Residuals as Diagnostics

Example 1: Here is the regression output for four different data sets. In each case we have just one x.

DATASET 1

The regression equation is
y1 = 3.00 + 0.500 x1

Predictor	Coef	Stdev	t-ratio	p
Constant	3.000	1.125	2.67	0.026
x1	0.5001	0.1179	4.24	0.002

s = 1.237 R-sq = 66.7% R-sq(adj) = 62.9%

DATASET 2

The regression equation is
y2 = 3.00 + 0.500 x2

Predictor	Coef	Stdev	t-ratio	p
Constant	3.001	1.125	2.67	0.026
x2	0.5000	0.1180	4.24	0.002

s = 1.237 R-sq = 66.6% R-sq(adj) = 62.9%

DATASET 3

The regression equation is
y3 = 3.00 + 0.500 x3

Predictor	Coef	Stdev	t-ratio	p
Constant	3.002	1.124	2.67	0.026
x3	0.4997	0.1179	4.24	0.002

s = 1.236 R-sq = 66.6% R-sq(adj) = 62.9%

DATASET 4

The regression equation is
y4 = 3.00 + 0.500 x4

Predictor	Coef	Stdev	t-ratio	p
Constant	3.002	1.124	2.67	0.026
x4	0.4999	0.1178	4.24	0.002

s = 1.236 R-sq = 66.7% R-sq(adj) = 63.0%

In each case the output is identical.
Whatever decision you are trying to make (eg. predicition) would be the same !!

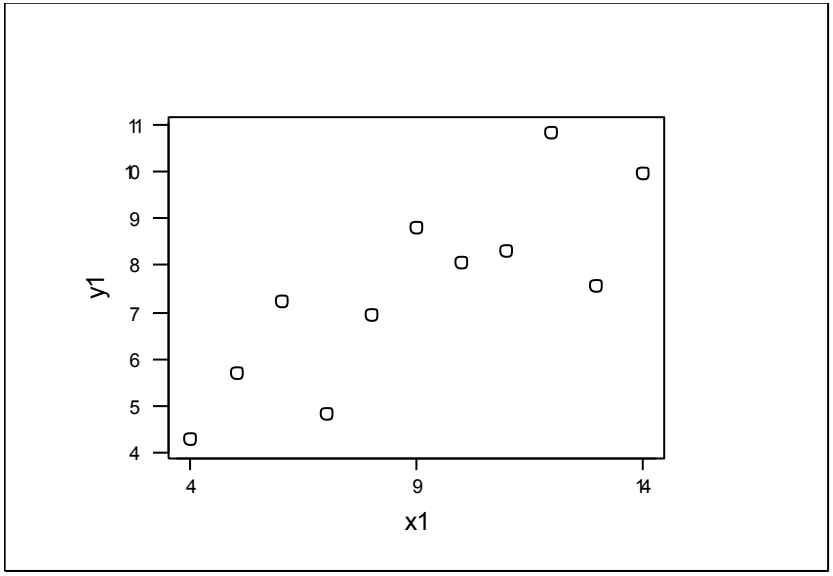
```
Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
-8.837 -2.242  0.062  2.417  8.916

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.3027     1.4274   0.212    0.832
x             3.3028     0.2765  11.946   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.359 on 198 degrees of freedom
Multiple R-squared:  0.4189,
  Adjusted R-squared:  0.4159
F-statistic: 142.7 on 1 and 198 DF,  p-value: < 2.2e-16
```

Regressao 1

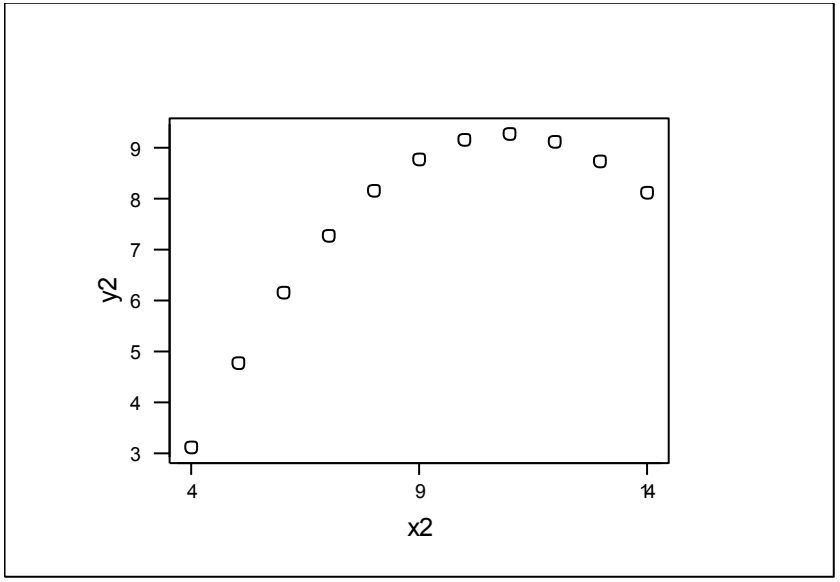
Regressao 2

```
Call:
lm(formula = y1 ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2641 -1.7536 -0.1168  1.5062  7.3514

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -25.0194     0.8993  -27.82   <2e-16 ***
x            10.1144     0.1742   58.07   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.116 on 198 degrees of freedom
Multiple R-squared:  0.9445,
  Adjusted R-squared:  0.9443
F-statistic:  3372 on 1 and 198 DF,  p-value: < 2.2e-16
```

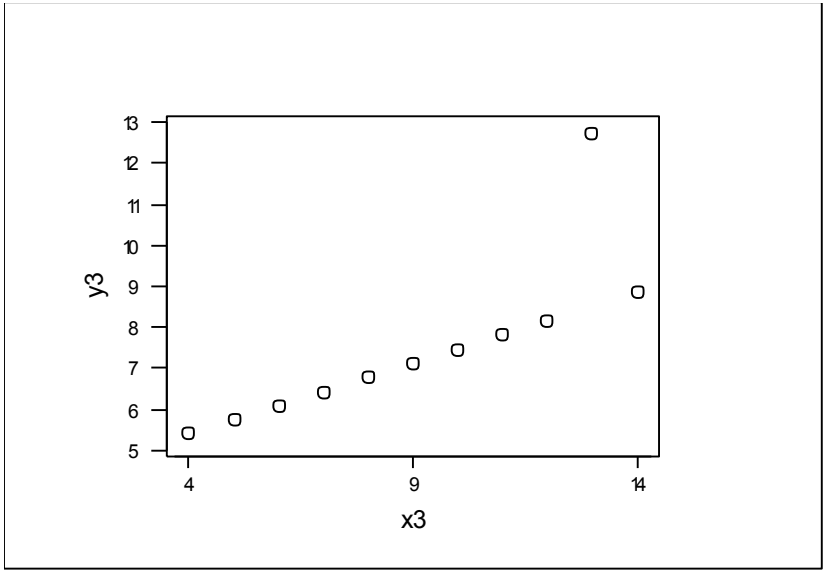
Data set 1:



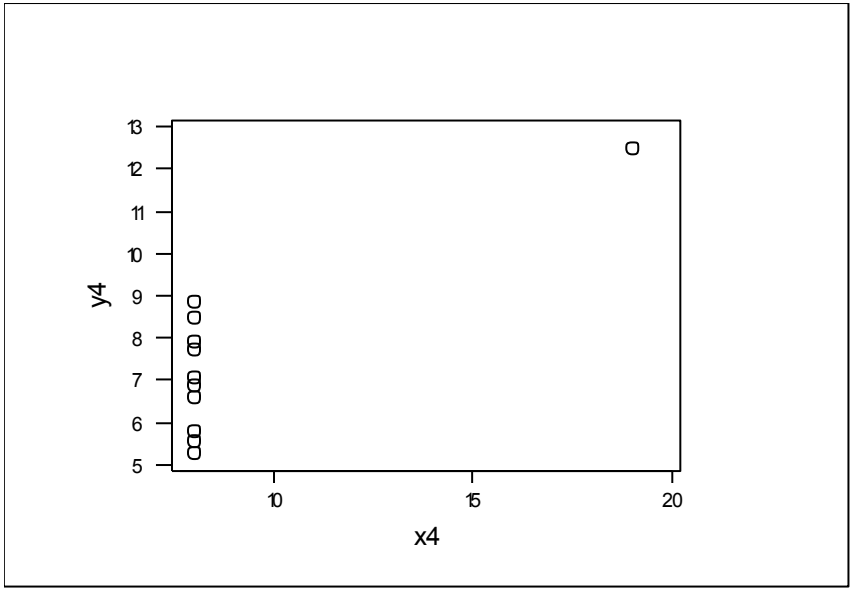
Data set 2:

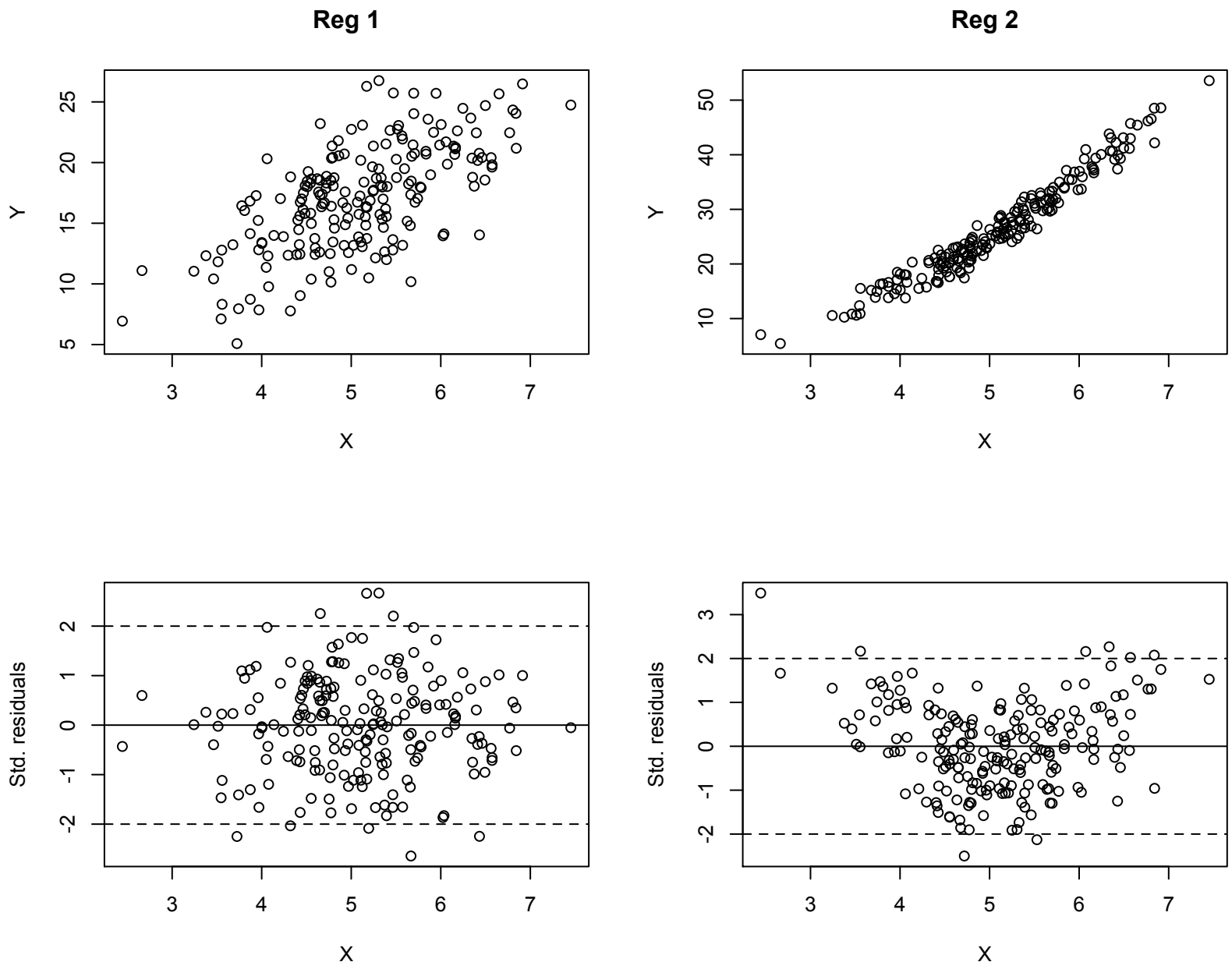


Data set 3:



Data set 4:




Reg 1
Reg 2
Y
X
Std. residuals
5
10
15
20
25
30
40
50
-2
-1
0
1
2
3
4
5
6
7

Moral of the Story

Only in the first case does the plot suggest that the simple linear regression model is a good way to think about the data.

In the other cases a blind use of the model would lead to bad decisions.

QUESTION:
So, how do you tell if the model is "a good way to think about your data"?

Plot the data!

ANOTHER QUESTION:

With more than one x, how do we "plot" the data?
How can we *diagnose* a problem with the regression model?

Basic idea: If the model is right then

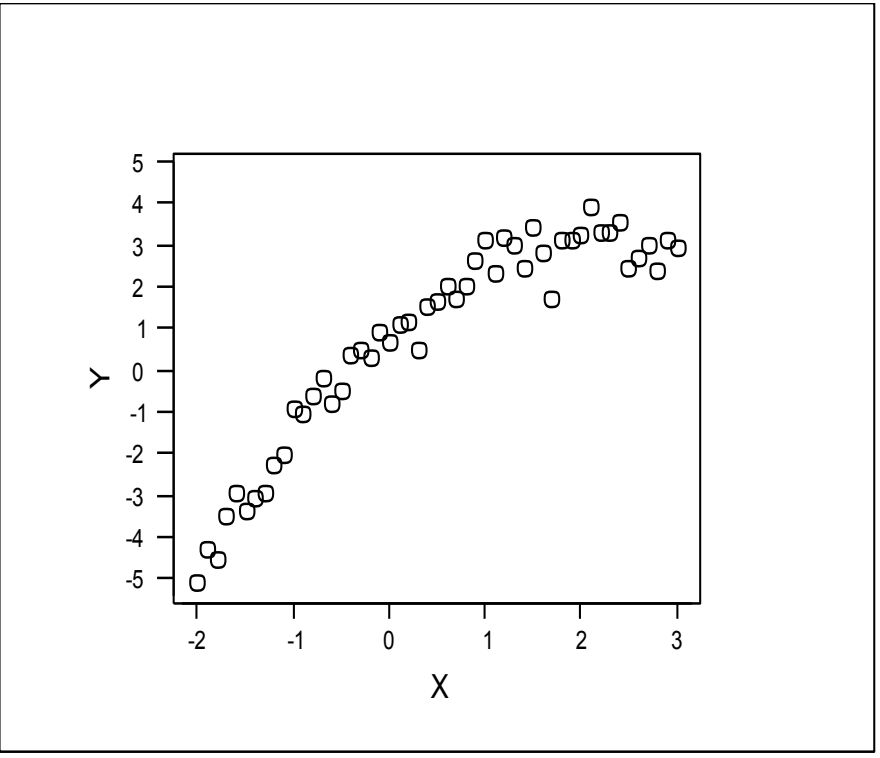
$$e_i \approx \varepsilon_i \sim N(0, \sigma^2)$$ ***independent of the x's !!!!***

The residuals should look i.i.d. normal;

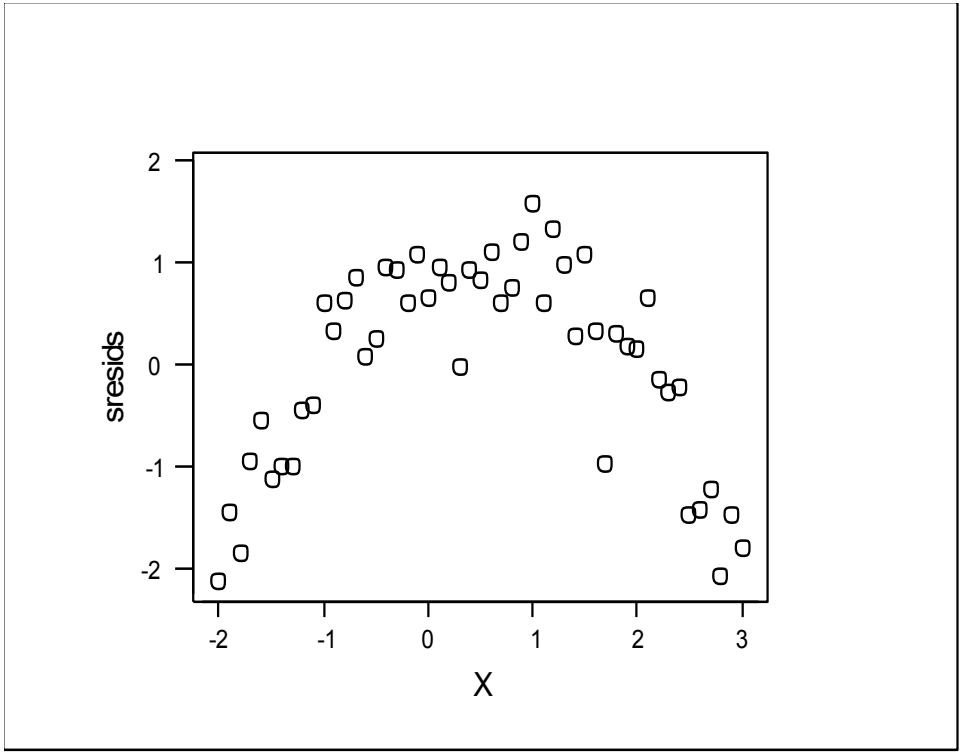
The residuals should be unrelated to the x's.

Example 2: nonlinear regression

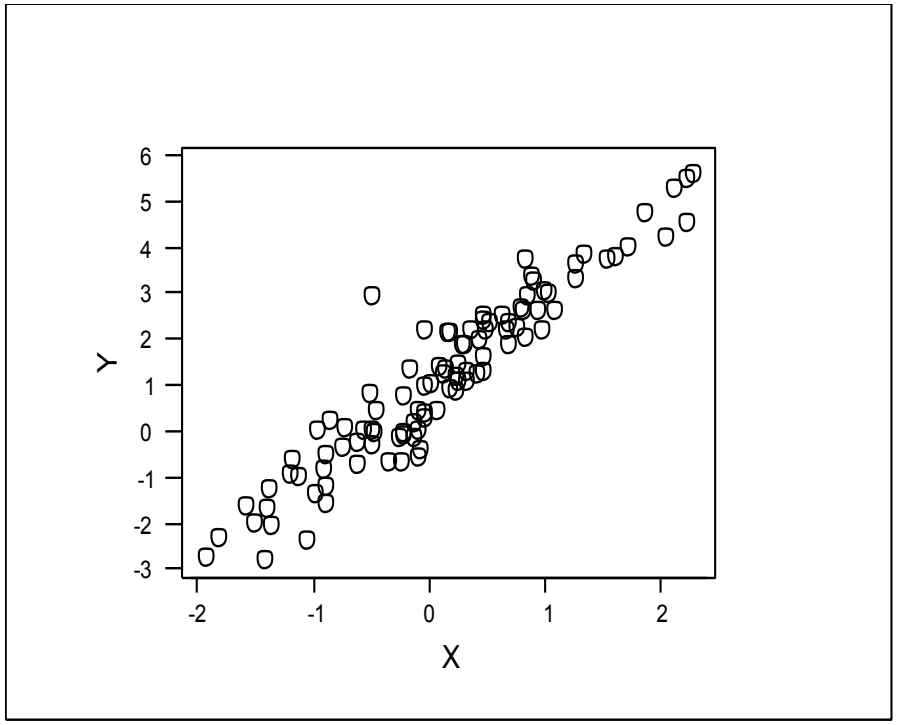
y versus x

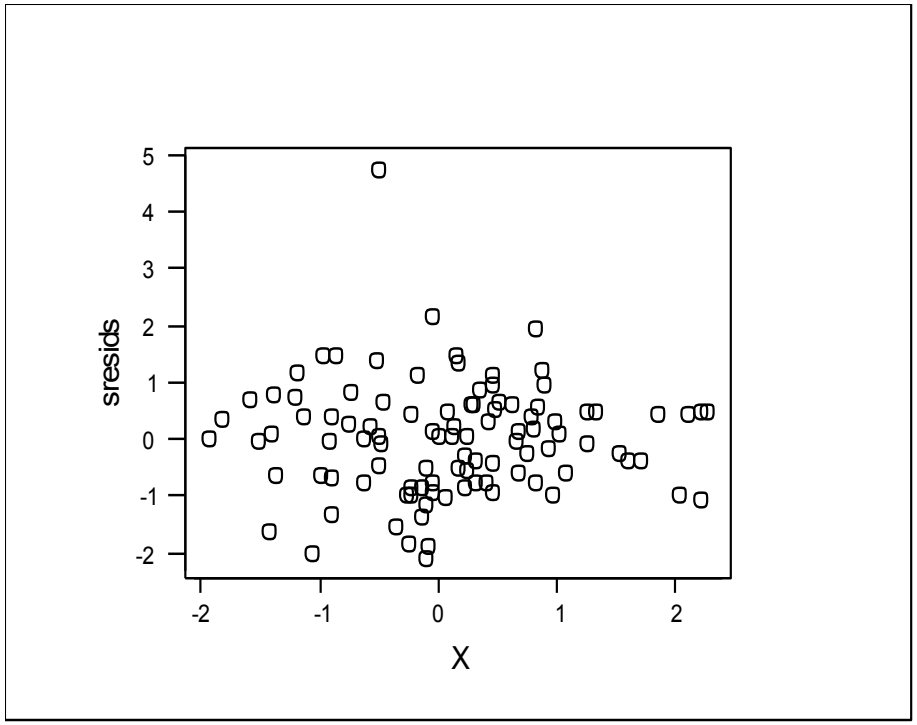


Residuals versus x (or fits)

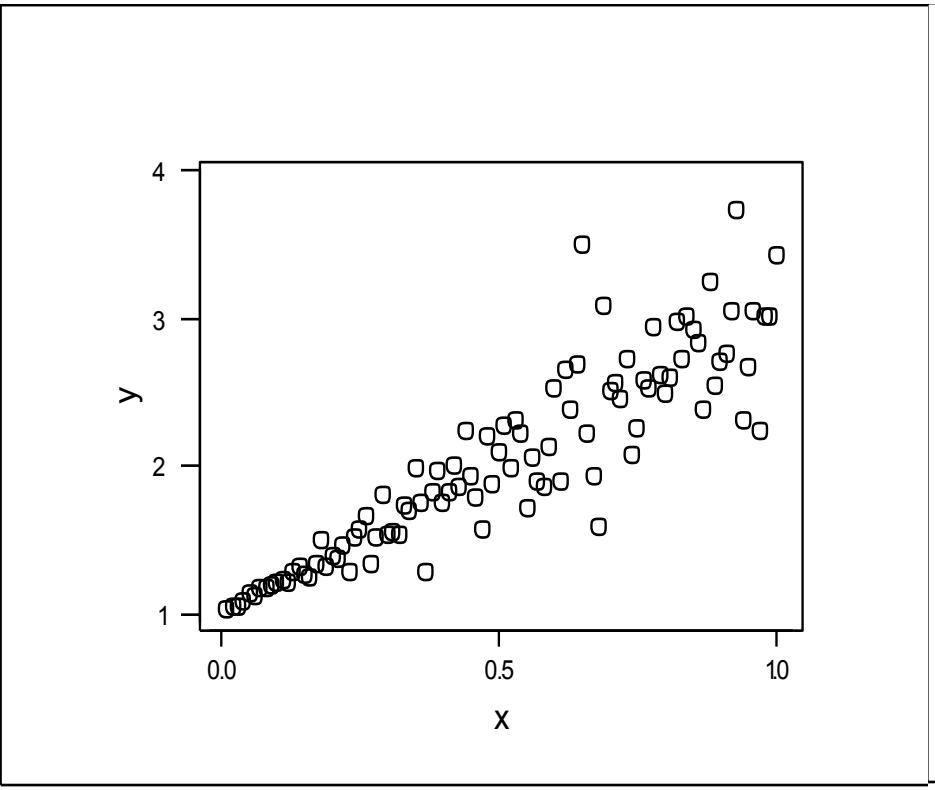


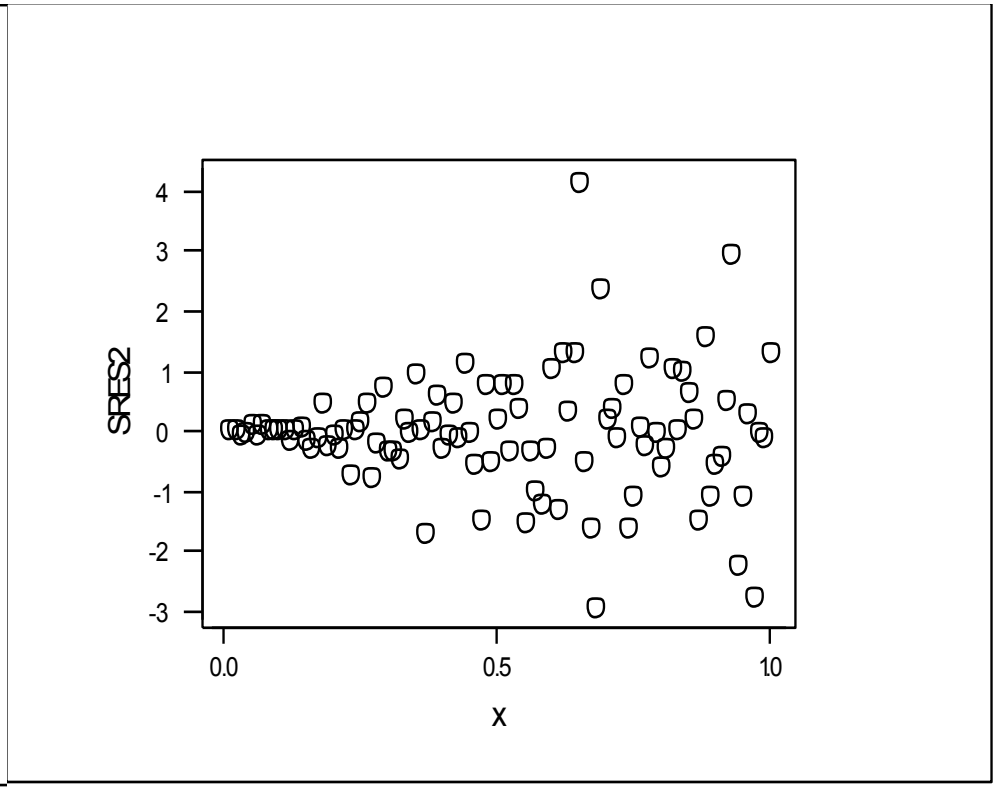
Example 3: outliers





Example 4: heteroskedasticity





Something wrong or peculiar!!

Example 2: Failure of basic assumption of linear relationship.
Example 3: A funny point, an outlier.
Example 4: The variance of errors increases with x, we have nonconstant variance: "**heteroskedasticity**".

We plot the **residuals versus each x**.
There should be nothing funny!!

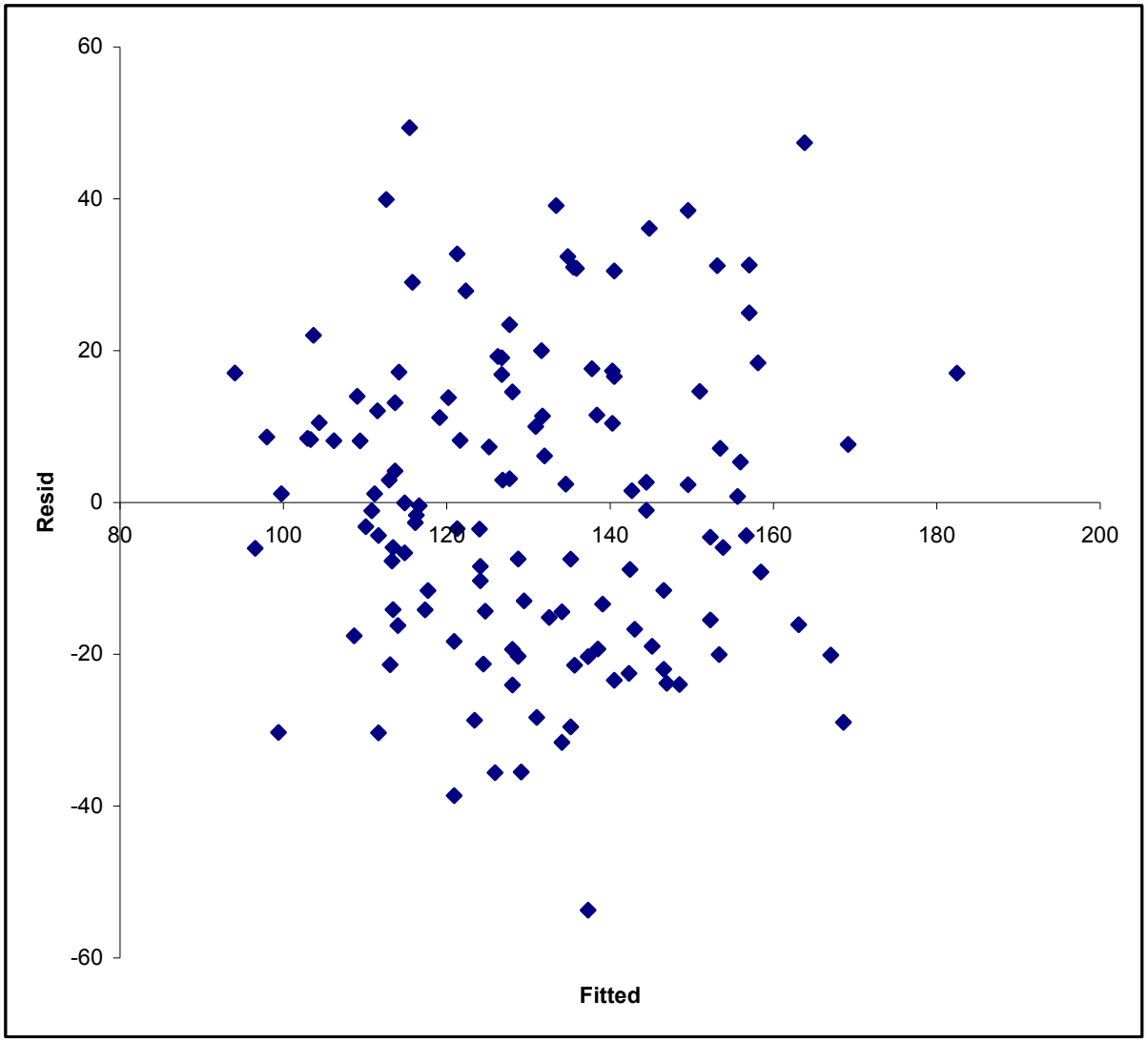
Since the fitted values are a function of the x's, we also plot the **residuals versus the fitted values** and again **there should be no relationship**.

Example 5

Here are residuals versus fitted values from the house price on size, number of bedrooms and number of bathrooms.

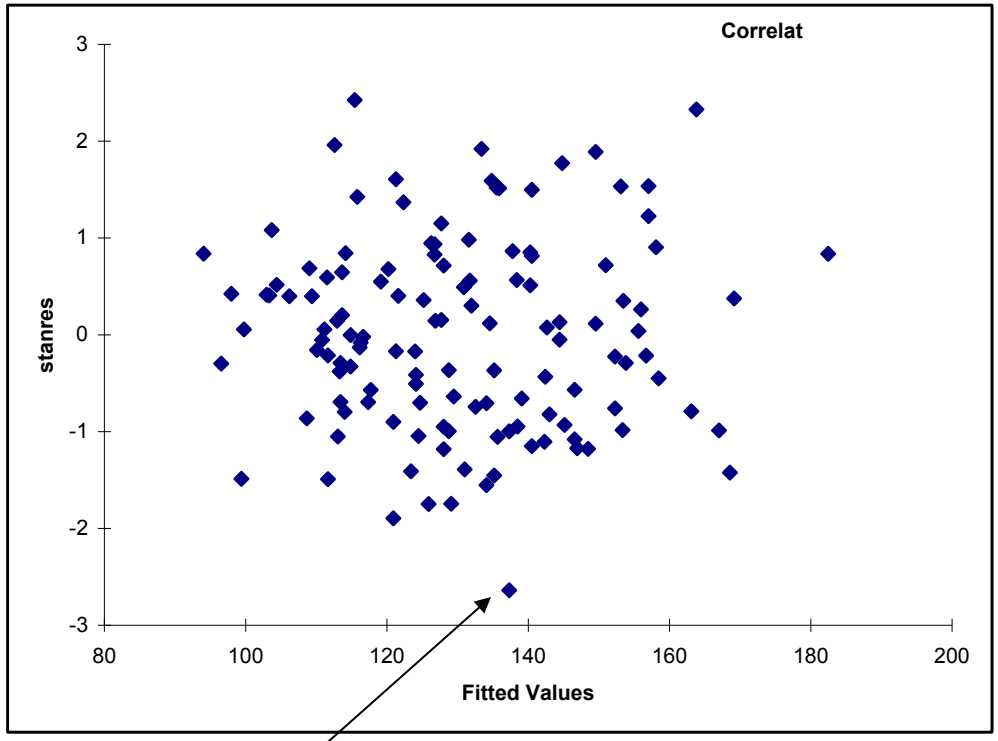
Looks pretty good!

Is there an outlier?



This plot is a good thing !!

This is a plot of standardize Residuals versus fitted values.



-2.64

If the model is right, then **standardized** residuals should look like i.i.d. normal draws independent of the x's (and fitted values).