Diagnostics: residual plots

ST552 Lecture 16

Charlotte Wickham

2019-02-15

Violations of assumptions

$Y = X\beta + \varepsilon$ $\quad \varepsilon \sim N(0, \sigma^2 I)$

ε_i iid $N(0, \sigma^2)$

In rough order of importance:

- **Systematic form of the model**, $\mathrm{E}(Y) = X\beta$. If violated, the parameters in the model may be meaningless, estimates may be biased. HW
- **Independence of errors**, ϵ_i independent of ϵ_j for all i and j. If violated, estimates are still unbiased, but standard errors are generally inappropriate. → midterm
- **Constant variance**, $\mathrm{Var}(\epsilon_i) = \sigma^2$ for all i. If violated, variance in predictions may not be properly quantified. → midterm
- **Normality**, $\epsilon \sim N()$. Can rely on CLT for large samples. If violated, prediction intervals are probably innappropriate.

Using the residuals to diagnose problems

- If our model is correct, $\epsilon \sim N(0, \sigma^2 I)$. But, we don't observe the errors.
- Usually, we use the residuals as our best guess for the errors, and examine them for problems with the assumptions.
- However, residuals by construction are not equal variance, or uncorrelated (you can try to standardize), but in practice the effects are small and ignored.
- We can't prove the assumptions are satisfied, but we can look for evidence of gross violations.

Graphical versus formal inferential methods

- I am a strong proponent of graphical methods over formal tests for assumption checking.
- Tests can only provide quantification of a deviation you are expecting, graphics reveal the unexpected.
- Tests tend to make you focus on statistical significance not practical significance.
  For example, a large sample of data that is just a little non-Normal, will tend to give tiny p-values in a test of Normality, but for our purposes it isn't really a problem.

Residual plots to examine

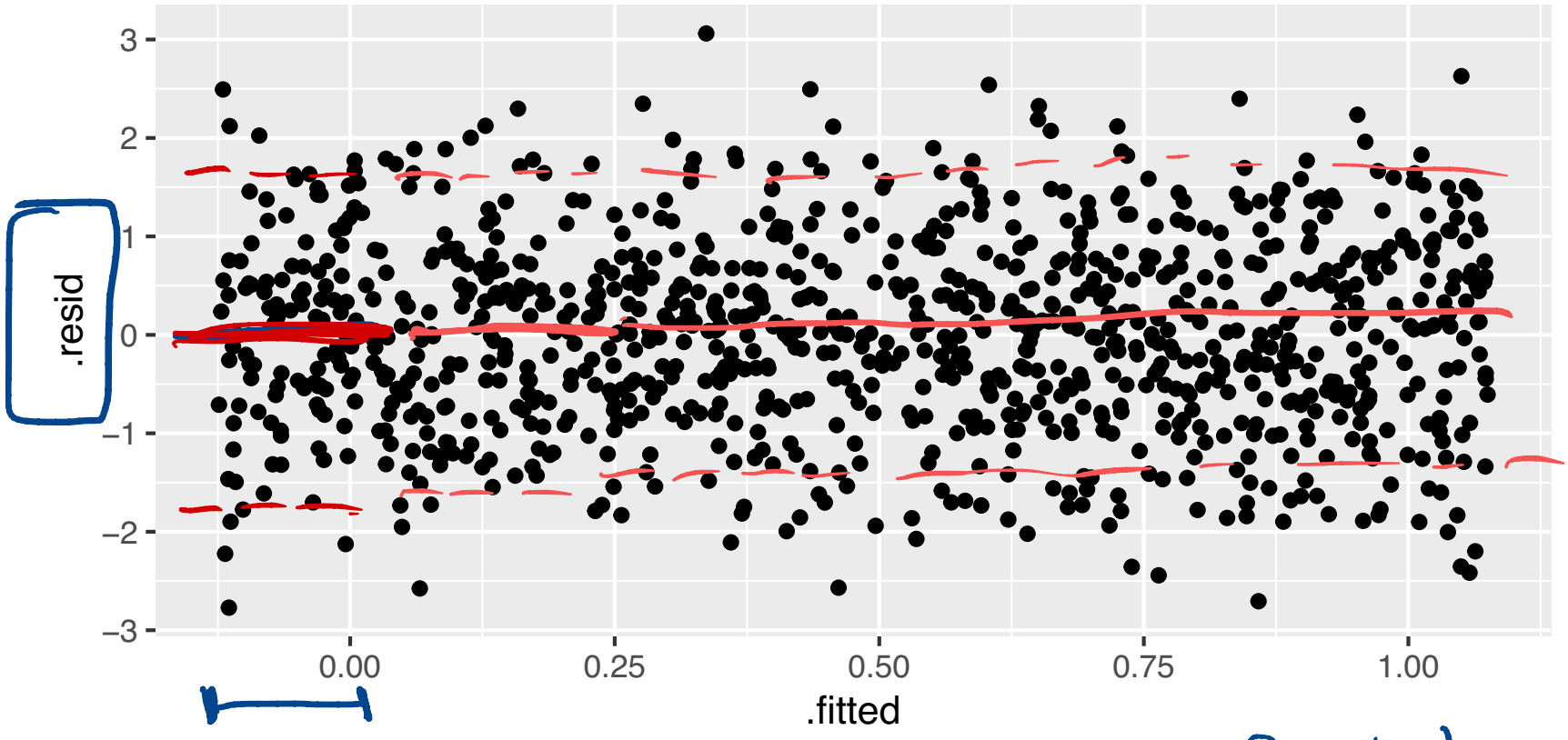
1. Residuals versus fitted values
3. Residuals versus explanatories (both those included and those excluded from the model)
2. Normal probability plot (Q-Q plot) of the residuals
- Anything else you can think of that might reveal structure in the residuals. For example, if measurements are made over time or space, look for temporal or spatial patterns in the residuals.

subset

What to look for

Residuals versus fitted or explantories

An even width band vertically centered around zero as you move left to right (always put the residuals on the y-axis).

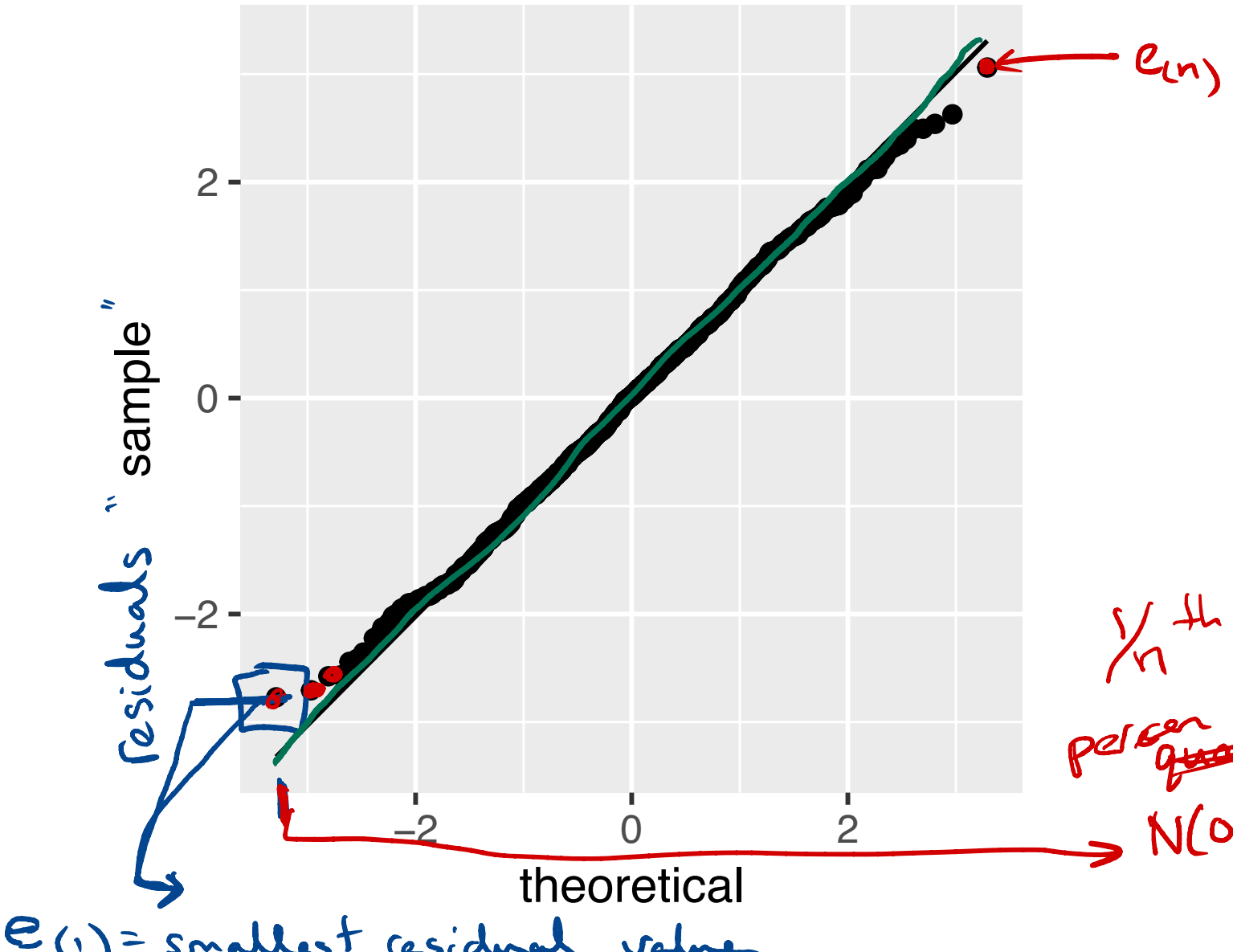


Q-Q plot of residuals → distribution

data

N(0, 1)

Points falling close to a straight line



Your turn: Part One

Handout (Charlotte will bring):

Part One Describe what you see in the residual plots that suggests a violation of assumptions.

Your turn: Part Two

Part Two The same five models are examined but in a random order, with a much smaller sample size. Can you match these diagnostics to those in Part Two??

Your turn: Part Three

Part Three Do you see any violations here?

Trick question

Generated satisfying assumptions

"Train your eyes"

Common problems and possible solutions

- Non-constant spread
  - transform response (background knowledge, trial & error, Box-Cox)
  - use more complicated models (glm, gee) gls
- Non-linearity
  - transform response
  - transform predictor
  - allow for curvature (add predictor^2, splines, gam)
  - use a non-linear model
- Non-normality
  - transform response
  - use more complicated models (glm) ← ST 623
- Structure when examined against an excluded variable - include it