Analytics 512: Solution Key for Homework 1

02/15/19

```
library(ISLR)
library(MASS)
set.seed(9955)
```

Through an oversight, problem 8 in chapter 3 was assigned twice. It will only count once (5 points). Then 15% be added to everybody's score for this problem set at the end of the semester.

ISLR Ch. 3 #3

The regression model can be written as

$$\begin{aligned} Y &= 50 + 20 \cdot GPA + .07 \cdot IQ + 35 \cdot gender + .01 \cdot GPA \cdot IQ - 10 \cdot gender \cdot GPA \\ &= 50 + 20 \cdot GPA + .07 \cdot IQ + (35 - 10 \cdot GPA) \cdot gender + .01 \cdot IQ \cdot GPA\,. \end{aligned}$$

(a) i., ii., iii., iv. This depends on the coefficient $35 - 10 \cdot GPA$. If this is positive, then females earn more on average, if it's negative, then males earn more on average. Therefore it is negative if $GPA > 3.5$ and males earn more on average in this case. Thus iii is true and iv is false.

(b) The predicted salary is 137.1.

(c) This is false. The magnitude of the interaction effect is between 2 and 4 for GPAs between 2 and 4 and typical IQ's $\approx$ 100.

ISLR Ch. 3 #10abc

Work with the Carseats data.

```
head(Carseats)
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age
## 1  9.50       138     73          11        276   120       Bad  42
## 2 11.22       111     48          16        260    83      Good  65
## 3 10.06       113     35          10        269    80    Medium  59
## 4  7.40       117    100           4        466    97    Medium  55
## 5  4.15       141     64           3        340   128       Bad  38
## 6 10.81       124    113          13        501    72       Bad  78
##   Education Urban  US
## 1        17   Yes Yes
## 2        10   Yes Yes
## 3        12   Yes Yes
## 4        14   Yes Yes
## 5        13   Yes  No
## 6        16    No Yes
```

(a) Predict Sales using Price, Urban, and US.

```
fit.10 <- lm(Sales ~ Price + Urban + US, data = Carseats)
summary(fit.10)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

(b) Interpretation:

- Baseline sales are ≈ 13,000.
- For each increase in price by $1, sales are predicted to decrease by about 54.
- If the store is in an urban location, sales are predicted to be lower by about 23. *However, this coefficient is not statistically different from 0.*
- If this store is in the US, Sales are predicted to be higher by about 1,200.

(c) The model is

$$Sales = 13.043 - 0.054 * price - 0.22 * Urban + 1.2 * US$$

where Sales are measured in 1,000, price is measured in $, Urban = 1 for stores that are in an urban location and 0 otherwise, and US = 1 if a store is located in the US and 0 otherwise.

ISLR Ch. 3 #15ab

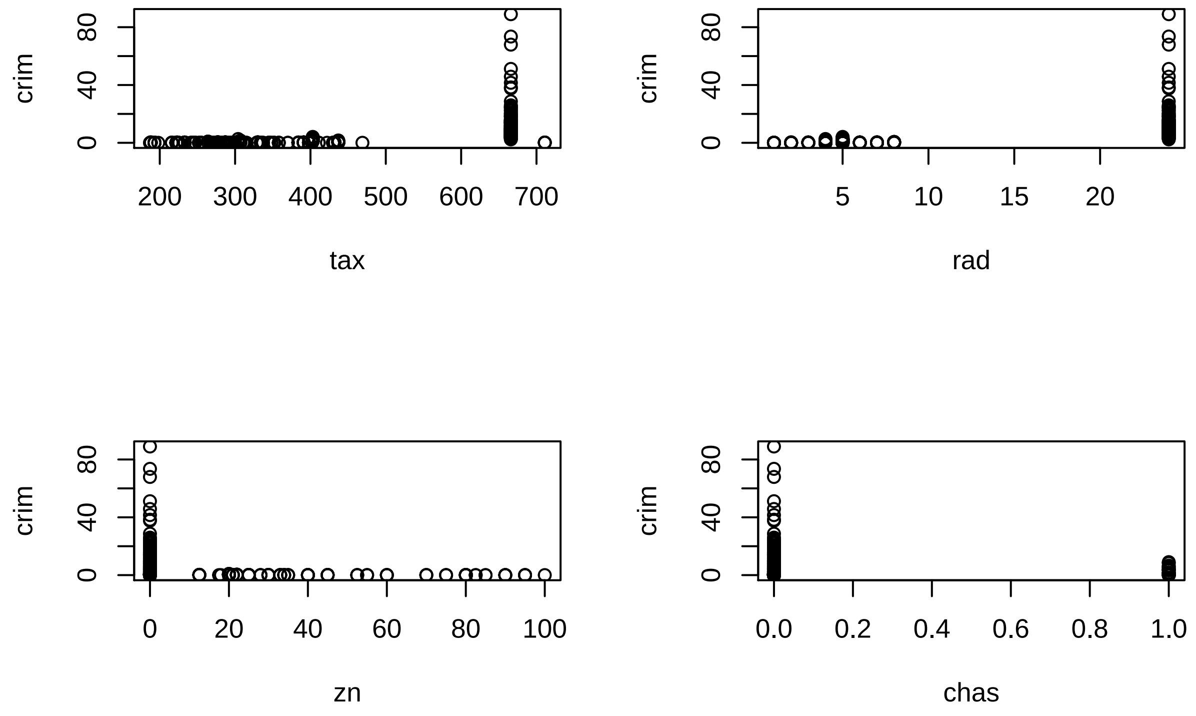
There are 13 possible predictors in the data set. We fit simple linear regression models for all 13 and record the corresponding p values in a data frame.

```
df.15 <- data.frame(predictors = names(Boston)[2:14], p.value = NA)
for (j in 1:13){
  my.lm <- lm(Boston$crim ~ Boston[,1+j])
  df.15$p.value[j] <- summary(my.lm)$coefficients[2,4]
}
df.15
```

```
##    predictors      p.value
## 1          zn 5.506472e-06
## 2       indus 1.450349e-21
## 3        chas 2.094345e-01
## 4         nox 3.751739e-23
## 5          rm 6.346703e-07
## 6         age 2.854869e-16
## 7         dis 8.519949e-19
## 8         rad 2.693844e-56
## 9         tax 2.357127e-47
## 10    ptratio 2.942922e-11
## 11      black 2.487274e-19
## 12      lstat 2.654277e-27
## 13       medv 1.173987e-19
```

All predictors, taken individually, are significant. Here are plots to illustrate this for the most significant and the least significant predictors. *It seems that a linear model is not appropriate for any of these relations.*

```
par(mfrow = c(2,2))
plot(crim ~ tax, data = Boston)
plot(crim ~ rad, data = Boston)
plot(crim ~ zn, data = Boston)
plot(crim ~ chas, data = Boston)
```



b) Here's a multiple regression model.

```
lm.all <- lm(crim ~ ., data = Boston)
summary(lm.all)
```

```
##
## Call:
## lm(formula = crim ~ ., data = Boston)
```

```
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -9.924 -2.120 -0.353  1.019 75.051 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  17.033228   7.234903   2.354 0.018949 *  
## zn            0.044855   0.018734   2.394 0.017025 *  
## indus        -0.063855   0.083407  -0.766 0.444294    
## chas         -0.749134   1.180147  -0.635 0.525867    
## nox         -10.313535   5.275536  -1.955 0.051152 .  
## rm            0.430131   0.612830   0.702 0.483089    
## age           0.001452   0.017925   0.081 0.935488    
## dis          -0.987176   0.281817  -3.503 0.000502 ***
## rad           0.588209   0.088049   6.680 6.46e-11 ***
## tax          -0.003780   0.005156  -0.733 0.463793    
## ptratio      -0.271081   0.186450  -1.454 0.146611    
## black        -0.007538   0.003673  -2.052 0.040702 *  
## lstat         0.126211   0.075725   1.667 0.096208 .  
## medv         -0.198887   0.060516  -3.287 0.001087 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.439 on 492 degrees of freedom
## Multiple R-squared:  0.454,  Adjusted R-squared:  0.4396 
## F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

There are only five significant predictors. These are **zn, dis, rad, black, medv**.

Xtra #10

Consider the built-in data set **cars** (see problem 1). Fit a linear regression model to the data. What are the three observations with the largest standardized residuals (in magnitude)? What are their leverages? Where are the three observation with the largest leverage? What are their standardized residuals? *Use the* ***R*** *function* ***influence.lm****.*

Solution.

Fit a linear model, extract the residuals r_i and leverages h_{ii} , and compute the standardized residuals. These are defined as $\frac{r_i}{s\sqrt{1-h_{ii}}}$ where s is the residual standard error.

```
lm.10 <- lm(dist ~ speed, data = cars)
res <- summary(lm.10)$residuals # residuals
lev <- influence(lm.10)$hat # leverages
s = summary(lm.10)$sigma # res. standard error
std.res <- res/s/sqrt(1-lev)
```

Now we can find the observations with the three largest standardized residuals and display their leverages.

```
order(abs(std.res),decreasing = T)[1:3] -> std.res.large
std.res.large
```

```
## [1] 49 23 35
```

```
lev[std.res.large]
```

```
##         49         23         35
## 0.07398540 0.02143066 0.02493431
```

The average leverage is $2/50 = 0.04$. Two of these observations don't have a large leverage.

Here are the three observations with the largest leverages and their standardized residuals.

```
order(lev,decreasing = T)[1:3] -> lev.large
lev.large
```

```
## [1]  1  2 50
```

```
std.res[lev.large]
```

```
##         1         2        50
## 0.2660415 0.8189327 0.2905345
```

The largest leverages occur at the extremes of the range of the predictor. Their standardized residuals are not large, since the standard deviation of the standardized residuals is near 1.

ISLR Ch. 3 #8 (5)

(a) Make a regression model and look at the summary.

```
fita = lm(mpg ~ horsepower, data = Auto)
summary(fita)
```

```
## 
## Call:
## lm(formula = mpg ~ horsepower, data = Auto)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -13.5710  -3.2592  -0.3435   2.7630  16.9240 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049 
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

The F statistic is very large, so there is a relationship between the predictor and the response. Since $R^2 \approx 0.6$, the relationship is fairly strong. Since the estimated slope is negative, so is the relationship.

Predicted mpg, confidence interval and prediction interval:

```
x = round(predict(fita,data.frame(horsepower = 98),interval="confidence"),2);
y = round(predict(fita,data.frame(horsepower = 98),interval="predict"),2)
x
```

```
##     fit   lwr   upr
## 1 24.47 23.97 24.96
```

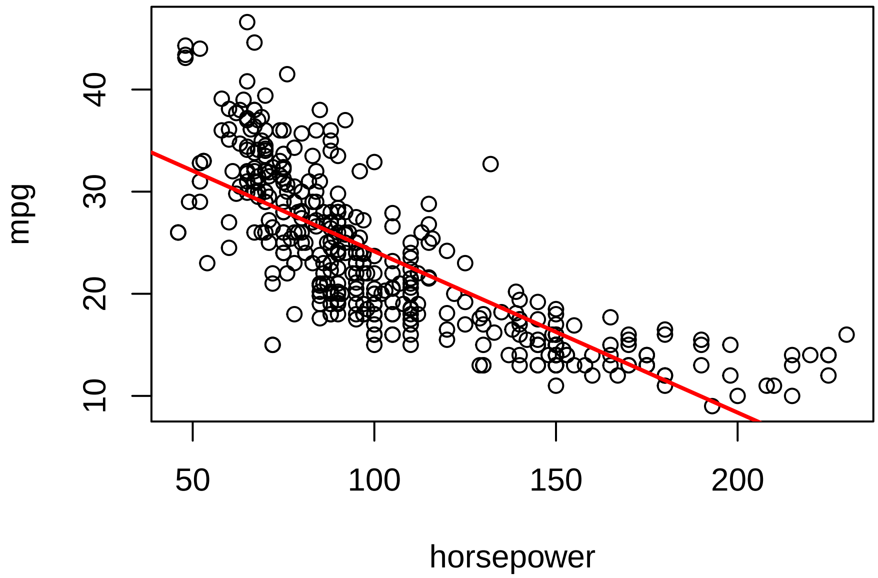
```
y
```

```
##     fit   lwr   upr
## 1 24.47 14.81 34.12
```

The predicted value for `horsepower = 98` is 24.47, the prediction interval ranges from 14.81 to 34.12, and the confidence interval ranges from 23.97 to 24.96.

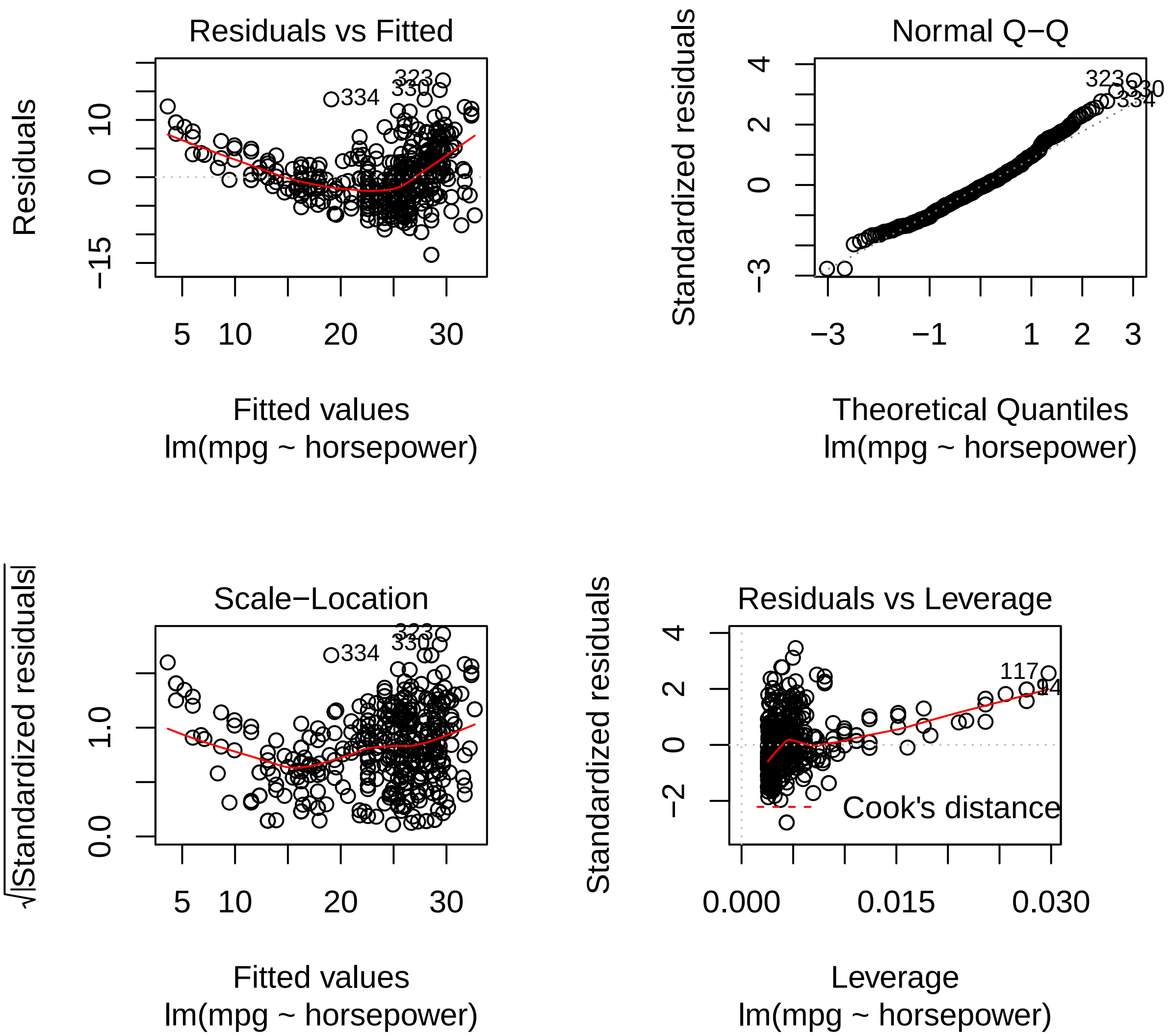
(b)

```
plot(mpg ~ horsepower, data = Auto)
abline(fita, col = 2, lwd = 2)
```



(c)

```
plot(fita)
```



There seems to be a nonlinear relation that is not captured by the linear model (see the "residuals versus fitted" plot, top left). In addition, there are a number of points with fairly large leverage (bottom right), coming from cars with large horsepower values.

Xtra #14 and 15 (5)

Problem 14

a) Simulate a time series X of length $N = 100$ from the formula

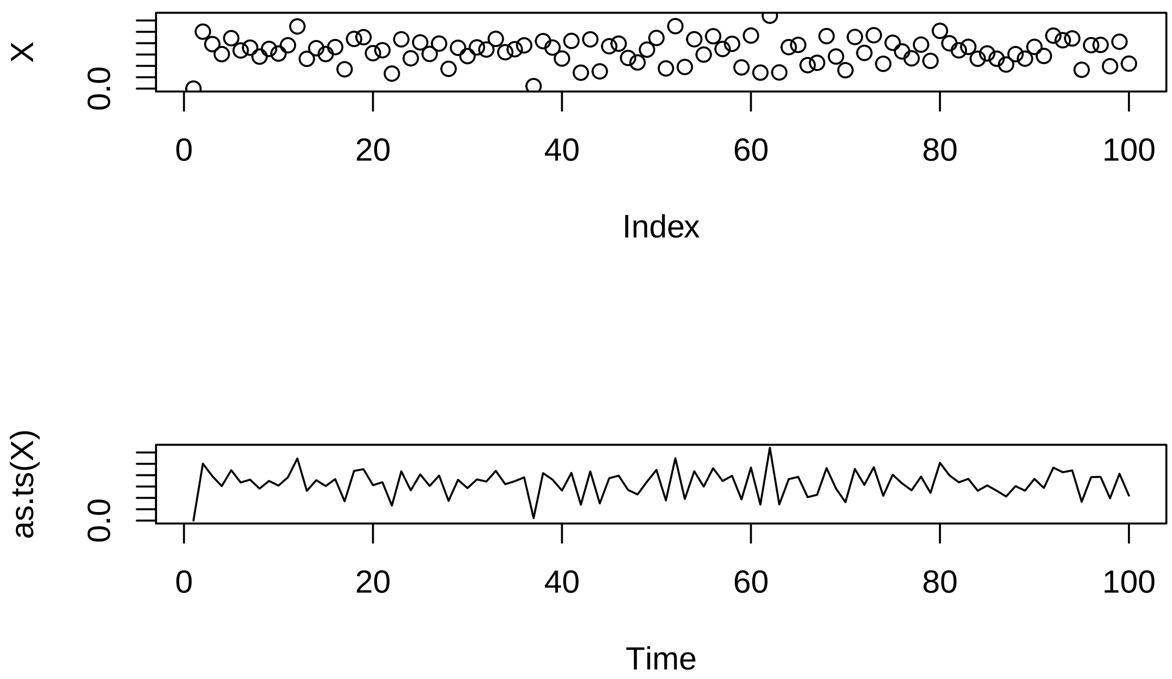
$$X_t = \beta_0 + \sum_{i=1}^{K} \beta_i X_{t-i} + \epsilon_t$$

using the lag $k = 1$, coefficients $\beta_0 = 1$ and $\beta_1 = -0.5$ and error terms $\epsilon_t \sim N(0, 0.2^2)$. The formula tells you how to make X_t for $t \geq k + 1$. Choose X_1 arbitrarily. Plot X as a vector. Convert X into a timeseries object with the function **as.ts()** and plot it again. Describe the plot.

Solution.

Chose $X_1 = 0$.

```
X <- rep(0,100)
for (j in 2:100){
  X[j] <- 1 - 0.5*X[j-1] + 0.2*rnorm(1)
}
par(mfrow = c(2,1))
plot(X)
plot(as.ts(X))
```

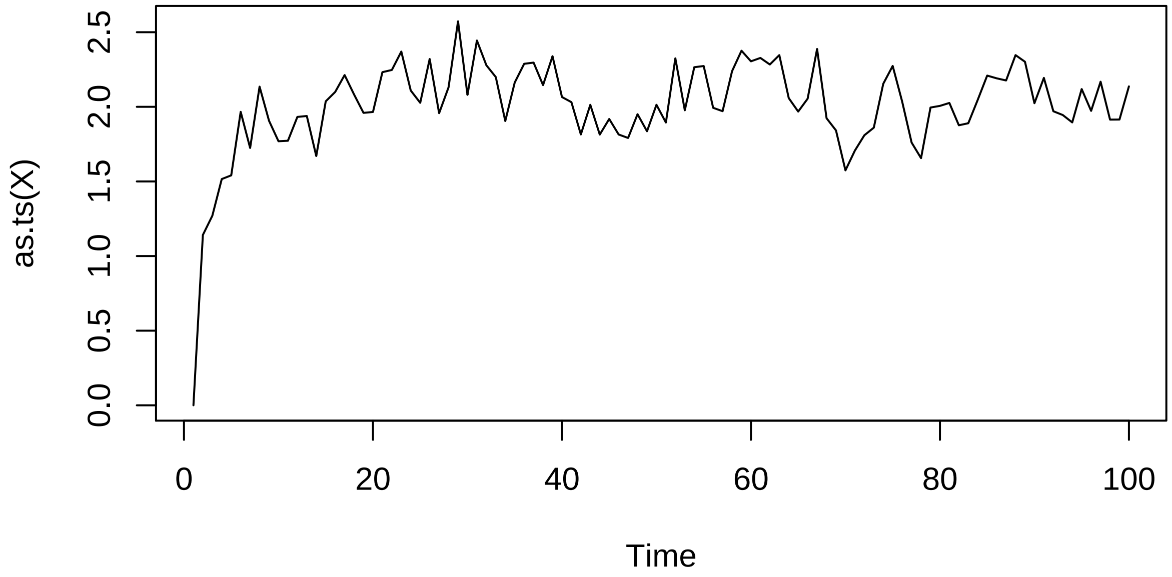


The plot shows regular oscillations (Up - down - up - down) about a value near 0.6. The time series plot is automatically done as a line plot.

b) Repeat part a) with $\beta_0 = 1$, $\beta_1 = +0.5$. How does the plot change?

Solution. Only the time series plot is shown.

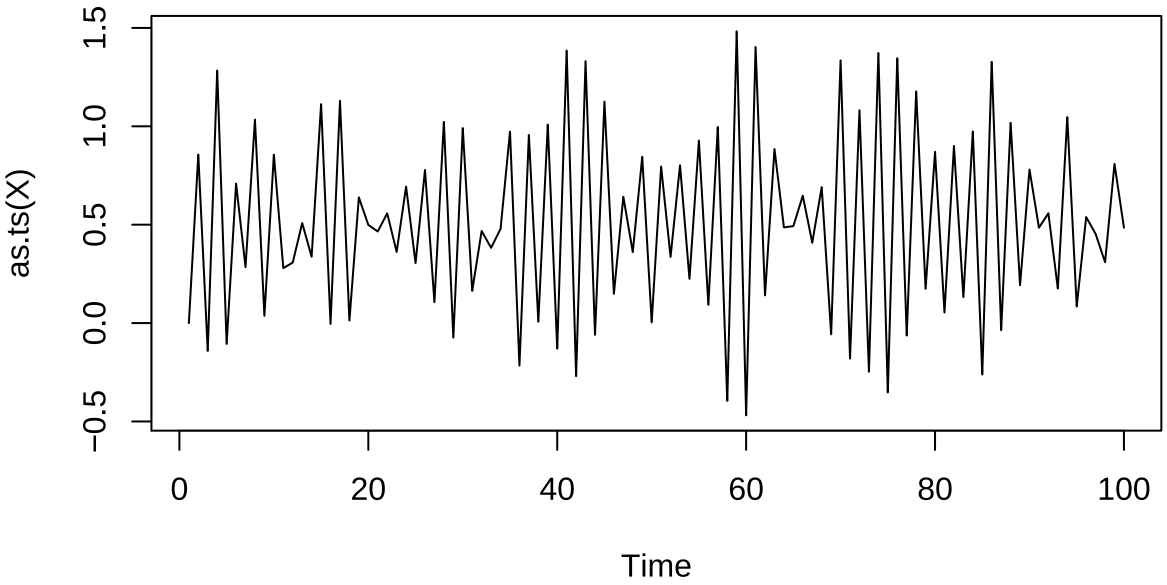
```
X <- rep(0,100)
for (j in 2:100){
  X[j] <- 1 + 0.5*X[j-1] + 0.2*rnorm(1)
}
plot(as.ts(X))
```



This plot shows observations about the value 2. The oscillation is less regular.

c) Repeat part a) with $\beta_0 = 1$, $\beta_1 = -0.9$. How does the plot change?

```
X <- rep(0,100)
for (j in 2:100){
  X[j] <- 1 - 0.95*X[j-1] + 0.2*rnorm(1)
}
plot(as.ts(X))
```



The plot shows very regular oscillations about a value near 0.5. The amplitides of the oscillations vary, becoming larger smaller every 10 to 20 observations.

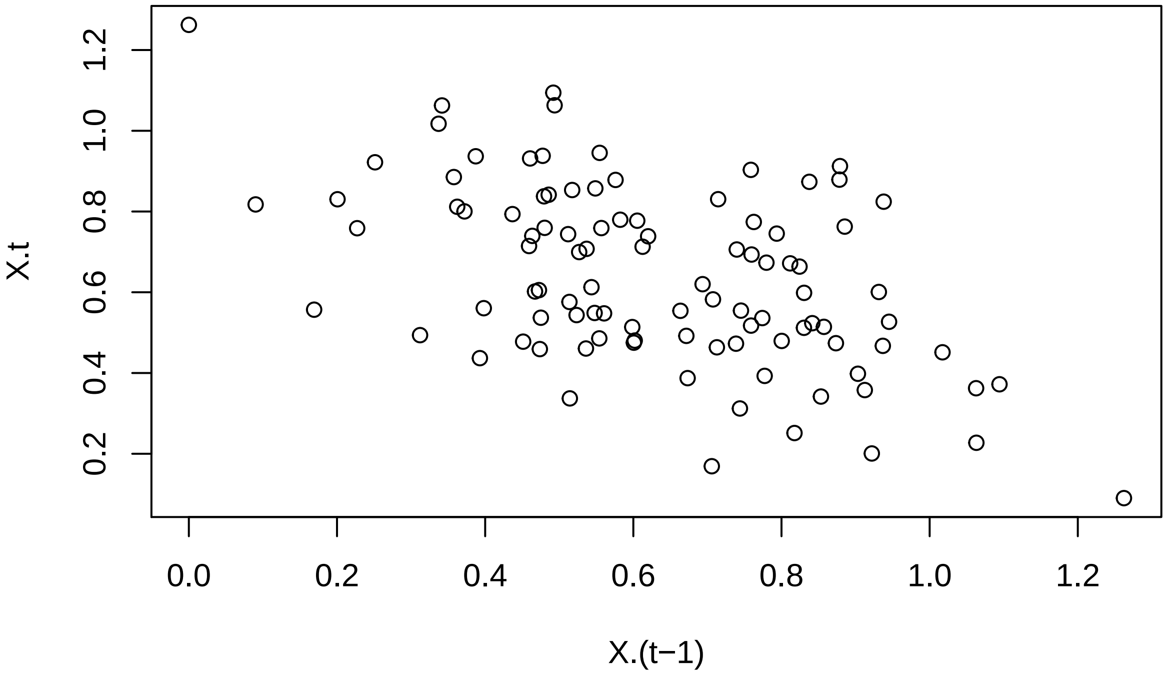
Problem 15

Simulate a time series X as in the previous problem ($N = 100$ observations, lag $k = 1$, $\beta_0 = 1$, $\beta_1 = -0.5$, $\epsilon_t \sim N(0, 0.2^2)$.

```r
X <- rep(0,100)
for (j in 2:100){
  X[j] <- 1 - 0.5*X[j-1] + 0.2*rnorm(1)
}
```

a) Make a scatterplot of X_t against x_{t-1} for $t = 2, \ldots, N$ and describe it.

```r
plot(X[1:99], X[2:100], xlab = "X.(t-1)", ylab = "X.t")
```



The plot shows a weak negative association.

b) Create a data frame of $N-1$ observations and 2 columns that contains (X_{t-1}, X_t) in row t. Use this to fit a linear model to predict X_t from X_{t-1}. Compare the estimated coefficients to the β_i. Also compare the residual standard error to the standard deviation of theϵ_t. Summarize your results and observations.

Make a data frame and fit a linear model:

```
df.15 <- data.frame(x1 = X[1:99], x2 = X[2:100])
lm.15 = lm(x2 ~ x1, data = df.15)
summary(lm.15)
```

```
##
## Call:
## lm(formula = x2 ~ x1, data = df.15)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43285 -0.13421 -0.01076  0.12145  0.39840
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.96057    0.05571  17.242  < 2e-16 ***
## x1          -0.50802    0.08282  -6.134 1.86e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1893 on 97 degrees of freedom
## Multiple R-squared:  0.2795, Adjusted R-squared:  0.2721
## F-statistic: 37.63 on 1 and 97 DF,  p-value: 1.857e-08
```

The estimates for the intercept and slope are close to β_0 and β_1. The residual standard error is close to the noise level that was used. The R^2 value is not very large.