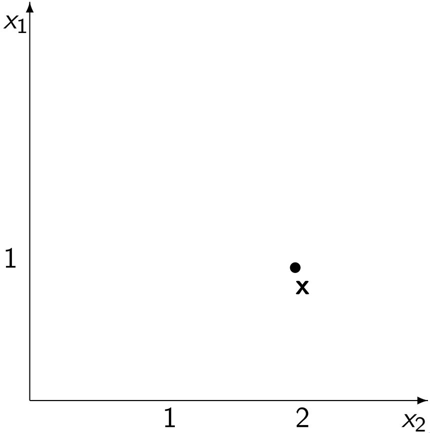
Geometric intuition of least squares

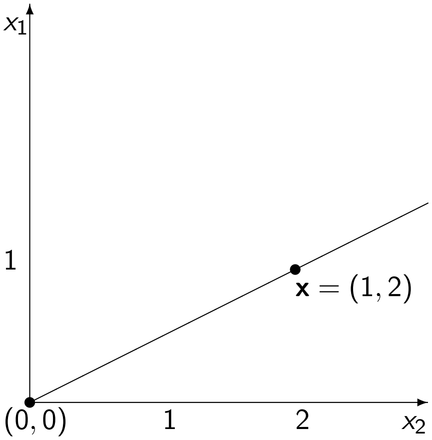
Consider the vector

$$\mathbf{x} = (1, 2)$$

A point in a two-dimensional space



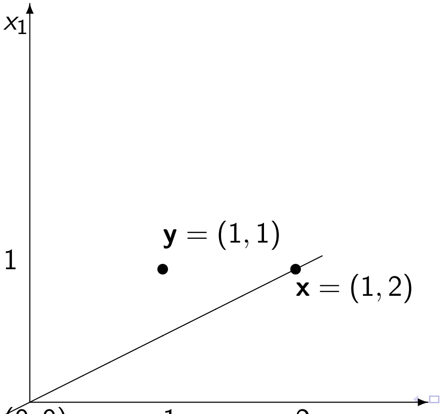
Linear combinations of $\mathbf{x}$:
multiply $\mathbf{x}$ with a constant β
generate a *line* that passes trough $\mathbf{x}$ and the origin $(0, 0)$.



The line: the *span* of $\mathbf{x}$.
What is regression?
Point $\mathbf{y}$ in the same space.
Model $\mathbf{y}$ as a linear function of $\mathbf{x}$

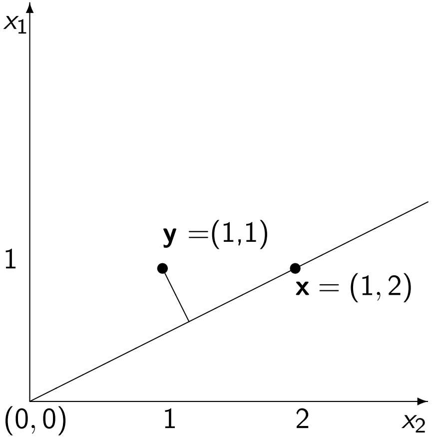
$$\mathbf{y} = \mathbf{x}\beta$$

Find a number $\hat{\beta}$ that "explains best" observations of $\mathbf{x}$ and $\mathbf{y}$.



Choose the $\mathbf{x}\beta$ that is *closest* to the observed $\mathbf{y}$.
Criterion: The usual (Euclidean) *distance* between $\mathbf{y}$ and $\mathbf{x}\beta$.
Minimize the distance with respect to β

$$\hat{\beta} = \text{argmin}_{\beta} \|\mathbf{y} - \mathbf{x}\beta\|,$$



OLS estimate

$\hat{\beta}$: The point on $\mathbf{x}\beta$ with the minimum distance to $\mathbf{y}$.
To find the minimum distance minimize

$$(\mathbf{y} - \mathbf{x}\beta)' (\mathbf{y} - \mathbf{x}\beta)$$

Optimization problem
First order condition:

$$\frac{\partial}{\partial \beta} = -2\mathbf{x}'(\mathbf{y} - \mathbf{x}\beta)$$

Set equal to zero and solve for β.

$$\begin{aligned} & \mathbf{x}'(\mathbf{y} - \mathbf{x}\beta) = 0 \\ \rightarrow \quad & \mathbf{x}'y - \mathbf{x}'\mathbf{x}\beta = 0 \\ \rightarrow \quad & \mathbf{x}'\mathbf{y} = \mathbf{x}'\mathbf{x}\beta \\ \rightarrow \quad & \left[\mathbf{x}'\mathbf{x}\right]^{-1} \mathbf{x}'\mathbf{y} = \beta \end{aligned}$$

The famed OLS (ordinary least squares) estimate of a linear regression model:

$$\hat{\beta} = \left[\mathbf{x}'\mathbf{x}\right]^{-1} \mathbf{x}'\mathbf{y}$$

In the example

$$\mathbf{x} = (1, 2)$$

$$\mathbf{y} = (1, 1)$$

What is the value of β?

$$\mathbf{x}'\mathbf{x} = 5$$

$$[\mathbf{x}'\mathbf{x}]^{-1} = \frac{1}{5}$$

$$\mathbf{x}'\mathbf{y} = 3$$

$$\hat{\beta} = \frac{3}{5}$$

The "closest" point to $\mathbf{y}$ on the line $\mathbf{x}\beta$ is

$$\mathbf{x}\hat{\beta} = (1\ 2) \cdot \frac{3}{5} = \begin{pmatrix} \frac{3}{5} & \frac{6}{5} \end{pmatrix}$$

Normal equation

$$\mathbf{x}'(\mathbf{y} - \mathbf{x}\beta) = \mathbf{0}$$

is called the *normal equation*.
What does this mean in terms of the model

$$\mathbf{y} = \mathbf{x}\beta + \mathbf{e}$$

Solve for **e**:

$$\mathbf{e} = \mathbf{y} - \mathbf{x}\beta$$

In other words, the first order condition

$$\mathbf{x}'(\mathbf{y} - \mathbf{x}\beta) = \mathbf{0}$$

can be written as

$$\mathbf{x}'\mathbf{e} = \mathbf{0}$$

or, the fitted errors are *orthogonal* to the data **x**.

This can also be thought about intuitively

$$\mathbf{y} - \mathbf{x}\hat{\beta} = \hat{\mathbf{e}}$$

is the *unexplained* part of the regression.
We have used the data $\mathbf{x}$ to find $\hat{\beta}$. What is left to explain is

$$\hat{\mathbf{e}} = \mathbf{y} - \mathbf{x}\hat{\beta}$$

If we have used $\mathbf{x}$ optimally to find β, $\mathbf{x}$ should have nothing left to explain of $\hat{\mathbf{e}}$, that is, $\mathbf{x}$ is *orthogonal* to $\hat{\mathbf{e}}$, or

$$\mathbf{x}'(\mathbf{y} - \mathbf{x}\beta) = \mathbf{x}'\mathbf{e} = \mathbf{0}$$

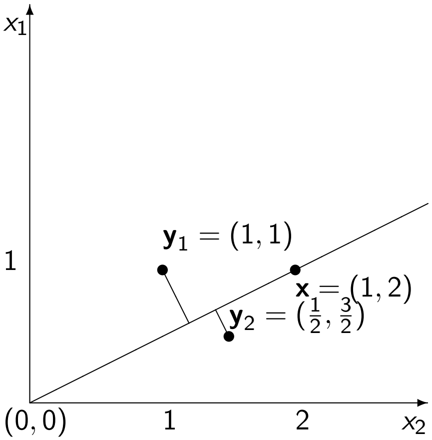
The normal equation

$$\mathbf{x}'(\mathbf{y} - \mathbf{x}\beta) = \mathbf{0}$$

has thus the interpretation that we use the information in $\mathbf{x}$ optimally to find $\hat{\beta}$

Measuring the fit of a regression.

Consider the picture with two $\mathbf{y}$ observations, $\mathbf{y}_1$ and $\mathbf{y}_2$



Can we conclude that $\mathbf{y}_2$ is closer than $\mathbf{y}_1$ to $\mathbf{x}\beta$?
Need a measure of the distance that is not sensitive to *scaling*.
Such a measure is the "R-squared" of a regression.
In geometric terms, look at the two points $\mathbf{y}_1$ and $\mathbf{y}_2$ above.
To compare the distance to $\mathbf{X}\beta$ of these two in a unit-free way: compare the *angles* that the line from the origin to $\mathbf{y}_1$ and $\mathbf{y}_2$ forms with $\mathbf{X}\beta$.
The *angle* θ between two vectors $\mathbf{x}$ and $\mathbf{y}$ was

$$\cos\theta = \frac{(\mathbf{x}, \mathbf{y})}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}$$

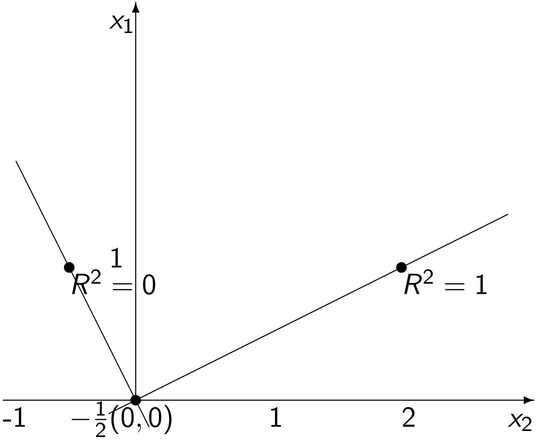
In a OLS problem the angle between $\mathbf{y}$ and the line $\mathbf{X}\beta$ is measured as

$$\cos\theta = \frac{\|\mathbf{y}'\hat{\beta}\|}{\|\mathbf{y}\|}$$

Taking the square of this, we find the (uncentered) R^2 of the regression,

$$R^2 = \cos^2\theta$$

Alternative way of interpreting R^2:
The *fraction* of the errors explained by the regression.
If $R^2 = 1$, then $\mathbf{y}$ is totally explained by the regression.
If $R^2 = 0$, the regression explains nothing.



Probably better known interpretation of R^2:

$$R^2 = \frac{\text{Explained square sum of errors}}{\text{Total square sum of errors}}$$

R^2 is one when everything is explained, since the sum of squares is also our criterion function.

Geometry of multivariate regressions

$$y = a + b_1 x_1 + b_2 x_2 + ... + b_k x_k + e$$

The dependent variable y is now a linear function of k independent variables $x_1, x_2, ...x_k$.
The factors b_i have the same interpretation as b in the univariate regressions:
They measure the marginal effect of a change in one of the explanatory variables, *holding everyting else constant*

$$\frac{dy}{dx_i} = \frac{d(a + b_1 x_1 + b_2 x_2 + ...b_i x_i + ... + b_k x_k + e)}{dx_i} = b_i$$

This is written in matrix form

$$\tilde{y}_i = a + bx_{1i} + bx_{2i} + \cdots + b_k x_{ki} + \tilde{e}_i$$

$$\tilde{\mathbf{y}} = \mathbf{X}\mathbf{b} + \tilde{\mathbf{e}}$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \dots & & & & \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} a \\ b_1 \\ \vdots \\ b_k \end{bmatrix} \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

The regression is formulated as

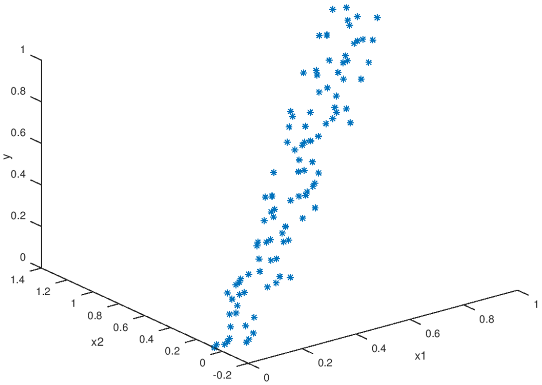
$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

Geometrically

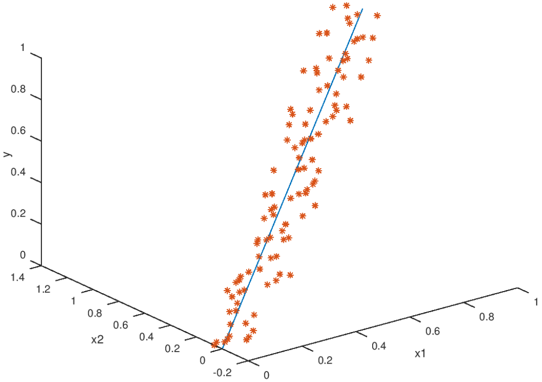
Regression line is a solution to minium distance problems - in several dimenstions.

Bivariate regression, y is a function of two variables (x_1, x_2)

A three-dimensional picture, with data a bunch of points (x_1, x_2) in this space



Regression is then drawing the “best fitting” line in this space.



Calculation of estimates

The minimum inference problem is

$$\mathbf{b} = \text{argmin}_{\mathbf{b}} (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$$

Calculation:
Step 1: the Normal Equation

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{0}$$

Step 2: The analytical solution

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Estimation using octave

$$\mathbf{y} = \mathbf{X}\mathbf{b}$$

We explain $\mathbf{y}$ as a linear function of $\mathbf{X}$.
Suppose we have two explanatory variables. Then $\mathbf{b}$ is a 2×1 matrix.
We simulate 100 observations of the model.
In simulating, we need to add some noise $\mathbf{e}$ to the data, to avoid a perfect fit.

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

Simulating the model

```
X=rand(100,2);
b = [2;1]
e = 0.25*randn(100,1);
y = X*b + e;
```

Estimating the model

```
> bhat = inv(X'*X)*X'*y
```

results in

```
> bhat = inverse(X'*X)*X'*y
bhat =
  2.00744
  0.99017
```

Alternatively

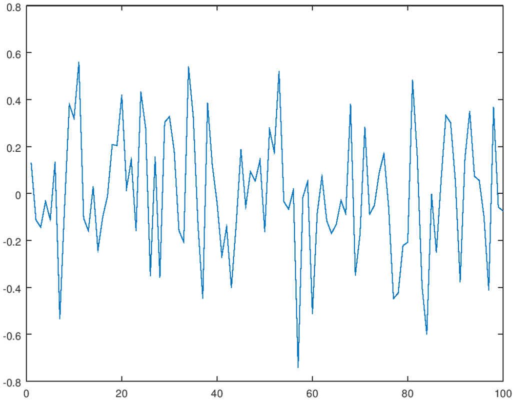
```
> ols(y,X)
ans =
  2.00744
  0.99017
```

Forecasting

```
> bhat
bhat =
  2.00744
  0.99017
> new_x = [1 1]
new_x =
  1   1
> forecast_y = new_x * bhat
forecast_y = 2.9976
```

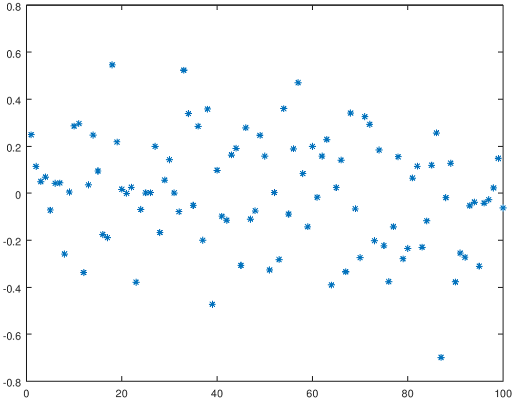
Residuals

```
ehat = y - X*bhat;
plot (ehat);
```



Remove lines, plots instead

```
>> plot(ehat,"*");
```



Detecting deviations from the assumed linear relationship

Residual $\hat{e} = y - \hat{a} - \hat{b}x$.
Plot residuals against other variables.
Should be: Centered at zero, No obvious relationships.

Simulated example: Linear model

True model:

$$x = [1, 2, 3, ...100]$$

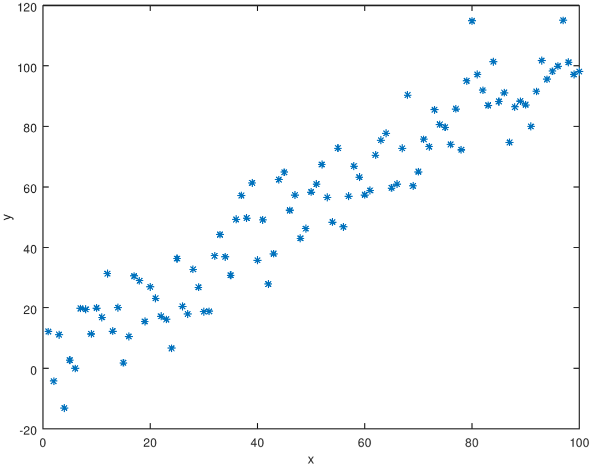
$$y = a + bx + e$$

$$a = 1$$

$$b = 1$$

$$e \sim N(0, 10^2)$$

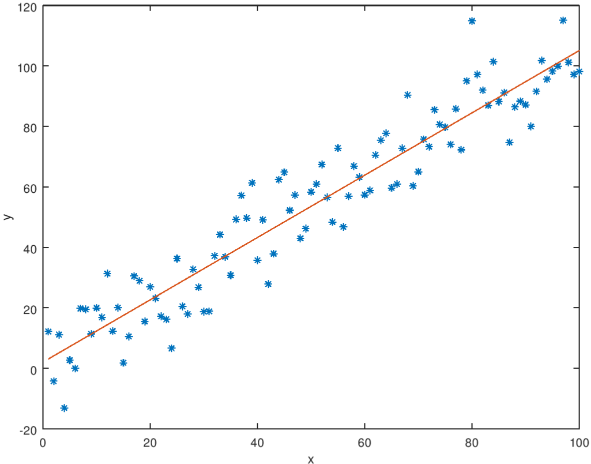
Plot data



Simulated example: Linear model

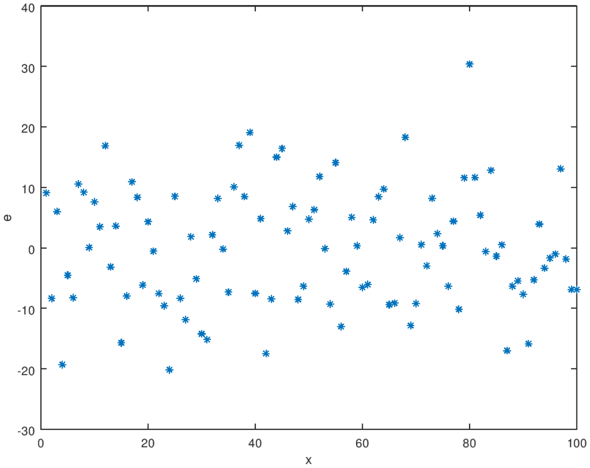
Fitted regression

```
b =
  -0.29247
   1.00990
```

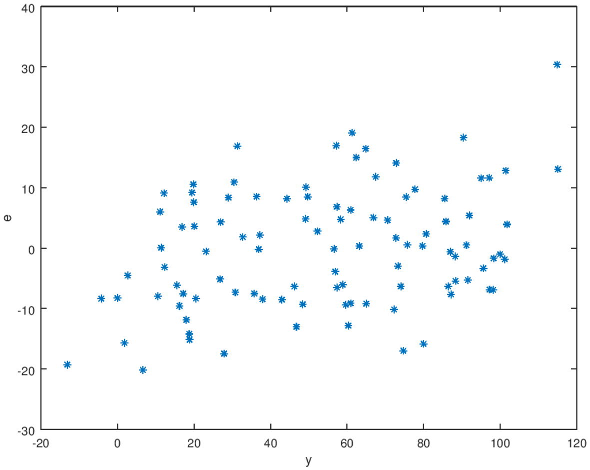


Plot residuals.

Against x



Plot residuals against y



Simulated example: nonlinear model

True model:

$$y = a + b\ln(x) + e$$

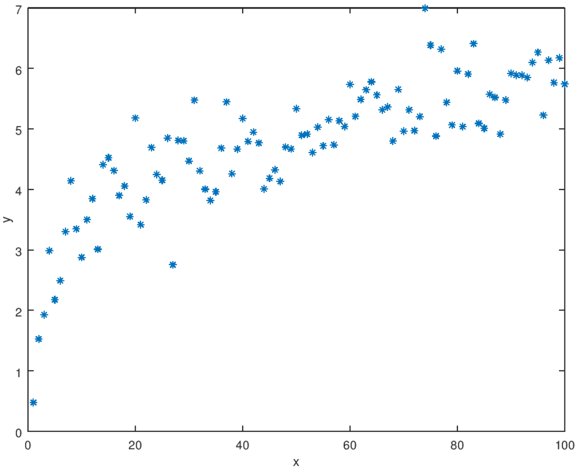
$$x = [1, 2, 3, ..., 100]$$

$$a = 1$$

$$b = 1$$

Simulate series,

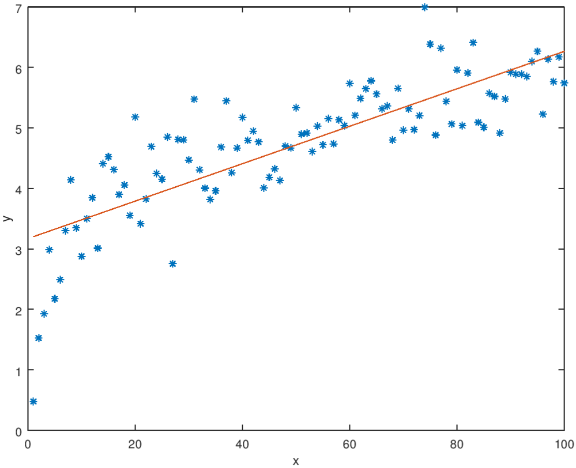
Plot observations



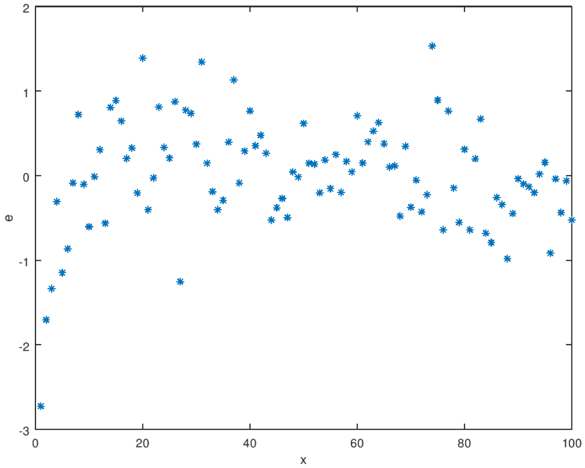
Fitted regression

$$y = a + bx + e$$

```
b =
   3.102127
   0.044478
```

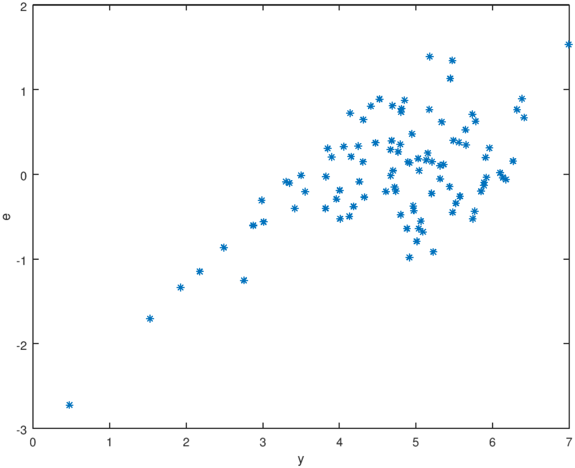


What do the residuals look like?



vs x

Residuals vs y

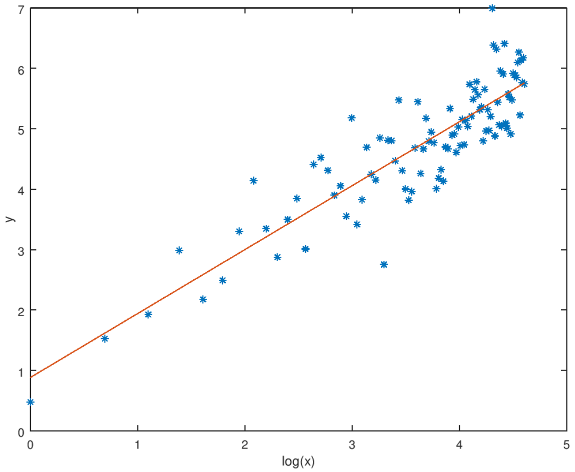


Aha, there is a problem.

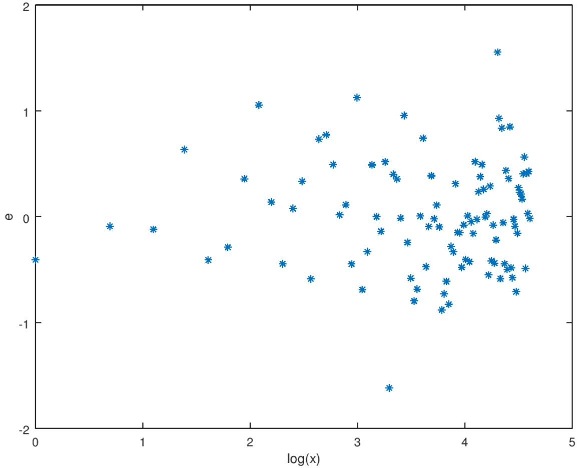
Solving this problem: Linearizing it.
Define $x' = log(x)$ and run regression with x' instead
Fitted regression

```
bhat =
   0.23685
   1.22831
```

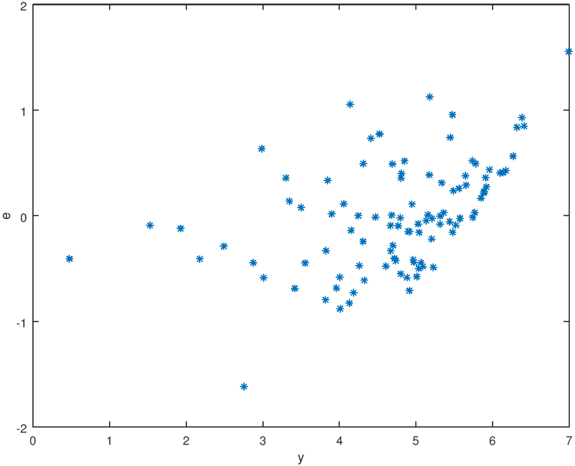
Correct model



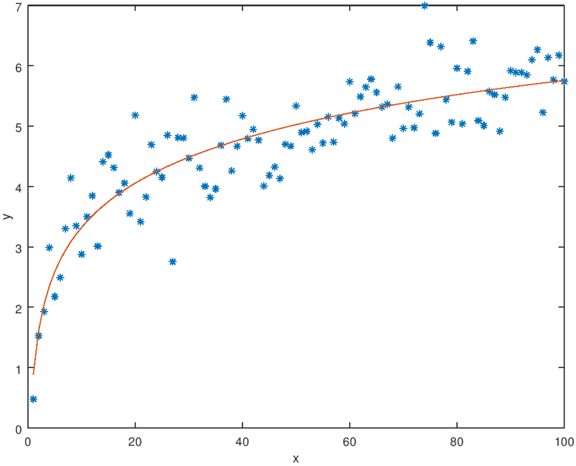
Residuals: against $x'(\ln x)$



Residuals: against y



Plotting the estimated relationship against x instead of $\ln(x)$.



Alternative nonlinear model

$$y = a + b\sin(0.1x) + e$$

$$a = 1$$

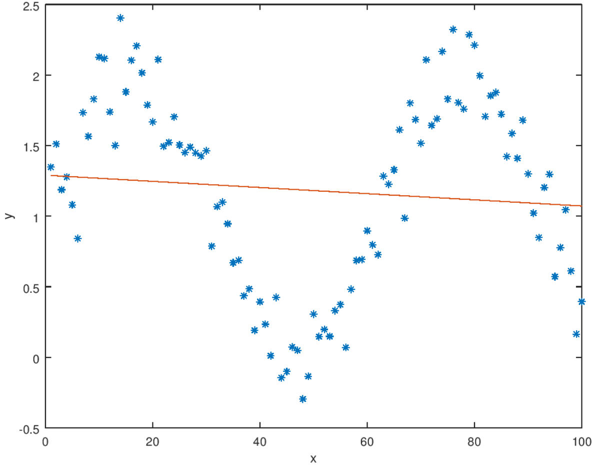
$$b = 1$$

Fitted regression

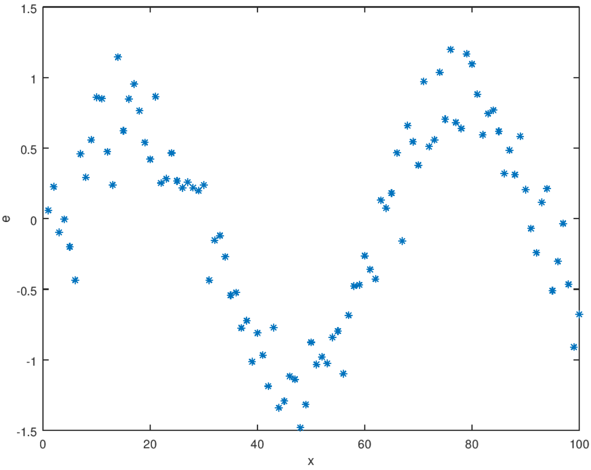
$$y = a + bx + e$$

```
b =
   1.4122e+00
  -5.9277e-04
```

Fitted regression



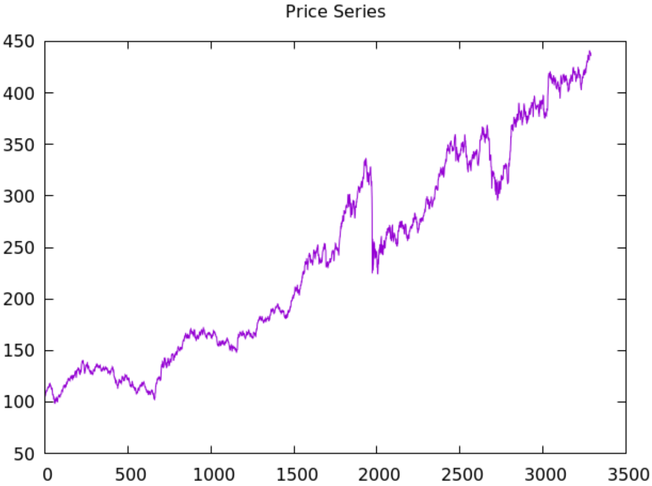
Residuals against x.



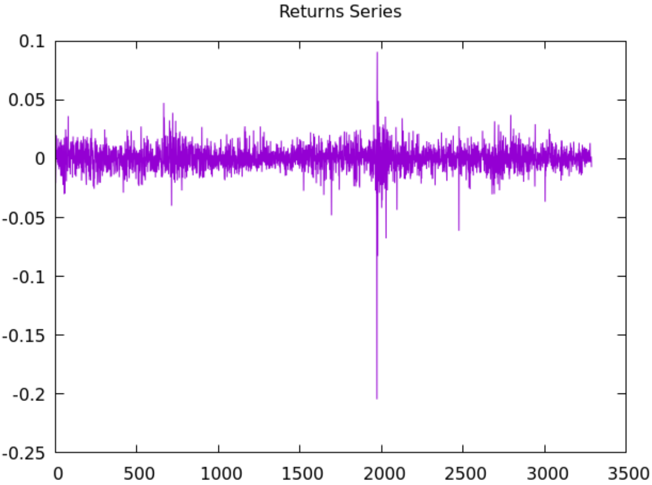
Problem

Residuals outliers are not always errors...

One are not always justified in simply throwing out any observations considered an outlier.



Difference the previous picture



Prediction example

$$\mathbf{y} = \mathbf{X}\mathbf{b}$$

Given an estimated set of parmeters $\widehat{\mathbf{b}}$, use it for prediction

Prediction example

At a large state university seven undergraduate students who are majoring in economics were randomly selected and surveyed. Two of the survey questions asked were:

- What was your grade-point average (GPA) in the preceding term?
- What was the average number of hours spent per week last term in the Orange and Brew?

The Orange and Brew is a favorite and only watering hole on campus.

Using the data below, estimate with ordinary least squares the equation

$$G = \alpha + \beta H$$

where G is GPA and H is hours per week in Orange and Brew. (The GPA is a numerical summary of grades with 4 as the largest number.)

What is the expected sign for β? Does the data support your expectations?

Student	GPA (G)	Hours per week in Orange and Brew (H)
1	3.6	3
2	2.2	15
3	3.1	8
4	3.5	9
5	2.7	12
6	2.6	12
7	3.9	4

Suppose that a freshman economics student has been spending 15 hours per week in the Orange and Brew during the first two weeks of class.

Predict his GPA for this year.

Solution

```
Data = [1 , 3.6 , 3 ; \
2 , 2.2 , 15 ; \
3 , 3.1 , 8 ; \
4 , 3.5 , 9 ; \
5 , 2.7 , 12 ; \
6 , 2.6 , 12 ; \
7 , 3.9 , 4 ]
y=Data(:,2);
x=Data(:,3);
X=[ones(7,1),x];
b=ols(y,X)
b =
   4.25727
  -0.13017
```

Thus, we estimated the parameters as $\alpha = 4.25727$ and $\beta = -0.13017$.

Now for prediction:

```
>> predicted=[1 15]*b
predicted =  2.3047
```

$$\textit{prediction} = \alpha + \beta 15 = 4.25727 - 0.13017 \cdot 15 = 2.3047$$

What if number of hours is 4?

```
>> predicted=[1 4]*b
predicted =  3.7366
```

A bit better

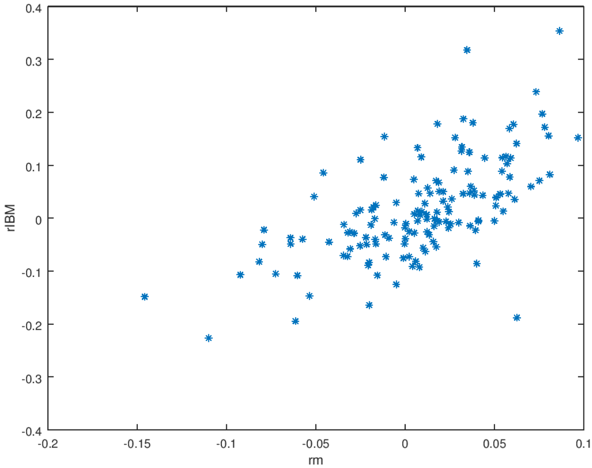
You want to find the risk of the stock IBM. To do so you think the "market model"

$$r_{it} = \alpha_i + \beta_i r_{mt} + e_{it}$$

is a reasonable description of the risk return relationship. To estimate the parameters α_i and β_i you need a history of stock returns and index returns. Collect monthly returns for IBM and a broad based US stock market index, for example the S&P 500. Take data for 1995:1 to 2006:12.

- Estimate the model.
- What is the R^2 in your estimation?

Plotting one against the other



Running the regression

```
>> r=ibm(:,2);
>> rm=sp500(:,2);
>> X=[ones(144,1) rm];
>> b=X\r
b =
   0.0041805
   1.4068184
```

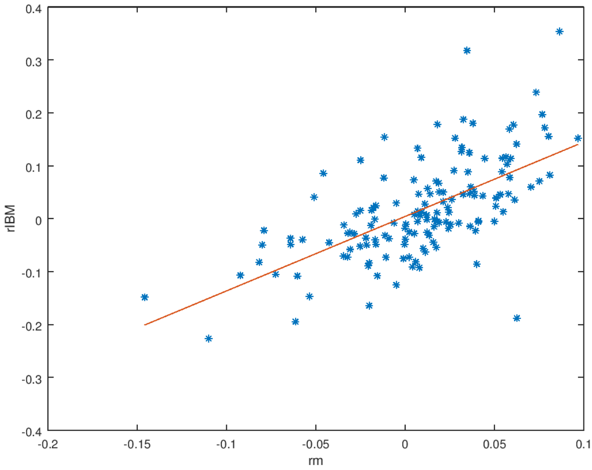
We estimate the parameters as

$$\hat{a} = 0.0041805$$

$$\hat{b} = 1.4068184$$

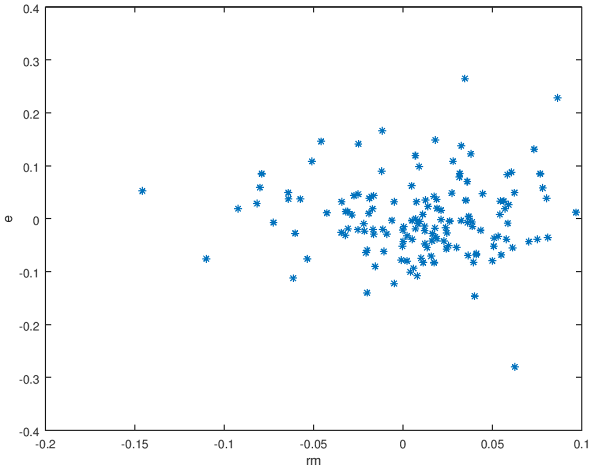
Next, plotting the observations and comparing it to the actual regrssion.

```
>> plot(rm,r,"*",rm,X*b)
```



Check for any obvious problems by calculating the residuals and plotting them against r_m:

```
>> plot(rm,e,"*");
```



Calculate the R^2

```
>> SSR=e'*e
SSR =  0.74944
>> TSS=(r-mean(r))'*(r-mean(r))
TSS =  1.2451
>> R2=1-SSR/TSS
R2 =  0.39811
```

The R^2 of the regression is 0.39811.

It is all optimization...

So far, all our analysis has actually solved an optimization problem to find the parameter estimates in a regression.
(minimum distance problem).
Remaining issue: Probability statements.
How can we evaluate an estimated coefficient – how "confident" are we that the true coefficient is close to what we have estimated.
To make such statements: Need additional assumptions.