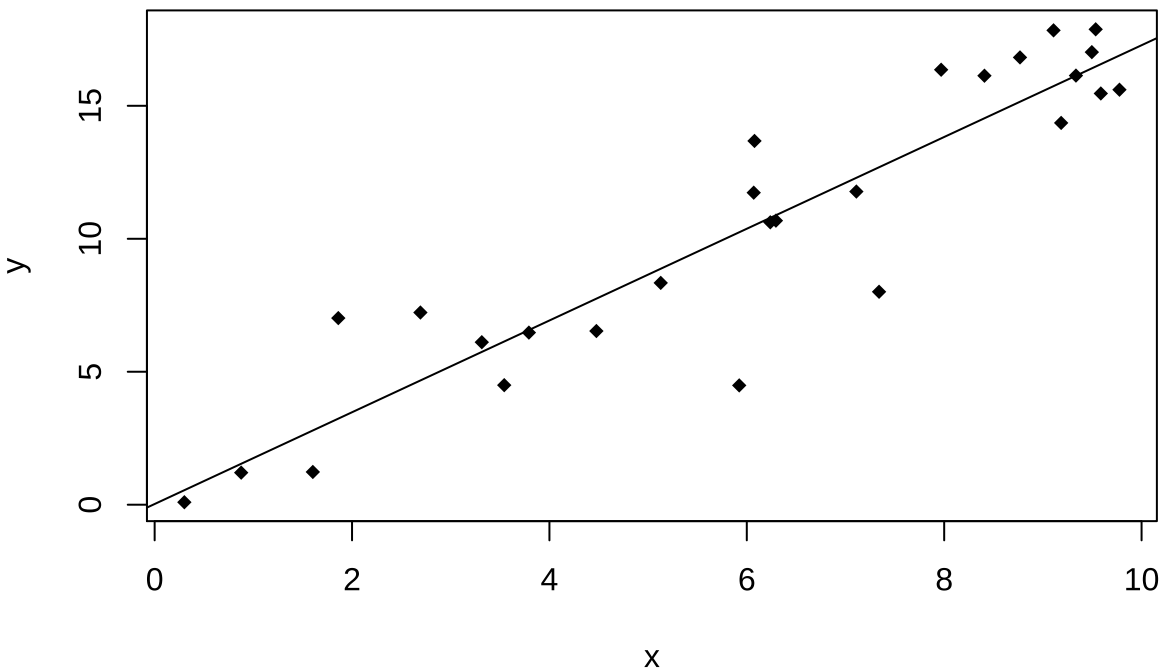
Assignment 1

1. Fit a straight line to the data point and produce a plot of the data with the line of best fit superimposed.

```
file1 <- "Ass1.txt"
data1<-read.table(file1, header=TRUE)
plot(data1$x, data1$y, pch=18, xlab = 'x', ylab = 'y')
x1<-data1$x
y<-data1$y

fit<-lm(y~x1)
abline(fit)
```



2. Give the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the intercept β_0 and and slope β_1 in a simple linear regression model. Report also the least-squares equation arising from a least squares fit.

```
coef(fit)
```

```
## (Intercept)          x1
##  0.02676552  1.72512091
```

```
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7577 -1.1595 -0.1691  1.5003  3.7808
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02677    0.96783   0.028    0.978
## x1           1.72512    0.14372  12.003 7.14e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.182 on 25 degrees of freedom
## Multiple R-squared:  0.8521, Adjusted R-squared:  0.8462
## F-statistic: 144.1 on 1 and 25 DF,  p-value: 7.145e-12
```

From above, we know the intercept $\beta_0 = 0.02676552$ and the slope $\beta_1 = 1.72512091$ as well as the least square estimates $\hat{\beta}_0 = 0.02677$ and $\hat{\beta}_1 = 1.72512$

Notice that the least-squares equation: $\hat{y^*} = E(Y|X \hat{=} x^*) = \hat{\beta}_0 + \hat{\beta}_1 x^* = 0.02677 + 1.72512091x^*$ is arised from a least square method with $\hat{\beta}_0$ and $\hat{\beta}_1$.

3. Give an estimate of the variance σ^2.

```r
(summary(fit)$sigma)**2
```

```
## [1] 4.761522
```

Thus, $s^2 = \frac{RSS}{n-2} = \frac{\sum e_i^2}{25} = 4.761522$

4. At level 5%, test $H_0 : \beta_1 = 0$ versus Ha : $\beta_1 \neq 0$. What is the p-value of your test ?

For hypothesis $H_0 : \beta_1 = 0$, test statistics is $T_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim t_{n-2} = \frac{1.72512091}{0.14372} = 12.003$ when H_0 is True. Use two-sided test, check if p-value is less than the level of significance $\alpha = 5\%$. From the summary function above, p-value $= 2 * P(T > T_0) = 2 * P(T > 12.003) = 7.145\text{e-}12 < 0.05$, where $T \sim t_{n-2}$. Therefore, reject H_0.

5. Use the least-squares equation to estimate the mean E(Y|X = 2.5) . Find a 95% confidence interval for E(Y|X = 2.5). Is 0 a feasible value for E(Y|X = 2.5). Give a reason to support your answer.

E(Y|X $= 2.5) = \beta_0 + \beta_1(x^*) = 0.02676552 + 1.72512091 * (2.5) = 4.34$

```r
predict(lm(fit), newdata=list(x1 = 2.5), interval="confidence", level=.95)
```

```
##        fit      lwr      upr
## 1 4.339568 2.974702 5.704434
```

Confidence Interval $= (\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{n-2}(1 - \alpha/2, n-2) * s\sqrt{\frac{1}{n} + \frac{(x^*-\bar{x})^2}{S_{xx}}}) = (2.974702, 5.704434)$

0 is not a feasible value because $0 < 2.974702$ which is the lower bound of the confidence interval for E(Y|X $= 2.5$).

6. Find a 95% prediction interval for y at x = 2.5

```r
predict(lm(fit), newdata=list(x1 = 2.5), interval="prediction", level=.95)
```

```
##        fit        lwr      upr
## 1 4.339568 -0.3572184 9.036354
```

Prediction Interval $= (\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{n-2}(1 - \alpha/2, n-2) * s\sqrt{1 + \frac{1}{n} + \frac{(x^*-\bar{x})^2}{S_{xx}}}) = (-0.3572184, 9.036354)$

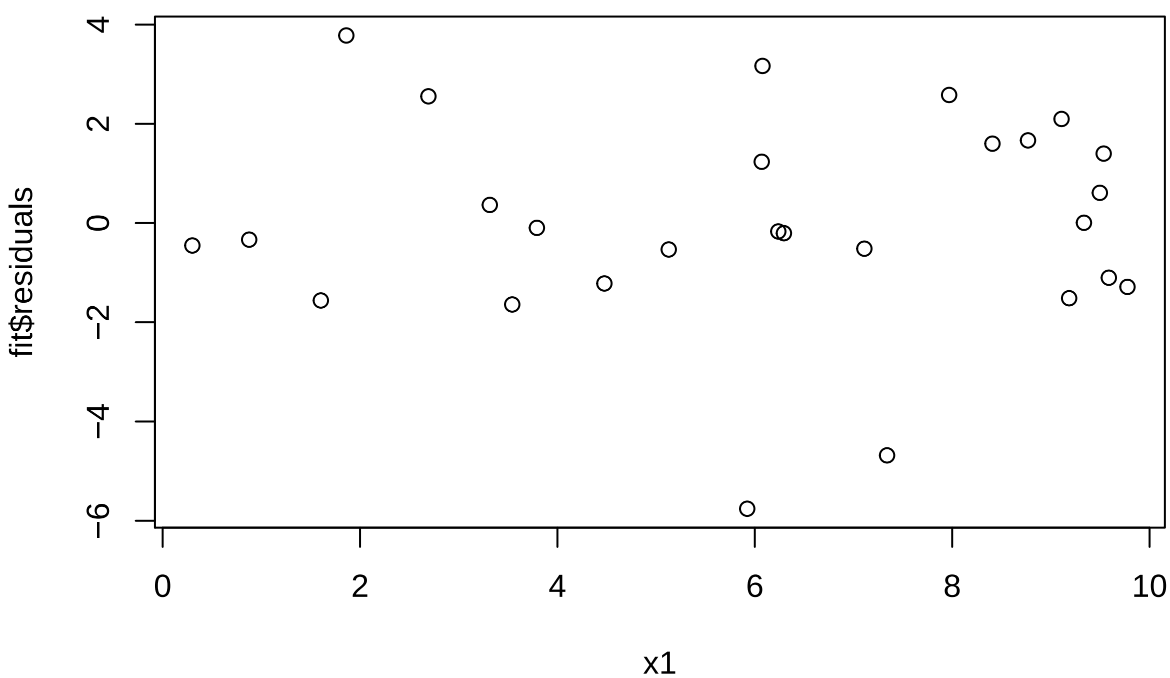
Bonus Questions:

1. plot(against the x values) the residuals e_i, i = 1, . . . , n, from the fit

```r
residuals(fit)
```

```
##            1            2            3            4            5 
## -1.286547523  0.609486332 -0.169069749 -1.101499671 -1.515001066 
##            6            7            8            9           10 
## -0.452026917  1.400206616  3.166949596  2.097833244  3.780784466 
##           11           12           13           14           15 
## -1.217406709  0.365458665  1.600417880 -0.096424008  2.554483870 
##           16           17           18           19           20 
##  0.006230798 -5.757656663 -0.204401641  2.581745959 -1.641143781 
##           21           22           23           24           25 
##  1.666374353 -0.333696561 -0.516375181 -1.560134340  1.235750563 
##           26           27 
## -0.532185541 -4.682152990
```

```r
plot(x1, fit$residuals)
```



2. Comment on the adequcy of the straight line model, based on the residual plot, that is, comment on weather the assumption of the least squares fitted and how they relate to the residual errors e_{i} are met by the observed data.

```r
summary(residuals(fit))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## -5.7577 -1.1595 -0.1691  0.0000  1.5003  3.7808
```

The assumption of least squares method states the error terms are independent random variables with zero mean and constant variance and each e_i are uncorrelated. On question 1, we get the plot of residuals as a function of x. We can see that the residuals are negative as well as positive scattered randomly around zero. Also, the vertical width of the scatter does not appear to increase or decrease across x values, so we can assume that the variance in the error terms is constant. By checking with the summary of the plot, we know the residuals average to zero and have constant variance. Thus, we can conclude the linear regression line model is adequate since the assumptions are met.