Section 5.4
Residuals

A **residual** value is the difference between an actual observed y value and the corresponding predicted y value, $\hat{y}$. Residuals are just errors.

Residual error = observed value – predicted value

Example 1: A least-squares regression line was fitted to the weights (in pounds) versus age (in months) of a group of many young children. The equation of the line is $\hat{y} = 16.6 + 0.65t$, where $\hat{y}$ is the predicted weight and t is the age of the child. A 20-month old child in this group has an actual weight of 25 pounds. What is the residual weight, in pounds, for this child?

observed

predicted: $\hat{y} = 16.6 + 0.65(20) = 29.6$

Residual: $25 - 29.6 = \boxed{-4.6}$

The plot of the residual values against the x values can tell us a lot about our LSRL model. Plots of residuals may display patterns that would give some idea about the appropriateness of the model. The sum of the residuals will always be zero, so they'll always be centered about the x-axis.

- If the **functional form of the regression model is incorrect**, the residual plots constructed by using the model will often **display a pattern**. The pattern can then be used to propose a more appropriate model.

- When a residual plot shows **no pattern**, it indicates that the proposed model is a **reasonable fit to a set of data**.

YES

Figure 1 shows a horn-shaped pattern (linear model is not a reasonable fit for the data). Figure 2 shows a quadratic pattern (linear model is not a reasonable fit for the data). Figure 3 has no pattern (linear model is a reasonable fit for the data).

YES

NO

Figure 1

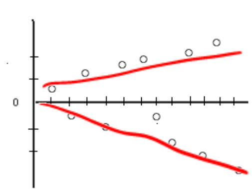


Figure 2

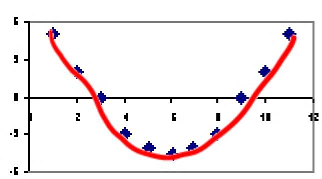
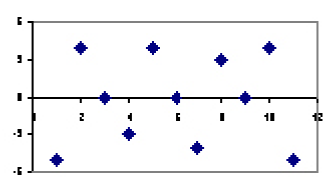


Figure 3



R command: resid( )

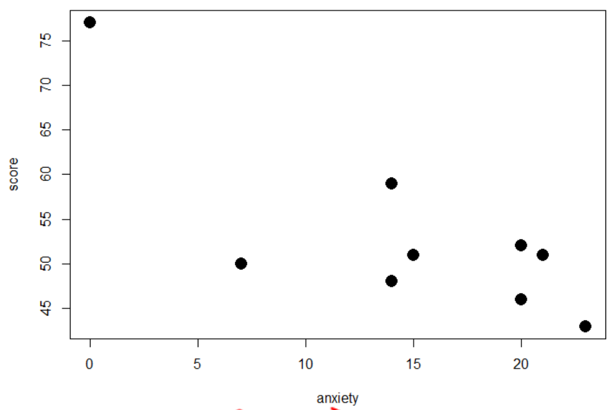
Example 2: The following data was collected comparing score on a measure of test anxiety and exam score.

Measure of test anxiety	23	14	14	0	7	20	20	15	21
Exam score	43	59	48	77	50	52	46	51	51

a. Construct a scatterplot.
Commands:

```
anxiety=c(23,14,14,0,7,20,20,15,21)
score=c(43,59,48,77,50,52,46,51,51)
plot(anxiety,score,cex=2,pch=16)
```

Result:



lm(y~x)

b. Find the LSRL and fit it to the scatter plot.
Commands:

lm(score ~ anxiety)

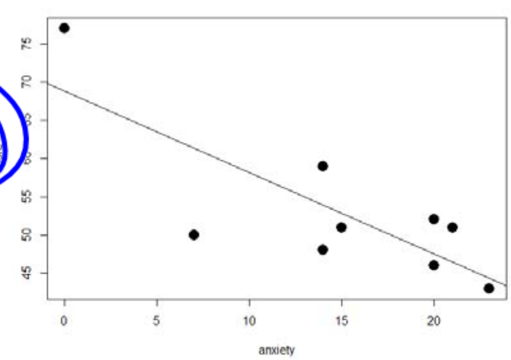
For the line:

abline(lm(score~anxiety))

Results:

```
Call:
lm(formula = score ~ anxiety)

Coefficients:
(Intercept)      anxiety
     68.838       -1.064
```



LSRL: $\hat{y} = 68.838 - 1.064x$

c. Find r and r^2.

Commands:

cor(anxiety, score)

cor(anxiety, score)^2

Answers:

-0.7877

0.6205 => 62.05% of the variation in scores is explained by LSRL

d. Does there appear to be a linear relationship between the two variables?

YES

e. Based on what you found, would you characterize the relationship as positive or negative? Strong or weak?

somewhat strong, negative

f. Find the values of the residuals and plot the residuals. What does this plot reveal?

Command:

resid(lm(score~anxiety))

Result:

```
        1         2         3         4          5         6         7
-1.371724  5.054435 -5.945565  8.161794 -11.391885  4.436996 -1.563004
        8         9
-1.881804  4.500756
```

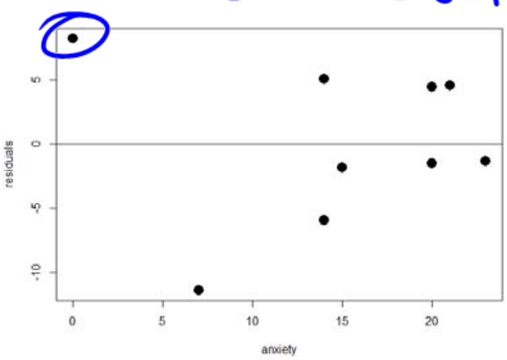
Commands:

residuals = resid(lm(score~anxiety))

plot(anxiety, residuals, cex=2, pch=16)

abline(0,0)

Results:



Plot of the residuals is pretty random, so the plot reveals the the LSRL is a good model for the data

Since the residuals show how far the data falls from the LSRL, examining the values of the residuals will help us to gauge how well the LSRL describes the data. The sum of the residuals is always 0 so the plot will always be centered around the x-axis.

An **outlier** is a value that is well separated from the rest of the data set. An outlier will have a large absolute residual value.

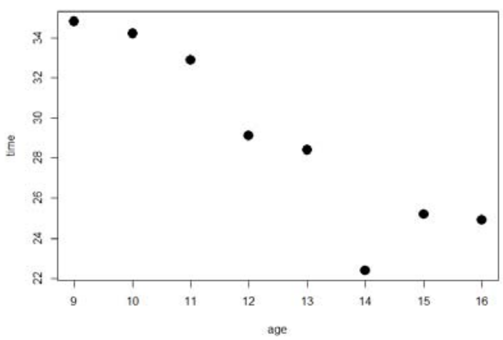
An observation that causes the values of the slope and the intercept in the line of best fit to be considerably different from what they would be if the observation were removed from the data set is said to be **influential**. When the influential is removed, it makes your LSRL look better (fits the data better).

Example 3: Johnny keeps track of his best swimming times for the 50 meter freestyle from each summer swim team season. Here is his data:

x	Age(years)	9	10	11	12	13	14	15	16
y	Time (sec)	34.8	34.2	32.9	29.1	28.4	22.4	25.2	24.9

a. Construct a scatterplot.
Commands:
age=c(9,10,11,12,13,14,15,16)
time=c(34.8,34.2,32.9,29.1,28.4,22.4,25.2,24.9)
plot(age,time,cex=2,pch=16)

Result:



b. Find the LSRL and fit it to the scatter plot.
Commands:

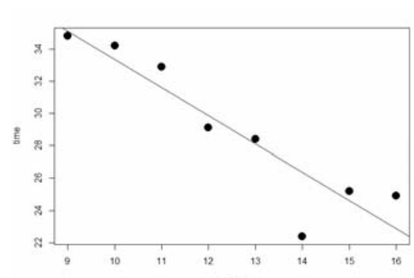
lm(time ~ age)

abline(lm(time ~ age))

Results:

```
Call:
lm(formula = time ~ age)

Coefficients:
(Intercept)          age
     50.788       -1.744
```



LSRL: $\hat{y} = 50.788 - 1.744x$

strong negative association

c. Find r and r^2.

Commands: Answers:

cor(age,time) -0.9196

cor(age,time)^2 0.8457 => 84.57% of the variation in time is explained by LSRL

d. Construct a residual plot then determine if the LSRL is a good model for his data.

Command:

plot(age,resid(lm(time~age)),cex=2,pch=16)

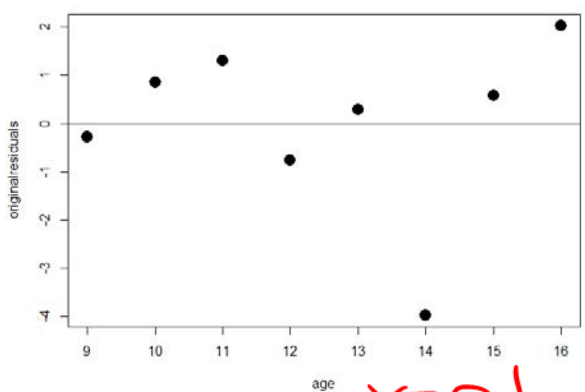
abline(0,0)

Result:

```
         1          2          3          4          5          6          7 
-0.2916667  0.8523810  1.2964286 -0.7595238  0.2845238 -3.9714286  0.5726190 
         8 
 2.0166667 
```

Commands:

Results:



YES! at age 14

e. Is there an influential point (i.e. a point that is an outlier and has a significant impact on the line of best fit)? If so, identify it, and remove it from your data.

i. Construct a scatterplot.

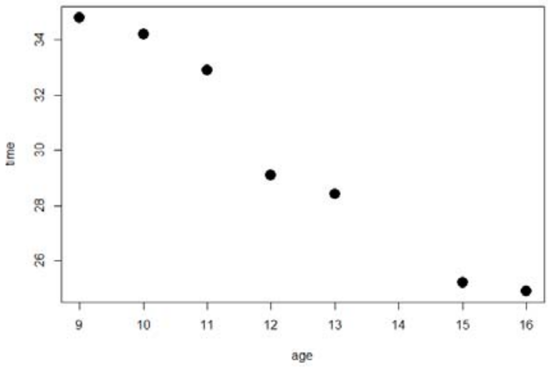
Commands:

age=c(9,10,11,12,13,15,16)

time=c(34.8,34.2,32.9,29.1,28.4,25.2,24.9)

plot(age,time,cex=2,pch=16)

Result:



ii. Find the LSRL and fit it to the scatter plot.

Commands:

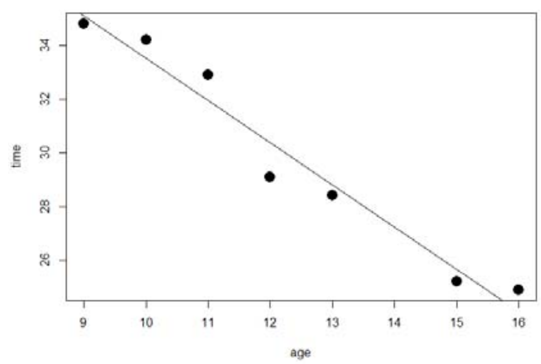
lm(time~age)

Results:

```
Call:
lm(formula = time ~ age)

Coefficients:
(Intercept)          age
     49.234       -1.571
```

abline(lm(time~age))



LSRL:

$\hat{y} = 49.234 - 1.571x$

iii. Find r and r^2.

Commands:

cor(age,time)

cor(age,time)^2

Answers:

NEW	OLD
-0.9795	-0.9196
0.9594	0.8457

Better!

iv. Construct a residual plot then determine if the LSRL is a good model for his data.
Command:

resid (lm (time ~ age))

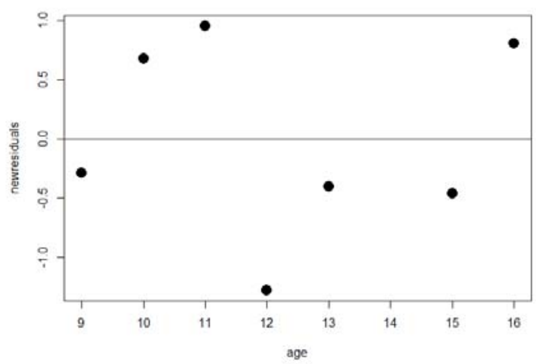
Result:

```
         1          2          3          4          5          6          7
-0.2916667  0.6797101  0.9510870 -1.2775362 -0.4061594 -0.4634058  0.8079710
```

Commands:

plot (age, resid(lm(time ~ age)), cex = 2, pch = 16)
abline (0,0)

Results:



There are many possible justifications for removing the point (14, 22.4) from the data that Johnny collected. The most likely reasons are suspicion that the data point was collected incorrectly or perhaps outside factors, such as the length of the pool being incorrectly measured or a defect in the timer used.