HW 1

Exercises

Exercise 1 (Learning the mechanics.)**.**

Use the method of least squares to fit a straight line to these six data points:

x	1	2	3	4	5	6
y	2	4	5	4	2	7

(a) What are the least squares estimates of β_0 and β_1? Compute manually.
(b) Plot the data points and graph the least squares line on the scatterplot.

Solution:

(a)

- With hand:
  - $\bar{x} = 3.5$, $\bar{y} = 4$.

$x - \bar{x}$	-2.5	-1.5	-0.5	0.5	1.5	2.5
$y - \bar{y}$	-2	0	1	0	-2	3

  - $S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = 9$, $S_{xx} = \sum(x_i - \bar{x})^2 = 17.5$.
  - $\hat{\beta}_1 = 9/17.5 \approx 0.5143$, $\hat{\beta}_0 = \bar{y} - \beta_1\bar{x} = 4 - 9/17.5 \times 3.5 = 2.2$.

- With R (manually)

```r
x <- c(1, 2, 3, 4, 5, 6)
y <- c(2, 4, 5, 4, 2, 7)
xbar <- mean(x)
ybar <- mean(y)
n <- length(x)
```

```
sxy <- sum((x - xbar) * (y - ybar)) / (n - 1)
sxx <- sum((x - xbar) ^ 2) / (n - 1)
b1hat <- sxy / sxx
b0hat <- ybar - b1hat * xbar
print(b1hat)
```

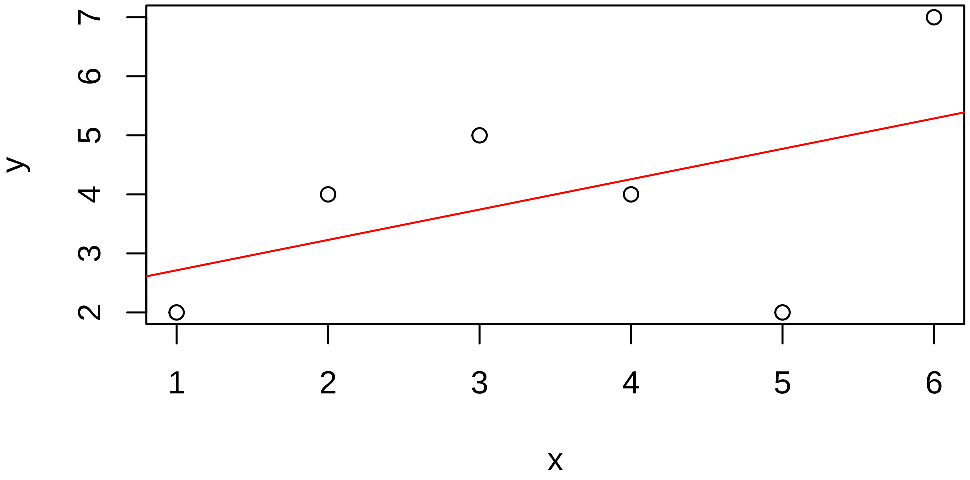
```
[1] 0.5142857
```

```
print(b0hat)
```

```
[1] 2.2
```

(b) Ploting with R code.

```
plot(x, y)
abline(coef = c(b0hat, b1hat), col='red')
```



Exercise 2 (Predicting home sales price.)**.**

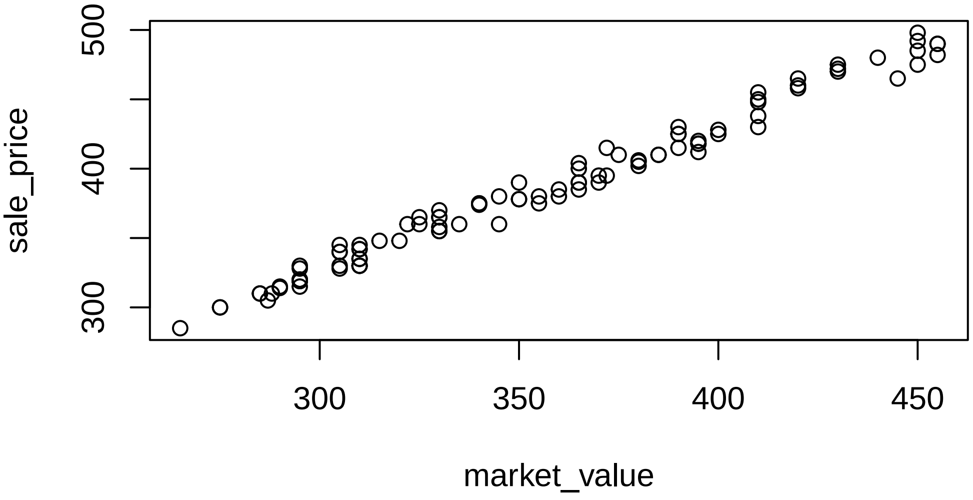
Real estate investors, homebuyers, and homeowners often use the appraised (or market) value of a property as a basis for predicting sale price. Please look at the provided dataset MARKET.csv. All the money are in 1000 dollars.

(a) Propose a simple linear model to relate the appraised market value x to the sale price y.

(b) A scatterplot of the data is shown below. Does it appear that a straight-line model will be an appropriate fit to the data?

(c) A R simple linear regression printout is also shown below. Find the equation of the best-fitting line through the data on the printout.

(d) Interpret the y-intercept of the least squares line. Does it have a practical meaning for this application? Explain.
(e) Interpret the slope of the least squares line.
(f) Over what range of x is the interpretation meaningful?
(g) Use the least squares model to estimate the mean sale price of a property appraised at $300,000.



```
Call:
lm(formula = sale_price ~ market_value)

Residuals:
    Min      1Q  Median      3Q     Max 
-14.674  -5.480  -1.287   6.300  13.409 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  10.72069    5.01930   2.136   0.0352 *  
market_value  1.05305    0.01399  75.256   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.234 on 98 degrees of freedom
Multiple R-squared:  0.983, Adjusted R-squared:  0.9828 
F-statistic:  5663 on 1 and 98 DF,  p-value: < 2.2e-16
```

Solution:

(a) $y = \beta_0 + \beta_1 x + \epsilon$.
(b) Yes, the data appears to demonstrate a straight-line relationship.

(c) $\widehat{y} = 10.721 + 1.053x$, or sale_price=10.721+1.053*market_value.

(d) The estimated intercept $\widehat{\beta}_0 = 10.721$ represents the expected sale price when the market value $x = 0$. Because a market value of 0 is not observed in the data and is not realistic in this application, the y-intercept should not be interpreted in a practical sense. Instead, it serves primarily as a mathematical component of the fitted regression line.

(e) The estimated slope $\widehat{\beta}_1 = 1.053$ represents the expected change in the sale price for a one-unit increase in the market value. Specifically, for each \$1000 increase in the market value, the sale price is expected to increases by \$1053.

(f) About the range, within \$200k to \$500k, where most of the data points are clustered, could be the interpretation meaningful.

(g) When the market value is 300k, we know that $x = 300$. Then sale_price=10.721+1.053*300=326.621. So the sales price is estimated to be around \$326621.

Exercise 3.

A study shows that during a certain sport the mean heart rate y and the maximal oxygen uptake x might have relations. The dataset SPORTHR.csv shows y (expressed as a percentage of maximum heart rate) and x (VO2max). The data are shown in the table.

```
player VO2max meanHR
     1    140   68.2
     2    150   71.1
     3    160   74.4
     4    170   76.5
     5    180   78.8
     6    185   80.1
     7    190   82.4
     8    200   84.6
```

(a) Find the equation of the least squares line.

(b) Give a practical interpretation (if possible) of the y-intercept of the line.

(c) Give a practical interpretation (if possible) of the slope of the line.

Solution:

(a)

```r
attach(df)
fit <- lm(meanHR~VO2max, data=df)
fit$coefficients
```

```
(Intercept)      VO2max
 30.6546403   0.2697185
```

So the equation is $\widehat{y} = 30.6546 + 0.2697x$.

(b) Since $x = 0$ corresponds to a maximal oxygen uptake of zero, which is physiologically impossible, the y-intercept doesn't have a practical interpretation.

(c) For each unit increase in the value of VO2Max, the mean HR is estimated to increase by 0.2697185.

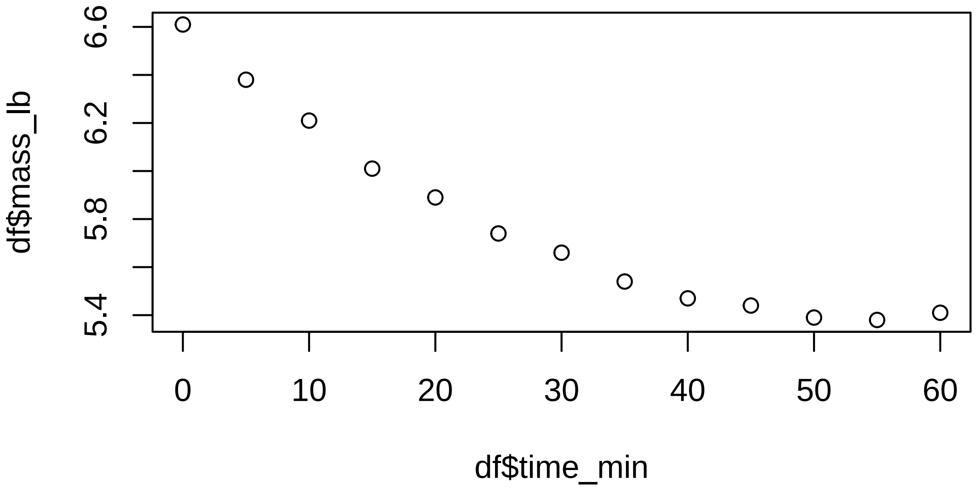
Exercise 4 (Spreading rate of spilled liquid.)**.**

A researcher studied the rate at which a spilled liquid will spread across a surface. The mass (in pounds) of the spill after a period of time ranging from 0 to 60 minutes is recorded and shown below (based on the dataset `SPILLS.csv`). Do the data indicate that the mass of the spill tends to diminish as time increases? If so, how much will the mass diminish each minute?

```
time_min mass_lb
       0    6.61
       5    6.38
      10    6.21
      15    6.01
      20    5.89
      25    5.74
      30    5.66
      35    5.54
      40    5.47
      45    5.44
      50    5.39
      55    5.38
      60    5.41
```

Solution: First we plot the data.

```r
plot(df$time_min, df$mass_lb)
```



As time increase, the mass of the spill tends to diminish in a nonlinear way. The scatterplot in this problem clearly shows a significant nonlinear trend. Therefore, the linear model is not the best to describe the data in this scatter plot.

```
fit <- lm(mass_lb~time_min, data=df)
summary(fit)
```

```
Call:
lm(formula = mass_lb ~ time_min, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.13940 -0.10890 -0.03874  0.09995  0.23176 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.380220   0.071917  88.717  < 2e-16 ***
time_min    -0.020033   0.002034  -9.849 8.61e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1372 on 11 degrees of freedom
Multiple R-squared:  0.8981,	Adjusted R-squared:  0.8889 
F-statistic: 96.99 on 1 and 11 DF,  p-value: 8.609e-07
```

If we fit the linear model, the fitted regression line is $\hat{y}$=6.3802198 - 0.020033x. Since the slope is negative, and the corresponding p-value is very small, there is evidence that the mass of the spill tends to decrease as time increases. Therefore the spill tends to diminish as time increases. Fo each minute increased, the mass is expcted to decrease by 0.020033 lb.

Exercise 5 (Sweetness of orange juice.)**.**

To study the sweetness of orange juices, researchers collect some data on the sweetness index (y) and the amount of pectin (x) in the orange juice (in g/L). The dataset is `ORANGEJUICE.csv`.

sample	pectin	sweetness
1	100	6.72
2	120	6.41
3	140	6.58
4	160	6.11
5	180	6.33
6	200	5.98
7	220	6.21
8	240	5.87
9	260	6.03
10	280	5.72
11	300	5.95
12	320	5.61
13	340	5.84
14	360	5.53
15	380	5.77
16	400	5.46
17	420	5.69
18	440	5.38
19	460	5.62
20	480	5.31
21	500	5.55
22	520	5.24
23	540	5.49
24	560	5.21

(a) Find the values of SSE, s^2, and s for this regression.
(b) Estimate σ^2, the variance of the random error term in the model.
(c) Estimate σ, the standard deviation of the random error term in the model.
(d) Explain why it is difficult to give a practical interpretation to s^2, the estimate of σ^2.
(e) Give a practical interpretation of the value of s.

Solution:

(a)

```r
fit <- lm(sweetness ~ pectin, data=df)
summary(fit)
```

```
Call:
lm(formula = sweetness ~ pectin, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.23416 -0.16737  0.02391  0.10927  0.28027 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.7104507  0.0886669   75.68  < 2e-16 ***
pectin      -0.0027072  0.0002478  -10.93 2.35e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.168 on 22 degrees of freedom
Multiple R-squared:  0.8444,    Adjusted R-squared:  0.8373 
F-statistic: 119.4 on 1 and 22 DF,  p-value: 2.352e-10
```

```r
anova(fit)
```

```
Analysis of Variance Table

Response: sweetness
          Df Sum Sq Mean Sq F value    Pr(>F)    
pectin     1 3.3712  3.3712  119.38 2.352e-10 ***
Residuals 22 0.6213  0.0282                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So SSE=3.3712437, s^2=0.0282387 and s=0.1680438.

(b) σ^2 is estimated by s^2, which is 0.0282387.

(c) σ is estimated by s, which is 0.1680438.

(d) The units of measure for s^2 are square units. It is very difficulty to interpret units such as \$^{2}, minutes^2, etc.

(e) We would expect most of the observed values to fall within $2s$ or 2(0.1680438)=0.3360877 units of the least squares line.