STAT 443
Assignment 2 Solution

1. Suppose that a future outcome that we want to predict is described by a random variable Y with the following probability mass function:

$$P(Y=-2)=p_{-2}, P(Y=-1)=p_{-1}, P(Y=0)=\frac{1}{3}, P(Y=1)=\frac{1}{6}, P(Y=2)=\frac{1}{6}.$$

If possible, give examples of distributions of the above form (that is, you have to specify the numbers p_{-2} and p_{-1}) for which:

(i) $E(Y) < \text{median}(Y)$

(ii) $E(Y) = \text{median}(Y)$

(iii) $E(Y) > \text{median}(Y)$

In each case briefly justify your answer. Based on your findings, which predictor (mean or median) you believe is more suitable when the distribution of Y is asymmetric and more extreme values may occur?

Solution:
In order to meet the definition of the p.m.f, p_{-2} and p_{-1} has to satisfy the following condition for (i) to (iii):

$$p_{-2}+p_{-1}=\frac{1}{3} \tag{1}$$

and $p_{-2}, p_{-1} \geqslant 0$. The expectation can be written as:

$$E(Y) = -2\cdot p_{-2} + -1\cdot p_{-1} + 0\cdot\frac{1}{3} + 1\cdot\frac{1}{6} + 2\cdot\frac{1}{6}$$
$$E(Y) = -2\cdot p_{-2} + -1\cdot p_{-1} + \frac{1}{2}$$

And the median of Y is 0 for any selection of p_{-2} and p_{-1}.
(i).

$$E(Y) = (-2)\cdot p_{-2} + (-1)\cdot p_{-1} + \frac{1}{2} < 0$$
$$2\cdot p_{-2} + 1\cdot p_{-1} > \frac{1}{2}$$

For example, we can set $p_{-2} = \frac{1}{4}$ and $p_{-1} = \frac{1}{12}$.
(ii).

$$E(Y) = -2 \cdot p_{-2} + -1 \cdot p_{-1} + \frac{1}{2} = 0$$
$$2 \cdot p_{-2} + 1 \cdot p_{-1} = \frac{1}{2}$$

Because p_{-2} and p_{-1} need to satisfy the equation (1), we know that $p_{-2} = p_{-1} = \frac{1}{6}$.

(iii).

$$E(Y) = -2 \cdot p_{-2} + -1 \cdot p_{-1} + \frac{1}{2} > 0$$
$$2 \cdot p_{-2} + 1 \cdot p_{-1} < \frac{1}{2}$$

When Y is asymmetric and more extreme values may occur, the expectation is a better choice compared to the median since the median will not change no matter how the tail behavior of Y changes. It can't capture the asymmetric property of Y.

2. Suppose that a random variable Z is defined in the following way:

$$Z := Z_1 + Z_2$$

where Z_1 and Z_2 are independent and $Z_1 \sim N(0, 9/10)$, $Z_2 \sim N(0, 1/10)$.

(i) find the best prediction of Z in the MS-sense. What is the variance of the prediction error? Show your calculations.

(ii) the same as in (i) above but now you can assume that we know that $Z_1 = 0.1$.

Solution:
(i). Under the MS-sense, the best prediction of Z is $E[Z] = 0$ when we have no additional information about Z. So variance of the prediction error can be calculated in the following way:

$$Var[Z - E[Z]] = Var[Z] = Var[Z_1] + Var[Z_2] = 1$$

(ii). When we know that $Z_1 = 0.1$, then the best predictor of Z is $E[Z|Z_1 = 0.1] = 0.1$. The variance of the prediction error is:

$$Var[Z - E[Z|Z_1 = 0.1]|Z_1 = 0.1] = Var[0.1 + Z_2 - 0.1|Z_1 = 0.1]$$

Due to the fact that Z_1 and Z_2 are independent, we can have:

$$Var[0.1 + Z_2 - 0.1|Z_1 = 0.1] = Var[Z_2] = 0.1$$

3. Suppose that a random variable Y follows a discrete distribution according to

$$P(Y = y_i) = p_i,\ i = 1, 2, \ldots, K,$$

where we assume that the probabilities $p_i, \ldots, p_K$ are all different.
We want to find the best predictor of Y using the zero-one loss function L_{01} defined in class.
Show that the value f that minimizes

$$E[L_{01}(Y, f)]$$

is given by

$$f = Arg\{\max_{i \in 1,\ldots,K} p_i\}$$

Solution:
Since $E[L_{01}(Y, f)] = P(Y \neq f)$, minimization of the expected value is equivalent to minimization $P(Y \neq f)$. Also we know that:

$$P(Y \neq f) = 1 - P(Y = f).$$

So the problem is equivalent to maximization of $P(Y = f)$ with respect to f. Then the best predictor for Y is the value of Y which has the highest probability. i.e.:

$$\hat{f} = Arg \max_{y \in \Omega} P(y)$$

4. In order to predict a random variable Y we want to use another variable X. We assume that

$$P(Y = y_i|X = x) = p_i(x),\ i = 1, \ldots, K,$$

where for each x the probabilities $p_1(x), \ldots, p_K(x)$ are all different. We also assume that X is a discrete random variable wit the following probability mass function

$$P(X = x) = m(x),\ x \in S_X := 1, \ldots, M.$$

When we use the zero-one loss function, the best predictor of Y in terms of X is defined as the function f that minimizes

$$E[L_{01}(Y, f(X))]. \tag{2}$$

Show that f that minimizes (2) is given by the Bayes classifier

$$\hat{f}(x) = Arg\{\max_{j \in 1,\ldots,K} p_j(x)\}$$

Hint: use the law of total expectation

$$E[L_{01}(Y, f(X))] = \sum_{x \in S_X} E[L_{01}(Y, f(x))|X = x]m(x),$$

and the Problem 3.

Solution:
Because of the hint and property mentioned in Problem 3, we have $E[L_{01}(Y, f(X))] = \sum_{x \in S_X} E[L_{01}(Y, f(x))|X = x]m(x) = \sum_{x \in S_X} P(Y \neq f(x)|X = x)m(x)$.
Now we wish solve the following optimization problem:

$$Arg \min_f \sum_{x \in S_X} P(Y \neq f(x)|X = x)m(x).$$

We can rewrite the inside part as:

$$\begin{aligned}
\sum_{x \in S_X} P(Y \neq f(x)|X = x)m(x) &= \sum_{x \in S_X} [1 - P(Y = f(x)|X = x)]m(x) \\
&= \sum_{x \in S_X} m(x) - \sum_{x \in S_X} P(Y = f(x)|X = x)m(x) \\
&= 1 - \sum_{x \in S_X} P(Y = f(x)|X = x)m(x)
\end{aligned}$$

So we wish to find the best predictor $f(x)$ which:

$$\max_{f(x)} \sum_{x \in S_X} P(Y = f(x)|X = x)m(x)$$

Since we can choose f independently for each x (there is no constraint on the values of $f(x_1), f(x_2), \dots$), we can put the max inside the summation.

$$\sum_{x \in S_X} m(x) \max_{f(x)} P(Y = f(x)|X = x)$$

Then we can conclude that the best predictor $\hat{f}(x)$ is the value of Y which has the greatest probability given x:

$$\hat{f}(x) = Arg\{\max_{j \in 1,\dots,K} p_j(x)\}.$$

5. Obtain the historical data from **mp3_y.txt**, which consists of the total number Y_t of a new type of MP3 player sold each month at a particular chain stores for 72 consecutive months $t = 1, \ldots, 72$. The goal is to use the historical data $Y_1, \ldots, Y_{72}$ to predict player sales in future months.

(a) Provide a scatter plot of player sales vs. time, and use this to develop a linear model, using *only an appropriate subset* of the historical data, which would be useful for forecasting future sales. Give reasons for your choice of subset.

(b) Perform a least squares fit on your model (using either **lm** or **lsfit**) and provide a summary of the output. Provide a scatter plot of the data (restricted to your chosen subset) with the least squares (regression) line superimposed.

(c) Do *residual analysis*: plot the residual vs. time, provide a QQ-polt to check for normality, and an autocorrelation plot (acf) to look for correlation between the residuals. Discuss your results: do i.i.d. $N(0, \sigma^2)$ errors seem reasonable here?

(d) Discuss the overall results of your regression analysis *assuming the OLSA holds* (even if you rejected them in the previous question). Does the model appear useful in explaining the average changes in player sales over the subset of time that you are using? What is the predictive power of the model?

(e) Again assuming that OLSA holds, write down your parameter estimates with 95% confidence intervals.

(f) Based on OLS assumptions, write down the general prediction equation (i.e. the estimated mean function) and predict player sales for months 73 and 78. Write down a 95% prediction interval for your predictions for the months 73 and 78.

Solution:
For practice.

6. The monthly revenue of a telephone company for 11 consecutive years (beginning January 1986 and ending December 1996) is contained in the file **telephone_y.txt**.

(a) Provide a plot of the data vs. time, and develop a suitable multivariate causal model using all the data, for forecasting the time series. You may wish to transform the data first and to include both the trend and seasonal components in your model. We will call this Model I. (If you transform your data, include a plot of the transformed data.)

(b) Perform least squares on Model I (using either **lm** or **lsfit**). You may wish to "borrow" some of the columns of the reduced design matrix **hotel_x** we used in class to build your design matrix. Provide a copy of the output from your procedure.

(c) *Assuming* the OLS assumptions are true, discuss the overall usefulness of Model I in (i) explaining and (ii) prediction the behavior of telephone revenues (that is, discuss the F-test and R^2-values). Discuss the usefulness of the individual variables. Are there any variables you might consider dropping from the model? (why or why not?)

(d) Perform a *residual analysis* to assess the validity of the OLS assumptions. Turn in the residual plot ϵ_t^* vs. t, the sample **acf** plot, and the sample QQ-plot. Discuss each plot individually. What is your conclusion?

(e) Another possibility for building a forecast model for this time series is to throw away some of the initial historical data. Construct a multivariable causal model using *only* the last 5 years of data, which includes a trend component that is *quadratic* in t (why?) as well as seasonal components (it should not be necessary to transform the data). *Explicitly write down this model, defining all parameters and variables.* We will call this Model II.

(f) Carry out part (b) for Model II.

(g) Carry out part (c) for Model II.

(h) Carry out part (d) for Model II.

(i) Assuming the OLS assumption are true for Model II, construct 95% prediction intervals for telephone revenues for March of 1997 and 1998.

(j) The data for 1997 is store in **telephone_future.txt**. Compute the SOS of forecast errors for each of Model I and Model II (for definitions, see the first set of slides we used in class). Which model performs best?

Solution:

(a). We use the log transformation to transform the data. So the model we use is:

$$log(Y_t) = \beta_0 + \beta_1 \cdot t + \sum_{t=2}^{12} \beta_t \cdot X_{t,k} + \epsilon_t$$

Where $X_{t,k}$ are indicator variables which equal to 1 if t correponds to month k and 0 otherwise.

(b). The output see below.

(c). Based on the summary provided by R, the R^2 values indicates that the model explains around 99% volatility of the data. The T-test and F-test suggest that we should reject the null hypothesis that the parameters are 0 under 95% and 99% confidence level. So we should not drop any variables in the model.

(d).The output and code see below. Based on the residual vs. time plot, we can find a pattern for the residuals. The QQ-plot also suggest the violation of the normal assumption. The acf plot indicate that the independence assumption for the residuals are violated. In conclusion, the model we used is not good enough to meet the OLSA.

(e). When we decompose the data into trend and seasonal component, the trend looks like a quadratic curve. It suggests that the trend can be expressed in terms of a quadratic function of t. So the model we use is:

$$Y_t = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2 + \sum_{t=3}^{13} \beta_t \cdot X_{t,k} + \epsilon_t$$

Where $X_{t,k}$ are indicator variables which equal to 1 if t correponds to month k and 0 otherwise.

(f). The output is provided below.

(g). Based on the summary provided by R, the R^2 values indicates that the model explains around 99.5% volatility of the data. The T-test and F-test suggest that we should reject the null hypothesis that the parameters are 0 under 95% and 99% confidence level. So we should not drop any variables in the model.

(h). The output and code is provided below. Based on the residual vs. time plot, we can't find a pattern for the residuals. The QQ-plot is improved a lot compared to the previous one which suggests that the normal distribution is a good approximation for the residuals. The acf plot indicate that the independence assumption for the residuals is satisfied. In conclusion, the model we used meets the OLSA.

(i). The 95% prediction interval for March of 1997 is $(63644.07, 66454.89)$ and the 95% prediction interval for March of 1998 is $(72419.49, 75760.3)$.

(j). The SOS of Model I for the upcoming 12 months is

$$SOS_{Model\ I} = 322448346$$

The SOS of Model II for the upcoming 12 months is

$$SOS_{Model\ II} = 1116618.129$$

Based on the values of SOS for Model I and II, we can concluded that the Model II performs better than Model I in predicting the future values.

R code:

Listing 1: R code for Problem 6

```
##### (a) #####
data <- scan("telephone_y.txt") #load data
par(mfrow=c(1,2))
plot(data, type = "l", main = "Original data",
ylab = "revenue", xlab = "time") #original data vs time plot
plot(log(data), type = "l", main = "log data",
ylab = "revenue", xlab = "time") #log transformation data vs time plot


plot(decompose(ts(log(data), frequency = 12),type = "additive"))
```

```
##### (b) #####
x <- matrix(scan("reduce_matrix.txt"),ncol =12,byrow = TRUE)
x1 <- x[,1]
x2 <- x[,2]
x3 <- x[,3]
x4 <- x[,4]
x5 <- x[,5]
x6 <- x[,6]
x7 <- x[,7]
x8 <- x[,8]
x9 <- x[,9]
x10 <- x[,10]
x11 <- x[,11]
x12 <- x[,12]
data.ls <- lm(log(data) ~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12)
summary(data.ls)
#
# Call:
# lm(formula = log(data) ~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12)
#
# Residuals:
#     Min       1Q   Median       3Q      Max
# -0.069626 -0.022735  0.003128  0.024598  0.085276
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) 1.004e+01  1.007e-02 997.420  < 2e-16 ***
# x1          6.465e-03  6.947e-05  93.062  < 2e-16 ***
# x2          1.617e-01  1.291e-02  12.522  < 2e-16 ***
# x3          1.126e-01  1.291e-02   8.720 1.98e-14 ***
# x4          1.899e-01  1.292e-02  14.706  < 2e-16 ***
# x5          1.409e-01  1.292e-02  10.904  < 2e-16 ***
# x6          1.159e-01  1.292e-02   8.972 5.05e-15 ***
# x7          4.717e-02  1.292e-02   3.650  0.00039 ***
# x8          1.001e-01  1.292e-02   7.744 3.53e-12 ***
# x9          1.021e-01  1.293e-02   7.897 1.58e-12 ***
# x10         1.313e-01  1.293e-02  10.154  < 2e-16 ***
```

```
# x11          1.583e-01  1.293e-02  12.244  < 2e-16 ***
# x12          7.432e-02  1.294e-02   5.745 7.20e-08 ***
# ---
# Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
#
# Residual standard error: 0.03029 on 119 degrees of freedom
# Multiple R-squared:  0.9871,  Adjusted R-squared:  0.9858
# F-statistic: 759.8 on 12 and 119 DF,  p-value: < 2.2e-16


#Alternative R commands that can be used to solve Q6(b):
t = 1:length(data)
data.log.ts <- ts(log(data), start = c(1986,1),
                  end = c(1996,12), frequency = 12)
month = factor(cycle(data.log.ts))
data.ls <- lm(data.log.ts~t+month)
summary(data.ls)


##### (c) #####


##### (d) #####
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
plot(data.ls$residuals, xlab = "time", ylab = "residuals")
 #plot residuals vs time
qqnorm(data.ls$residuals) #QQ-plot of residual
qqline(data.ls$residuals)
acf(data.ls$residuals, main = "ACF of residuals")# autocorrelation plot


##### (e) #####
sub_data <- data[73:132] #load data
par(mfrow=c(1,1))
plot(sub_data, type = "l", main = "Original data",
     xlab = "time", ylab = "revene")
     #original data vs time plot
#original data vs time plot
plot(decompose(ts(sub_data, frequency = 12),type = "additive"))
```

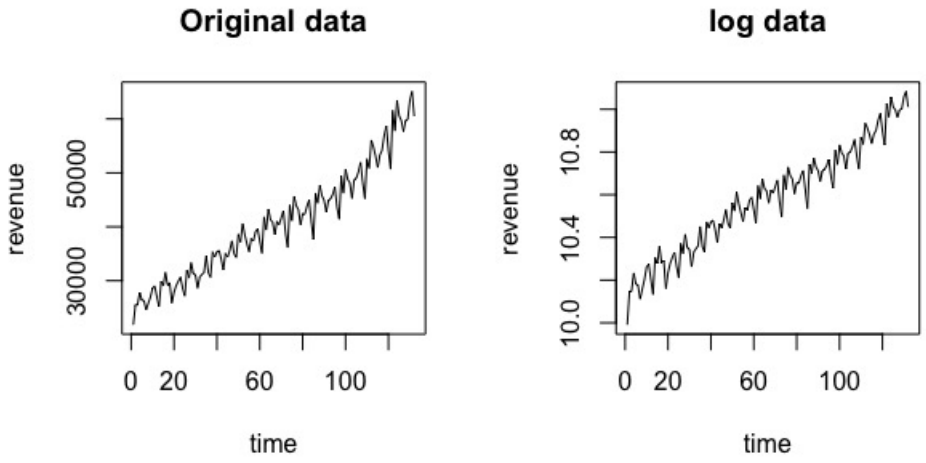
```
##### (f) #####
sub_x <- x[1:60,1:12]
sub_x <- matrix(cbind(sub_x[1:60,1],c((1:60)^2),
sub_x[1:60,2:12]),ncol = 13)
sub_x1 <- sub_x[,1]
sub_x2 <- sub_x[,2]
sub_x3 <- sub_x[,3]
sub_x4 <- sub_x[,4]
sub_x5 <- sub_x[,5]
sub_x6 <- sub_x[,6]
sub_x7 <- sub_x[,7]
sub_x8 <- sub_x[,8]
sub_x9 <- sub_x[,9]
sub_x10 <- sub_x[,10]
sub_x11 <- sub_x[,11]
sub_x12 <- sub_x[,12]
sub_x13 <- sub_x[,13]
sub_data <- matrix(sub_data, ncol = 1)
sub_data.ls <- lm(sub_data ~ sub_x1+sub_x2+sub_x3+sub_x4+sub_x5+sub_x6
+sub_x7+sub_x8+sub_x9+sub_x10+sub_x11+sub_x12+sub_x13)
summary(sub_data.ls)

# Call:
#   lm(formula = sub_data ~ sub_x)
#
# Residuals:
#   Min        1Q    Median        3Q       Max
# -1299.09   -373.13    -48.29    325.96   1644.66
#
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
# (Intercept) 3.616e+04  3.345e+02 108.108  < 2e-16 ***
#   sub_x1       5.873e+01  1.787e+01   3.287  0.00194 **
#   sub_x2       5.034e+00  2.835e-01  17.752  < 2e-16 ***
#   sub_x3       8.218e+03  3.722e+02  22.079  < 2e-16 ***
#   sub_x4       5.208e+03  3.723e+02  13.987  < 2e-16 ***
#   sub_x5       9.498e+03  3.725e+02  25.498  < 2e-16 ***
#   sub_x6       7.226e+03  3.727e+02  19.389  < 2e-16 ***
#   sub_x7       5.884e+03  3.730e+02  15.776  < 2e-16 ***
#   sub_x8       3.131e+03  3.732e+02   8.387 8.00e-11 ***
```

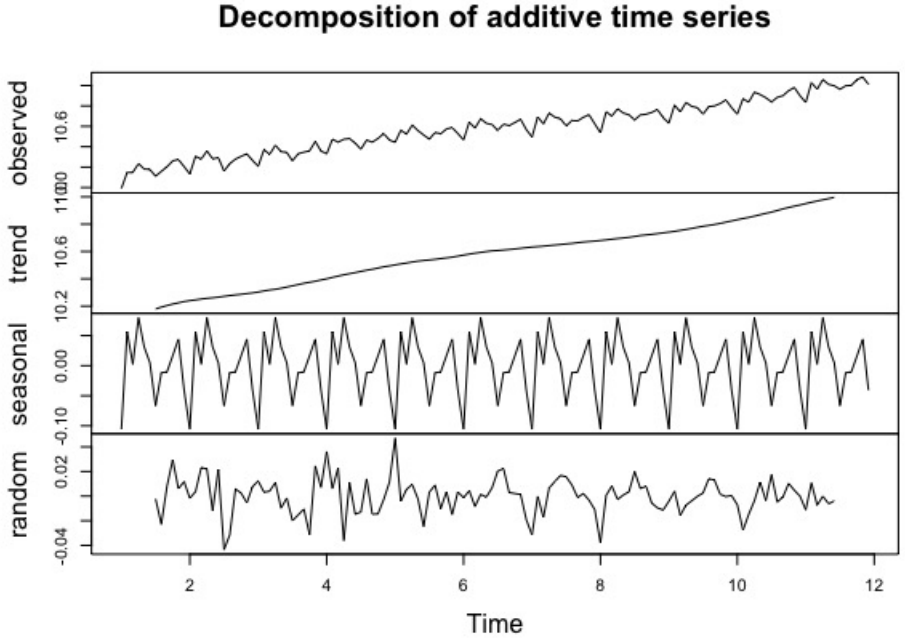
```
#    sub_x9       5.135e+03  3.736e+02  13.745  < 2e-16 ***
#    sub_x10      5.115e+03  3.740e+02  13.677  < 2e-16 ***
#    sub_x11      6.643e+03  3.744e+02  17.744  < 2e-16 ***
#    sub_x12      7.776e+03  3.749e+02  20.742  < 2e-16 ***
#    sub_x13      3.273e+03  3.754e+02   8.718 2.64e-11 ***
#    ---
#    Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
#
# Residual standard error: 588.5 on 46 degrees of freedom
# Multiple R-squared:  0.9945,  Adjusted R-squared:  0.993
# F-statistic: 644.6 on 13 and 46 DF,  p-value: < 2.2e-16


#Alternative R commands that can be used to solve Q6(f):
t = 1:length(sub_data)
t2 = c(1:length(sub_data))^2
sub_data.ts <- ts(sub_data, start = c(1986,1),
                  end = c(1996,12), frequency = 12)
month = factor(cycle(sub_data.ts))
sub_data.ls <- lm(sub_data.ts~t+t2+month)
summary(sub_data.ls)




##### (g) #####


##### (h) #####
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
plot(sub_data.ls$residuals, xlab = "time", ylab = "residuals")
       #plot residuals vs time
qqnorm(sub_data.ls$residuals) #QQ-plot of residual
qqline(sub_data.ls$residuals)
acf(sub_data.ls$residuals, main = "ACF of residuals")
       # autocorrelation plot


##### (i) #####
new_data_97_3 <- data.frame(sub_x1=63,sub_x2=63^2,sub_x3=0,sub_x4=1,
```

```
                                sub_x5=0,sub_x6=0,sub_x7=0,sub_x8=0,sub_x9=0,
                                sub_x10=0,sub_x11=0,sub_x12=0,sub_x13=0)
new_data_98_3 <- data.frame(sub_x1=75,sub_x2=75^2,sub_x3=0,sub_x4=1,
                                sub_x5=0,sub_x6=0,sub_x7=0,
                                sub_x8=0,sub_x9=0,sub_x10=0,sub_x11=0,
                                sub_x12=0,sub_x13=0)
predict(sub_data.ls, new_data_97_3,interval = "prediction",level = 0.95)
predict(sub_data.ls, new_data_98_3,interval = "prediction",level = 0.95)


#Alternative R commands that can be used to solve Q6(i):
new_data <- data.frame(t=length(sub_data)+c(3,15),
                       t2=()length(sub_data)+c(3,15))^2,
                       month = factor(c(3,3)))
predict(sub_data.ls, new_data,interval = "prediction",level = 0.95)



##### (j) #####
new_x1_1 <- c(133:144)
new_x1 <- c(61:72)
new_x2 <- c(61:72)^2
new_x3 <- c(0,1,0,0,0,0,0,0,0,0,0,0)
new_x4 <- c(0,0,1,0,0,0,0,0,0,0,0,0)
new_x5 <- c(0,0,0,1,0,0,0,0,0,0,0,0)
new_x6 <- c(0,0,0,0,1,0,0,0,0,0,0,0)
new_x7 <- c(0,0,0,0,0,1,0,0,0,0,0,0)
new_x8 <- c(0,0,0,0,0,0,1,0,0,0,0,0)
new_x9 <- c(0,0,0,0,0,0,0,1,0,0,0,0)
new_x10 <- c(0,0,0,0,0,0,0,0,1,0,0,0)
new_x11 <- c(0,0,0,0,0,0,0,0,0,1,0,0)
new_x12 <- c(0,0,0,0,0,0,0,0,0,0,1,0)
new_x13 <- c(0,0,0,0,0,0,0,0,0,0,0,1)

new_data_model_1 <- data.frame(x1=new_x1_1,x2=new_x3,x3=new_x4,x4=new_x5,
                               x5=new_x6,x6=new_x7,x7=new_x8,x8=new_x9,
                               x9=new_x10,x10=new_x11,x11=new_x12,
                               x12=new_x13)
new_data_model_2 <- data.frame(sub_x1=new_x1,sub_x2=new_x2,sub_x3=new_x3,
                               sub_x4=new_x4,sub_x5=new_x5,sub_x6=new_x6,
```

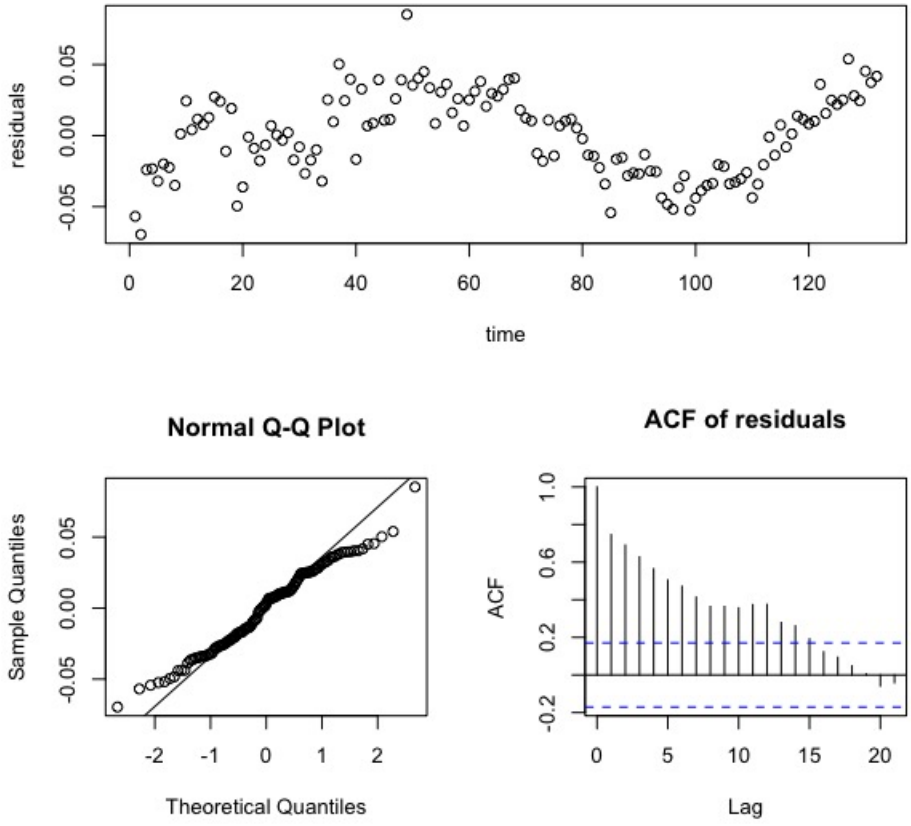
```
                                    sub_x7=new_x7, sub_x8=new_x8, sub_x9=new_x9,
                                    sub_x10=new_x10, sub_x11=new_x11,
                                    sub_x12=new_x12, sub_x13=new_x13)
model_1_pred <- exp(predict(data.ls, new_data_model_1))
model_2_pred <- predict(sub_data.ls, new_data_model_2)
future_data <- scan("telephone_future.txt") #load data

model_1_pred_err <- sum((future_data-model_1_pred)^2)
model_2_pred_err <- sum((future_data-model_2_pred)^2)
```

Plots:
(a).

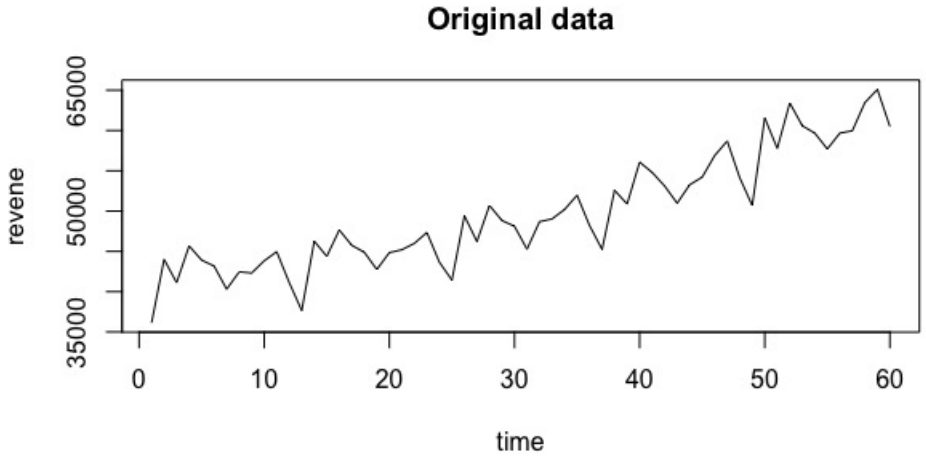


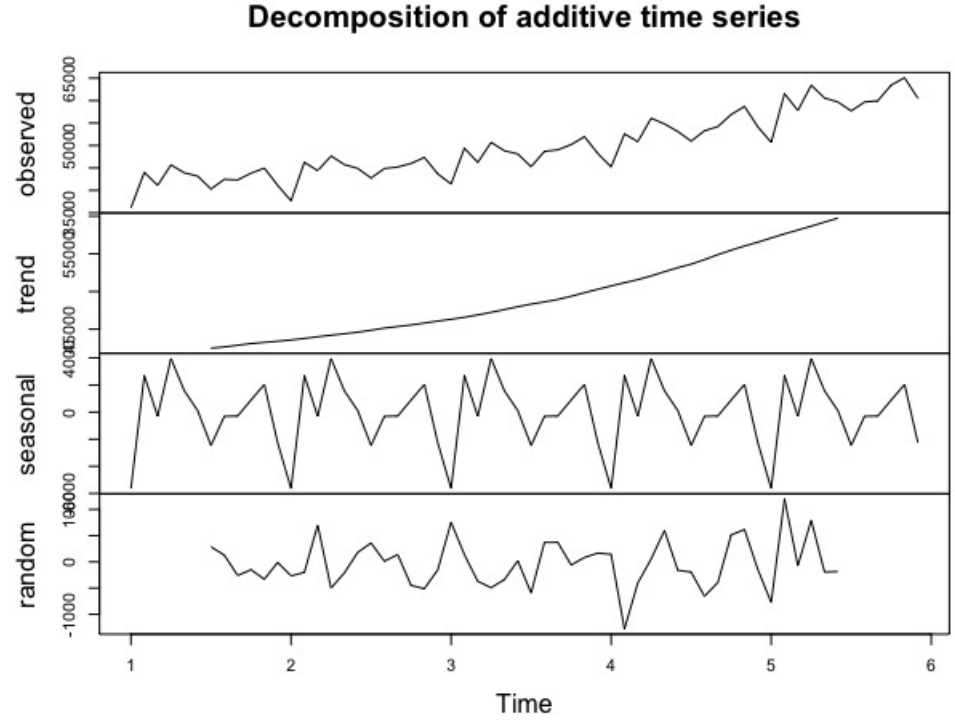

Decomposition of additive time series
observed
trend
seasonal
random
10.6
11.0
10.6
10.2
0.00
-0.10
0.02
-0.04
2
4
6
8
10
12
Time

(d).

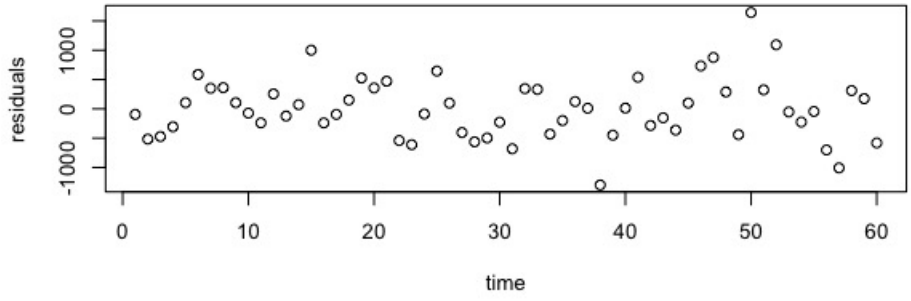

residuals
-0.05
0.00
0.05
0
20
40
60
80
100
120
time
Normal Q-Q Plot
Sample Quantiles
-0.05
0.00
0.05
-2
-1
0
1
2
Theoretical Quantiles
ACF of residuals
ACF
-0.2
0.2
0.6
1.0
0
5
10
15
20
Lag

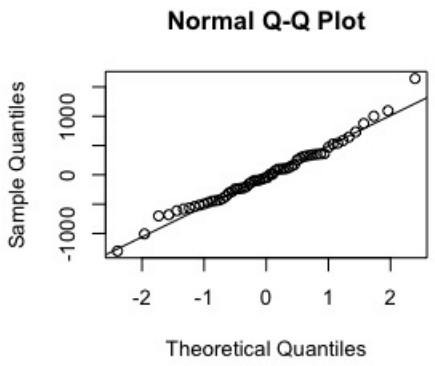
(e).

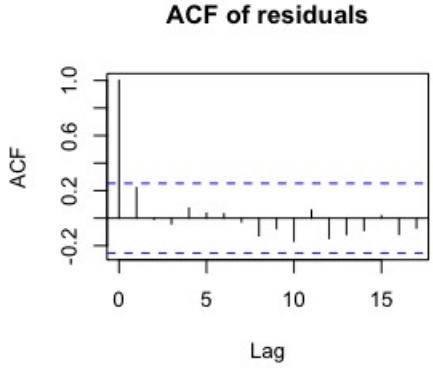

Original data
revene
35000
50000
65000
0
10
20
30
40
50
60
time


Decomposition of additive time series
observed
trend
seasonal
random
Time

(h).


residuals
-1000
0
1000
0
10
20
30
40
50
60
time


Normal Q-Q Plot
Sample Quantiles
-1000
0
1000
-2
-1
0
1
2
Theoretical Quantiles


ACF of residuals
ACF
-0.2
0.2
0.6
1.0
0
5
10
15
Lag