Homework 1

Solution

Question 1. Page 102, Problem 3.6

(10 pts) Please refer to Question 3.6 (Solution).xlsx

Question 2. Predicting home sales price

(a) (5 pts) Write down the straight_line model.

Solution: $y = \beta_0 + \beta_1 x + \epsilon$

(b) (5 pts)

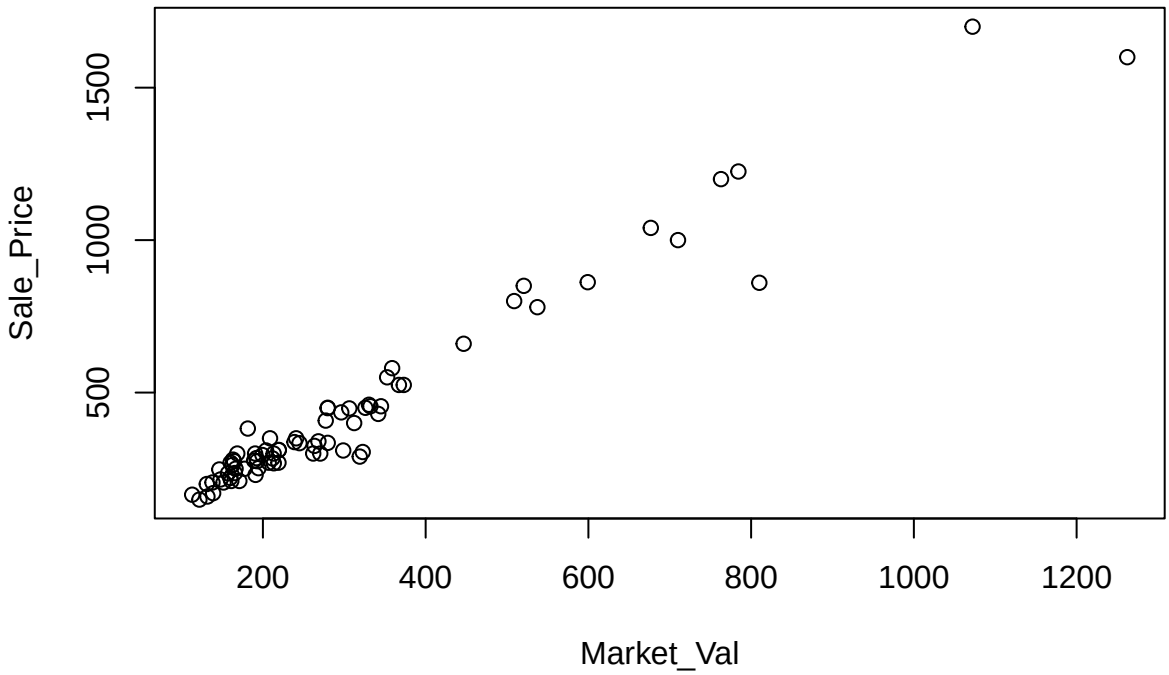
- Draw the scatter plot of Sale_Price vs Market_val in R.

```
### Read the data set in .csv
setwd("/cloud/project/STAT 3113 Data Sets")
df_tampalms = read.csv("TAMPALMS.csv", header=TRUE, sep=',', dec='.')
names(df_tampalms)
```

```
## [1] "Property"   "Market_Val" "Sale_Price"
```

```
attach(df_tampalms)

### Plot the scatter plot
### Market_Val is independent variable, Sale_Price is dependent variable.
plot(Market_Val, Sale_Price)
```



- (5 pts) Does it appear that a straight-line model will be an appropriate fit to the data?

Solution: Yes, the data appears to demonstrate a straight-line relationship.

(c) (5 pts)

-Fit the regression model in R.

```
### Fit the model and find the summary and anova table

fit_tampalms = lm(Sale_Price ~ Market_Val)
summary(fit_tampalms)
```

```
##
## Call:
## lm(formula = Sale_Price ~ Market_Val)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -282.168  -24.830    1.808   29.790  188.787
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.35864   13.76837   0.099    0.922
## Market_Val   1.40827    0.03693  38.132   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.76 on 74 degrees of freedom
## Multiple R-squared:  0.9516, Adjusted R-squared:  0.9509
## F-statistic:  1454 on 1 and 74 DF,  p-value: < 2.2e-16
```

```
anova(fit_tampalms)
```

```
## Analysis of Variance Table
```

```
##
## Response: Sale_Price
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## Market_Val   1 6874024 6874024    1454 < 2.2e-16 ***
## Residuals   74  349842    4728
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Write down the equation of the best-fitting line through the data.

Solution:

$\hat{y} = 1.36 + 1.41x$

or

Sale_Price = 1.36 + 1.41 * Market_Val

(d) (5 pts) Interpret the y-intercept of the least squares line. Does it hava a practical meaning for this application? Explain.

Solution:

$\hat{\beta}_0 = 1.36$

Interpretation: Since $x = 0$ (no market value) is not in the range of the observed values of market values, the y-intercept does not have a practical interpretation. That is, when $x = 0$ (no market value), then the sales price has no practical meaning.

(e) (5 pts)

- Interpret the slope of the least squares line.

Solution:

$\hat{\beta}_1 = 1.41$

Interpretation:

For each \$1,000 increase in the market value, the sale price is expected to increase \$1,410.

- Over what range of x is the interpretation meaningful?

Answer: About the range, within \$100,000 to \$1,000,000, where most of the data points are clustered, could be the interpretation meaningful.

(f) (5 pts) Use the least squares model fitted to estimate the mean sale price of a property appraised at \$300,000.

When the Market_Val is \$300,000, we know $x = 300$ as the unit of Market_Val is thousands of dollars. Then

Sale_Price = 1.36 + 1.41 * 300 = 424.36.

The sales price is estimated to be around \$ 424,360.

Question 3. Game performance of water polo players

(a) (5 pts) Find the equation of the least squares line.

```
setwd("/cloud/project/STAT 3113 Data Sets")
df_polo = read.csv("POLO.csv", header=TRUE, sep=',', dec='.')
names(df_polo)
```

```
## [1] "HR."    "VO2Max"
```

```
attach(df_polo)

fit_polo = lm(HR.~VO2Max)
summary(fit_polo)
```

```
##
## Call:
## lm(formula = HR. ~ VO2Max)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3739 -3.6785  0.4989  2.6764  7.9520
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -27.237     19.771  -1.378   0.2175
## VO2Max         0.558      0.113   4.940   0.0026 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.269 on 6 degrees of freedom
## Multiple R-squared:  0.8027, Adjusted R-squared:  0.7698
## F-statistic: 24.41 on 1 and 6 DF,  p-value: 0.002603
```

```
anova(fit_polo)
```

```
## Analysis of Variance Table
##
## Response: HR.
##           Df Sum Sq Mean Sq F value   Pr(>F)
## VO2Max     1 677.45  677.45  24.406 0.002603 **
## Residuals  6 166.55   27.76
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The estimated regression equation is:

Solution: $\hat{y} = -27.2 + 0.558x$

(b) (5 pts) Give a practical interpretation (if possible) of the y-intercept of the line.

Solution: $\hat{\beta}_0 = -27.2$

Since 0 is not in the range of observed values of VO2Max, the y-intercept does not have a practical interpretation.

(c) (5 pts) Give a practical interpretaiton (if possible) of the slope of the line.

Solution: $\hat{\beta}_1 = 0.558$

For each unit increase in the value of VO2Max, the mean HR is estimated to increase by 0.558.

Question 4. Spreading rate of spilled liquid

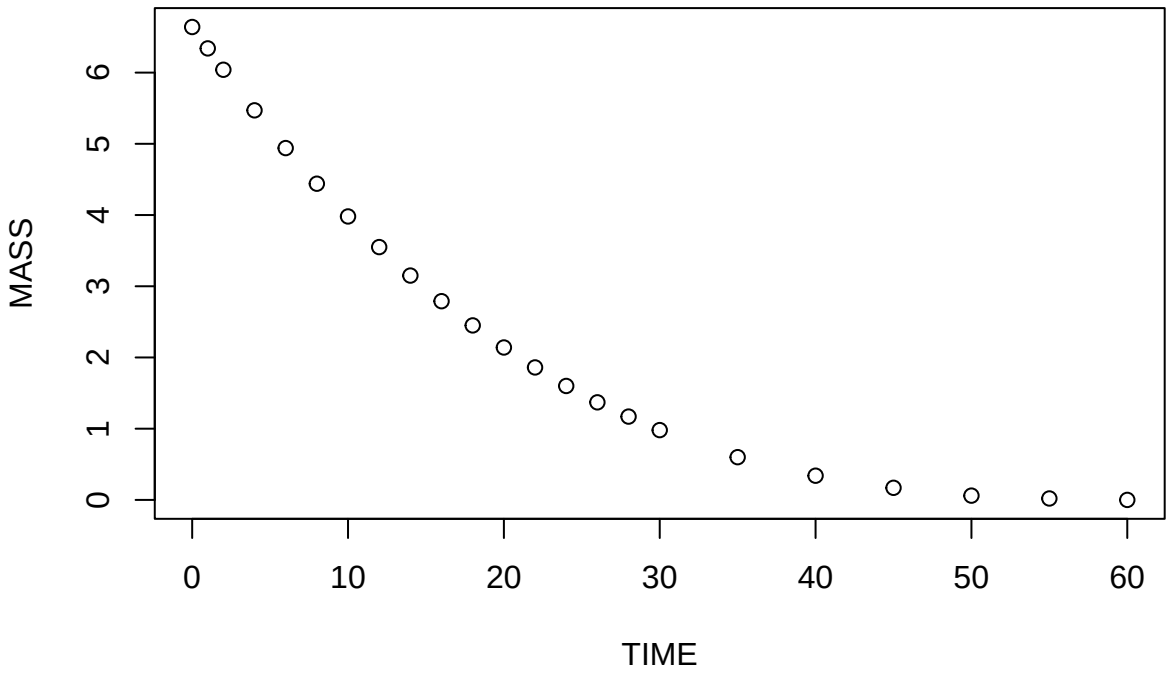
```
### Read the data set in .csv and draw scatter plot

setwd("/cloud/project/STAT 3113 Data Sets")
```

```r
df_liquid = read.csv("LIQUIDSPILL.csv", header=TRUE, sep=',', dec='.')
names(df_liquid)
```

```
## [1] "TIME" "MASS"
```

```r
attach(df_liquid)
plot(TIME, MASS)
```



- (5 pts) Do the data indicate that the mass of the spill tends to diminish as time increase?

Solution: 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.

```r
### Fit the simple linear regression model
fit_liquid = lm(MASS~TIME)
summary(fit_liquid)
```

```
##
## Call:
## lm(formula = MASS ~ TIME)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8861 -0.7593 -0.3024  0.6229  1.6207
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.22070    0.29598   17.64 4.55e-14 ***
## TIME        -0.11402    0.01032  -11.05 3.26e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8573 on 21 degrees of freedom
```

```
## Multiple R-squared:  0.8533, Adjusted R-squared:  0.8464
## F-statistic: 122.2 on 1 and 21 DF,  p-value: 3.26e-10
```

```
anova(fit_liquid)
```

```
## Analysis of Variance Table
##
## Response: MASS
##           Df Sum Sq Mean Sq F value   Pr(>F)
## TIME       1 89.794  89.794  122.19 3.26e-10 ***
## Residuals 21 15.433   0.735
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- (5 pts) If the data indicate the trend, how much will the mass diminish each minutes?

Solution: If we fit the linear model, the fitted regression line is $\hat{y} = 5.221 - 0.114x$. Since the coefficient of time is negative, there is evidence that the mass of the spill tends to decrease as time increases. For each minute increase in time, the mean mass is estimated to diminish by .1140 pounds.

Question 5. Sweetness of orange juice

(a) (10 pts) Fit the model and find the values of SSE, s^2, and s for this regression.

```
setwd("/cloud/project/STAT 3113 Data Sets")
df_ojuice = read.csv("OJUICE.csv", header=TRUE, sep=',', dec='.')
names(df_ojuice)
```

```
## [1] "Run"        "SweetIndex" "Pectin"
```

```
attach(df_ojuice)
fit_ojuice = lm(SweetIndex~Pectin)
summary(fit_ojuice)
```

```
##
## Call:
## lm(formula = SweetIndex ~ Pectin)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54373 -0.11039  0.06089  0.13432  0.34638
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.2520679  0.2366220  26.422   <2e-16 ***
## Pectin      -0.0023106  0.0009049  -2.554   0.0181 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.215 on 22 degrees of freedom
## Multiple R-squared:  0.2286, Adjusted R-squared:  0.1936
## F-statistic:  6.52 on 1 and 22 DF,  p-value: 0.01811
```

```
anova(fit_ojuice)
```

```
## Analysis of Variance Table
##
## Response: SweetIndex
```

```
##           Df Sum Sq  Mean Sq F value  Pr(>F)
## Pectin     1 0.3014 0.301402  6.5204 0.01811 *
## Residuals 22 1.0169 0.046224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- SSE = 1.0169
- $s^2 = 0.046$
- $s = 0.215$

(b) (5 pts) Estimate σ^2, the variance of the random error term in the model.

Solution: $\hat{\sigma}^2 = 0.046$

(c) (5 pts) Estimate σ, the standard deviation of the random error term in the model.

Solution: $\hat{\sigma} = 0.215$

(d) (5 pts) Explain why it is difficult to give a practical interpretation to s^2, the estimate of σ^2.

Solution: The units of measure for s^2 are square units. It is very difficulty to interpret units such as $\$^2$, minutes^2, etc.

(e) (5 pts) Give a practical interpretation of the value of s.

Solution: We would expect most of the observed values to fall within $2s$ or 2(0.215)=0.43 units of the least squares line.