Chapter 2. Looking at the Relationships

2.1 Relationships

Often, researchers are interested in discovering relationships between two variables.

Explanatory Variable and Response Variable

In research, we wish to determine how varying amount of an **explanatory variable** affects the value of a **response variable**

Note: Algebra class $\begin{cases} \text{Explanatory Variable} = \\ \text{Response Variable} = \end{cases}$

< Ex > Identify the explanatory variable and response variable for each case.

(a) A sample of students drank different numbers of cans of beer. Thirty minutes later, their blood alcohol levels were measured.

- explanatory variable =
- response variable =

Relationship:

(b) A study says that children exposed to lead are more likely to suffer tooth decay.

Lead Poisoning

Tooth Decay

Relationship:

2.2 Scatterplots

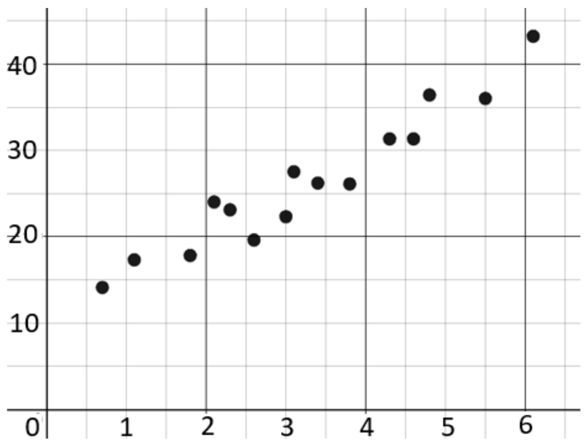
The explanatory variable is also called the "**predictor**" in regression analysis.

Scatterplot

A **scatterplot** is a graph that shows the relationship between two quantitative variables measured on the same individuals/objects.

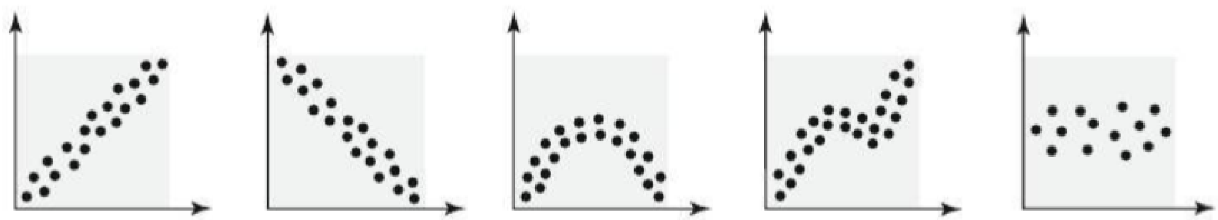
< Ex > Fire Damage: Data Set with 15 fires

$$\begin{cases} y\text{: damage (in \$1,000) from a fire} \\ x\text{: distace (in miles) from the nearest fire statiion} \end{cases}$$



Distance (x)	Damage (y)
3.4	26.2
1.8	17.8
4.6	31.3
2.3	23.1
3.1	27.5
5.5	36.0
0.7	14.1
3.0	22.3
2.6	19.6
4.3	31.3
2.1	24.0
1.1	17.3
6.1	43.2
4.8	36.4
3.8	26.1

Type of Relationship



Positive and Negative Relationship

- **Positively Related**: Whenever the value of x increases, the value of y also increases.
- **Negatively Related**: Whenever the value of x increases, the value of y decreases.

< Ex > Positively or Negatively Related?

(a) amount of time on a treadmill and calories burned

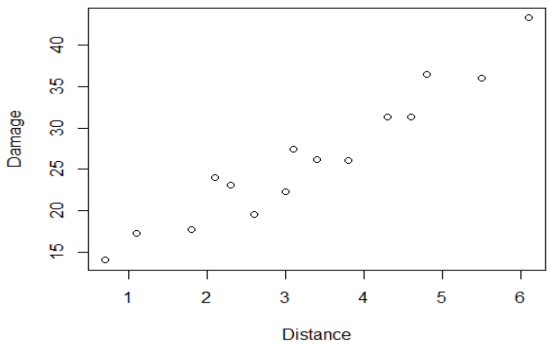
(b) speed of a car and amount of time to reach the destination

< Ex > Scatterplots

(a) Fire Damage and Distance

x : distance (miles) from a nearest fire station
y : fire damage (\$1000)

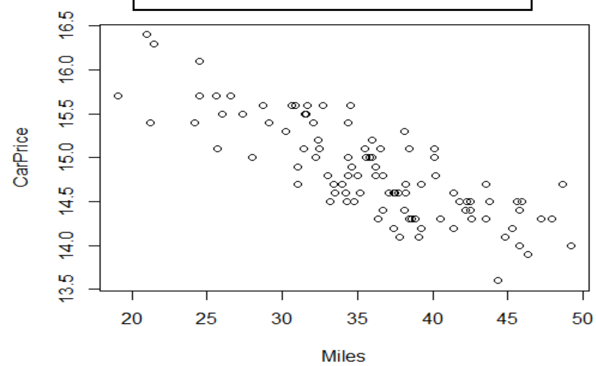
```
> plot(Damage~Distance)
```



(b) Used Car Sale Price and Mileage

x : mileage of a car (1000 miles)
y : sale price of a car (\$1000)

```
> plot(CarPrice~Miles)
```

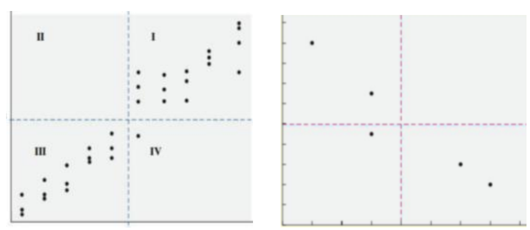


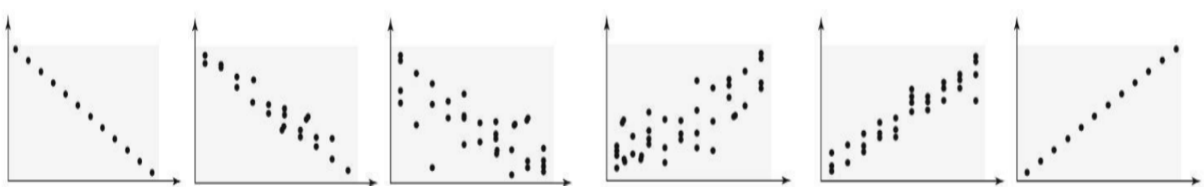
2.3 Correlation

(Sample) Correlation Coefficient, r

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

where s_x: sample standard deviation of x
s_y: sample standard deviation of y





Properties of Correlation Coefficient, r

- It is a unitless and measures the strength of a linear relationship between x and y
- The correlation coefficient (the value) is always between _________ and ________
- r is close to zero : there is _________________________________ between x and y

< Ex > Fire Damage: Correlation

x : distance (miles) from a nearest fire station
y : fire damage ($1000)

```
> cor(Distance, Damage)
[1] 0.9609777
```

Correlation Coefficient:

< Ex > Used Car Sale Price: Correlation

x : mileage of a car (1000 miles)
y : sale price of a car ($1000)

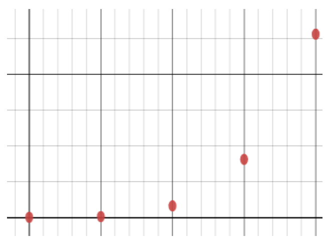
```
> cor(CarPrice, Miles)
[1] -0.805168
```

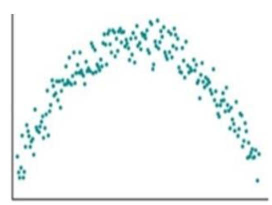
Correlation Coefficient:

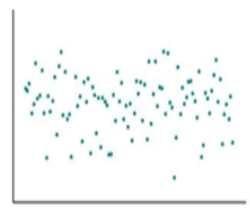
< Ex > What would be the correlation coefficient?

Data Set

x	y
0	0
1	1
2	16
3	81
4	256







<u>Correlation Does Not Imply Cause-and-Effect (Correlation ≠ Causation)</u>

Of data used in a study are observational, we cannot conclude the two variables have a causal relationship. No matter how strong the correlation, there is no way to conclude that one variable (x) causes the other variable (y).

< Ex > Study of "Teenage Birthrate" and "Homicide Rate" : The correlation is 0.9987

2.4 Least-Squares Regression

We want to find the line (model) which describes this linear relationship.

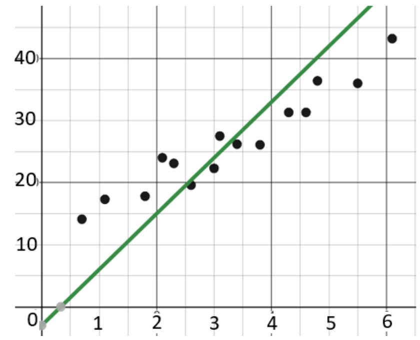
<u>Least Squares Regression Line</u>

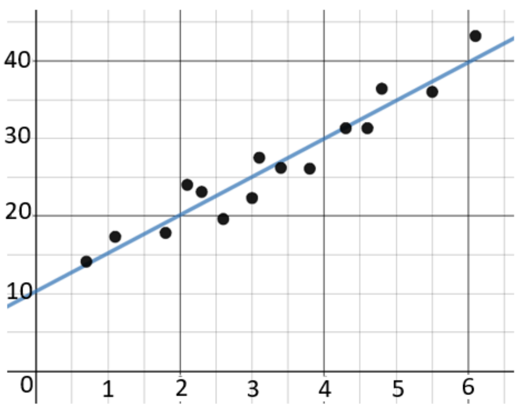
The least-square regression line is the line that minimizes the sum of the squared residuals.

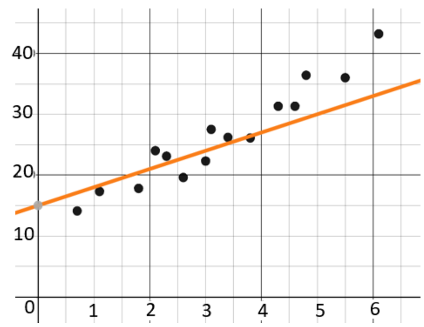
< Ex > Fire Damage: Data Set with 15 fires

$\begin{cases} y\text{: damage (in \$1,000) from a fire} \\ x\text{: distace (in miles) from the nearest fire statiion} \end{cases}$

What is the best line that fits the data the best?







Least Squares Regression Line

The equation of the least-square regression line is given by $\boldsymbol{\hat{y} = b_0 + b_1 x}$

- Slope of the Regression Line: $b_1 = r \cdot \dfrac{s_y}{s_x} = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$
- y intercept of the Regression Line: $b_0 = \bar{y} - b_1\bar{x}$
- $\hat{y}$: predicted value (predicted y value using the regression equation)
- residual $= y - \hat{y}$: vertical distance

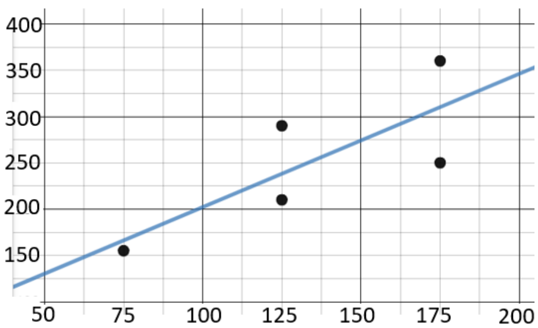
< Ex > Data: a sample of 5 lots $\begin{cases} y\text{: lot sale price (in \$1,000)} \\ x\text{: lot size (in 100 square footage)} \end{cases}$

```
> summary( lm(Price~Size) )

Coefficients:
            Estimate Std. Error t value Pr(>|t
(Intercept)  58.2143    95.2447   0.611    0.5
Size          1.4429     0.6799   2.122    0.1

Residual standard error: 56.88 on 3 degrees of
Multiple R-squared:  0.6002,   Adjusted R-squared:  0.4669
F-statistic: 4.504 on 1 and 3 DF,  p-value: 0.1239
```

```
> lm(Price~Size)

Call:
lm(formula = Price ~ Size)

Coefficients:
(Intercept)         Size
     58.214        1.443
```

(a) Write the Regression Line

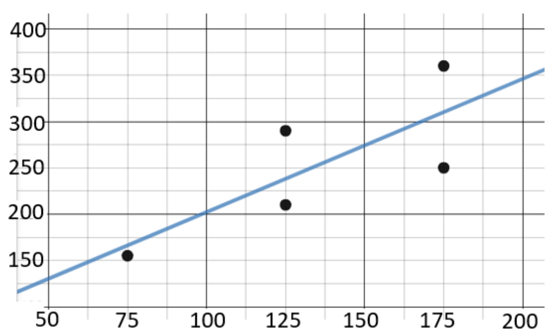
(b) Calculate the predicted value and the residual of a lot that is 7,500-sq.ft. in size.



Data		Predicted Value	Residual
Size (x)	Price (y)	$\hat{y}$	$y - \hat{y}$
75	155		
125	290		
125	210	238.2	−28.2
175	360	310.2	49.8
175	250	310.2	−60.2

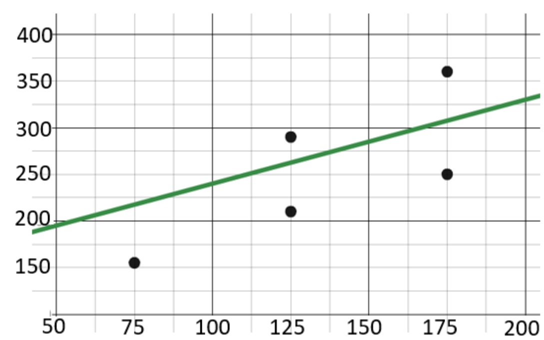
(c) Compare the Sum of the Squared Residuals

Regression Line:



Sum of the Squared Residuals =

Other Line: $\hat{y} = 150 + 0.9x$



Sum of the Squared Residuals = 13481.25

Note: The least-squares regression line is the line that minimizes the residuals sum of squares.

===

<< Side Note: Calculus >> Least Squares Method to find regression coefficients b_0 and b_1

The residual sum of squares (SSE) is written as

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

We need to find b_0 and b_1 such that SSE is minimized. Thus,

$$\frac{\partial \text{SSE}}{\partial b_0} = \frac{\partial \sum_{i=1}^{n}[y_i - (b_0 + b_1 x_i)]^2}{\partial b_0} = 0 \quad \rightarrow \quad \text{Solve for } b_0$$

$$\frac{\partial \text{SSE}}{\partial b_1} = \frac{\partial \sum_{i=1}^{n}[y_i - (b_0 + b_1 x_i)]^2}{\partial b_1} = 0 \quad \rightarrow \quad \text{Replace } b_0 = \bar{y} - b_1\bar{x} \text{ and Solve for } b_1$$

==

< Ex > Fire Damage: Data with 15 fires $\begin{cases} y\text{: damage (in \$1,000) from a fire} \\ x\text{: distance (in miles) from the nearest fire statiion} \end{cases}$

(a) Write the equation of the regression Line.

```
> lm(Damage~Distance)

Coefficients:
(Intercept)     Distance
     10.278        4.919
```

(b) Predict the damage from a fire that occurs 3 miles away from the nearest fire station, using the regression line. (See page 4 for a scatterplot and the regression line)

Predicted Fire Damage: $_________________

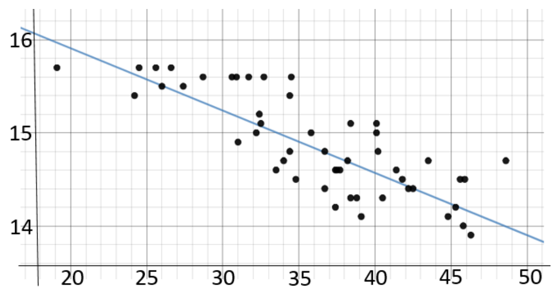
(c) What is the actual damage from a fire that occurred 3 miles away from the nearest fire station?

Portion of the Data

Distance (x)	**Damage (y)**
⋮	⋮
0.7	14.1
3.0	22.3
2.6	19.6
⋮	⋮

(d) What is the residual for a fire that occurs 3 miles away the nearest fire station? What tis the meaning of this residual?

< Ex > Used Car Price: Data with 100 cars $\begin{cases} y\text{: sale price (in \$1,000) of a car} \\ x\text{: mileage (in 1,000 miles) of a car} \end{cases}$



Portion of the Data

Miles (x)	**CarPrice (y)**
⋮	⋮
40.1	15.1
32.4	15.2
⋮	⋮

(a) Write the equation of the regression Line.

```
> lm(CarPrice~Miles)

Coefficients:
(Intercept)        Miles
   17.24873     -0.06686
```

(b) You drove your car 48,000 miles and want to sell it. How much money can you expect?

Interpreting the Slope and the y-Intercept of the Regression Line

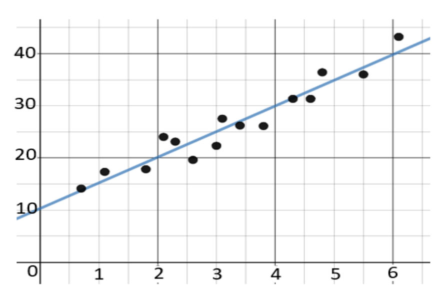
- Slope b_1: predicted change in y for every one unit increase in x
- y-intercept b_0: predicted value of y when $x = 0$

< Ex > Fire Damage: Data with 15 fires

```
> lm(Damage~Distance)

Coefficients:
(Intercept)      Distance
     10.278         4.919
```

$\begin{cases} y\text{: damage (in \$1,000) from a fire} \\ x\text{: distance (in miles) from the nearest fire statiion} \end{cases}$

(a) Interpret the slope of the regression line.



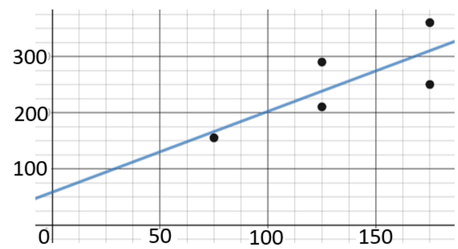
(b) Interpret the y-intercept of the regression line.

Note: To interpret the y-intercept, you should ask a question "Is 0 a reasonable value for the predictor (x)?" If the answer is no, we do not interpret the y-intercept.

< Ex > Data: 5 lots $\begin{cases} y\text{: lot sale price (in \$1,000)} \\ x\text{: lot size (in 100 square footage)} \end{cases}$

Regression Line: $\widehat{\text{Price}} = 58.2 + 1.44 \cdot \text{Size}$

```
> lm(Price~Size)

Coefficients:
(Intercept)         Size
     58.214        1.443
```



Is the y-intercept meaningful in this case?

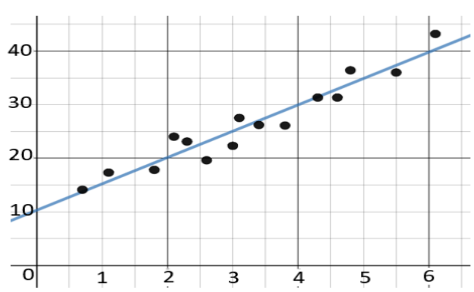
< Ex > Fire Damage: Data with 15 fires

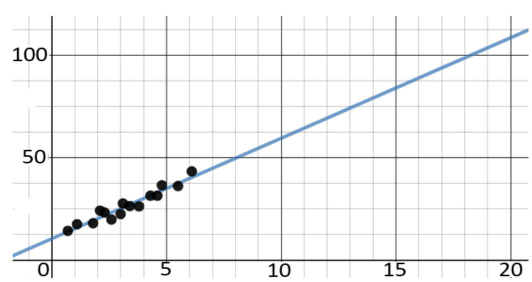
$\begin{cases} y\text{: damage (in \$1,000) from a fire} \\ x\text{: distance (in miles) from the nearest fire statiion} \end{cases}$

The predicted damage from a fire that occurs 30 miles away from the nearest fire station is

$\widehat{\text{Damage}} = 10.28 + 4.92 \cdot (20) = 108.68$ That is, the fire damage would be \$108,680

However, this prediction may not be accurate. Explain why.





Extrapolation: Reaching Beyond the Data

Use of a regression line for predictions far outside the scope of the model.

Assessing the Model

We need to assess whether the model is good or bad to make a prediction.

R-Squared Value (of Coefficient of Determination)

The R-squared value measures the proportion of total variation in the response variable that is explained by the regression line

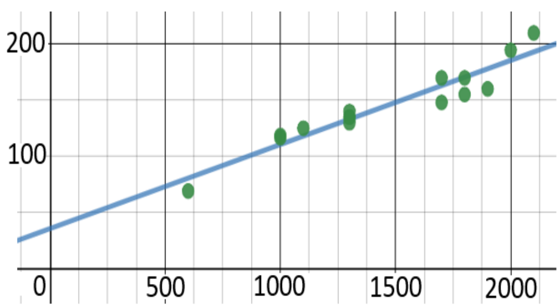
< Ex > We want to predict the sale price of a house. We have two potential predictors.

- Response Variable y : Sale Price of a House
- Two Predictors: $\begin{cases} \text{Size of a house (in sq. ft.)} \\ \text{Number of Bathrooms in a House} \end{cases}$

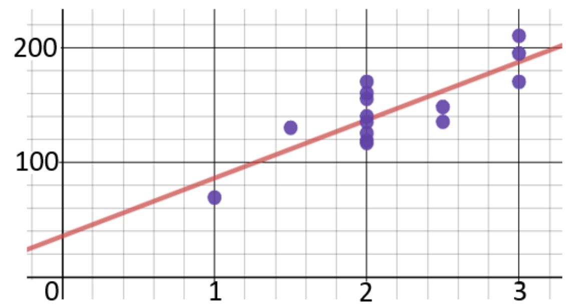
(a) Which would be a better model (better predictor of a house price)? ____________________

(b) A data set contains 16 houses. Find the R-squared value for each case

Predictor: House Size



Predictor: Number of Bathrooms



R-square Value:

R-square Value:

```
> summary(lm(Price~Size))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.546687  10.225378   3.476   0.0041 **
Size         0.075005   0.006732  11.142 5.06e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.93 on 13 degrees of freedom
Multiple R-squared:  0.9052,	Adjusted R-squared:  0.8979
F-statistic: 124.1 on 1 and 13 DF,  p-value: 5.061e-08
```

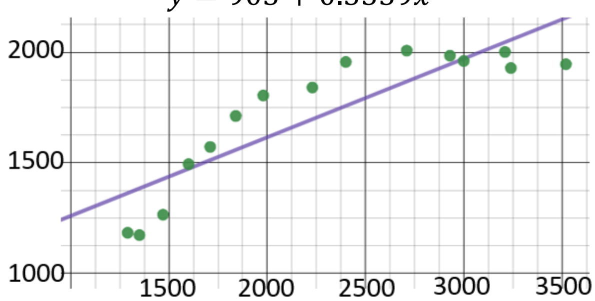
```
> summary(lm(Price~Bath))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   35.470     21.648   1.639 0.125276
Bath          50.577      9.697   5.216 0.000167 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.19 on 13 degrees of freedom
Multiple R-squared:  0.6766,   Adjusted R-squared:  0.6518
F-statistic:  27.2 on 1 and 13 DF,  p-value: 0.0001665
```

< Ex > Data with 15 houses

$$\begin{cases} y\text{: monthly electrical usage (in kilowatthours) of a house} \\ x\text{: size (in square footage) of a house} \end{cases}$$

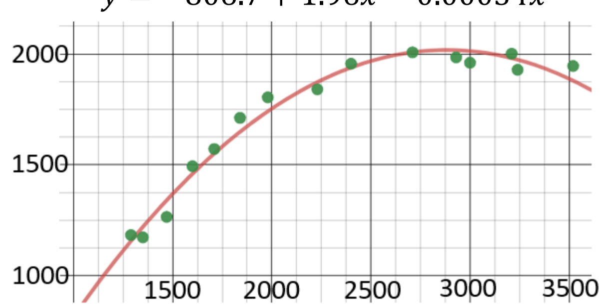
Fit a Linear Model

$\hat{y} = 903 + 0.3559x$



Fit a Quadratic Model

$\hat{y} = -806.7 + 1.96x - 0.00034x^2$



R-square Value:

R-square Value:

```
> summary( lm(Usage ~ Size) )
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 903.01147  132.13594   6.834 1.20e-05 ***
Size          0.35594    0.05477   6.498 2.01e-05 ***
---
Residual standard error: 155.3 on 13 degrees of freedom
Multiple R-squared:  0.7646,   Adjusted R-squared:  0.7465
F-statistic: 42.23 on 1 and 13 DF,  p-value: 2.009e-05
```

```
> summary( lm(Usage ~ Size + SizeSq) )
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.067e+02  1.669e+02  -4.834 0.000409 ***
Size         1.962e+00  1.525e-01  12.861 2.23e-08 ***
SizeSq      -3.404e-04  3.212e-05 -10.599 1.90e-07 ***
---
Residual standard error: 50.2 on 12 degrees of freedom
Multiple R-squared:  0.9773,   Adjusted R-squared:  0.9735
F-statistic: 258.1 on 2 and 12 DF,  p-value: 1.375e-10
```

2.5 Cautions about Correlation and Regression

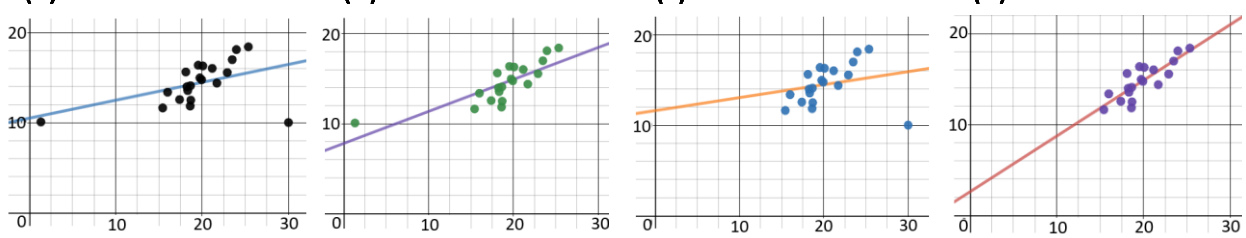
Outliers and Influential Observations in Regression

An **outlier** is an observation that lies outside the overall pattern of the other observations. Points that are outliers in the y direction of a scatterplot have large regression residuals, but other outliers need not have large residuals.

An observation is **Influential** if removing it would markedly change the result of the statistical calculation (e.g. slope, y-intercept, the correlation coefficient, etc.). Points that are outlies in the x direction of a scatterplot are often influential.

< Ex > Influential Observations?

(a) 21 Observations **(b)** 20 Observations **(c)** 20 Observations **(d)** 19 Observations



(a)	(b)	(c)	(d)
Correlation: 0.45	Correlation: 0.79	Correlation: 0.22	Correlation: 0.83
$\hat{y} = 10.5 + 0.2x$	$\hat{y} = 7.8 + 0.36x$	$\hat{y} = 11.6 + 0.14x$	$\hat{y} = 2.6 + 0.61x$

2.6 Data Analysis for Two-Way Table

We want to identify **association between two categorical/quualitative data**

< Ex > Titanic Data with 1,309 passengers (A portion of the data is shown below)

Contingency Table or Two-Way table

Survival Status	Passenger Ticket Class			Total
	First	Second	Third	
Died	123	158	528	809
Survived	200	119	181	500
Total	323	277	709	1309

- Row Variable: Survival Status
- Column Variable: Ticket Class

Class	Status
First	Survived
First	Survived
Third	Died
First	Died
First	Died

Joint Distribution and Marginal Distribution

For each cell, we compute a proportion by dividing the cell entry by the total sample size. The collection of these proportions is the **Joint Distribution**

When we examine he distribution of a single variable in a two-say table, we are looking at a **Marginal Distribution**

< Ex > Titanic Data with 1,309 passengers

(a) Find the joint distribution.

	Passenger Ticket Class		
Survival Status	First	Second	Third
Died			
Survived			

```
> titanic.table = table(Status, Class)
> titanic.table
          Class
Status     First Second Third
  Died       103    146   370
  Survived   181    115   131
> prop.table(titanic.table)
          Class
Status          First     Second      Third
  Died     0.09847036 0.13957935 0.35372849
  Survived 0.17304015 0.10994264 0.12523901
```

(b) Find the marginal distribution of the passenger ticket class.

Passenger Ticket Class		
First	Second	Third

```
> prop.table( table(Class) )
Class
    First    Second     Third
0.2715105 0.2495220 0.4789675
```

(c) Find the marginal distribution of the passenger ticket class.

Survival Status	
Died	Survived

```
> prop.table( table(Status) )
Status
     Died  Survived
0.5917782 0.4082218
```

(d) Find the conditional distribution of survival status for the first class passengers.

Survival Status for the First-Class Passengers	
Died	Survived

```
> prop.table(titanic.table, margin=2)
          Class
Status         First    Second     Third
  Died     0.3626761 0.5593870 0.7385230
  Survived 0.6373239 0.4406130 0.2614770
```

`margin`: index, or vector of indices to generate margin for

- `margin=1` : row
- `margin=2` : column

```
> prop.table(titanic.table, margin=1)
          Class
Status         First    Second     Third
  Died     0.1663974 0.2358643 0.5977383
  Survived 0.4238876 0.2693208 0.3067916
```

(e) Find the conditional distribution of survival status for the first class passengers.

Survival Status for the Third-Class Passengers	
Died	Survived

<u>Conditional Distribution</u>

When we condition on the value of one variable and calculate the distribution of the other variable, we obtain the **Conditional Distribution**

```
> barplot( prop.table(titanic.table, margin=2), beside=TRUE )
```