Miscellanous topics on linear models

Gilles Guillot

gigu@dtu.dk

October 8, 2013

1. Analysis of residuals
2. Variable selection

Assumptions on residuals

- In the model $y_i = \sum_j \beta_j x_{ij} + \varepsilon_i$, we make the assumption that $(\varepsilon_i) \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.
- All the tests that we make subsequently rely on this assumption.
- The residuals are not observed. The assumption that $(\varepsilon_i) \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ can not be tested directly

Estimated residuals

If we have an estimate $(\widehat{\beta_j})$ of (β_j) we have also an estimate of the residuals:

$$\widehat{\varepsilon_i} = y_i - \sum_j \widehat{\beta_j} x_{ij}$$

How does the distribution of $(\widehat{\varepsilon_i})$ relate to that of (ε_i)?

Distribution of estimated residuals

It can be shown that

$$V(\widehat{\varepsilon_i}) = (1 - h_{ii})\sigma^2$$

where

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})}$$

This shows that even if the *true* redisuals have the same variance, the *estimated* residuals will not.

Normalized residuals

- We denote by $\widehat{\sigma^2}$ the unbiased estimate of the residual variance (e.g. $\widehat{\sigma^2} = \frac{1}{n-p-2}\sum_i (y_i - \hat{y}_i)^2$ for a multivariate regression with p explanatory variables)
- We estimate $V(\widehat{\varepsilon_i})$ by $\widehat{V(\widehat{\varepsilon_i})} = (1 - h_{ii})\widehat{\sigma^2}$

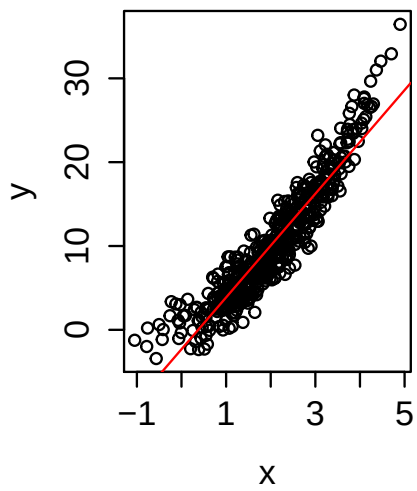
Normalized residuals

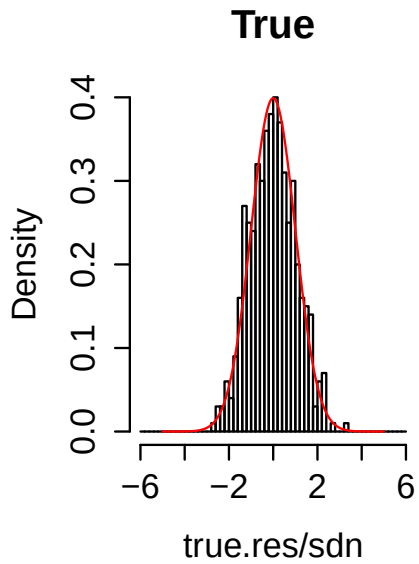
We define the normalized residuals as

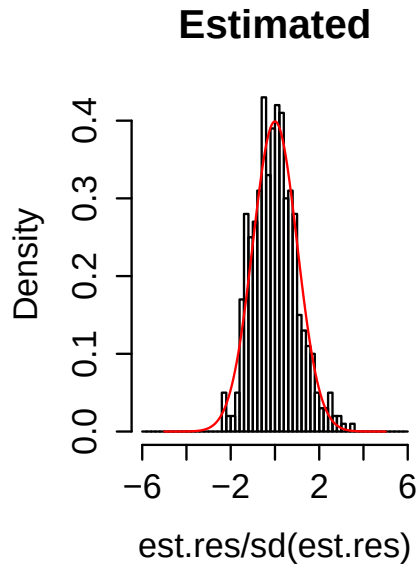
$$r_i = \frac{\widehat{\varepsilon_i}}{(\widehat{V(\widehat{\varepsilon_i})})^{1/2}} \quad i = 1, ..., n$$

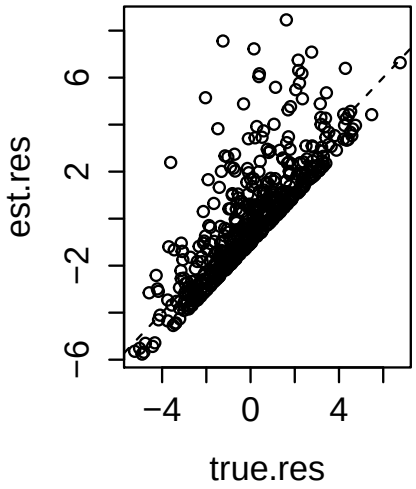
Diagnostic based on the normalized residuals

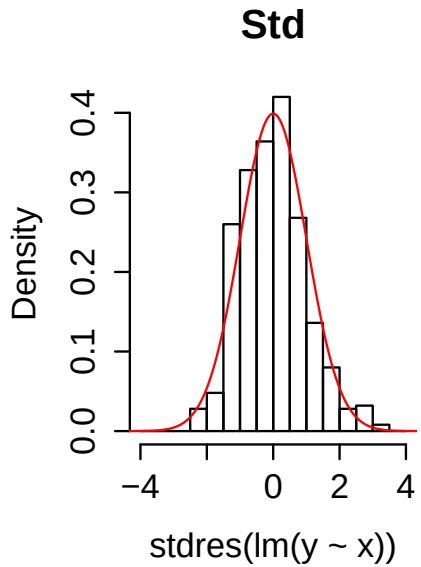
- It can be shown that the normalized residuals r_i's are almost i.i.d $\mathcal{N}(0,1)$.
- If there is any sign that (r_i) departs from an i.i.d $\mathcal{N}(0,1)$ sample, the p-value of all tests should not be trust too much.

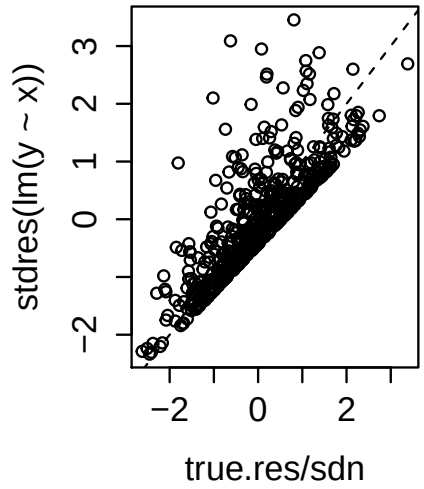

y
x
−1
1
3
5
0
10
20
30


True
Density
true.res/sdn
−6
−2
2
6
0.0
0.1
0.2
0.3
0.4


Estimated
Density
est.res/sd(est.res)
−6
−2
2
6
0.0
0.1
0.2
0.3
0.4


est.res
true.res
−6
−2
2
6
−4
0
4


Std
Density
stdres(lm(y ~ x))
−4
0
2
4
0.0
0.1
0.2
0.3
0.4


stdres(lm(y ~ x))
true.res/sdn
−2
0
1
2
3
−2
0
2

Variable selection problem

Example: in the study of marathon running times we contemplate $\mathcal{M}_1 : Time \sim Age$ versus $\mathcal{M}_2 : Time \sim Weight$

Two models with same number of variables

Using R^2 or $\mathcal{L}$ is OK to compare models with the same number of variables

Variable selection problem

In the study of marathon running times we contemplate $\mathcal{M}_1 : Time \sim Age$ versus $\mathcal{M}_2 : Time \sim Age + Weight$

Nested models

- Chosing the model achieving the highest R^2 or the highest likelihood $\mathcal{L}$ is **not** a good idea.
  - In the case above, $\mathcal{M}_1$ is nested in $\mathcal{M}_2$, i.e. $\mathcal{M}_1$ is a particular case of $\mathcal{M}_2$.
  - For nested models, by construction $R^2_{\mathcal{M}_1} \leq R^2_{\mathcal{M}_2}$ and $\mathcal{L}^2_{\mathcal{M}_1} \leq \mathcal{L}^2_{\mathcal{M}_2}$

Variable selection problem

In the study of marathon running times we contemplate $\mathcal{M}_1 : Time \sim Age$ versus $\mathcal{M}_2 : Time \sim Height + Weight$

Non-nested models of different dimensions

- If $\mathcal{M}_1$ is not nested in $\mathcal{M}_2$ but contain less variables than $\mathcal{M}_2$, using R^2 or $\mathcal{L}$ still incurs a risk of choosing $\mathcal{M}_2$

The model selection 'Graal'

We need a method for selecting a model among a set of competing models of different dimensions.

We want this method to be

- theoretically well grounded
- easy to implement
- yields meaningfull results in practice

In many situations, such a method does not exist.

The problem with R^2_{adj}

$$R^2 = 1 - \frac{RSS}{SST}$$

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)}$$

Even though R^2_{adj} includes a penalty term linear in the number of parameters, in practice it often tends to favour models that overfit the data.

The Akaike information criterion (AIC)

Definition

The AIC is defined as

$$AIC(\mathcal{M}) = 2K - 2\log\mathcal{L}$$

where K is the number of parameters and $\mathcal{L}$ is the maximum likelihood achieved by the model.
For a multivariate regression with p variables $K = p + 2$

When contemplating two models, we chose the one that achieves the lowest AIC.

Combinatorial issues with variable selection

- For a data set with up to p explanatory variables, there are 2^p sub-models.
- Brute force strategy (evaluating the maximum likelihood for all models) often unfeasible

Heuristic n case of combinatorial issues: step-wise selection

- In practice one has to use heuristics i.e. approximate solutions.
- One of them is the step-wise selection procedure.
- It consists in exploring a small subset of all possible models and select the best of them in such a way that the final model is not 'too bad'.

Step-wise selection procedure

Backward selection

- starts from the full model (all variables)
- discard variable than incurs the smallest increase of AIC
- iterate step above
- Chose model achieving lowest AIC among those visited

Forward selection

- starts from the null model (containing only the constant mean)
- add the variable than incurs the largest decrease of AIC
- iterate step above
- Chose model achieving lowest AIC among those visited

Stepwise variable selection with R

The R function step()

```
## Fit a starting model (usually null model or full model)
starting.model = lm(data=...,formula=...)

## perform stepwise selection
step(object=starting.model, direction=...,scope=...)
```

Argument direction='forward' or 'backward'
Argument scope is a list with components lower and upper
Note there is an internal stopping rule which may prevent exploring all the models specified via the scope argument

Example of stepwise variable selection with R

```
full = lm(data=health,
          formula=(death.rate ~ 1 + doctor + hospital + income
                   + pop.density)
step.back = step(full,direction='backward',
                 scope=list(lower=(death.rate ~ 1),
                            upper=(death.rate ~ .)))
```

Example of stepwise variable selection with R cont'

```
step.back
Start:  AIC=54.65
death.rate ~ 1 + doctor + hospital + income + pop.density

              Df Sum of Sq    RSS    AIC
- hospital     1    1.6763 124.75 53.369
- doctor       1    2.9139 125.99 53.892
<none>                     123.07 54.652
- income       1    5.0825 128.16 54.797
- pop.density  1    9.6144 132.69 56.639

Step:  AIC=53.37
death.rate ~ doctor + income + pop.density

              Df Sum of Sq    RSS    AIC
<none>                     124.75 53.369
- doctor       1    5.1882 129.94 53.529
- income       1    6.1544 130.91 53.921
- pop.density  1    8.3192 133.07 54.791
```