Regression Model Development and Yet Another Regression Function

Werner Stahel
Seminar für Statistik, ETH Zürich

useR 2008, Dortmund

Data Analysis ... needs a system that gives MORE support.

I show you such a system for regression models

Regression ... is 80% of statistics that is worthwile.

Model checking and often model development is needed.

Overview of talk: Example, conclusions

Example Blasting

for tunnel excavation.

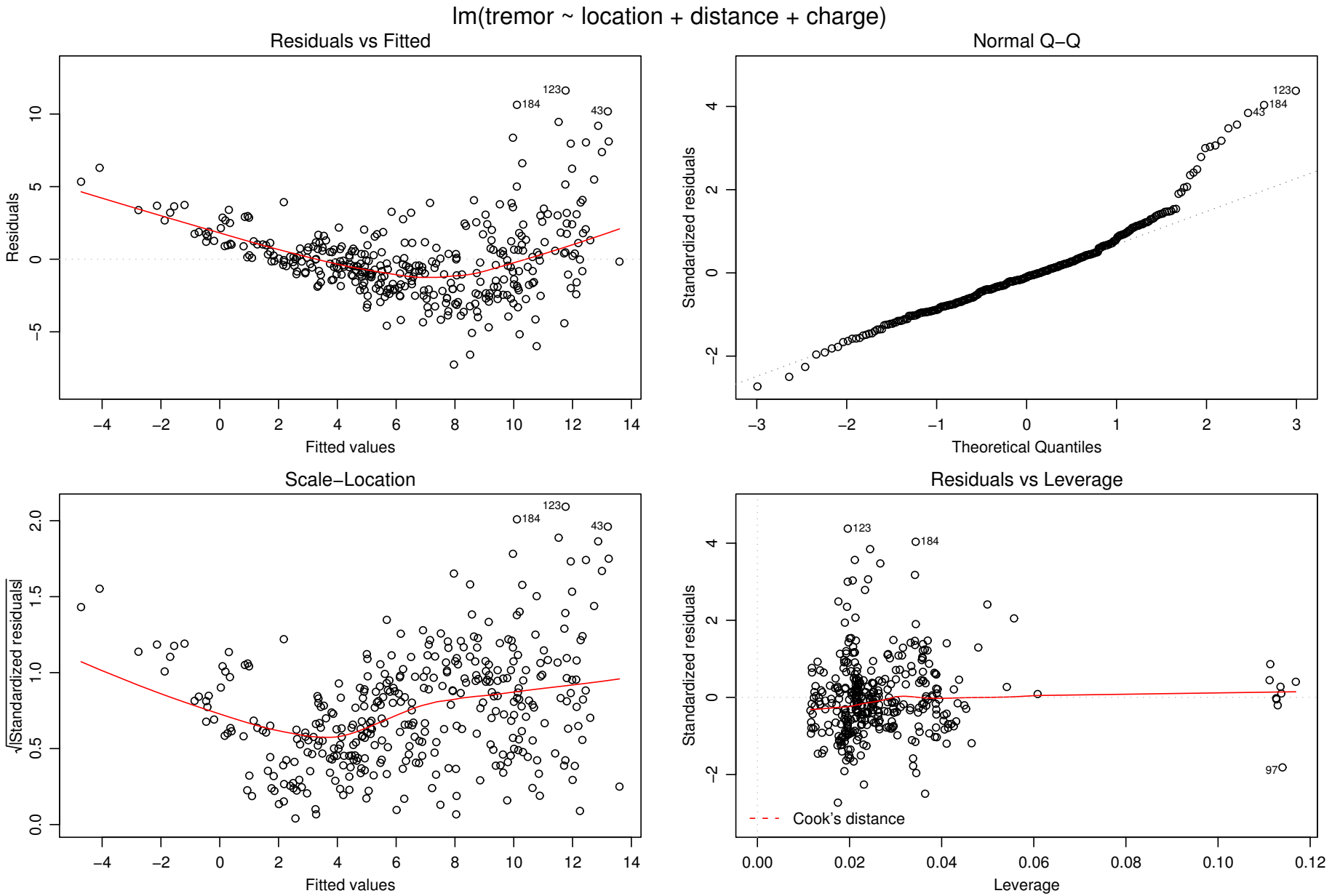
Tremor in a house (meas.site) must not exceed a threshold.
Need a forecast of

`tremor`	target variable, from
`distance`	between blasting site and measurement location
`charge`	and
`location`	house (factor)

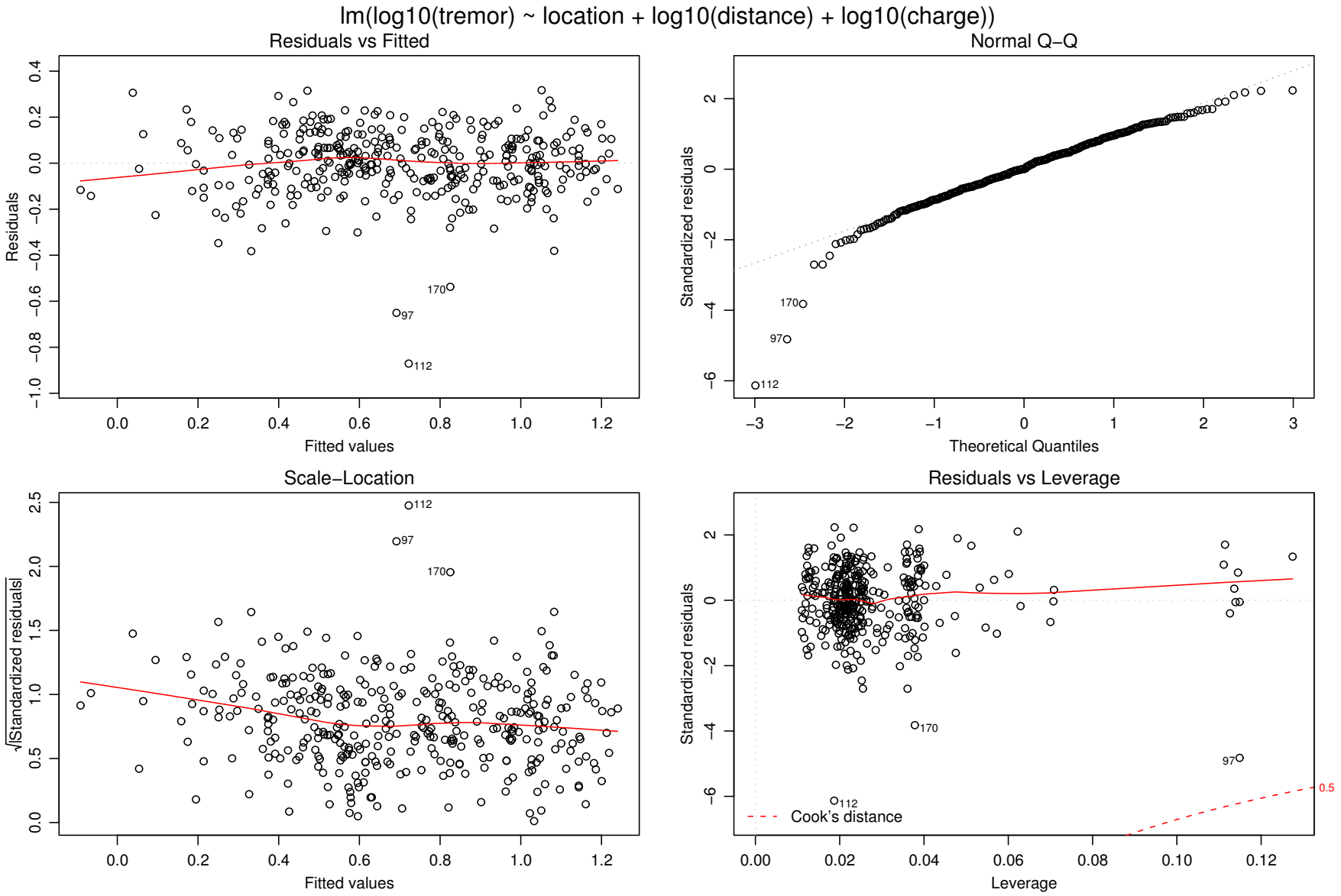
```
> showd(d.blast)
    charge distance tremor location
1    0.952       44   2.93     loc5
98   0.952       69   1.76     loc1
195  0.952      108   0.62     loc6

316  7.072       93   5.15     loc3
50   5.493       36  21.34     loc2
171  5.285       46  10.19     loc8
388  3.952       77   3.95     loc3
```

Yes, this is the data I wanted. $\longrightarrow$ Regression!

```
r.blast.lm0 <- lm(tremor~location+distance+charge,
  data=d.blast)
plot(r.blast.lm0)
```

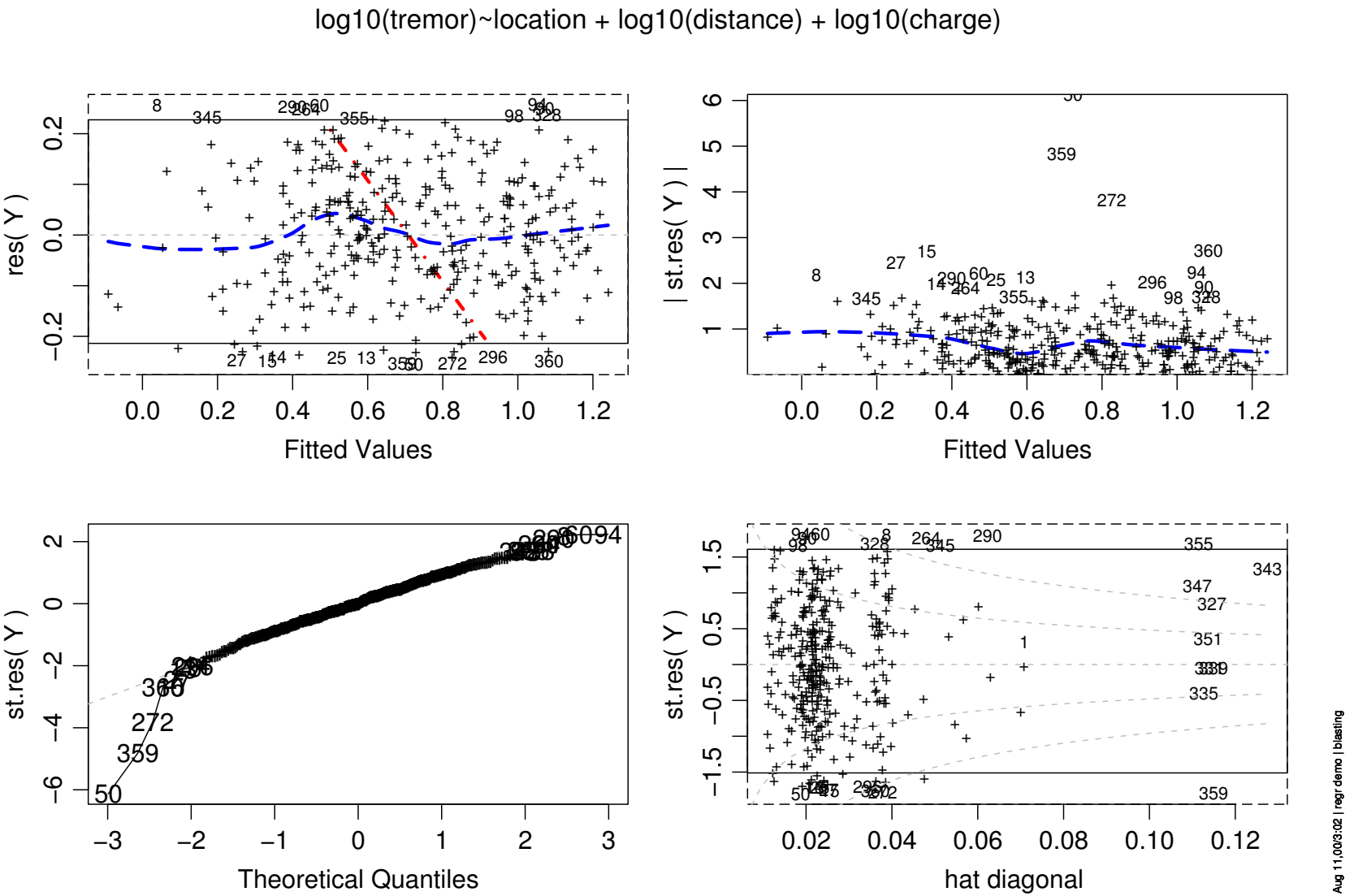

lm(tremor ~ location + distance + charge)
Residuals vs Fitted
Residuals
Fitted values
123
184
43
Normal Q–Q
Standardized residuals
Theoretical Quantiles
Scale–Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
97

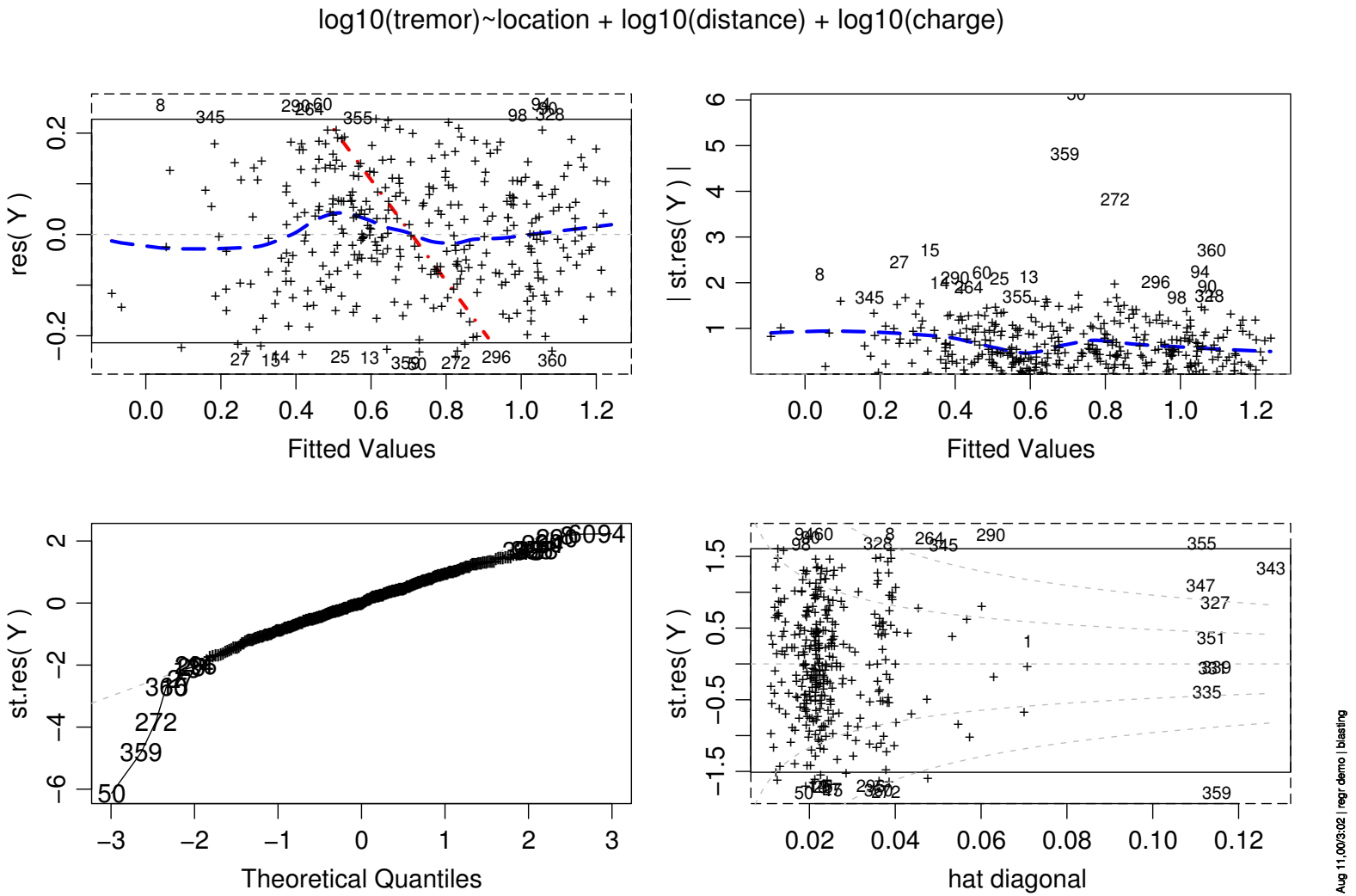
Wrong model! Take logs of `tremor`, `distance`, `charge`!



Using regr!

```
r.blast <- regr(log10(tremor)~location+
  log10(distance)+log10(charge), data=d.blast)
```

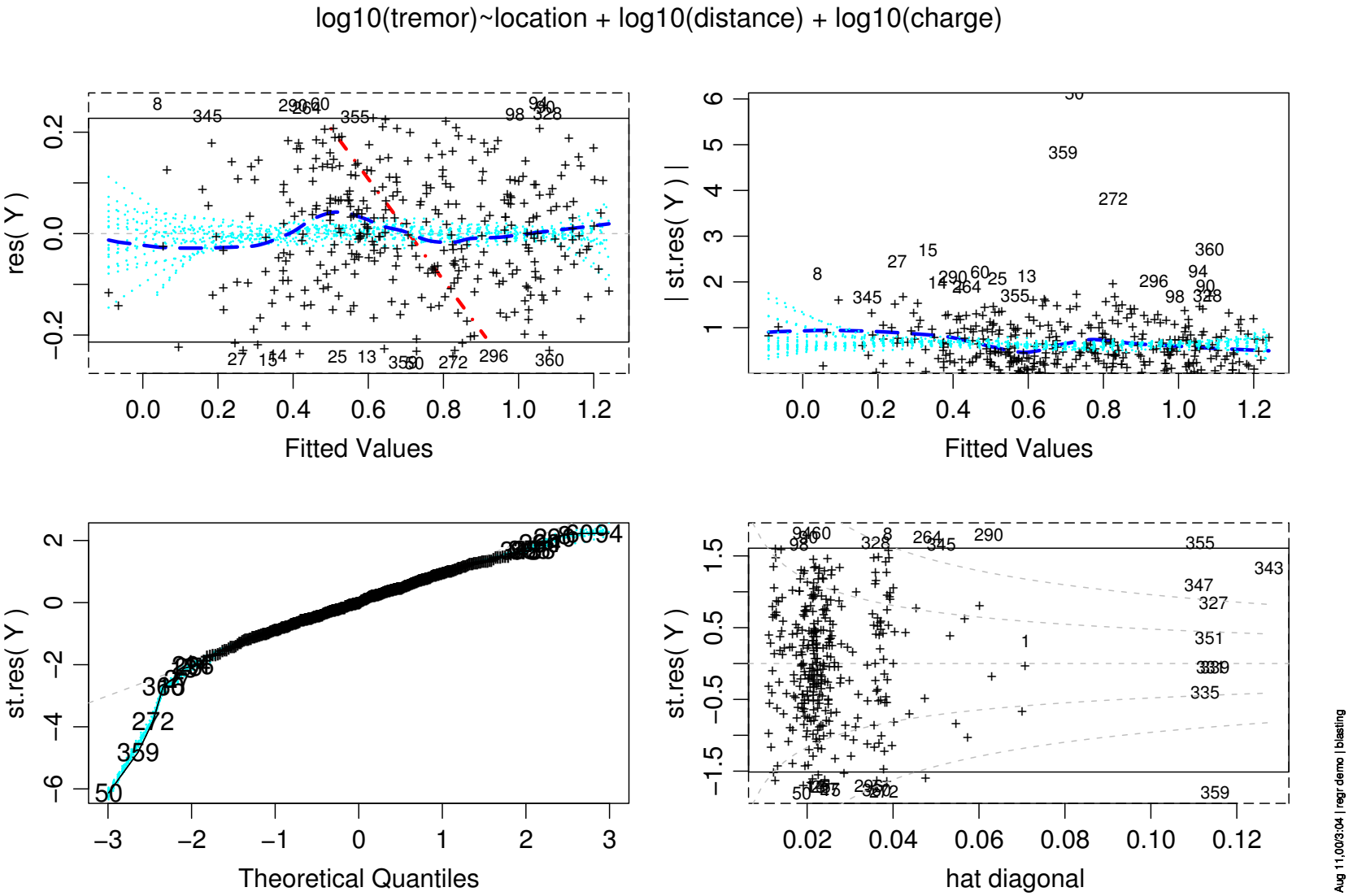




– Reference line $Y =$ constant
– show outliers in plot margin
– stamp: project, step, date

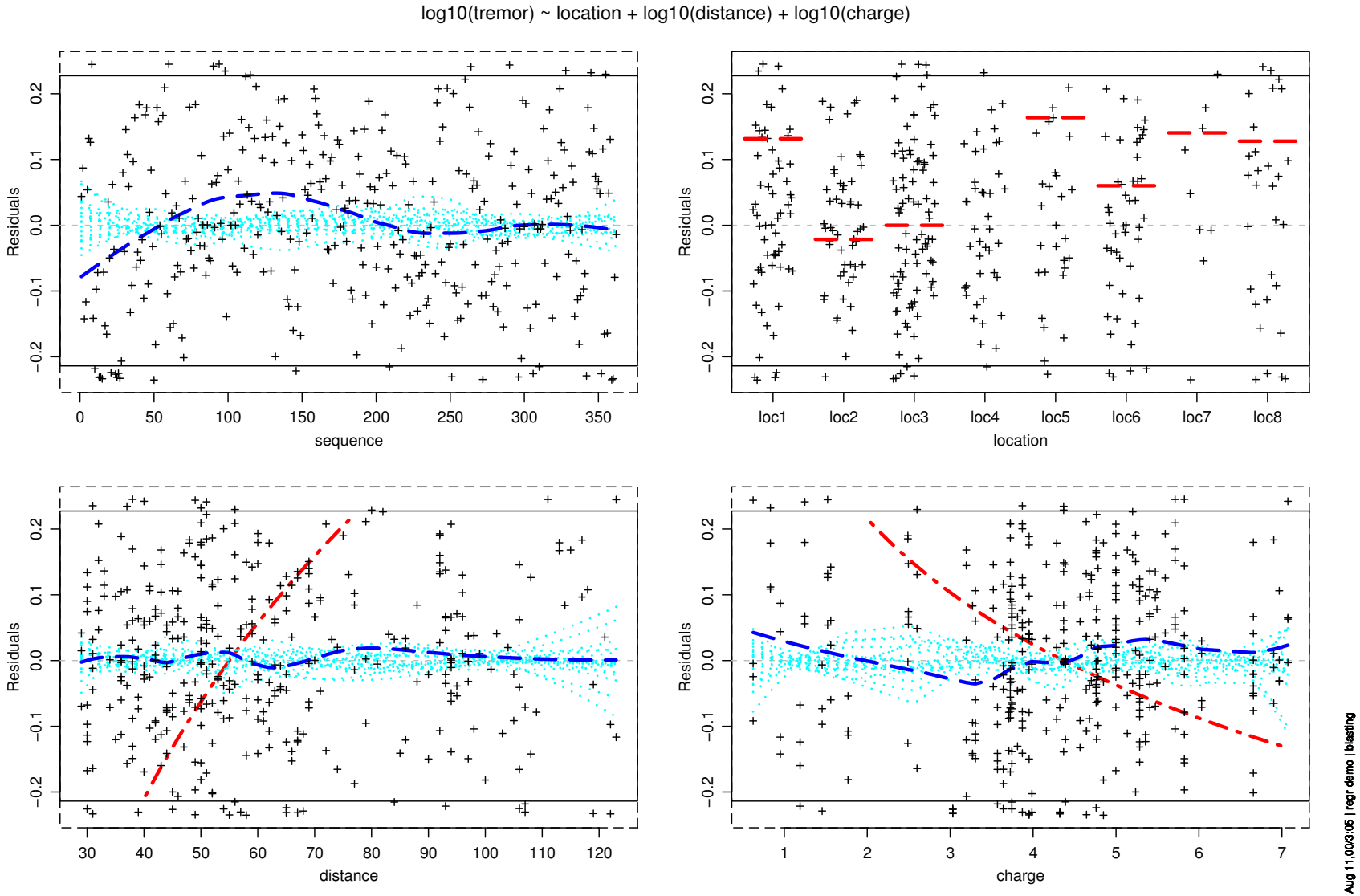
Deviations “significant”?

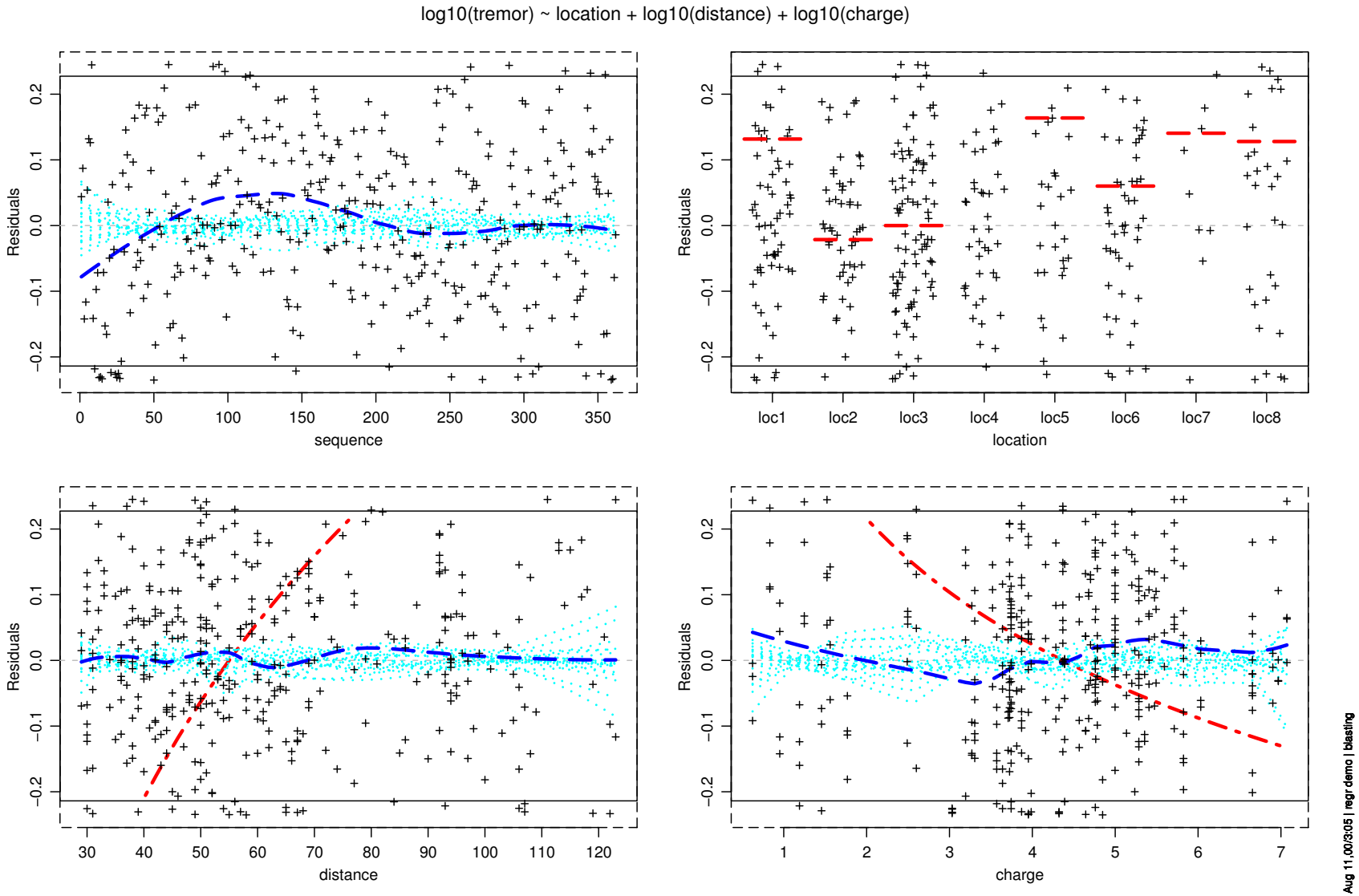
Deviations "significant"?



Residual analysis: always include plotting against X's!

Residual analysis: always include plotting against X's!





– Reference line $Y \approx$ constant (Resid + comp.effect = const.)

– Factors: use jittering

Numerical Results

```
summary(r.blast.lm1)

lm(formula = log10(tremor) ~ location + log10(distance)
    + log10(charge), data = d.blast)
Residuals: ...
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.96387    0.11090   26.72  < 2e-16 ***
locationloc2     0.15306    0.02701    5.67  3.0e-08 ***
locationloc3     0.13169    0.02592    5.08  6.1e-07 ***
locationloc4    -0.16185    0.03018   -5.36  1.5e-07 ***
locationloc5    -0.03211    0.03287   -0.98    0.329
...
log10(distance) -1.51830    0.06423  -23.64  < 2e-16 ***
log10(charge)    0.63558    0.03944   16.12  < 2e-16 ***
```

Is `location` significant? $\longrightarrow$ call `drop1`!

Output, continued:

```
Residual standard error: 0.143 on 352 degrees of freedom
  (26 observations deleted due to missingness)
Multiple R-squared: 0.795,Adjusted R-squared: 0.79
F-statistic:  152 on 9 and 352 DF,  p-value: <2e-16
```

"Residual standard error"? – oh my!

```
regr(formula = log10(tremor) ~ location + log10(distance)
    + log10(charge), data = d.blast)

                  coef stcoef signif   R2.x df p.value
(Intercept)      2.964  0.000   13.6     NA  1       0
location            NA     NA   10.5 0.0522  7       0
log10(distance) -1.518 -0.788  -12.0 0.2767  1       0
log10(charge)    0.636  0.410    8.2 0.0526  1       0

Coefficients for factors:
$location
    loc1     loc2     loc3     loc4     loc5     loc6 ...
 0.00000  0.15306  0.13169 -0.16185 -0.03211  0.07161 ...

St.dev.error:  0.143   on 352 degrees of freedom
Multiple R^2:  0.795   Adjusted R-squared: 0.79
F-statistic:    152   on 9 and 352 d.f.,  p.value:    0
```

A fair analysis of a model fit needs:

```
> r.lm <- lm(...)
> summary(r.lm)
> drop1(r.lm)
> plot(r.lm)
> plot(1:length(residuals(r.lm)), residuals(r.lm), xlab=...)
> termplot(r.lm, ...)
```

or

```
> ( r.regr <- regr(...) )
> plot(r.regr)
```

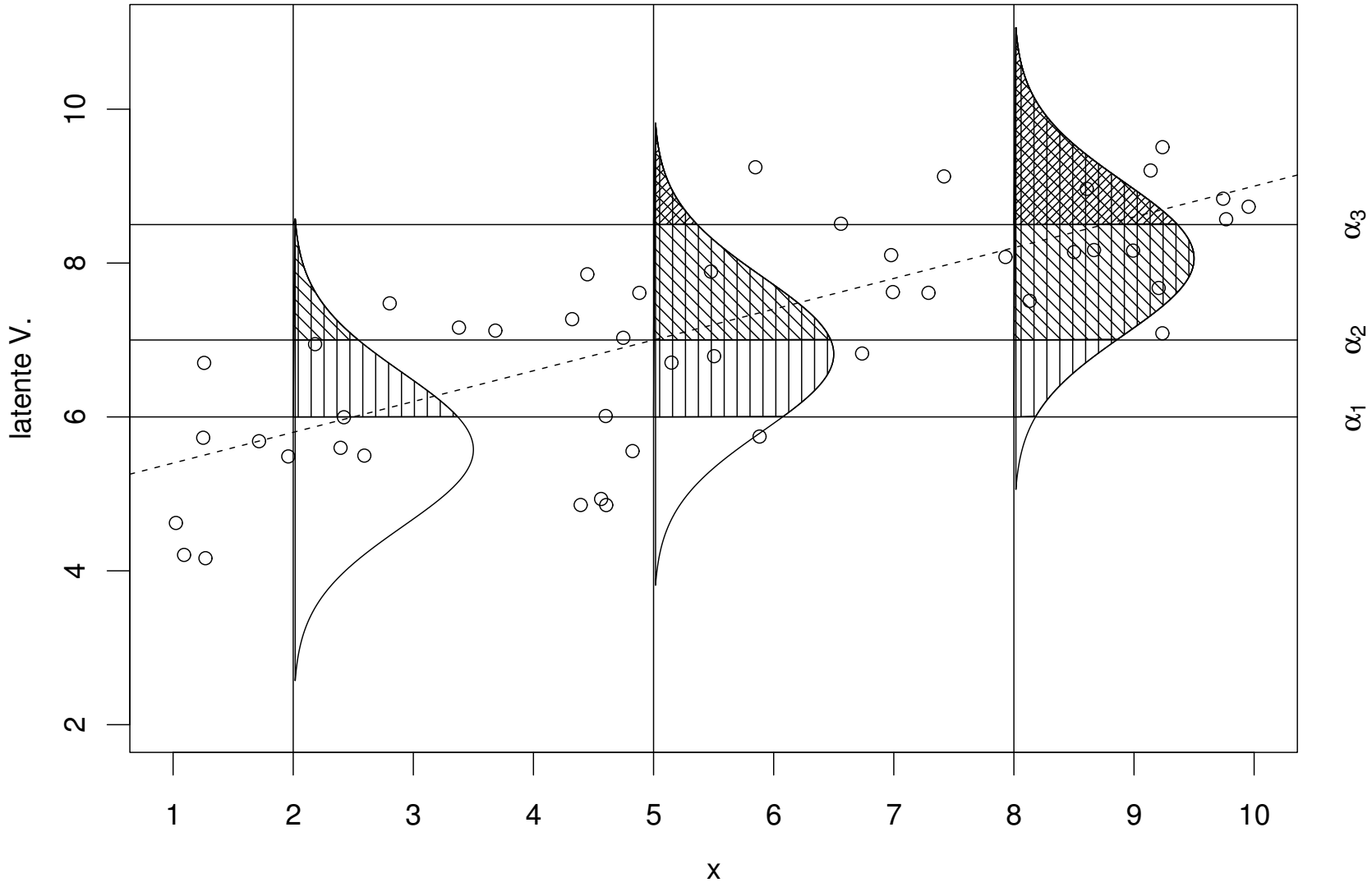
... and you get more from it!

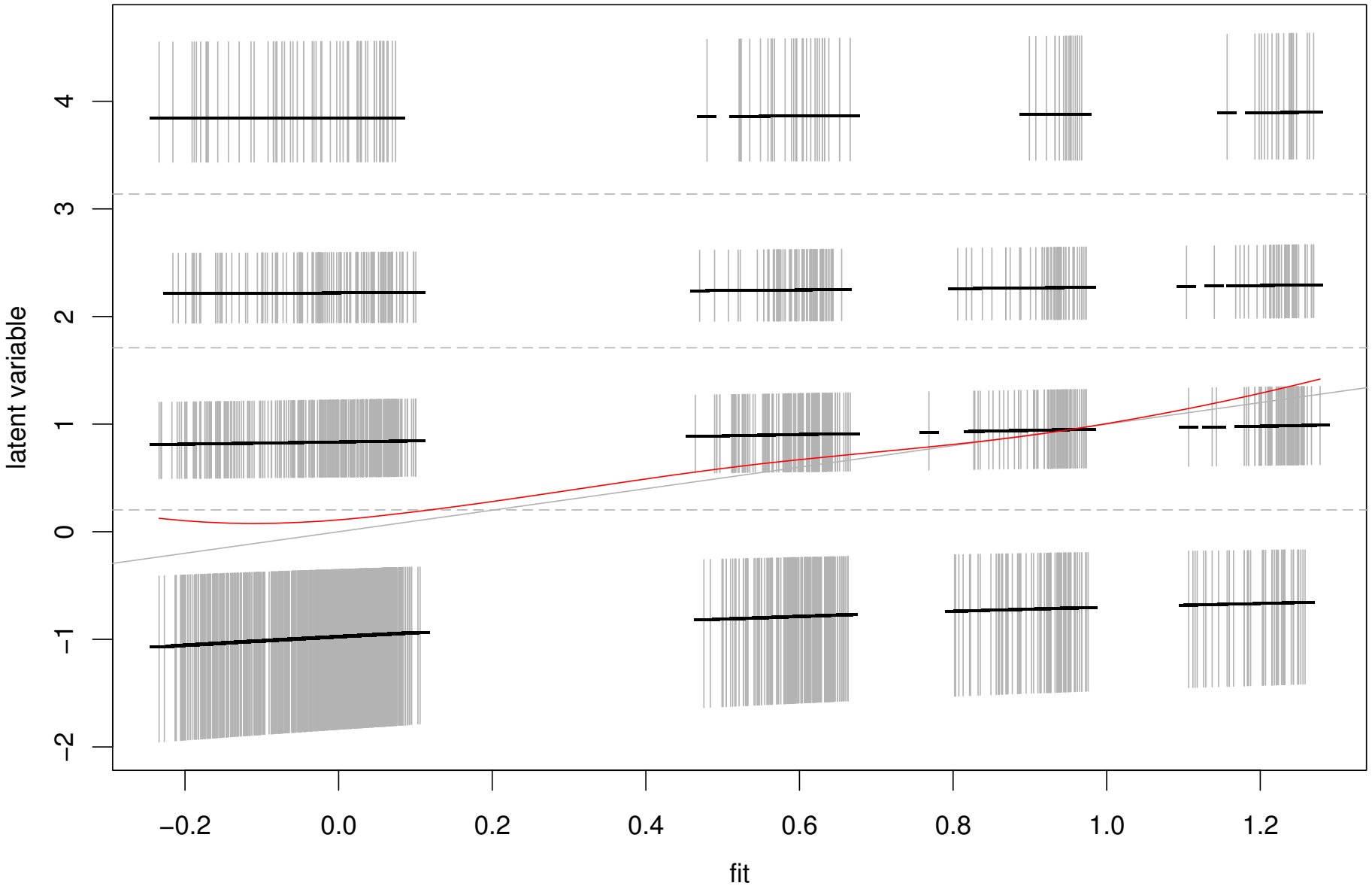
More Features

- Same wrapper function for `glm`, `polr`, `multinom`, `roblm`
- Additional information
- `plot.regr` includes specific features for `polr`, multivariate, ...
- Model selection function for adaptive lasso method: `lassoselect`
- Utilities ...

A goody: Residuals for ordered response variables

The model for `polr` relies on a latent variable




latent variable
4
3
2
1
0
−1
−2
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
fit

Messages

Package `regr`

- Model building and model checking are best done following (flexibly) a strategy.
  The strategy should be easily performed with the help of user oriented functions.

- The information needed may be more condensed and consistent between models:
  – numerical summary: More useful columns in table of terms
  – graphical: reference lines, random variation, outlier treatment

- Documentation of data sets, graphs, ...

- The package is still in development. Can be obtained from me (email).