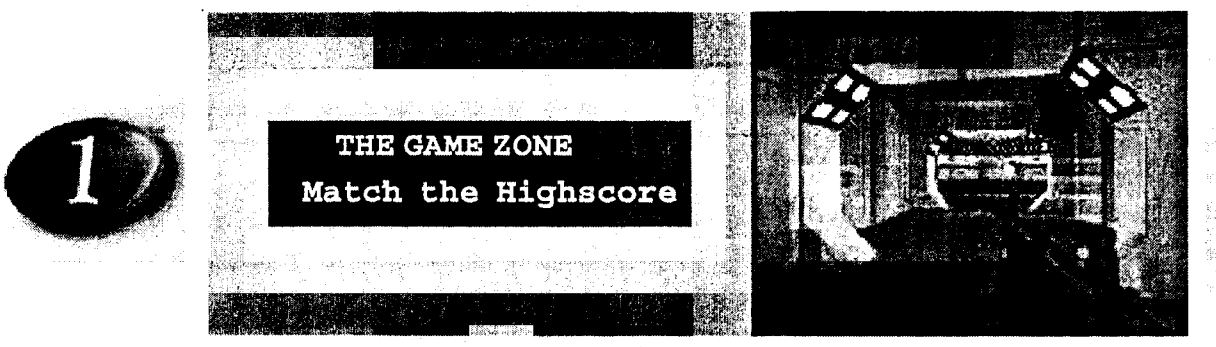
Regression Correlation

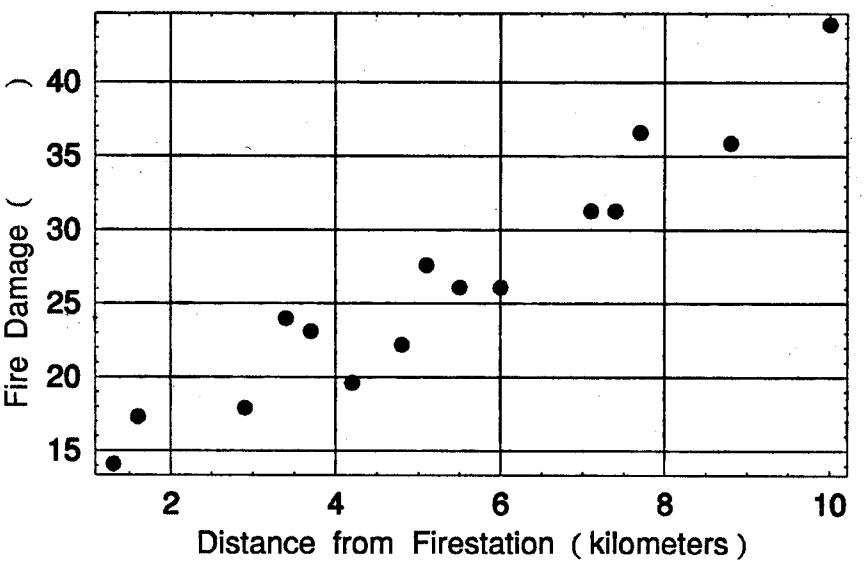
Linear Regression

Get Palette!



1.1 The GAME - Minimize the Failure

The graphic below shows data concerning fire damage compared to the distance from the fire station.



Is there any relationship between these two variables? If yes, can we use this relationship to predict the fire damage if the house is 7.5km away form the firestation?

The Game:

The ambition of this game is to get a feeling for the method of linear regression.
Predict: How big is the damage in various distances?

Therefore two players should enter graphically a line which is most suitable to **all** datapoints. This lines can be used to predict fire damage to a not given distance, e.g. 7.5 km. Try to beat your neighbour in finding the best model (smallest error) for the given sets of data points!

Instructions to the game:

Step 1: Hit the **NEW GAME** button to create a new game board.

Step 2: Both players have to enter two coordinates to draw their assumed regression line.

Hint: You can CTRL - click on the graphic to enter the coordinates.

Step 3: Show the lines by hitting the button **SHOW LINES.**

Step 4: Press **GO ! ! ! ! !**. The button calculates the score = standard error of estimate $= \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$ of your regression line. The winner has **lower** scores.

NEW GAME **Click to start a new game!**

Questions About the Game

1. Which conditions must satisfy the data points, that there is no error (score = 0)?

Answer: It is only possible, if all points lie exactly on the guessed line. Although, it is not possible in this game.

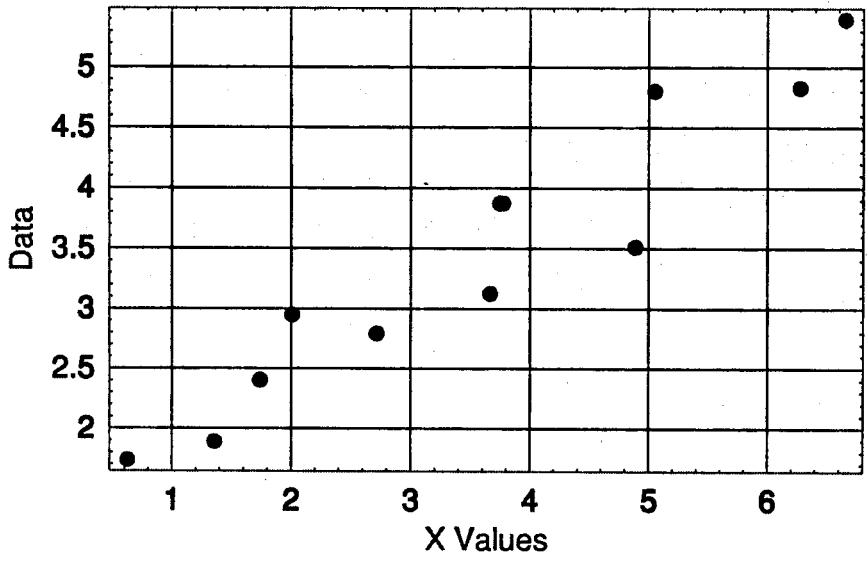
2

DEFINITION
Linear Regression

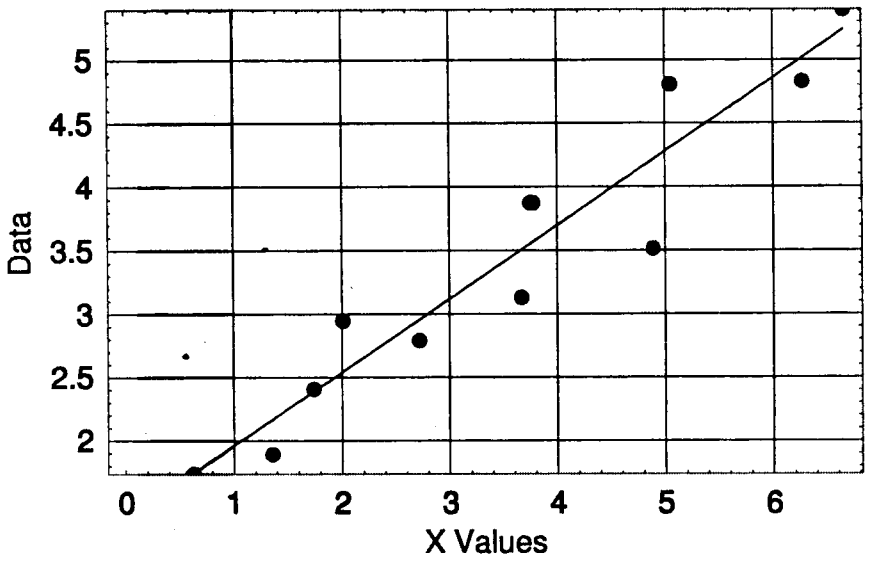
2.1 Graphical Explanation

Open / Close

The purpose of the regression analysis is to find a tendency between two variables X and Y. **X** is called the **predictor variable, Y** is called the **response variable** (data corresponding to X):

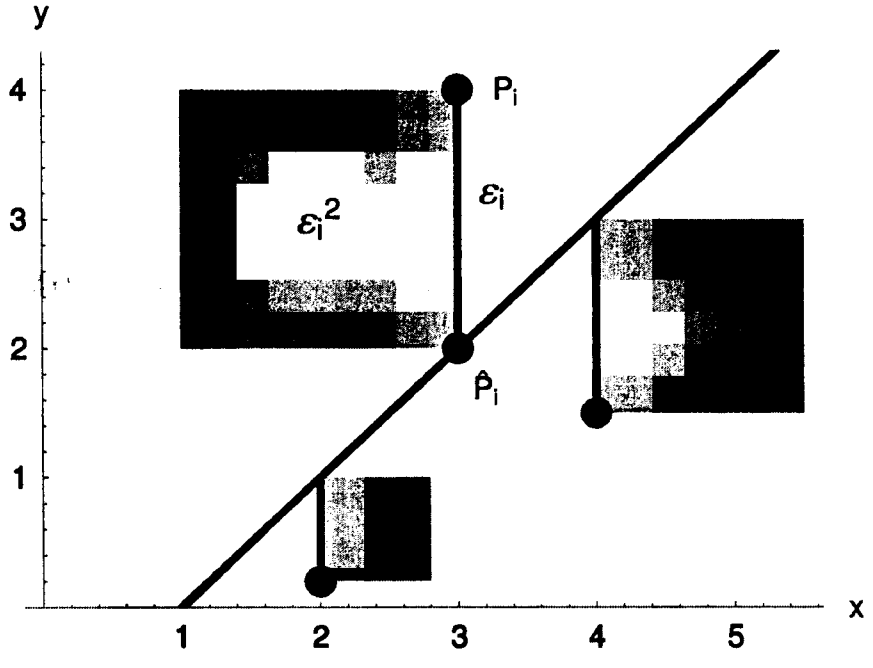


The **linear regression** tries to find a **line**, called the line of regression, which describes the relationship between these two variables X and Y best:



The method for finding the regression line is called the method of least squares:

The difference ε_i of the true datapoints $\mathbf{P(x_i,y_i)}$ and the corresponding predicted points on the regression line $\mathbf{\hat{P}(x_i,\hat{y}_i)}$ gets squared. By partial derivations the **sum of squares** (SSE) could be minimized. The line which SSE gives this minimum is called the line of regression or least-squares line.



2.2 Formulas

Open / Close

The line of regression is defined by the deterministic model:

$$\hat{y} = \beta_0 + \beta_1 x$$

$\hat{y}$ ... predicted y value on the regression line to a given x
β_0 ... the intersection of the regression line and the y-axes
β_1 ... the slope of the regression line

The difference between

β_1 and β_0 are calculated by the formulas

$$\beta_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

$$\beta_0 = \frac{\sum y_i}{n} - \beta_1 \frac{\sum x_i}{n}$$

x_i ... the x-coordinate of the i-th data point
y_i ... the y-coordinate of the i-th data point
n ... number of data points

Property of the regression line $\hat{y} = \beta_0 + \beta_1 x$:
The sum of its squared errors, $SSE = \Sigma (y_i - \hat{y}_i)^2$ is smaller than for any other straight-line model. The sum of errors, $SE = \Sigma (y_i - \hat{y}_i)$ equals 0.
$(y_i - \hat{y}_i)$, the differences between the observed y_i and the predicted values $\hat{y}_i$ are called residuals.

Abbreviations in textbooks are, SS... sum of squares, $\bar{x} = \Sigma x_i/n$, $\bar{y} = \Sigma y_i/n$,
$SS_{xy} = \Sigma (x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n$
$SS_{xx} = \Sigma (x_i - \bar{x})(x_i - \bar{x}) = \Sigma x_i^2 - (\Sigma x_i)^2/n$

$$\beta_1 = \frac{SS_{xy}}{SS_{xx}}, \quad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

Sometimes a scatter diagram clearly indicates the existence of a linear relationship between x and y, but it can happen that the points are widely scattered around the regression line.

One method to measure the spread of a set of points from the least-squares line is the **standard error of estimate S_e:**

$$S_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

y_i ... the y-coordinates of the given datapoints
$\hat{y}_i$ ... the predicted y-coordinate on the regression line
n ... number of data points, $n \geq 3$

The nearer the scatter points lie to the regression line, the smaller S_e will be.

Abbreviations in textbooks are SSE... sum of squared error
$SS_{yy} = \Sigma (y_i - \bar{y})(y_i - \bar{y}) = \Sigma y_i^2 - (\Sigma y_i)^2/n$
$SSE = \Sigma (y_i - \hat{y})^2 = SSyy - \beta_1 SS_{xy}$

$$S_e = \sqrt{\frac{SSE}{n-2}}$$

S_e is often called Root Mean Square Error or **Root MSE**.

Interpretation: 95% of the observed y values will lie within the interval $\hat{y} \pm 2 S_e$ for each x.

2.3 Estimating the Mean Value of y Confidence Interval

Open / Close

All points of the data can be described by the following probabilistic model:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

ε ... the random error component.

The random error ε is the vertical difference between the observed value y and the predicted value $\hat{y} = \beta_0 + \beta_1 x$.

4 basic assumptions about the distribution of ε are necessary.
- ε is normally distributed
- $\mu = 0$, this means, the mean value of y for a given x equals $\hat{y} = \beta_0 + \beta_1 x$.

- same variance σ^2 of ε for all x values, σ is estimated by $S_e = \sqrt{\frac{(y_i - \hat{y})^2}{n-2}}$
- each y value has its own ε value.

A **c confidence interval** for the **mean value** of y is given by the formula:

$$\hat{y} - t_c S_e \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{SS_{xx}}} \leq y \leq \hat{y} + t_c S_e \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{SS_{xx}}}$$

$\hat{y}$ ... the predicted value of y on the regression line for the specified x value
t_c ... critical value from the Student´s t distribution for a c confidence level using n - 2 degrees of freedom
$\bar{x}$ ... $\Sigma x_i/n$, mean of x values
n ... number of data points
S_e ... $\sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$, standard error of estimate
SS_{xx} ... $\sum x_i^2 - \frac{(\sum x_i)^2}{n}$

Interpretation of the confidence interval:
An x value is given. The formula above estimates the mean value of y for a **very large number** of experiments.
E.g. a department store spends x = $ 12,000.- on advertising. Which confidence interval encloses the **mean** monthly sales y for that x?

2.4 Predicting a new individual y value Prediction Interval

Open / Close

The prediction interval **estimates a new individual y value** for a given x-value without being interested in the mean of this y value. E.g. a department store spends x = $ 12,000.- on advertising next month.
Which prediction interval encloses the store's sales revenue for that month?

A **c prediction interval** for y is given by the formula:

$$\hat{y} - t_c S_e \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SS_x}} \leq$$

$$y \leq$$

$$\hat{y} + t_c S_e \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SS_x}}$$

$\hat{y}$ ... the predicted value of y on the regression line for the specified x value

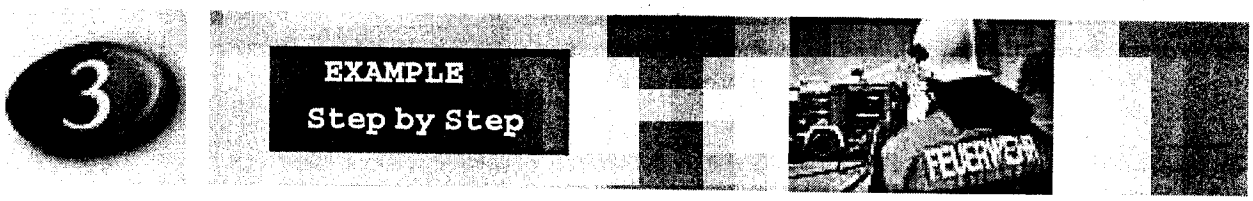
t_c ... critical value from the Student´s t distribution for a c confidence level using n - 2 degrees of freedom

$\bar{x}$ ... $\Sigma x_i/n$, mean of x values

n ... number of data points

S_e... $\sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$, standard error of estimate

SS_x... $\sum x_i^2 - \frac{(\sum x_i)^2}{n}$



3.1 Data File: Predicting Fire Damage

Open / Close

The Fire Station Example The file **firestation.dat** (column 1: distance in km, column 2: damage in $1000) provides data describing the relationship between the distance from fire station in kilometers (x-data, predictor variable) and the fire damage (y-data, response variable).

(a) Read the file and plot the data.

(b) Determine the least-squares line for y as response variable. Plot the linear model with the data. How many $ is the standard error of this liner model?

(c) Predict the fire damage, if the house is 7.5 km away from the fire station.

Solution:

(a) Read in the data from firestation.dat. x-values are the distance in km, y-values are the damage in $1000.

Click the Import button and select the file.

Input ▷

```
data = ReadList[ "D:\\MathDesktopLineFeed\\
     MathDesktopStatistics\\English\\RegressionCorrelation
     \\RegressionCorrelSampleFiles\\FIRESTATION.dat" ,
   Expression] [[1]]
(* Fileformat y or {x,y}:
   {1,-2.2,3} or { {1,2}, {-2.3,3} } *)
```

```
{{5.5, 26.1}, {2.9, 17.9}, {7.4, 31.3}, {3.7, 23.1}, {5.1, 27.6},
 {8.8, 35.9}, {1.3, 14.1}, {4.8, 22.2}, {4.2, 19.6}, {7.1, 31.3},
 {3.4, 24}, {1.6, 17.3}, {10, 43.9}, {7.7, 36.6}, {6, 26.1}}
```

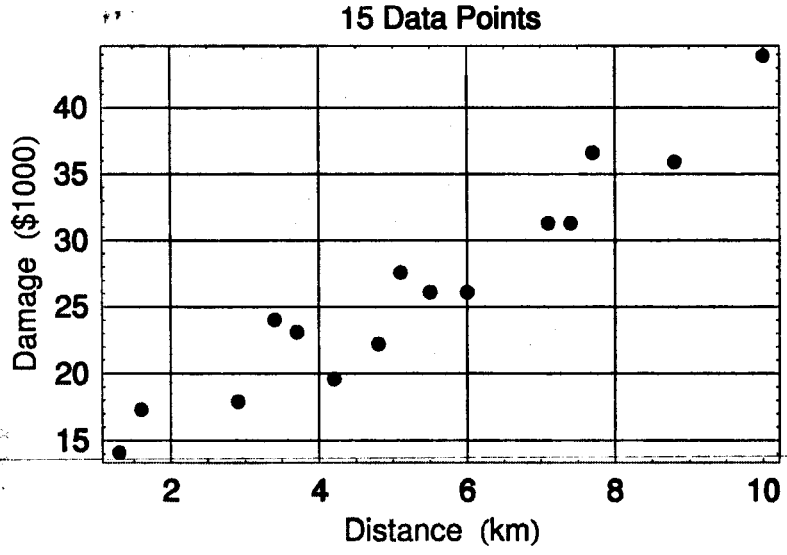
More details ▼.

How do the data look like? Visualize the data. Label the x-axis with Distance (km) and the y-axis with Damage ($1000).

Use the Plot Data button.
Edit the FrameLabel.

Input ▷

```
MDPlotData[data ,
FrameLabel → {"Distance  (km)", "Damage  ($1000)"},
 PlotStyle → {Red, PointSize[0.02]}];
```



More details ▾ .

(b) Determine the least-squares line corresponding to the data. The line is stored in regrLineY[x].

Click the Regr Line button to calculate the linear regression.

```
Switch to X[y];
Clear[x];
regrLineY[x_] = Fit[data, {1, x}, x]
```

```
10.1567 + 3.07735 x
```

Answer: The regression line reads 10.1567+3.07735 x

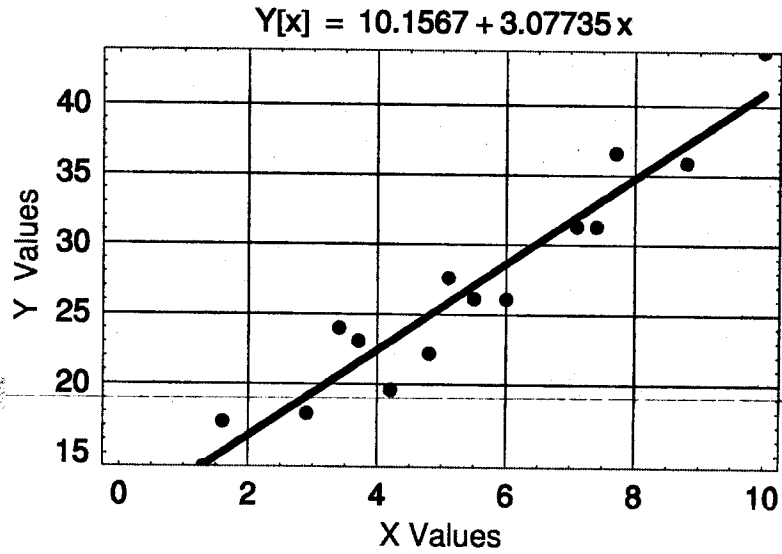
Graph the data together with the least-squares line regrLineY[x]

Use the Plot Regr Line button.

```
Both Least Squares Lines;
Clear[x]; regrLineY[x_] = Fit[data, {1, x}, x];
```

Input ▷

```
MDSPlotDataRegressionLineY[data, {x, regrLineY[x]},
 PointStyle → {Red, PointSize[0.02]},
 PlotStyle → {{DarkGreen, Thickness[0.01]}}];
```



More details ▼.

The standard error S_e of the model measures the spread of the damage costs in thousand $ about the least-squares line. .

Click the Report button.
The value of S_e with other square sums is calculated with the Report button.

Input ▷
```
Switch to X[y];
MDSLinearRegressionReportY[data];
```

Linear Regression Report Y[x]

S_e	$\bar{x}$	$\bar{y}$	SS_{xy}
$\sqrt{\frac{\Sigma(y_i-\hat{y}_i)^2}{n-2}}$	$\frac{\Sigma x_i}{n}$	$\frac{\Sigma y_i}{n}$	$\Sigma x_i\, y_i - \frac{\Sigma x_i\, \Sigma y_i}{n}$
2.344	5.3	26.4667	281.27
SS_{xx}	Slope β_1	yint β_0	x - range
$\Sigma(x_i^2) - \frac{(\Sigma x_i)^2}{n}$	$\frac{SS_{xy}}{SS_{xx}}$	$\bar{y} - \beta_1 \bar{x}$	$\{x_{min}, x_{max}\}$
91.4	3.07735	10.1567	{1.3, 10.}

Answer: The standard error is $2344.

More details ▼.

(c) Finally, predict the fire damage in a distance of 7.5 km.
Note: x values **outside** the x-range of the data may lead to tremendous error of of estimation of y.
Take a look at the report. The x-range of the data is always included. In this case the x-range is {1.3,10} km.

The Regr[□] button calculates the costs.

Input ▷
```
Switch to X[y];
Clear[x];
regrLineY[x_] = Fit[data, {1, x}, x];
regrLineY[7.5]  (* xvalue MUST be in data x-range *)
```

```
33.2368
```

Answer: The mean value for fire damage is $ 33,236 , if the firestation is 7.5 km away.

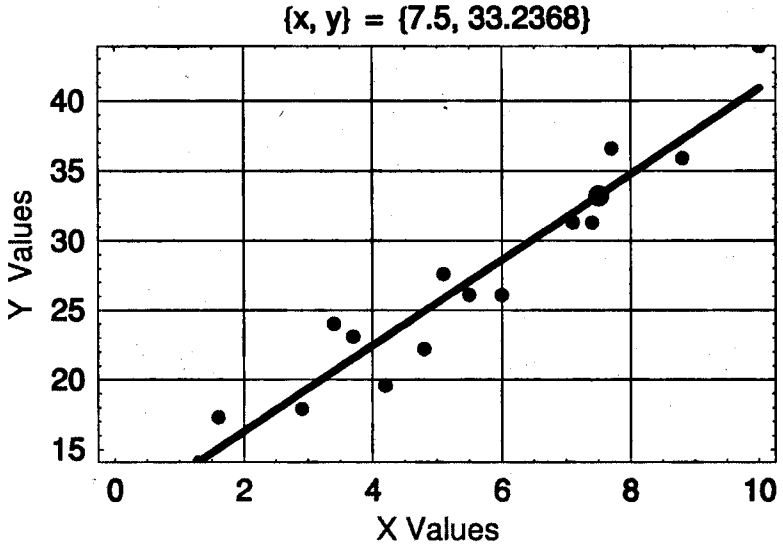
Visualize your result.

Click the Point button and enter 7.5.
The x,y coordinates are saved in coord.

```
Switch to X[y] ;
xvalue = 7.5; (* xvalue MUST be in data x-range *)
```

Input ▷

```
MDSPlotDataRegressionLineYPt[data, xvalue,
 Epilog → {Blue, PointSize[.03]}];
```



More details ▼.

3.2 Quality of Predicting

Open / Close

Determine Quality Example Using the line of regression, it is possible to predict fire damage in relation to the distance. But assurances want to know, between which y values the fire damage in thousand $ will vary for a given distance x, using a 98% significance level. Use the data from the file firestation.dat.

(a) Read the datafile.

(b) Find a 98% confidence interval for the mean damage in 7.5 km distance.
Calculate a prediction interval for the damage in 7.5 km distance.

(c) Determine the least-squares line for the damage as independent variable x and the distance as dependent variable y. A newspaper reports a damage of $ 35 000,- in which distance did the fire burn?
Plot both regression lines. For which distance and which amount of money are the predictions most reliable?

Solution:

(a) Read in the data from firestation.dat.

Click the Import button and select the file FIRESTATION.dat

Input ▷

```
data = ReadList[ "D:\\MathDesktopLineFeed\\
     MathDesktopStatistics\\English\\RegressionCorrelation
     \\RegressionCorrelSampleFiles\\FIRESTATION.dat" ,
   Expression] [[1]]
(* Fileformat y or {x,y}:
   {1,-2.2,3} or { {1,2}, {-2.3,3} } *)
```

```
{{5.5, 26.1}, {2.9, 17.9}, {7.4, 31.3}, {3.7, 23.1}, {5.1, 27.6},
 {8.8, 35.9}, {1.3, 14.1}, {4.8, 22.2}, {4.2, 19.6}, {7.1, 31.3},
 {3.4, 24}, {1.6, 17.3}, {10, 43.9}, {7.7, 36.6}, {6, 26.1}}
```

More details ▼.

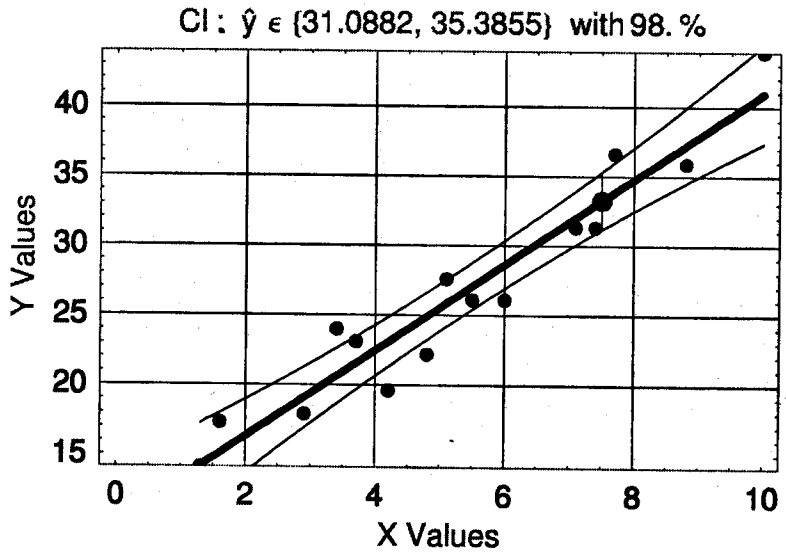
(b) The damage in 7.5 km distance is the y value on the regression line for x = 7.5. For many fires in 7.5 km distance the mean damage in 1000$ lies in an interval according to the 98% confidence interval.

Use the y Confidence Interval button. Enter 7.5 and **change** 0.95 to 0.98.
The blue curves are the borders of the confidence interval.

Input ▷

```
xvalue = 7.5; (* xvalue MUST be in data x-range *)
ConfLevel = 0.98 ;

MDSLinearRegressionCI[data, xvalue, ConfLevel];
```



Answer: With a probability of 98% the mean fire damage of a house, which is 7.5 km away from the fire station, will be between 31,088 $ and 35,385 $.

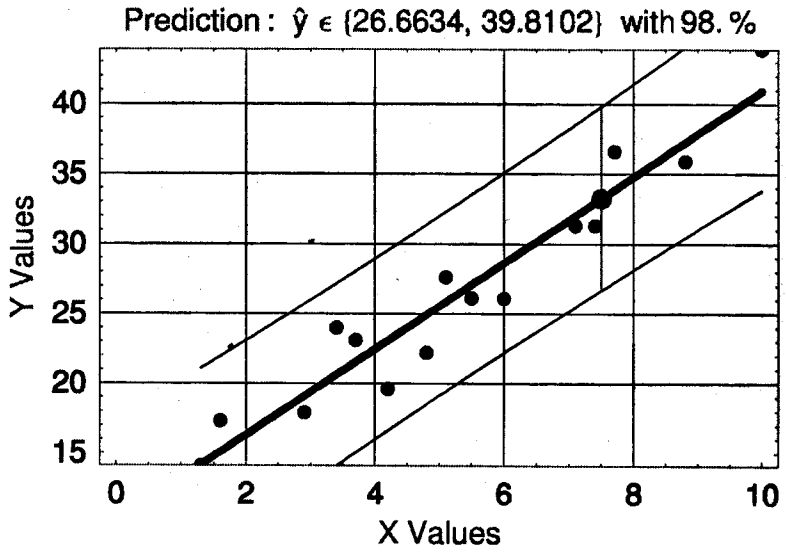
To predict the damage for a particular distance for a single event, x = 7.5 km, calculate the **98% prediction interval.**

Use the y Prediction Int button. Enter 7.5 and **change** 0.95 to 0.98.

Input ▷

```
xvalue = 7.5; (* xvalue MUST be in data x-range *)
ConfLevel = 0.98 ;

MDSLinearRegrPredictionInt[data, xvalue, ConfLevel];
```



Answer: With a probability of 98% the predicted fire damage of a house, which is 7.5 km away from the fire station, will be between 26,663 $ and 39,810 $.

More details ▾ .

(c) Now the least squares are formed with respect to the y-axis, the costs of the damage. The regression line is stored in regrLineX[y].

Use the Regr Line button. Click **Switch** to X[y].

```
Switch to Y[x] ;
```

Input ▷
```
Clear[y];
regrLineX[y_] =
 Fit[Transpose[RotateLeft[Transpose[data]]], {1, y}, y]
```

```
-2.64486 + 0.300184 y
```

When a newspaper reports a fire damage of $ 35 000,- in which distance did the fire burn?

Type regrLineX[35] because the units are 1000 $.

Input ▷
```
regrLineX[35]
```

```
7.86157
```

Answer: In a distance of 7.8 km the damage is about $ 35,000.

More details ▼.

Visualize both regression lines, regrLineY[x] and regrLineX[y]. The intercept of the two lines gives you the distance and amount of money where the predictions are most reliable.

Use the Plot Regr Line button and **click** Both Least-Squares Lines.
In order to draw regrLineX[y] the line was converted to regrLineX[x] by the Solve command. {x̄, ȳ} is the intercept.

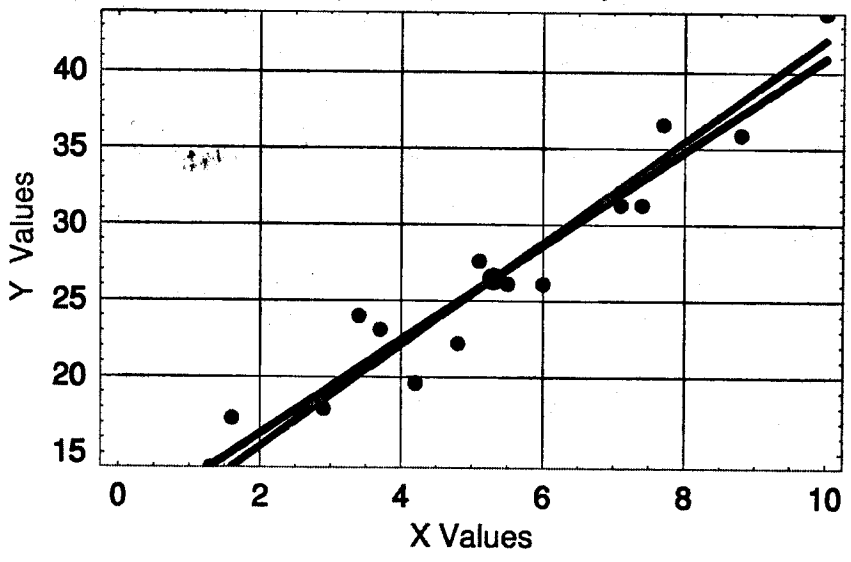
```
Switch to Y[x] ;
MDSPlotDataRegressionLineXY[data,
 PointStyle → {Red, PointSize[0.02]},
 PlotStyle → { {DarkGreen, Thickness[0.01]},
 {CadetBlue, Thickness[0.01]} } ];
```

Input ▷

$$Y[x] = 10.1567 + 3.07735\,x$$

$$X[y] = -2.64486 + 0.300184\,y$$

$$\{\bar{x}, \bar{y}\} = \{5.3, 26.4667\}$$



Answer: The intercept of the two lines is {5.3, 26.46}. The prediction of damage for 5.3 km or the prediction of distance for the damage of $ 26,4667 is most reliable.

More details ▼.

SUMMARY & INTERNET Resources

SUMMARY

The problem of linear regression is to find the "best" linear function representing given {x,y} data.

The method widely used is the **least-squares method** $\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}$. It says the line we fit to the data points must be such that the sum of squares of the vertical distances from the points to the line be made as small as possible.

This line, called **line of regression**, can be used to predict values of the predicted variable to the given predictor variable.

One method to measure the the spread of the data points about the least-square line is the **standard error of estimate**. The nearer the scatter points lie on the least-square line the smaller the standard error will be.

All data points can be described by a **probabilistic model:** $\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \varepsilon$. ε is the random error component.

4 basic assumptions about the distribution of ε are necessary.

- ε is normally distributed
- $\mu = 0$, this means, the mean value of y for a given x value equals $\beta_0 + \beta_1$ x.
- same variance $\sigma^2$of ε for all x values
- each y value has its own ε value.

The variability $\sigma^2$of the random error ε is usually unknown. Therefore the standard error S_e is used to estimate σ.

INTERNET

A list of servers with a server description providing **Internet resources** is given at:

www.deltasoft.at/english/mdlink.htm

SOURCES