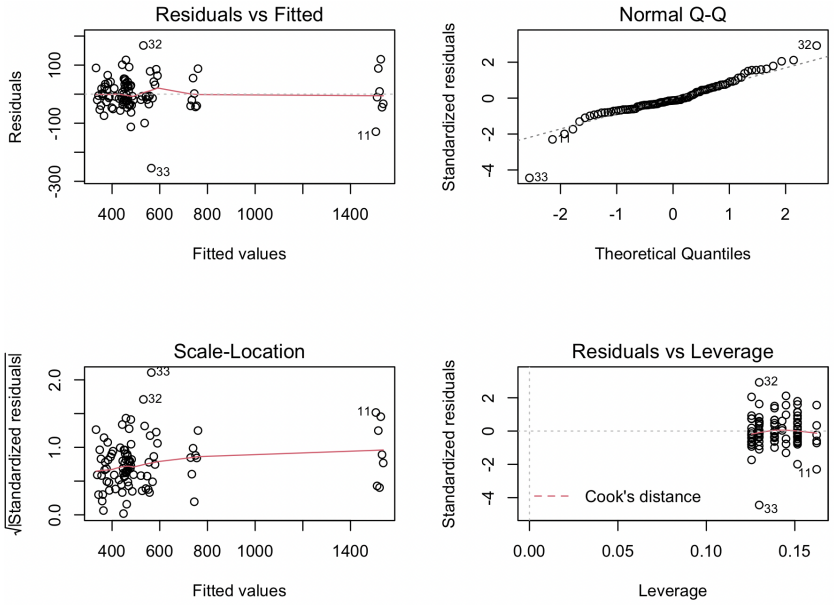

Figure 3: Diagnostics for the model in part (a).

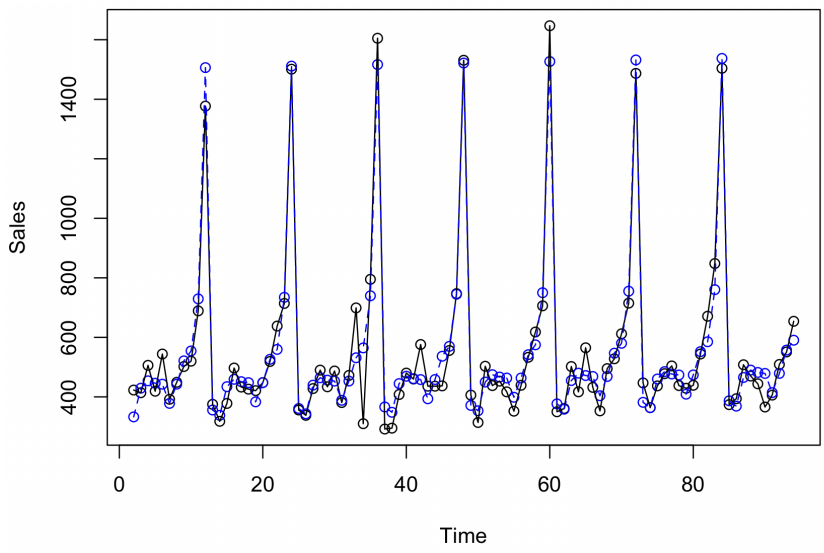

Figure 4: Marginal model plot for Time in part (a).

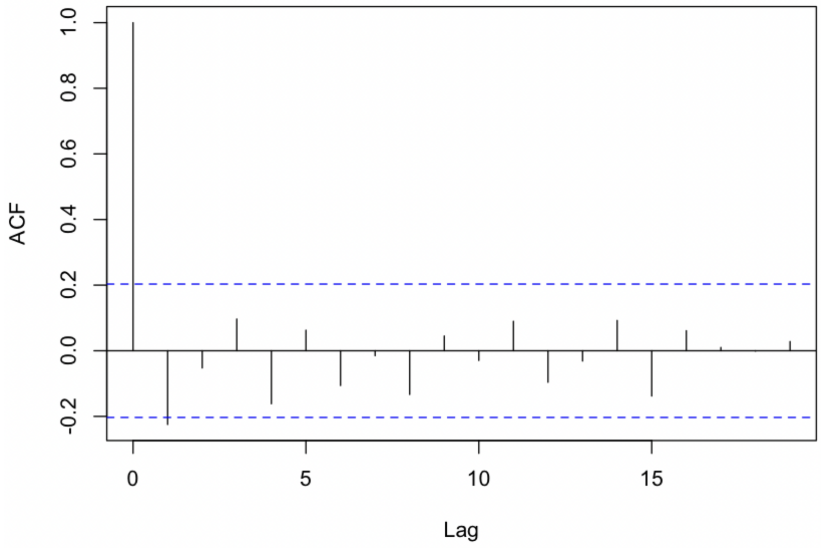

Figure 5: Plot of autocorrelations of residuals for model in part (a).

(b) Here is the model summary output:

```
Call:
lm(formula = Sales ~ Time + Month_2 + Month_3 + Month_4 + Month_5 +
    Month_6 + Month_7 + Month_8 + Month_9 + Month_10 + Month_11 +
    Month_12 + Advert + Lag1Advert, data = dta)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-240.283  -28.203   -6.308   28.352  167.624

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  392.3619    40.8122   9.614 6.99e-15 ***
Time           0.4322     0.2377   1.818 0.072853 .
Month_2      -45.8840    37.8276  -1.213 0.228799
Month_3       26.6907    40.5869   0.658 0.512720
Month_4       81.0497    33.5470   2.416 0.018030 *
Month_5       59.3321    36.7394   1.615 0.110361
Month_6       59.3564    36.4203   1.630 0.107184
Month_7      -12.4035    37.4233  -0.331 0.741203
Month_8       60.0235    35.9118   1.671 0.098648 .
Month_9      129.0218    37.2445   3.464 0.000867 ***
Month_10     165.7306    35.4326   4.677 1.20e-05 ***
Month_11     324.7993    43.5922   7.451 1.08e-10 ***
Month_12    1149.1212    39.4806  29.106  < 2e-16 ***
Advert         2.2999     2.5743   0.893 0.374385
Lag1Advert    -4.5174     2.6625  -1.697 0.093744 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.55 on 78 degrees of freedom
Multiple R-squared:  0.9653,Adjusted R-squared:  0.9591
F-statistic:   155 on 14 and 78 DF,  p-value: < 2.2e-16
```

Notice that neither of the new variables are statistically significant, which means they can likely be removed without losing fit quality. In fact, the p-value for a partial F test comparing model (b) and the nested model (a) gives a p-value of 0.1215. The default diagnostic plots are shown in Figure 6 and do not reflect anything new relative to the plot for model (a). The same data points as in (a) are potential outliers here. The marginal model plots for variables Time, Advert, and Lag1Advert are shown in Figure 7. They are messy but do not reveal any major issues (the curves follow each other closely). The autocorrelation function does not look qualitatively different from the one in part (a).

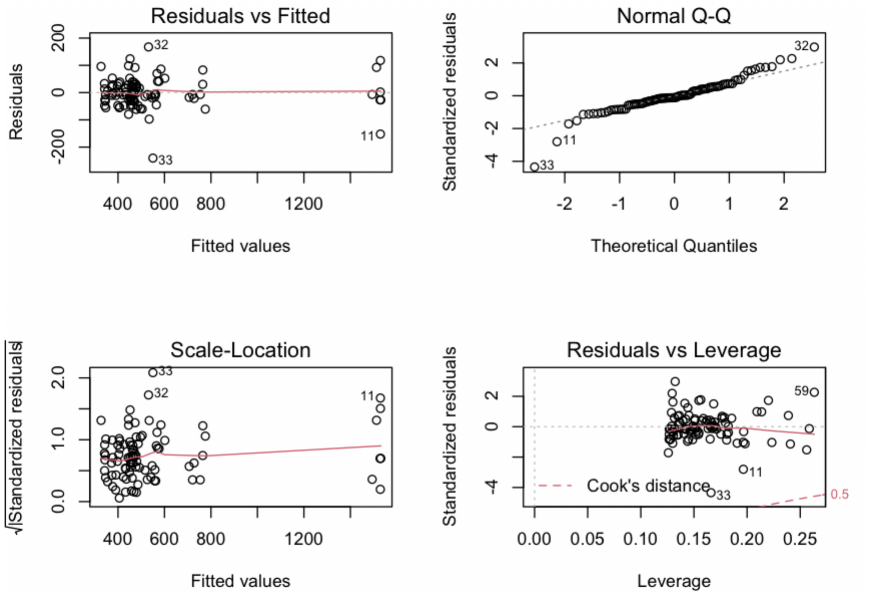

Figure 6: Default diagnostic plots for model in part (a).

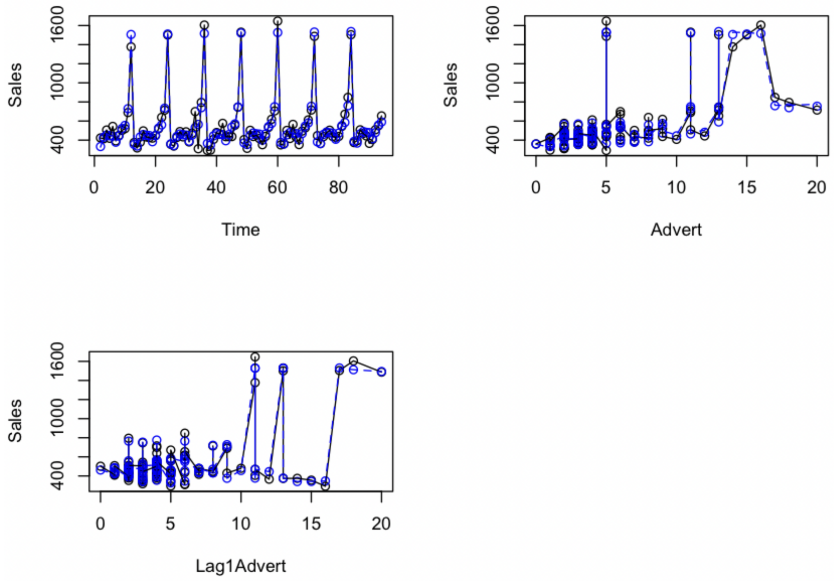

Figure 7: Marginal model plots for part (b).

I went ahead and tried a GLS model with AR(1) correlation. Here is (a subset of) the model summary output:

```
Generalized least squares fit by REML
  Model: Sales ~ Time + Month_2 + Month_3 + Month_4 + Month_5 + Month_6 +      Month_
  Data: dta
```

```
       AIC      BIC    logLik
  939.5951 979.6592 -452.7976

Correlation Structure: AR(1)
 Formula: ~Time
 Parameter estimate(s):
       Phi
-0.2326567

Coefficients:
                Value Std.Error   t-value p-value
(Intercept)  393.8984  38.19317 10.313320  0.0000
Time           0.4394   0.18868  2.328626  0.0225
Month_2      -45.9027  39.54279 -1.160836  0.2492
Month_3       24.1860  40.33222  0.599669  0.5505
Month_4       81.9327  33.50145  2.445645  0.0167
Month_5       58.6588  36.43199  1.610091  0.1114
Month_6       59.0634  35.94067  1.643357  0.1043
Month_7      -13.4383  37.22104 -0.361040  0.7190
Month_8       59.8914  35.48374  1.687854  0.0954
Month_9      127.9248  37.14578  3.443859  0.0009
Month_10     165.2170  35.89174  4.603205  0.0000
Month_11     321.0325  43.76953  7.334610  0.0000
Month_12    1153.8282  41.26176 27.963623  0.0000
Advert         2.5135   2.49796  1.006227  0.3174
Lag1Advert    -5.0511   2.58994 -1.950290  0.0547

Standardized residuals:
        Min          Q1         Med          Q3         Max
-3.94958774 -0.48372908 -0.06461214  0.48492736  2.77561225

Residual standard error: 60.45548
Degrees of freedom: 93 total; 78 residual
```

All code is shown at the end of this document.

```
####
#### Simulation to explore actual coverage probabilities of nominal 95% confidence
#### intervals for a mean response in the presence of AR(1) correlation but where we
#### assume there is no correlation and just use usual least squares.
####

n <- 50
M <- 1000
sig2_nu <- 0.5
rho_seq <- seq(from = 0.1, to = 0.9, by = 0.1)
beta_0 <- 0.5
beta_1 <- 1.5

set.seed(393844)

cvg <- matrix(NA, nrow = length(rho_seq), ncol = M)
for(j in 1:length(rho_seq)) {
  cat(j)
  for(m in 1:M) {
    ## Simulate error terms e_t.
    nu <- rnorm(n, 0, sqrt(sig2_nu))
    ee <- rep(NA, n)
    sig2_e <- sig2_nu / (1 - rho_seq[j] ^ 2)
    ee[1] <- rnorm(1, 0, sqrt(sig2_e))
    for(i in 2:n) {
      ee[i] <- rho_seq[j] * ee[i - 1] + nu[i]
    }

    ## Simulate x.
    x <- runif(n, 0, 1)

    ## Create response.
    y <- beta_0 + beta_1 * x + ee

    ## Fit least squares model.
    fit <- lm(y ~ x)

    ## Nominal 95% CI for mean response when x = 0.5.
    CI <- predict(fit, newdata = data.frame("x" = 0.5), interval = "confidence")[2:3]
    mean_resp <- beta_0 + beta_1 * 0.5
    cvg[j, m] <- as.numeric(mean_resp >= CI[1] & mean_resp <= CI[2])
  }
}

## Simulation-based coverage probability estimates.
rowMeans(cvg)

####
#### Exercise 9.2.
####
```

```
require(car)

dta <- read.delim("bookstore.txt")
n <- nrow(dta)

## Month of the year.
mo <- rep(NA, n)
for(i in 1:n) {
  if(all(dta[i, 5:15] == 0)) {
    mo[i] <- 1
  } else if(dta[i, 5] == 1) {
    mo[i] <- 2
  } else if(dta[i, 6] == 1) {
    mo[i] <- 3
  } else if(dta[i, 7] == 1) {
    mo[i] <- 4
  } else if(dta[i, 8] == 1) {
    mo[i] <- 5
  } else if(dta[i, 9] == 1) {
    mo[i] <- 6
  } else if(dta[i, 10] == 1) {
    mo[i] <- 7
  } else if(dta[i, 11] == 1) {
    mo[i] <- 8
  } else if(dta[i, 12] == 1) {
    mo[i] <- 9
  } else if(dta[i, 13] == 1) {
    mo[i] <- 10
  } else if(dta[i, 14] == 1) {
    mo[i] <- 11
  } else if(dta[i, 15] == 1) {
    mo[i] <- 12
  }
}

boxplot(dta[, 1] ~ mo, ylab = "Sales", xlab = "Month")

##
## (a)
##

fit_a <- lm(Sales ~ Time + Month_2 + Month_3 + Month_4 + Month_5 + Month_6 + Month_7 +
  Month_8 + Month_9 + Month_10 + Month_11 + Month_12, data = dta)
summary(fit_a)

## Some high leverage points but none appear to be "bad."
par(mfrow = c(2, 2))
plot(fit_a)

## Residuals and influence measures. A few potential outlier.
ee <- rstandard(fit_a)
```

```
ii_outlier <- (1:n)[abs(ee) > 2]
dta[ii_outlier, ]

par(mfrow = c(1, 1))
h_ii <- hatvalues(fit_a)
plot(1:n, h_ii, xlab = "Sample Index", ylab = "Hat Values")

## Marginal model plot for Time. Looks good.
mmps(fit_a)

plot(dta$Time, dta$Sales, xlab = "Time", ylab = "Sales")
points(dta$Time, fit_a$fitted.values, col = "blue")
lines(loess(dta$Sales ~ dta$Time))
lines(loess(fit_a$fitted.values ~ dta$Time), col = "blue", lty = 2)

## Strong correlation between lag 12 values of Sales. But then look at autocorrelation
## function of model residuals. After adjusting for month, weak correlation at lag 1.
plot(acf(dta$Sales))
acf_a <- acf(fit_a$residuals)
plot(acf_a)

##
## (b)
##

fit_b <- lm(Sales ~ Time + Month_2 + Month_3 + Month_4 + Month_5 + Month_6 + Month_7 +
  Month_8 + Month_9 + Month_10 + Month_11 + Month_12 + Advert + Lag1Advert, data = dta)
summary(fit_b)

## Some high leverage points but none appear to be "bad."
par(mfrow = c(2, 2))
plot(fit_b)

## Residuals and influence measures. Same potential outliers show up as for model (a).
ee <- rstandard(fit_b)
ii_outlier <- (1:n)[abs(ee) > 2]
dta[ii_outlier, ]

## Marginal model plots. Messy but the two curves generally match well for each variable.
par(mfrow = c(2, 2))

plot(dta$Time, dta$Sales, xlab = "Time", ylab = "Sales")
points(dta$Time, fit_a$fitted.values, col = "blue")
lines(loess(dta$Sales ~ dta$Time))
lines(loess(fit_b$fitted.values ~ dta$Time), col = "blue", lty = 2)

oo <- order(dta$Advert)
plot(dta$Advert, dta$Sales, xlab = "Advert", ylab = "Sales")
points(dta$Advert, fit_a$fitted.values, col = "blue")
lines(loess(dta$Sales[oo] ~ dta$Advert[oo]))
lines(loess(fit_b$fitted.values[oo] ~ dta$Advert[oo]), col = "blue", lty = 2)
```

```
oo <- order(dta$Lag1Advert)
plot(dta$Lag1Advert, dta$Sales, xlab = "Lag1Advert", ylab = "Sales")
points(dta$Lag1Advert, fit_b$fitted.values, col = "blue")
lines(loess(dta$Sales[oo] ~ dta$Lag1Advert[oo]))
lines(loess(fit_b$fitted.values[oo] ~ dta$Lag1Advert[oo]), col = "blue", lty = 2)

## Autocorrelation. Lag 1 correlation is barely significant.
acf_b <- acf(fit_b$residuals)
plot(acf_b)

## Try GLS with AR(1) correlation structure.
require(nlme)

fit_b_gls <- gls(Sales ~ Time + Month_2 + Month_3 + Month_4 + Month_5 + Month_6 +
  Month_7 + Month_8 + Month_9 + Month_10 + Month_11 + Month_12 + Advert + Lag1Advert,
  correlation = corAR1(form=~Time), data = dta)
summary(fit_b_gls)

## In both models, Advert and Lag1Advert are not quite statistically significant. In
## addition, there is not significant evidence that the more complex model (without
## correlation) fits better based on a nested model F test. I will prefer the simpler
## model from (a).
anova(fit_a, fit_b)
```