MASM22/FMSN30: Linear and Logistic Regression, 7.5 hp
FMSN40: ... with Data Gathering, 9 hp

Lecture 6, spring 2019

Regression diagnostics

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

10/4-19

Problem areas in least squares

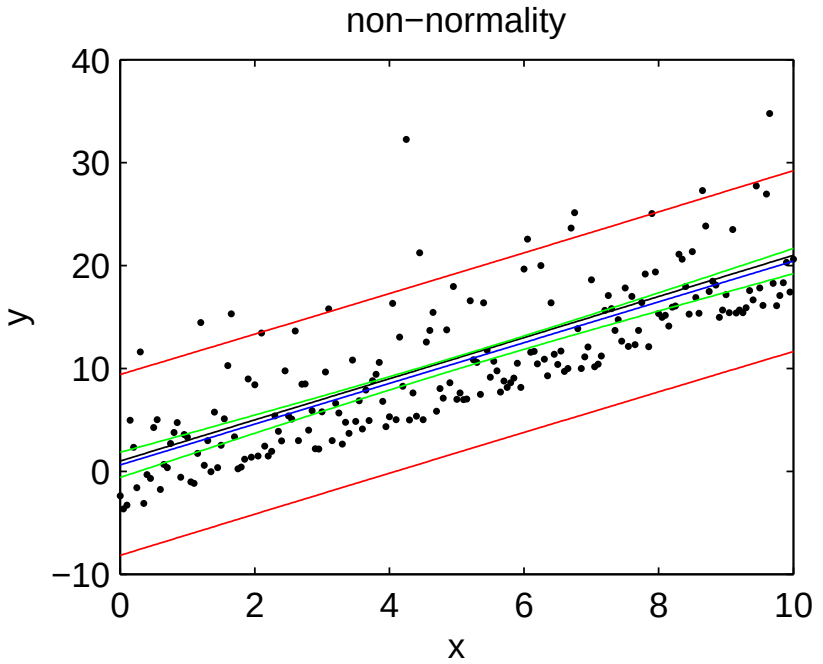
We assume:

1. additive errors ϵ_i
2. Gaussian errors
3. independent errors
4. homoscedastic errors (constant variance)

- When (3)–(4) hold and $\hat{\beta}$ from OLS, then $Var(\hat{\beta})$ minimal among all unbiased estimators of β.
- When (2) holds: least squares $\equiv$ maximum likelihood
- We do not need (2)–(4) to prove that $E(\hat{\beta}) = \beta$.
- What is tricky is verify (2)–(4).
- Assumptions allow construction of inference procedures. Not necessary to numerically compute least squares estimates.

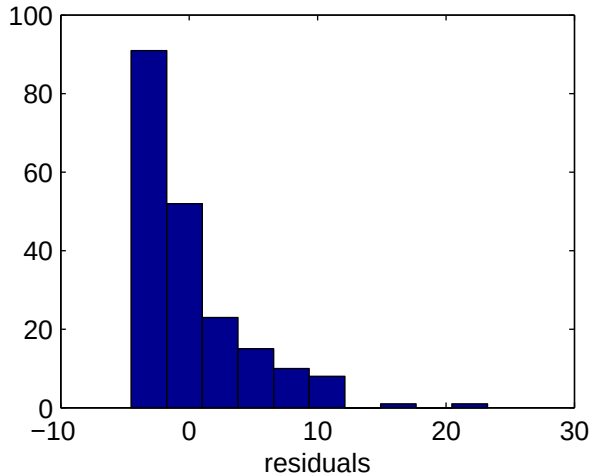
Non-normal ϵ_i

- Confidence and prediction intervals will be more or less wrong, particularly with skewed distributions.



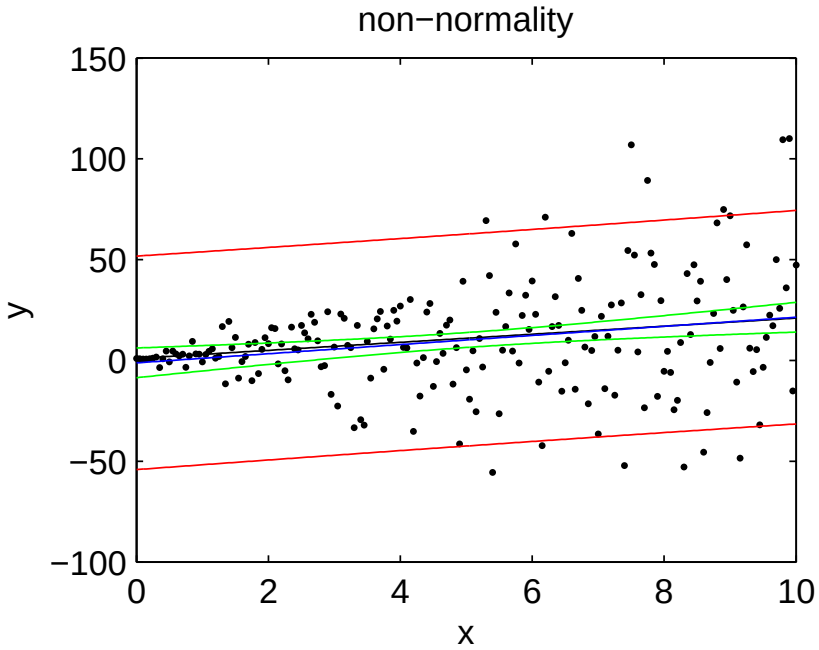
Found by: Histogram, qqplots etc. of residuals.
Solutions:

- Transformations, e.g. $\ln(Y_i)$
- Use other methods that can handle the true distribution (maximum-likelihood, bootstrap, etc.)



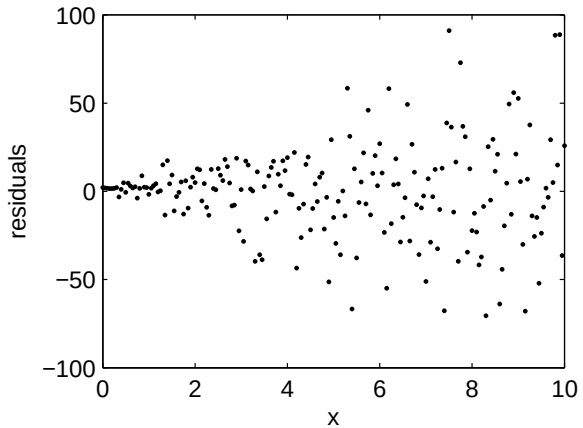
Heterogenous variance

- $V(\epsilon_i) \neq \sigma^2$ for all ϵ_i. Often larger variance with larger mean.
- Uncertain observations have too much influence on the estimates.
- Prediction intervals will be wrong.

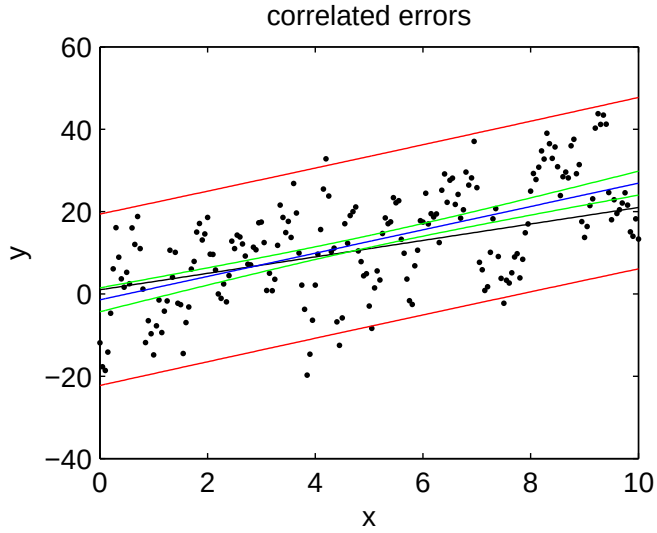


Found by: Plot of residuals against $\hat{Y}$.
Solutions:

- Transformations, e.g. $\ln(Y_i)$
- Weighted least squares (less weight to observations with larger variance).



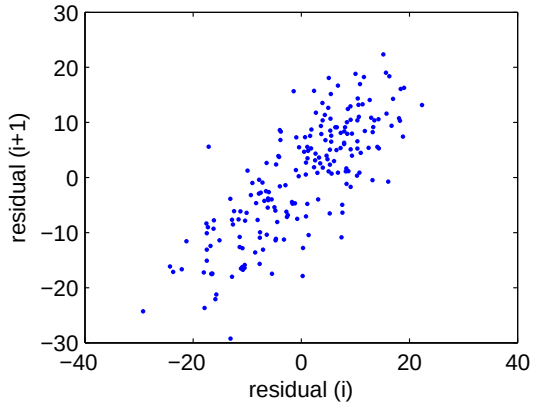
Correlated errors



- $C(\epsilon_i, \epsilon_j) \neq 0$ for some $i \neq j$ (e.g. for $j = i + 1$). Often in time-series data.
- Variance estimates ($V(\hat{\beta}_i)$) will be biased: too small (if positive correlation) or too large (if negative correlation).
- Confidence (and prediction) intervals will be too narrow or too wide.

Found by: Plot residuals against next residual. Autocorrelation plots.
Solutions:

- Time-series, e.g. AR-model, MA-model.
- generalized least squares



Influential points and outliers

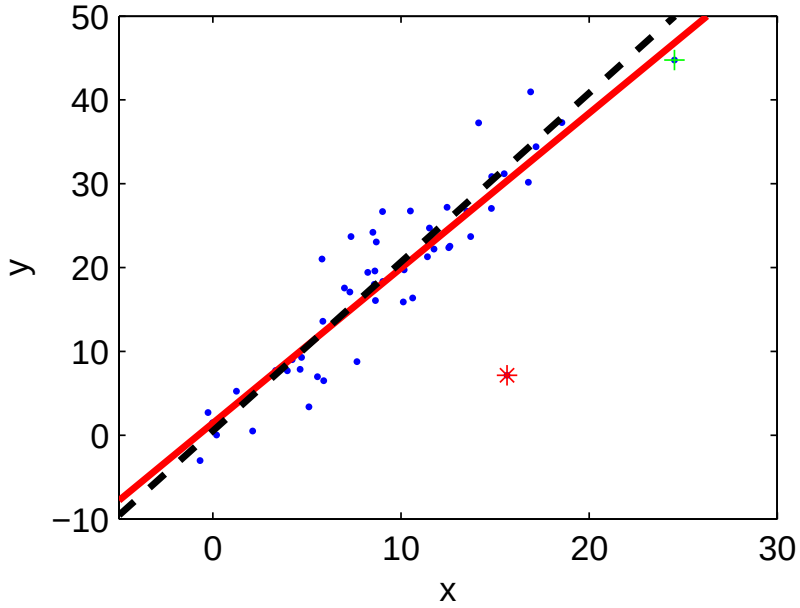
Individual observations, far from the others, that can have a large influence on the estimates of β and σ^2, and thus on predictions and statistical conclusions.

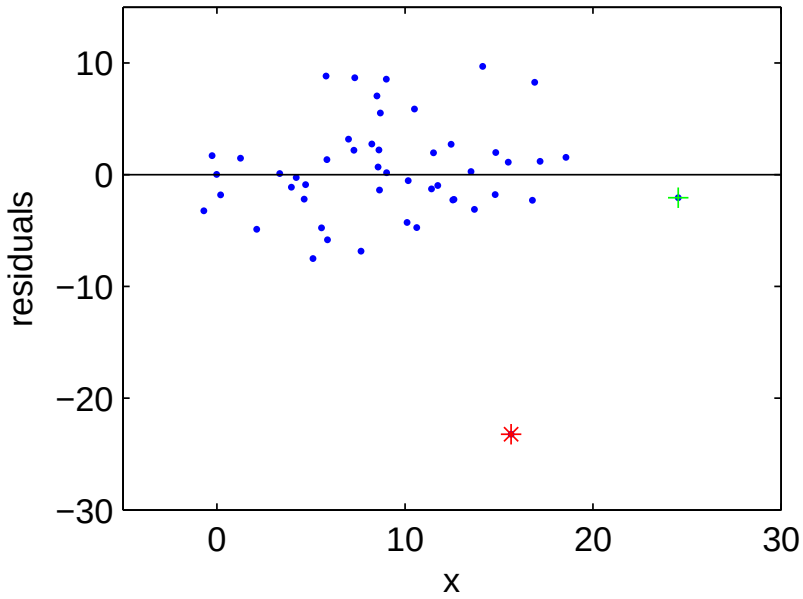
- Outlier: in some sense inconsistent with the rest (Y-wise).
- Outlier in residual: Unexpectedly large ($\pm$) residual
- Potentially influential point: outlier in the space spanned by the columns of $\mathbf{X}$.

Causes (and remedies):

- Faulty measurement equipment (correct it or leave it out)
- Coding error (correct it or leave it out)
- Wrong or inadequate model (refine the model)
- an "interesting" (and unexpected) measurement result escaping conventional models (revise theory/knowledge of the phenomenon at study). Might lead to a discovery!

Example: Outlier (red $*$) and potentially influential point (green $+$).
Estimated line with these points (red) and without them (black).





Find them using a combination of plots and influence measures.
Plot on the right shows residuals obtained using all data points.

Leverage (outliers w.r.t. X)

Given the usual multivariate regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ it is possible to write:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{P}\mathbf{Y}$$

where

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Denote with v_{ij} the generic element of $\mathbf{P}$.

We can then write

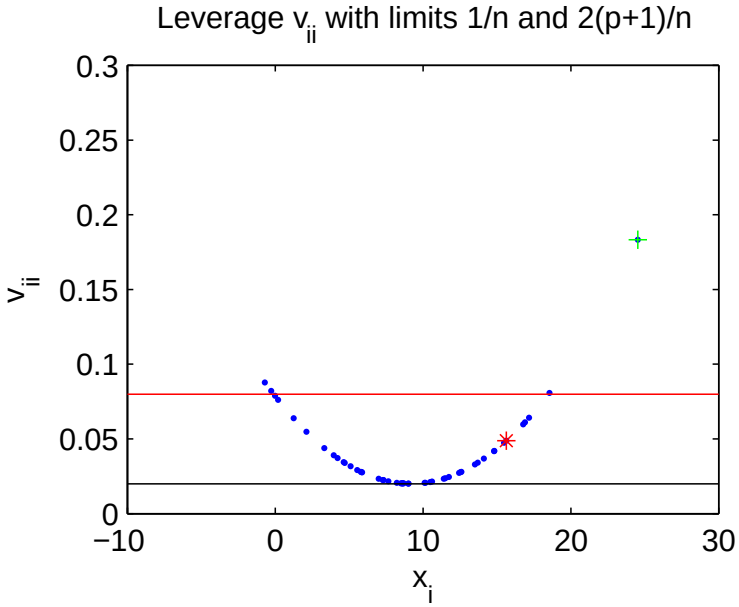
$$\hat{Y}_i = v_{i1}\,Y_1 + \cdots + v_{ii}\,Y_i + \cdots + v_{in}\,Y_n$$

and "leverage" (="ability to influence the estimates"), v_{ii}, measures the impact of Y_i on its own estimated value $\hat{Y}_i$.

Potentially influential points are those far from the centre of **X**-space. The leverage v_{ii} gives a distance from such centre. We have that

$$\frac{1}{n} \leq v_{ii} \leq \frac{1}{c} \text{ (model with intercept)}$$

$$0 \leq v_{ii} \leq \frac{1}{c} \text{ (model without intercept)}$$



where c ($\geq$ 1) is the number of observations with identical X-values.

Also, v_{ii} is minimal when X_i is the centre point. If $v_{ii} = 1/c$ then point i will force the estimated line through itself.

Leverage above $2(p+1)/n$ can be considered high. The points with most extreme X-values have the highest leverage, in one case very high.

A case having high leverage may not be actually influential!

Residual analysis (outliers w.r.t. Y)

In our model, we have assumed that $\epsilon_i \sim N(0, \sigma^2)$ and independent, i.e.

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) = N(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix})$$

- Is this true? How can we find out, since we cannot observe ϵ_i?
- When normality holds, residuals e_i should behave in a certain way. Check this instead.

Residuals

If $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ then the observed residuals,

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{PY} = (\mathbf{I} - \mathbf{P})\mathbf{Y}$$

have the following property:

$$\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{P}))$$

where $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = (v_{ij})$. Thus they will have unequal variances and be dependent ($Var(\mathbf{e})$ is not a diagonal matrix), the unequality and dependence determined by the structure of $\mathbf{X}$.

Because of different variances it is tricky to compare the residuals e_i.

So let's standardize those . . . (we'll see there are issues . . .)

Proofs

- **Normality**: Since $\mathbf{e}$ are linear combinations of $\mathbf{Y}$, which are multivariate normal, $\mathbf{e}$ will also be multivariate normal.
- **Zero mean**

$$\begin{aligned} E(\mathbf{e}) &= E((\mathbf{I}-\mathbf{P})\mathbf{Y}) = (\mathbf{I}-\mathbf{P})E(\mathbf{Y}) \\ &= (\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{X}\beta = (\mathbf{X}-\mathbf{X})\beta = \mathbf{0} \end{aligned}$$

- **Covariance matrix**
  Since $\mathbf{P}' = \mathbf{P}$ and $\mathbf{PP} = \mathbf{P}$ we get

$$\begin{aligned} Var(\mathbf{e}) &= Var((\mathbf{I}-\mathbf{P})\mathbf{Y}) = (\mathbf{I}-\mathbf{P})Var(\mathbf{Y})(\mathbf{I}-\mathbf{P})' \\ &= (\mathbf{I}-\mathbf{P})\sigma^2\mathbf{I}(\mathbf{I}-\mathbf{P}) = \sigma^2(\mathbf{I}-\mathbf{P}) \end{aligned}$$

Standardized residuals

Standardize the residuals, subtract the mean ($= 0$) and divide by the (estimated) standard deviation, to have variance approximately equal to 1:

$$r_i = \frac{e_i}{s\sqrt{1 - v_{ii}}}$$

where $v_{ii} = i$:th diagonal element of $\mathbf{P}$ and

$$s^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - (p + 1)}.$$

The r_i have similar variances ($\simeq 1$) but are still dependent and have no nice distribution: while e_i is normal and $(n-(p+1))s^2/\sigma^2$ is χ^2 they are not independent so r_i **will not be t-distributed**.

Studentized residuals

All e_i are included in s^2 so a large residual will contribute to a large s^2 affecting all the other standardized residuals. Reduce this influence by using

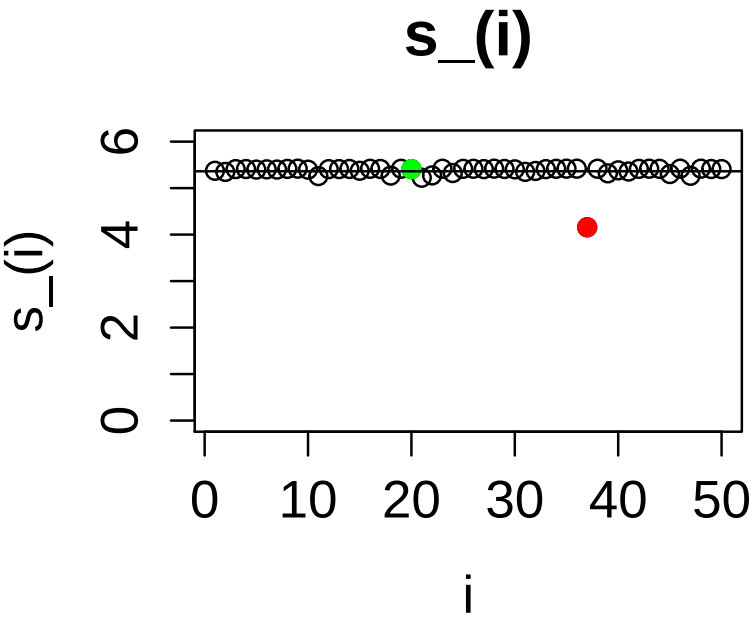
$$r_i^* = \frac{e_i}{s_{(i)}\sqrt{1 - v_{ii}}}$$

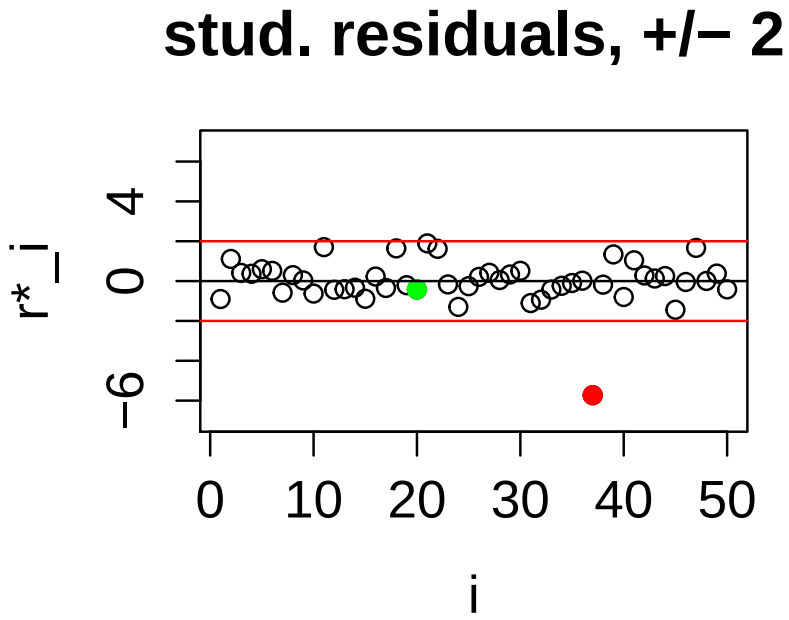
where $s^2_{(i)}$ is the variance estimate from a regression where observation i is excluded. Now e_i and $s_{(i)}$ are independent so that

$$r_i^* \sim t_{n-1-(p+1)} \qquad \text{when } \boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2).$$

The r_i^* are still not independent of each other, though.

Removing an outlier will reduce the σ-estimate and increase the studentized residual:





When $n \to \infty$, $t_n \to N(0,1)$ (a student's t variable has distribution approaching N(0,1) as n grows).

We can thus consider a studentized residual as suspiciously large when $|r_i^*| > 2$ (the $2 \approx 1.96$ is the 2.5 % quantile from a N(0,1)).

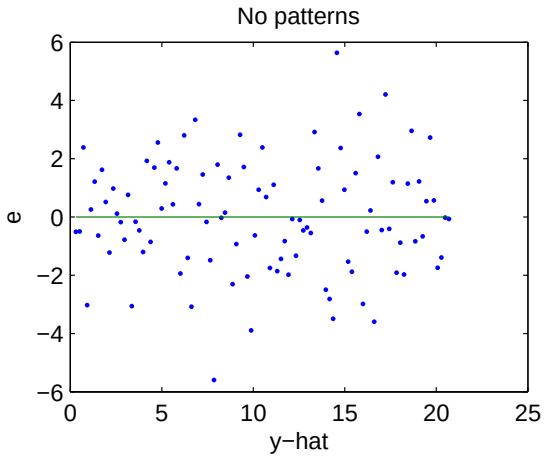
Plot!

All residuals should look like random variation around zero.
Nonrandom patterns indicate some inadequacy of the fitted model.

- e_i vs $\hat{Y}_i$: Finds structural inadequacies of the model ("need a quadratic term?").
- r_i^* vs $\hat{Y}_i$ finds points with unusually large residuals.
- r_i^* vs X_i finds outliers in the residuals and structural inadequacies ("which X_i needs a quadratic term?").
- e_{i+1} vs e_i finds correlation between successive residuals, or use `acf()`

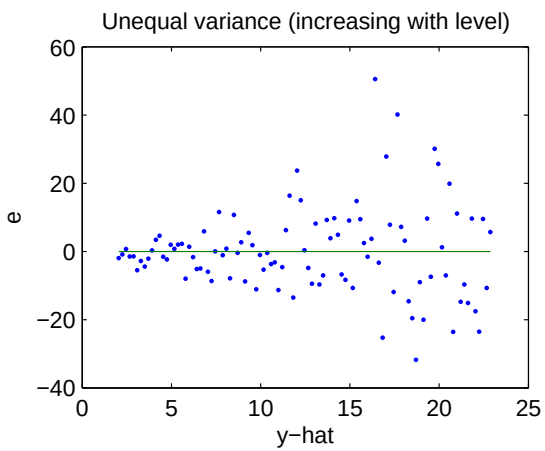
Correct model:

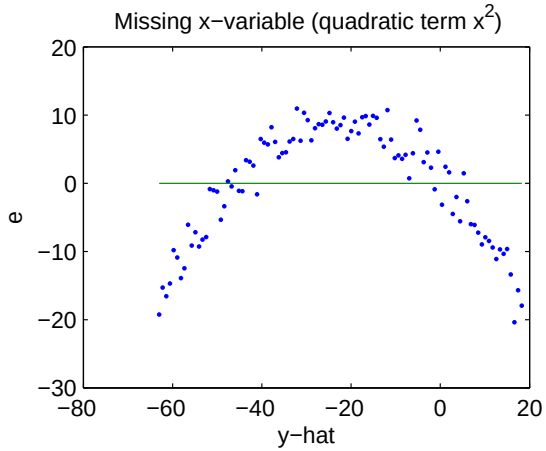
Random variation:

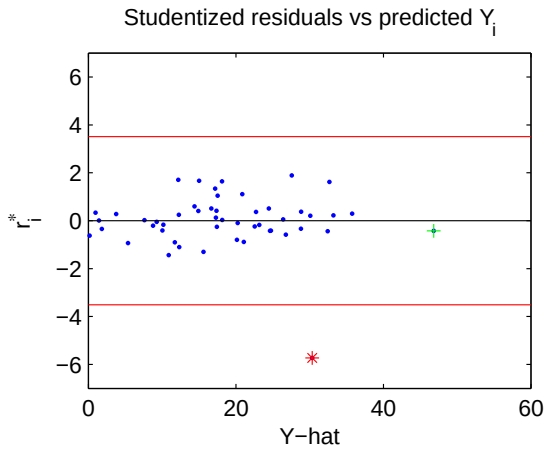


Problems:

Structural inadequasies or outliers:







Cook's distance

Do the potentially influential points actually have an influence?
What happens to the estimates if a point is removed?
Denote with $\hat{\beta}_{(i)}$ the estimate of β when point i is excluded and the corresponding prediction as $\hat{\mathbf{Y}}_{(i)} = \mathbf{X}\hat{\beta}_{(i)}$.
Cook's Distance, D_i measures the effect of observation i on $\hat{\beta}$.

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' \cdot \hat{Var}_{\hat{\beta}}^{-1} \cdot (\hat{\beta}_{(i)} - \hat{\beta})}{(p+1)} = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'(\mathbf{X}'\mathbf{X})(\hat{\beta}_{(i)} - \hat{\beta})}{(p+1)s^2}$$
$$= \frac{(\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}})'(\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}})}{(p+1)s^2} = \frac{r_i^2}{p+1} \cdot \frac{v_{ii}}{1 - v_{ii}}$$

[Here $\hat{Var}_{\hat{\beta}} \equiv \hat{Var}(\hat{\beta}) = s^2(\mathbf{X}'\mathbf{X})^{-1}$]

No unanimous consensus on how to use D_i: for some, point i can be considered to have a large influence on the estimates if $D_i > 1$ (for small/medium datasets), and $D_i > 4/n$ (large datasets).

Caution

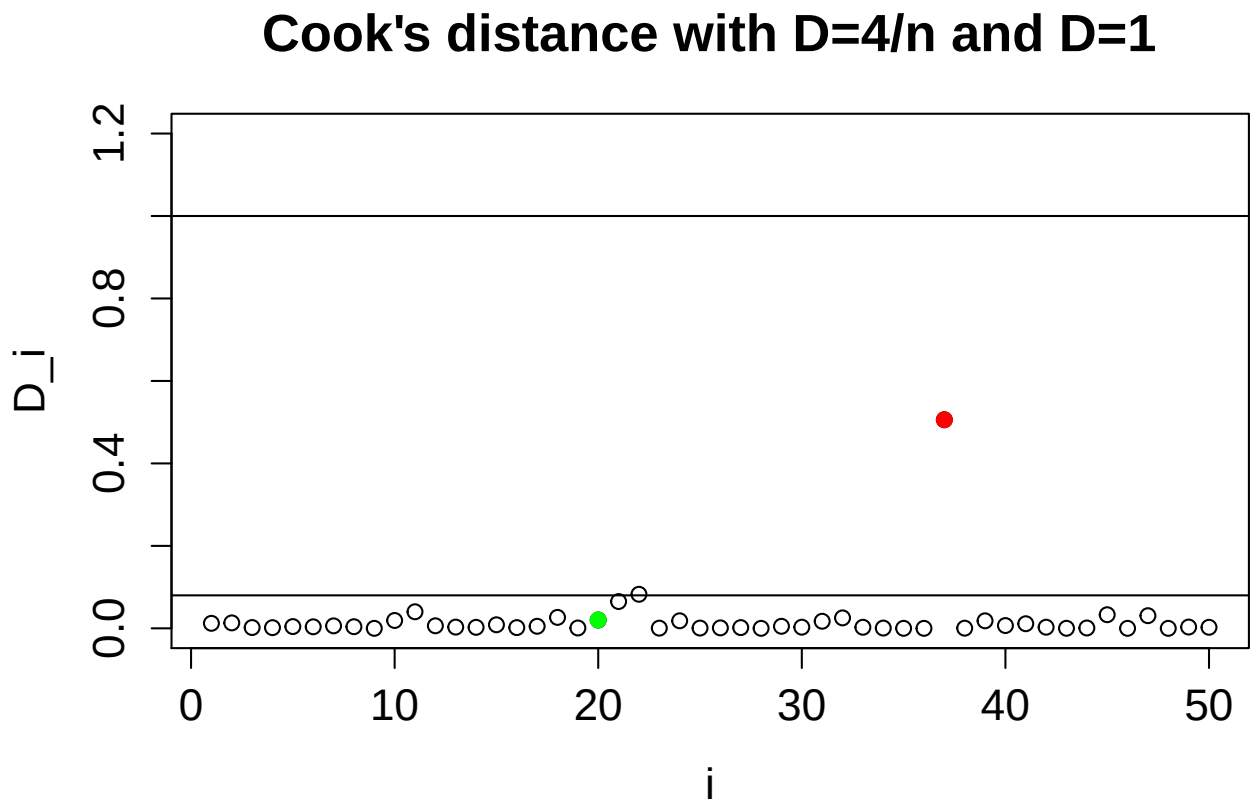
Don't be overzealous in comparing a quantity to an empirical threshold, e.g. automatically classify an observation according to $D_i \leqslant 1$.

Do not take these thresholds as absolute truth, when these are coming out of empirical experience.

Advice: use graphics to examine in closer details the points with "values of D that are substantially larger than the rest Thresholds should only be used to enhance graphical displays.

Going back to the example in the second slide:

For our small dataset, the outlier in red had more effect than the other observations but the potentially influential point had no particular effect.

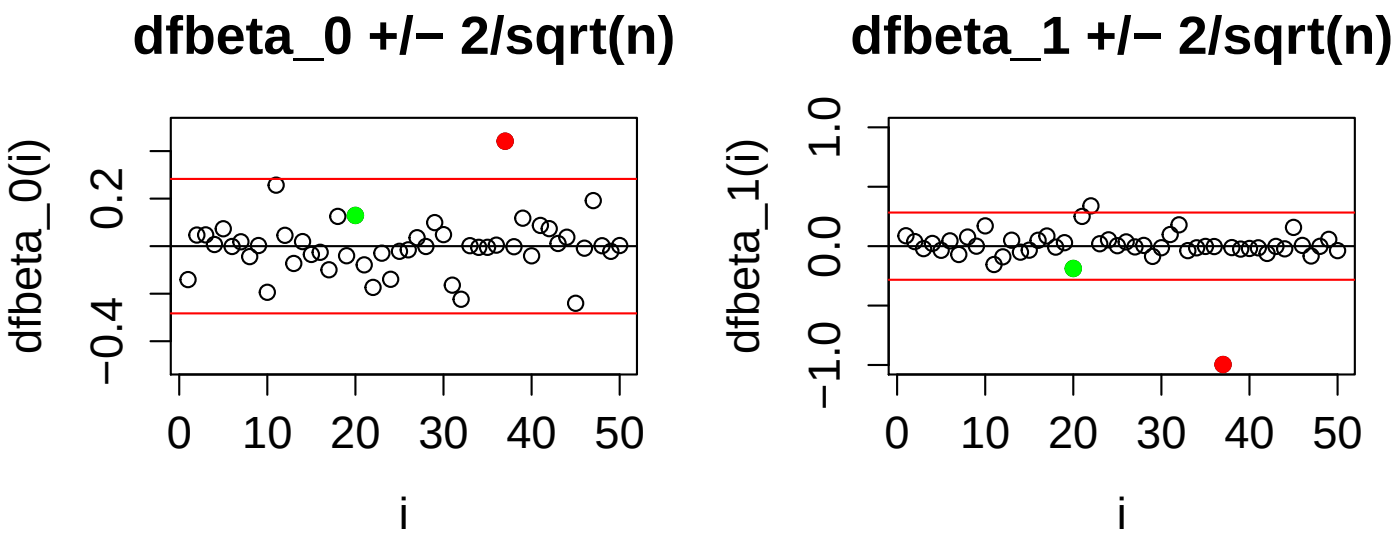


Influence on a specific parameter

The impact of an observation i on a specific element $\hat{\beta}_j$ of vector $\hat{\beta}$ can be assessed using DFBETAS:

$$DFBETA_j = \frac{|\hat{\beta}_j - \hat{\beta}_{j(i)}|}{s_{(i)}\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{jj}}}$$

The change in $\hat{\beta}_j$ $(j = 0, \ldots, p)$ can be considered large if its $DFBETA_j > 2/\sqrt{n}$ (or > 1). With the same words of caution as for D_i.



Summary

- Model validation, model diagnostics (influence analysis, residual analysis) is more like an *art*.
- We can't check for any possible thing that can go wrong. In particular, large datasets always have some "strange observation".
- And **our model might be correct even if some observation is not well represented/fitted**.
- What is important is to be aware of model assumptions, try to verify those, try to fix what can be fixed, spot anomalous/suspicious observations that might (badly) affect inferences and results.
- The previous methods are some "recipes" more than formal tests. Use them as a guiding tool but ultimately follow your judgement.

R functions

- leverages: `hatvalues(mymodel)`
- standardised residuals: `rstandard(mymodel)`
- studentised residuals: `rstudent(mymodel)`
- a number of influence measures are available using `influence()`:
  `infl <- influence(mymodel)` then extract fields
  `s_i <- infl$sigma` gives $s_{(i)}$
  `v <- infl$hat` is another way to obtain leverages v_{ii};
- `cooks.distance(mymodel)` the Cook's distances
- `dfbetas(mymodel)` gives the DFBETAS.