Modelling

Mandy Vogel

July 12, 2015

Table of Contents

Choosing the appropriate method

It is essential, therefore, that you can answer the following questions:

- Which of your variables is the response variable?
- Which are the explanatory variables?
- Are the explanatory variables continuous or categorical, or a mixture of both?
- What kind of response variable do you have: is it a continuous measurement, a count, a proportion, a time at death, or a category?

Choosing the appropriate method

Explanatory Variables are	
all continuous	Regression
all categorical	Analysis of variance (ANOVA)
both continuous and categorical	Analysis of covariance (ANCOVA)

Choosing the appropriate method

Response Variables	
(a) Continuous	Normal regression, ANOVA or ANCOVA
(b) Proportion	Logistic regression
(c) Count	Log-linear models
(d) Binary	Binary logistic analysis
(e) Time at death	Survival analysis

The best model is the model that produces the **least unexplained variation** (the minimal residual deviance), subject to the constraint that **all the parameters** in the model **should be statistically significant**.

Choosing the appropriate method

It is very important to understand that there is not *one* model; this is one of the common implicit errors involved in traditional regression and ANOVA, where the same models are used, often uncritically, over and over again. In most circumstances, there will be a large number of different, more or less plausible models that might be fitted to any given set of data.
And of course: We are looking for the BEST.

Maximum Likelihood

We define *best* in terms of maximum likelihood.

- given the data,
- and given our choice of model,
- what values of the parameters of that model make the observed data most likely?

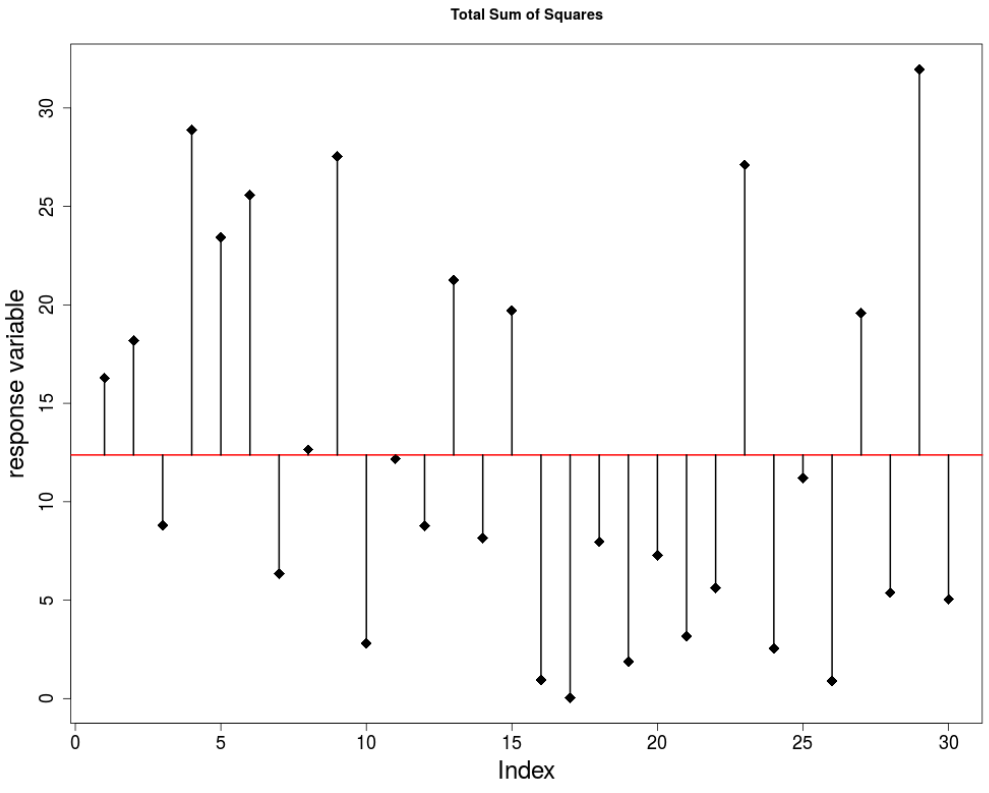
We judge the model on the basis how likely the data would be if the model were correct.

Ockham's Razor

For statistical modelling, the principle of parsimony means that:

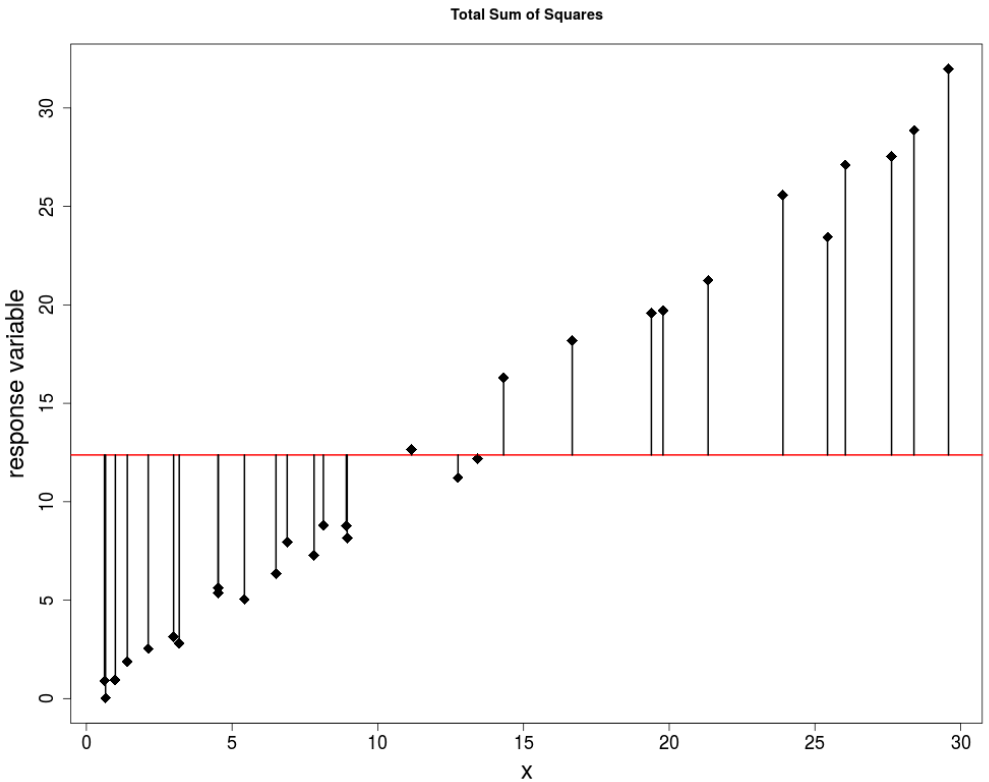
- models should have as few parameters as possible;
- linear models should be preferred to non-linear models;
- experiments relying on few assumptions should be preferred to those relying on many;
- models should be pared down until they are minimal adequate;
- simple explanations should be preferred to complex explanations.

The Null model



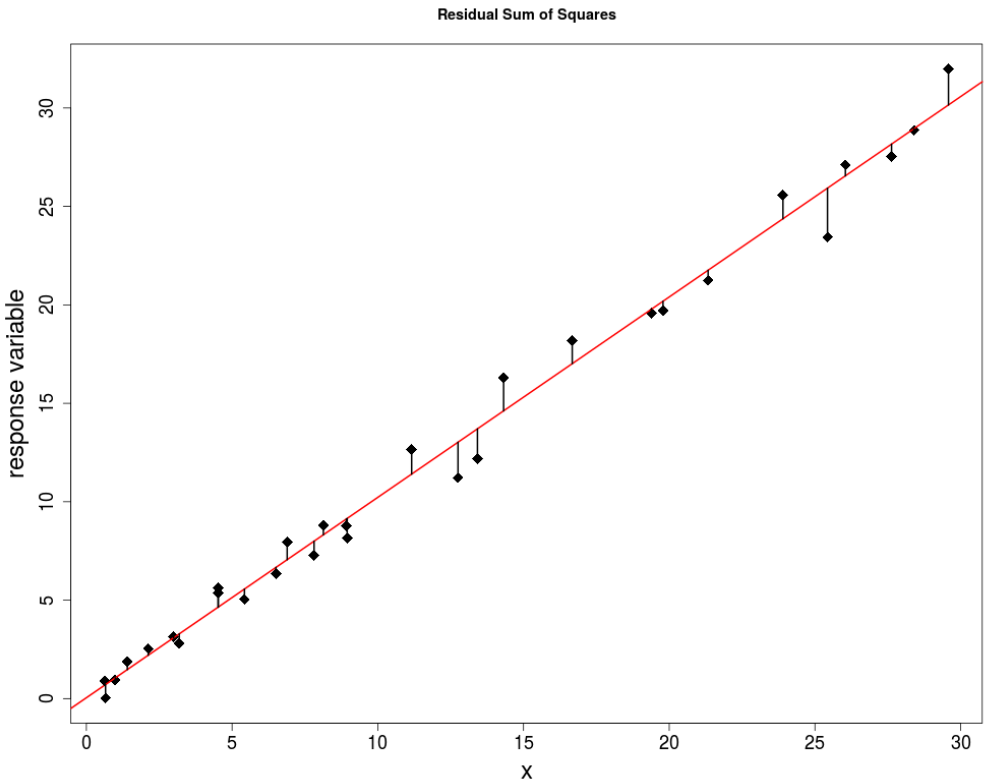
- Just one parameter, the overall mean $\bar{y}$
- Fit: none; $SSE = SSY$
- Degrees of freedom: $n - 1$
- Explanatory power of the model: none

Adding Information



- model with $0 \leq p' \leq p$ parameters
- Fit: less than the maximal model, but not significantly so
- Degrees of freedom: $n - p' - 1$
- Explanatory power of the model: $r^2 = \frac{SSR}{SSY}$

Adding Information



- model with $0 \leq p' \leq p$ parameters
- Fit: less than the maximal model, but not significantly so
- Degrees of freedom: $n - p' - 1$
- Explanatory power of the model: $r^2 = \frac{SSR}{SSY}$

General Formula Syntax

response variable $\sim$ explanatory variable(s)

one can read the tilde symbol as *is modelled as a function of*

Simple Linear Regression

$$\mathrm{lm}(y \sim x)$$

Output

```
Call:
lm(formula = y ~ x, data = df)

Coefficients:
(Intercept)            x
    -0.1253       1.0207
```

Simple Linear Regression

The model contains much more information. We can access them if we assign it to a variable.

Input

```
> m <- lm(y~x,data=df)
> summary(m)
```

Simple Linear Regression

Output

```
Call:
lm(formula = y ~ x, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.30324 -0.59581  0.00489  0.57777  1.42182

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1253     0.3658  -0.343    0.735
x             1.0207     0.0203  50.293   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.7412 on 28 degrees of freedom
Multiple R-squared: 0.9891, Adjusted R-squared: 0.9887
F-statistic:  2529 on 1 and 28 DF,  p-value: < 2.2e-16
```

Simple Linear Regression

With `aov(m)` you can fit an analysis of variance model for each stratum.

Input

```
> aov(m)
```

Output

```
Call:
   aov(formula = m)

Terms:
                         x Residuals
Sum of Squares  1389.7606   15.3842
Deg. of Freedom          1         28

Residual standard error: 0.7412399
Estimated effects may be unbalanced
```

Notation

symbol	meaning
$+$	indicates inclusion of an explanatory variable in the model (not addition)
$-$	indicates deletion of an explanatory variable from the model (not subtraction)
$*$	indicates inclusion of explanatory variables and interactions (not multiplication)
$/$	indicates nesting of explanatory variables in the model (not division)
$\|$	indicates conditioning (not "or"), so that $y \sim x\|z$ is read as y as a function of x given z.

$A : B$ means the two way interaction between A and B and $A : B : C : D$ means the four-way interaction between the four variables.

Notation 2

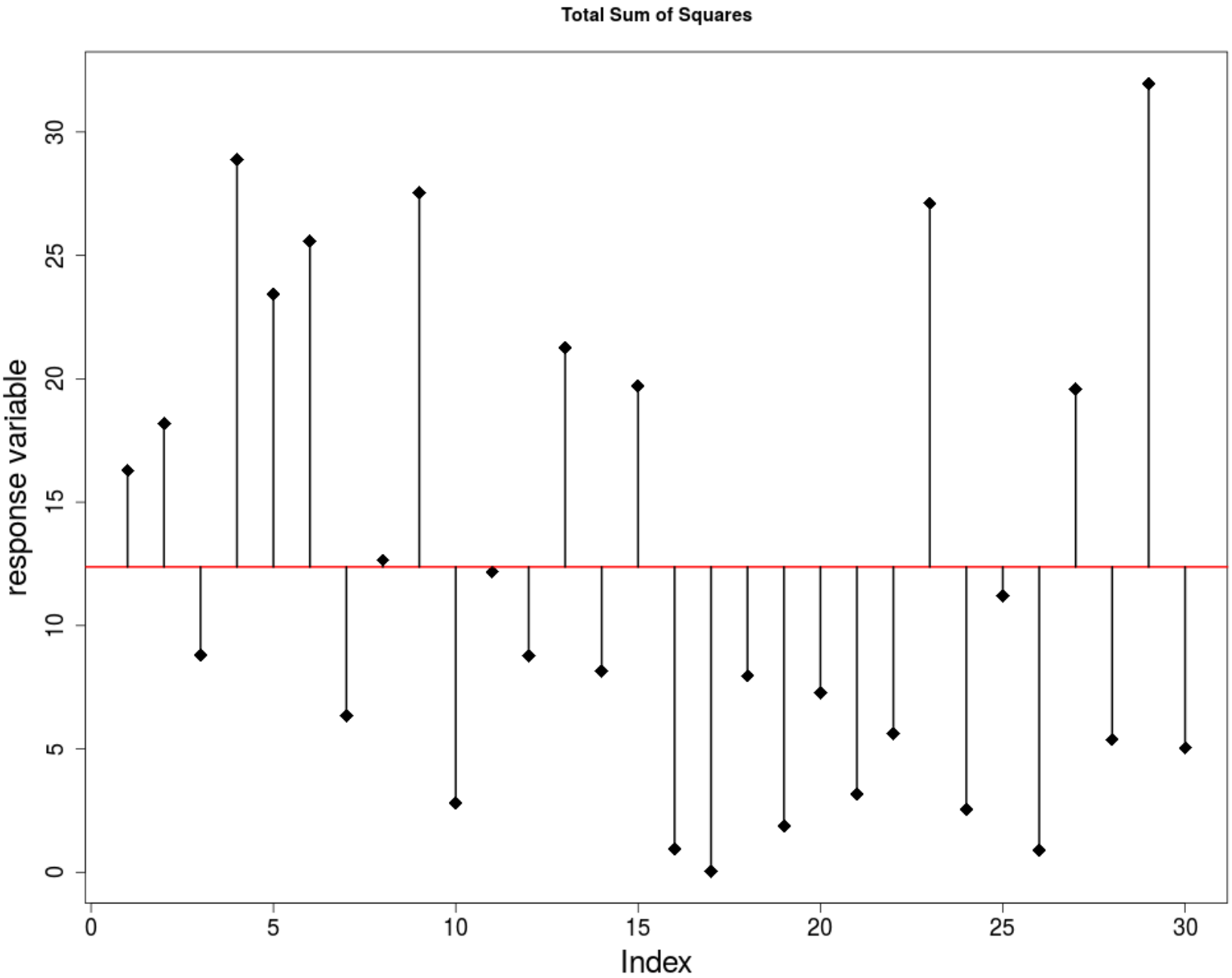
A term of the form `A/x` where `A` is a factor, gives "separate" regression lines of type $1 + x$ for different levels of `A`. In this case the intercept term is not needed (removed by `-1`).
Another important function term is the identity function `I(...)`. It is used to evaluate its argument with operators having their arithmetical meaning and returns the result.

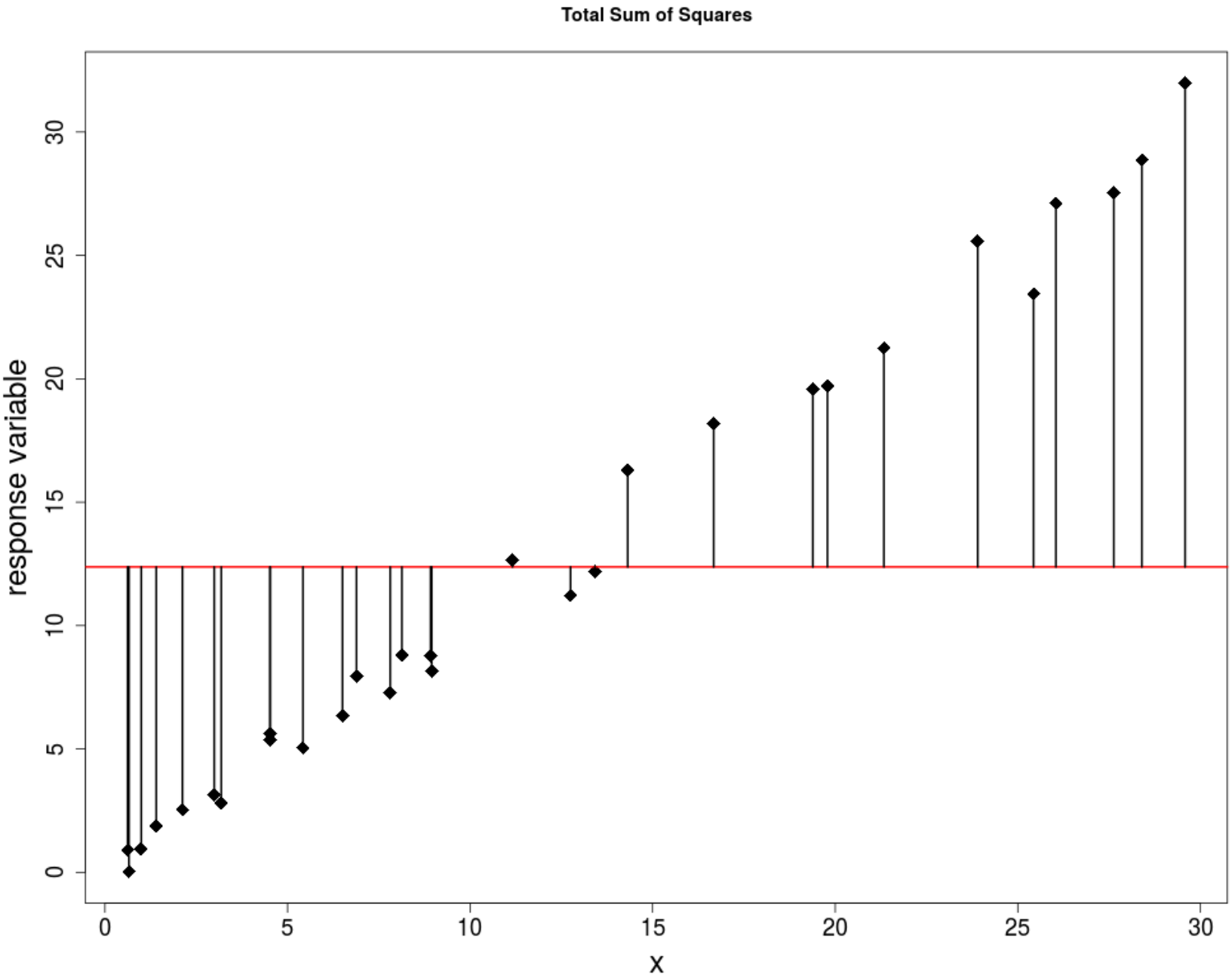
examples

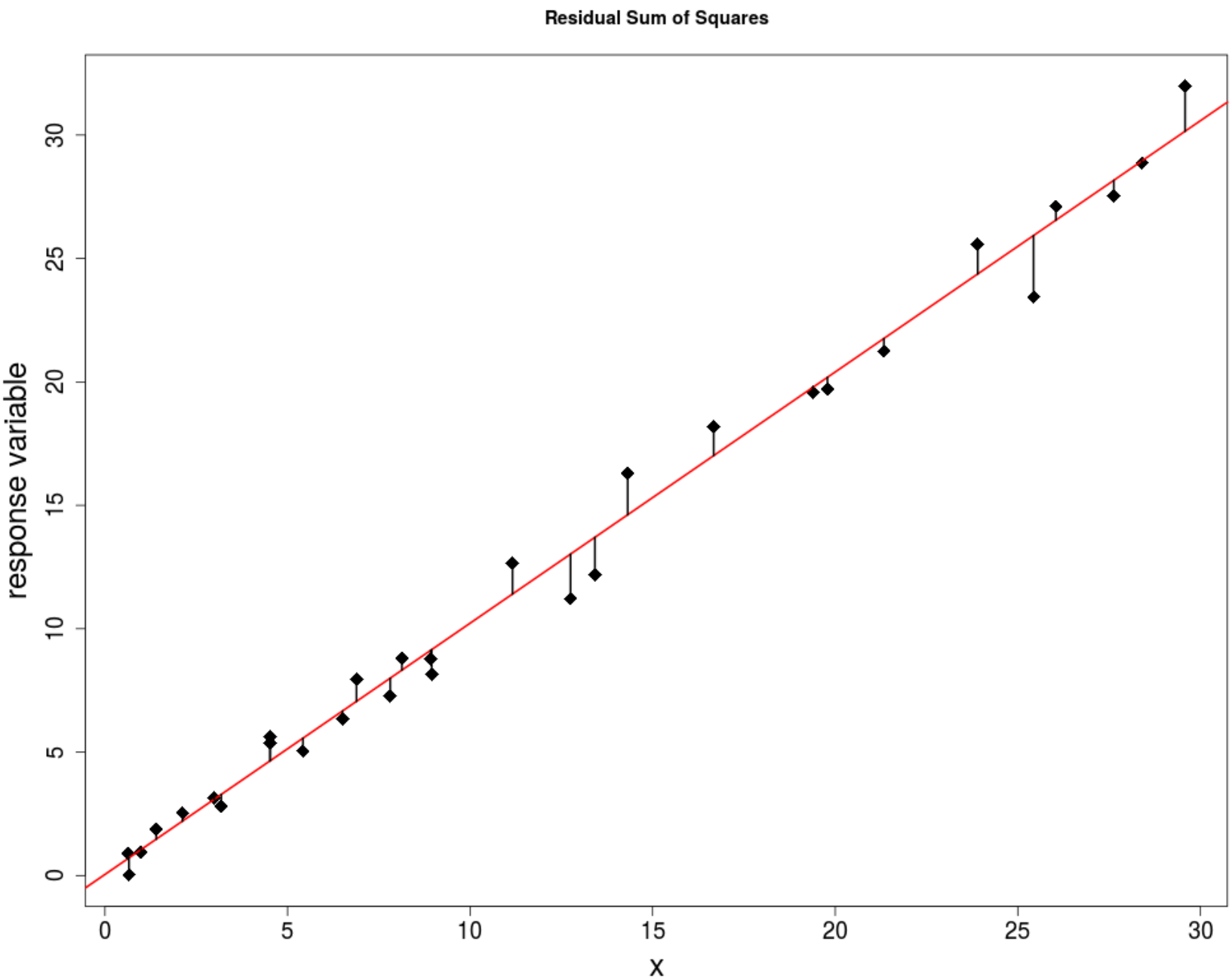
Model	`Formula`	Comments
Null	$y \sim 1$	is the intercept in regression models, but here it is the overall mean y
Regression	$y \sim x$	x is a continuous explanatory variable
Regression through origin	$y \sim x - 1$	Do not fit an intercept
One-way ANOVA	$y \sim sex$	sex is a two-level categorical variable
One-way ANOVA	$y \sim sex - 1$	as above, but do not fit an intercept (gives two means rather than a mean and a difference)
Two-way ANOVA	$y \sim sex + genotype$	genotype is a three-level catEegorical variable
Analysis of covariance	$y \sim x + sex$	A common slope for y against x but with two intercepts, one for each sex
Analysis of covariance	$y \sim x * sex$	Two slopes and two intercepts

examples

Model	`Formula`	Comments
Multiple regression	$y \sim x + z$	Two continuous explanatory variables, flat surface fit
Multiple regression	$y \sim x * z$	Fits an interaction term as well (x + z + x:z)
Multiple regression	$y \sim x + I(x^2) + z + I(z^2)$	Fit a quadratic term for both x and z
Multiple regression	$y \sim poly(x, 2) + z$	Fit a quadratic polynomial for x and linear z
Multiple regression	$y \sim (x + z + w)^2$	Fit three variables plus all their interactions up to two-way


Total Sum of Squares
response variable
Index
0
5
10
15
20
25
30


Total Sum of Squares
response variable
X
0
5
10
15
20
25
30


Residual Sum of Squares
response variable
X
0
5
10
15
20
25
30