@graemeleehickey

www.glhickey.com

graeme.hickey@liverpool.ac.uk

Checking model assumptions with regression diagnostics

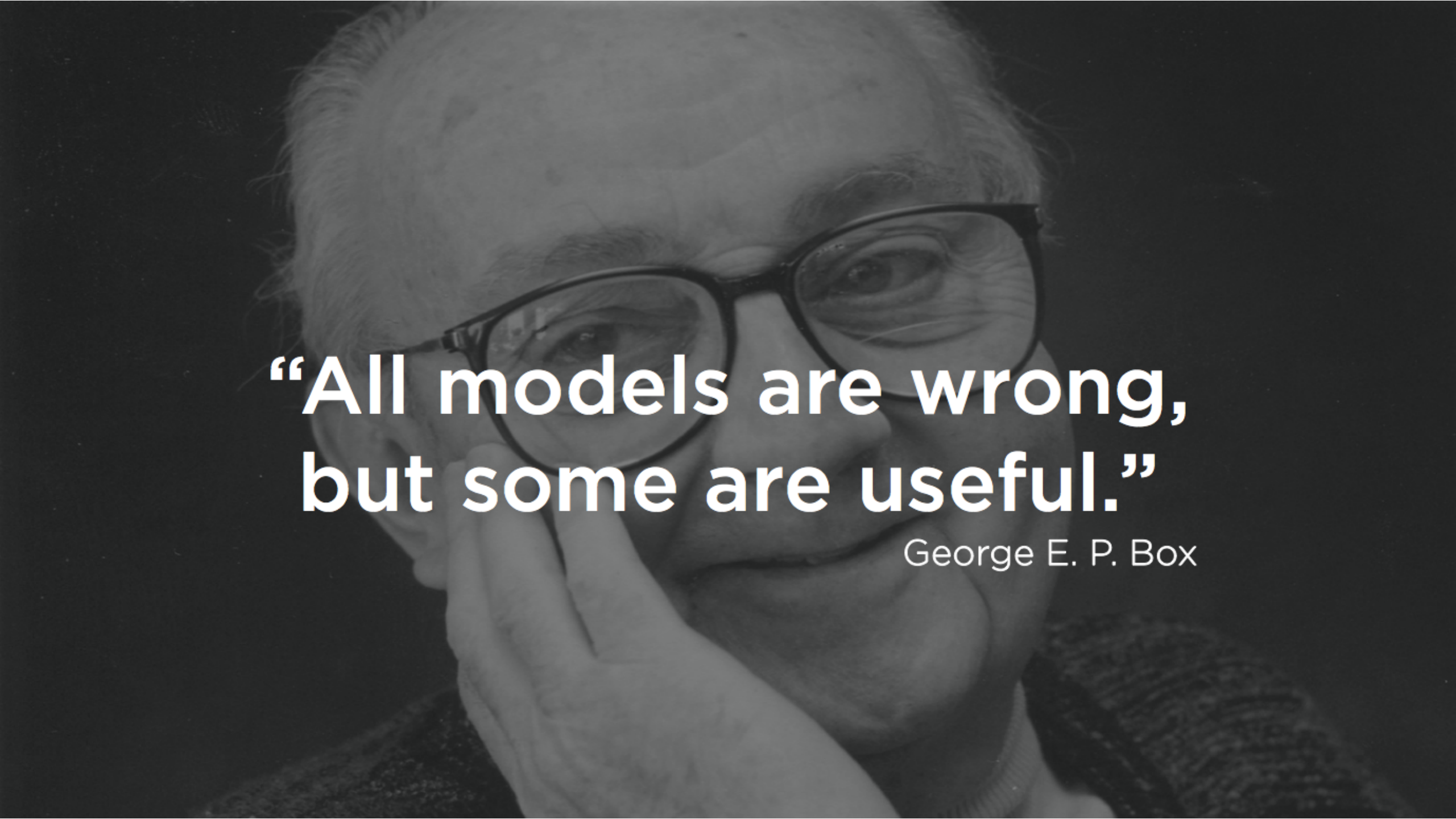
Graeme L. Hickey

University of Liverpool

UNIVERSITY OF LIVERPOOL

Conflicts of interest

- None
- Assistant Editor (Statistical Consultant) for EJCTS and ICVTS


“All models are wrong,
but some are useful.”
George E. P. Box

Question: who routinely checks model assumptions when analyzing data?

*(raise your hand if the answer is **Yes**)*

Outline

- Illustrate with multiple linear regression
- Plethora of residuals and diagnostics for other model types
- Focus is not to “what to do if you detect a problem”, but “how to diagnose (potential) problems”

My personal experience*

- Reviewer of EJCTS and ICVTS for 5-years
- Authors almost never report if they assessed model assumptions
- Example: only one paper submitted where authors considered sphericity in RM-ANOVA at first submission
- Usually one or more comment is sent to authors regarding model assumptions

* My views do not reflect those of the EJCTS, ICVTS, or of other statistical reviewers

Linear regression modelling

- Collect some data
  - y_i: the observed continuous outcome for subject i (e.g. biomarker)
  - $x_{1i}, x_{2i}, \ldots, x_{pi}$: p covariates (e.g. age, male, …)
- Want to fit the model
  - $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{i2} + \cdots + \beta_p x_{pi} + \varepsilon_i$
- Estimate the regression coefficients
  - $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$
- Report the coefficients and make inference, e.g. report 95% CIs
- But we do not stop there…

Residuals

- For a linear regression model, the residual for the i-th observation is

$$r_i = y_i - \hat{y}_i$$

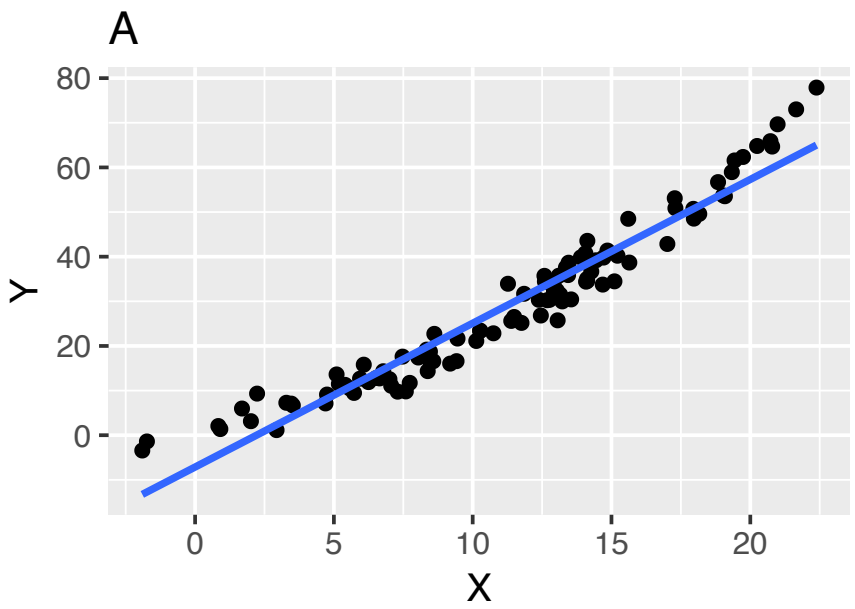
- where $\hat{y}_i$ is the predicted value given by

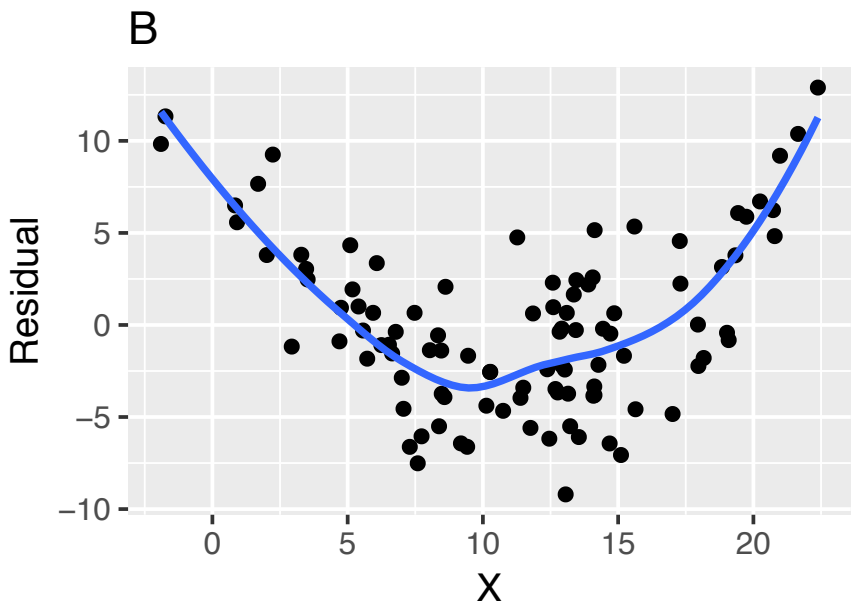
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{pi}$$

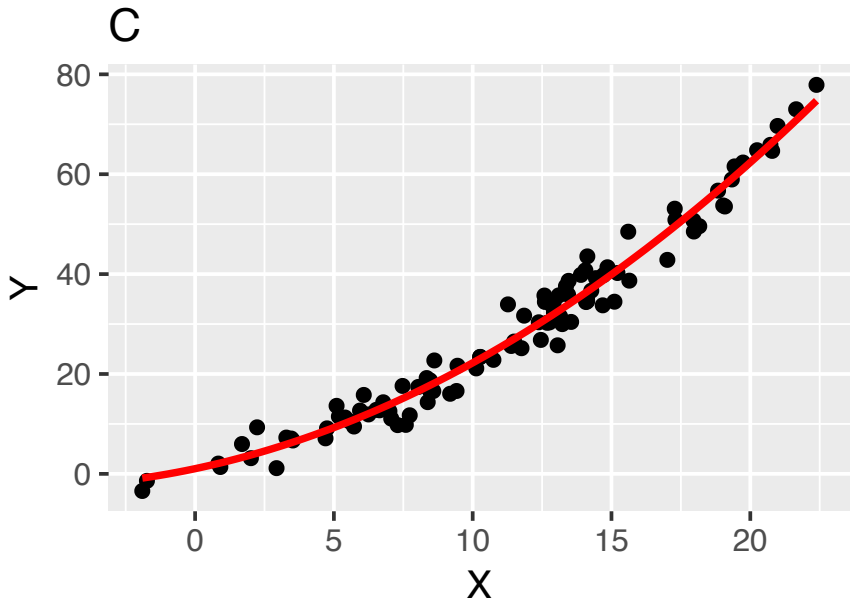
- Lots of useful diagnostics are based on residuals

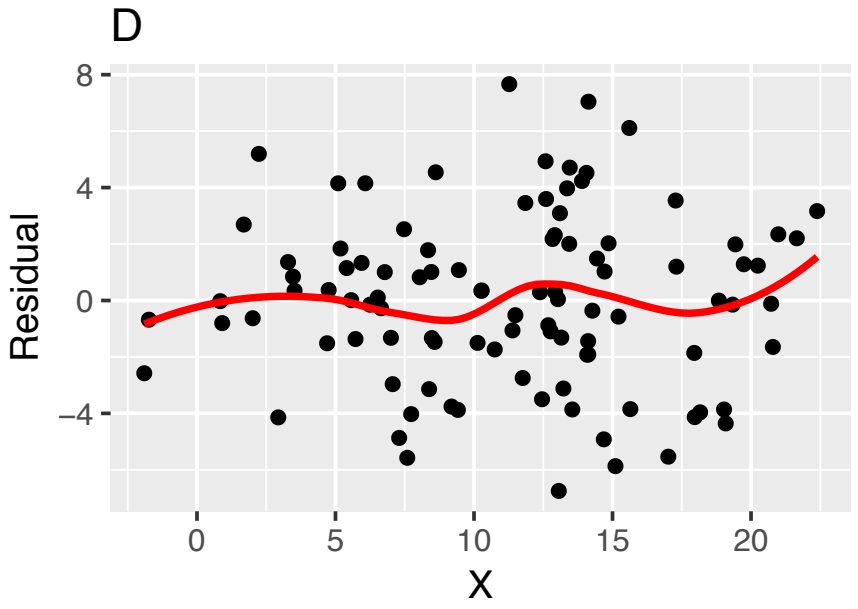
Linearity of functional form

- Assumption: scatterplot of (x_i, r_i) should not show any systematic trends
- Trends imply that higher-order terms are required, e.g. quadratic, cubic, etc.









Fitted model:

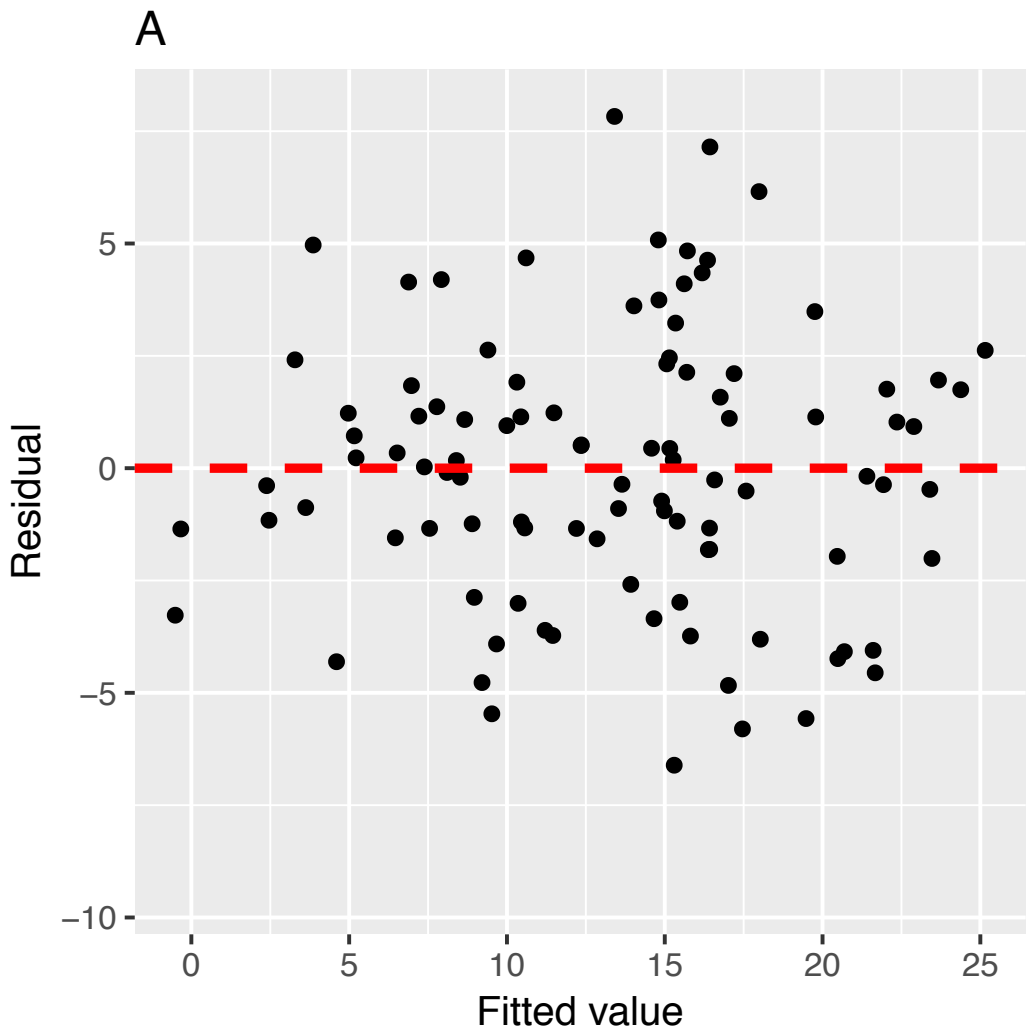
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

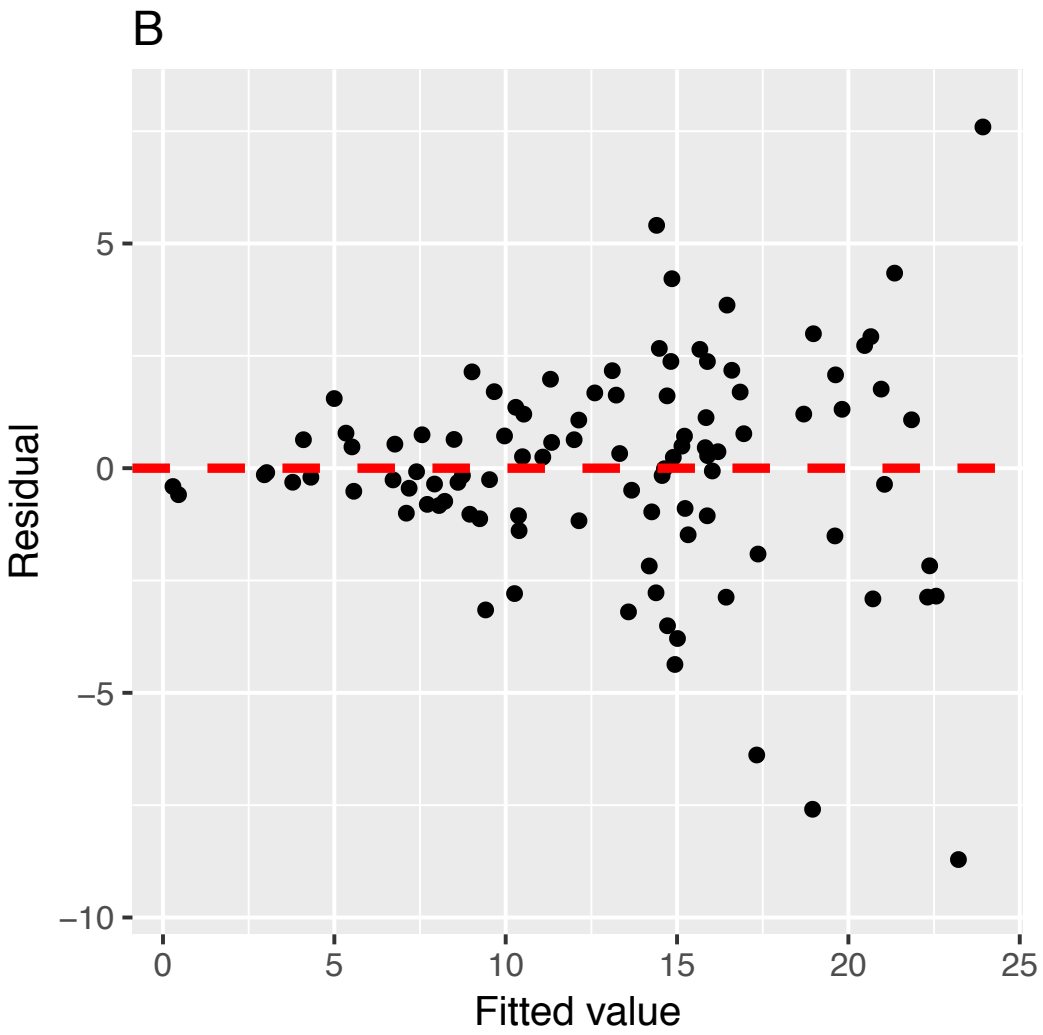
Homogeneity

- We often assume assume that $\varepsilon_i \sim N(0, \sigma^2)$
- The assumption here is that the variance is constant, i.e. homogeneous
- Estimates and predictions are robust to violation, but not inferences (e.g. *F*-tests, confidence intervals)
- We should not see any pattern in a scatterplot of $(\hat{y}_i, r_i)$
- Residuals should be symmetric about 0

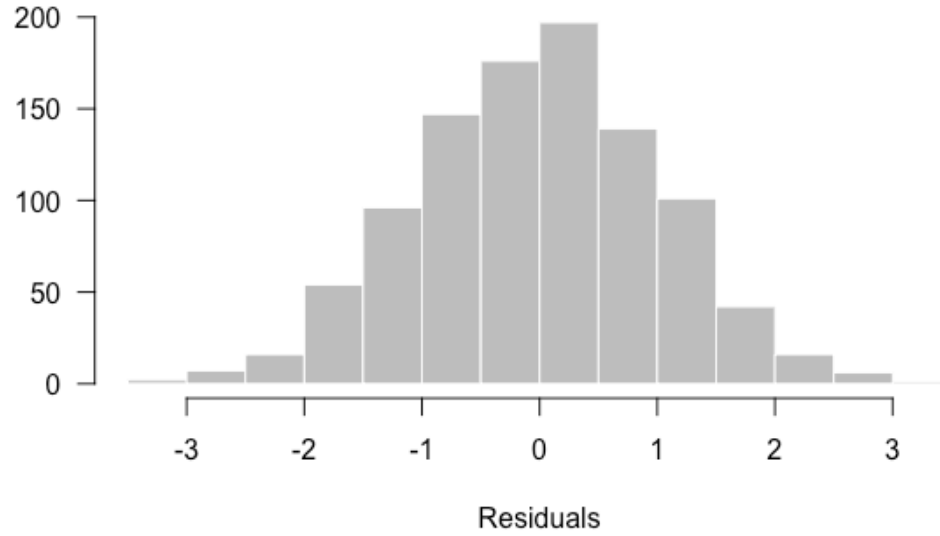
Homoscedastic residuals



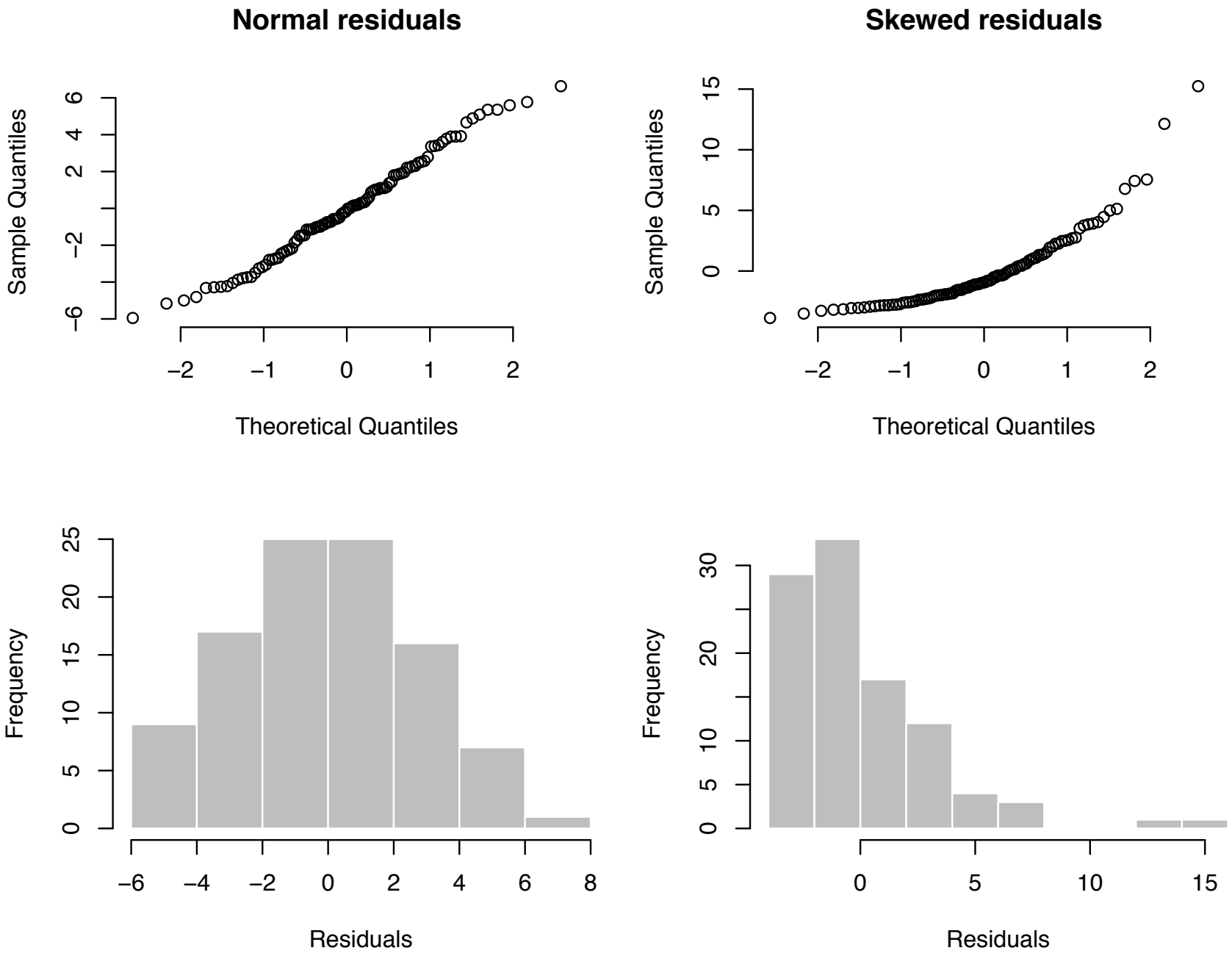
Heteroscedastic residuals



Normality



- If we want to make inferences, we generally assume $\varepsilon_i \sim N(0, \sigma^2)$
- Not always a critical assumption, e.g.:
  - Want to estimate the 'best fit' line
  - Want to make predictions
  - The sample size is quite large and the other assumptions are met
- We can assess graphically using a Q-Q plot, histogram
- **Note**: the assumption is about the errors, not the outcomes y_i


Normal residuals
Sample Quantiles
Theoretical Quantiles
Skewed residuals
Sample Quantiles
Theoretical Quantiles
Frequency
Residuals
Frequency
Residuals

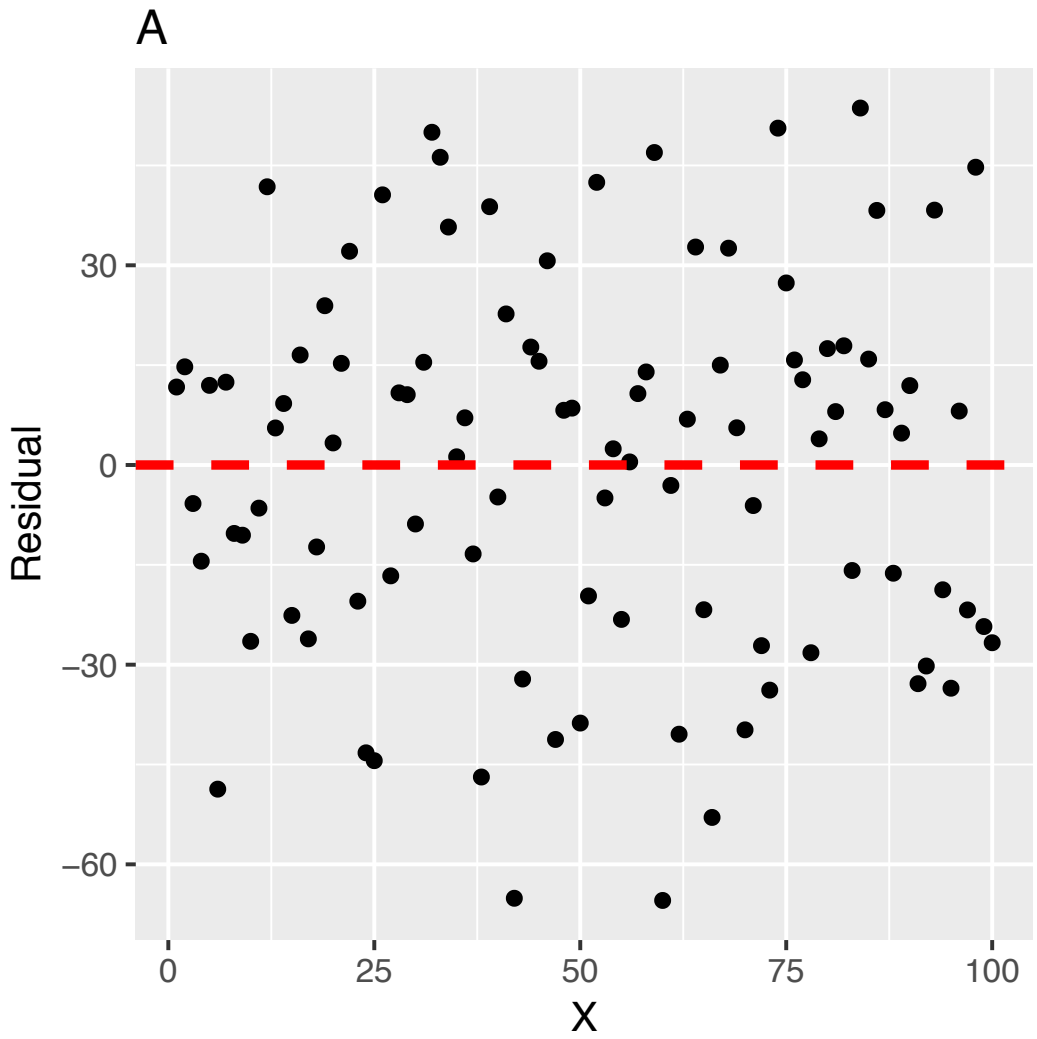
Independence

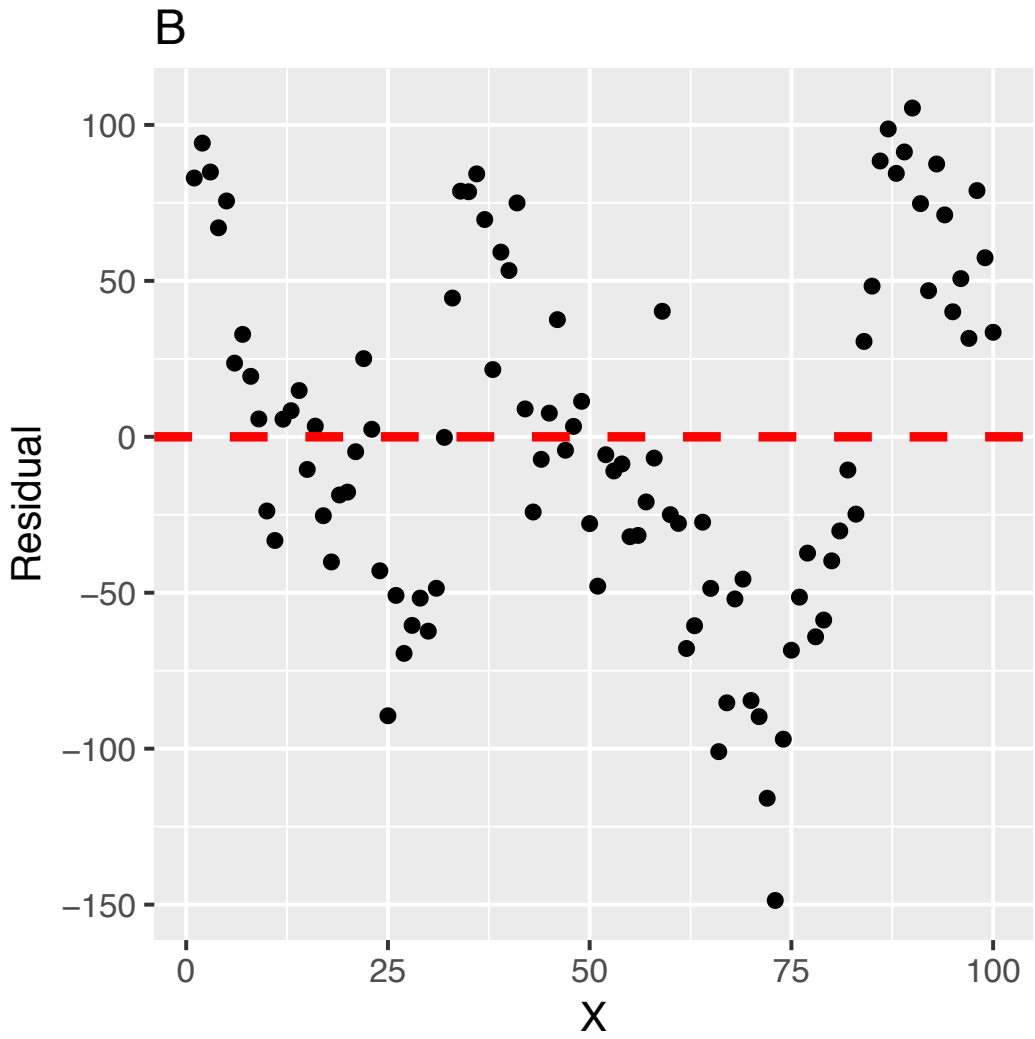
- We assume the errors are independent
- Usually able to identify this assumption from the study design and analysis plan
  - E.g. if repeated measures, we should not treat each measurement as independent

- If independence holds, plotting the residuals against the time (or order of the observations) should show no pattern

Independent

Non-independent





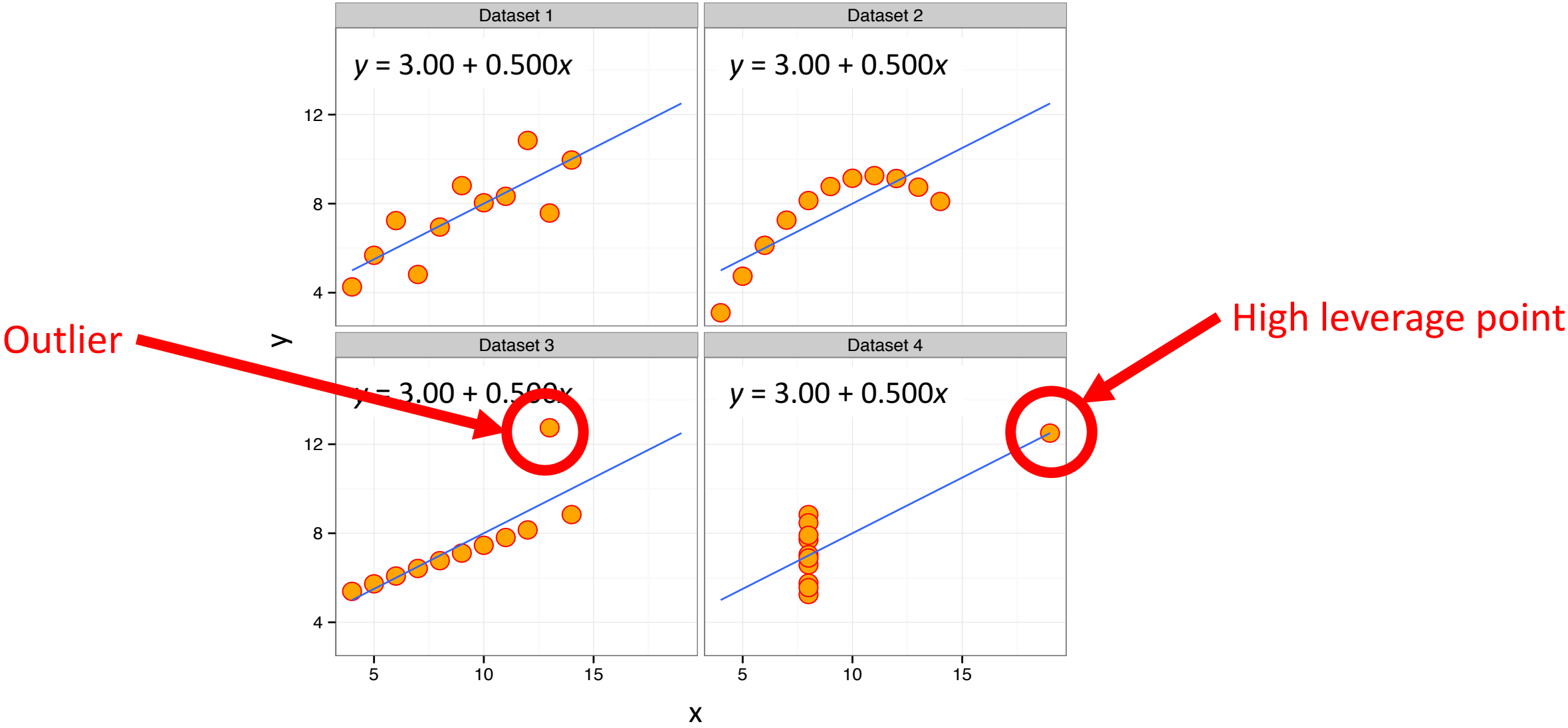
Multicollinearity

- Correlation among the predictors (independent variables) is known as collinearity (multicollinearity when >2 predictors)
- If aim is inference, can lead to
  - Inflated standard errors (in some cases very large)
  - Nonsensical parameter estimates (e.g. wrong signs or extremely large)
- If aim is prediction, it tends not to be a problem
- Standard diagnostic is the variance inflation factor (VIF)

$$VIF(X_j) = \frac{1}{1 - R_j^2}$$

Rule of thumb: VIF > 10 indicates multicollinearity

Outliers & influential points



Diagnostics to detect influential points

- DFBETA (or Δβ)
  - Leave out *i*-th observation out and refit the model
  - Get estimates of $\hat{\beta}_{0(-i)}, \hat{\beta}_{1(-i)}, \hat{\beta}_2(-i), \ldots, \hat{\beta}_{p(-i)}$
  - Repeat for $i = 1, 2, \ldots, n$

- Cook's distance *D*-statistic
  - A measure of how influential each data point is
  - Automatically computer / visualized in modern software
  - Rule of thumb: $D_i > 1$ implies point is influential

Residuals from other models

GLMs (incl. logistic regression)

- Deviance
- Pearson
- Response
- Partial
- Δβ
- …

Cox regression

- Martingale
- Deviance
- Score
- Schoenfeld
- Δβ
- …

Useful for exploring the influence of individual observations and model fit

Two scenarios

Statistical methods routinely submitted to EJCTS / ICVTS include:

1. Repeated measures ANOVA
2. Cox proportional hazards regression

Each has very important assumptions

Repeated measures ANOVA

- **Assumptions**: those used for classical ANOVA + sphericity
- **Sphericity**: the variances of the differences of all pairs of the within subject conditions (e.g. time) are equal

Patient	T0	T1	T2	T0 – T1	T0 – T2	T1 – T2
1	30	27	20	3	10	7
2	35	30	28	5	7	2
3	25	30	20	−5	5	10
4	15	15	12	0	3	3
5	9	12	7	−3	2	5
Variance				17.0	10.3	10.3

- It’s a questionable *a priori* assumption for longitudinal data

Mauchly's test

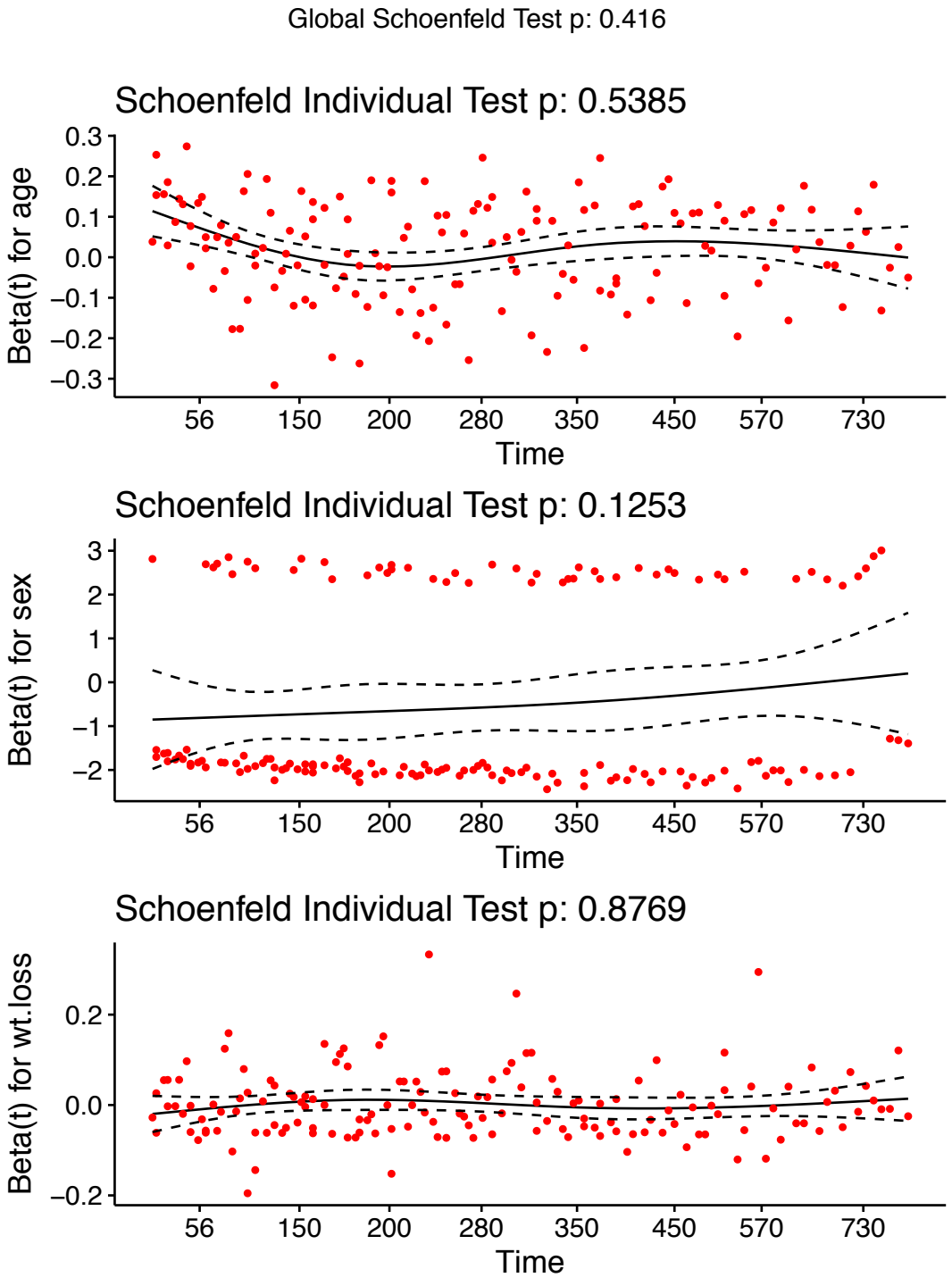
- A popular test (but criticized due to power and robustness)
  - H_0: sphericity satisfied (i.e. $\sigma^2_{T_0-T_1} = \sigma^2_{T_0-T_2} = \sigma^2_{T_1-T_2}$)
  - H_1: non-sphericity (at least one variance is different)
- If rejected, it is usual to apply a correction to the degrees of freedom (df) in the RM-ANOVA *F*-test
- The correction is ϵ x df, where ϵ = epsilon statistic (either Greenhouse-Geisser or Huynh-Feldt)
- Software (e.g. SPSS) will automatically report ϵ and the corrected tests

Proportionality assumption

- Cox regression assumes proportional hazards:
- Equivalently, the hazard ratio must be constant over time

- There are many ways to assess this assumption, including two using residual diagnostics:
  - Graphical inspection of the (scaled) Schoenfeld residuals
  - A test* based on the Schoenfeld residuals

* Grambsch & Therneau. *Biometrika*. 1994; 81: 515-26.



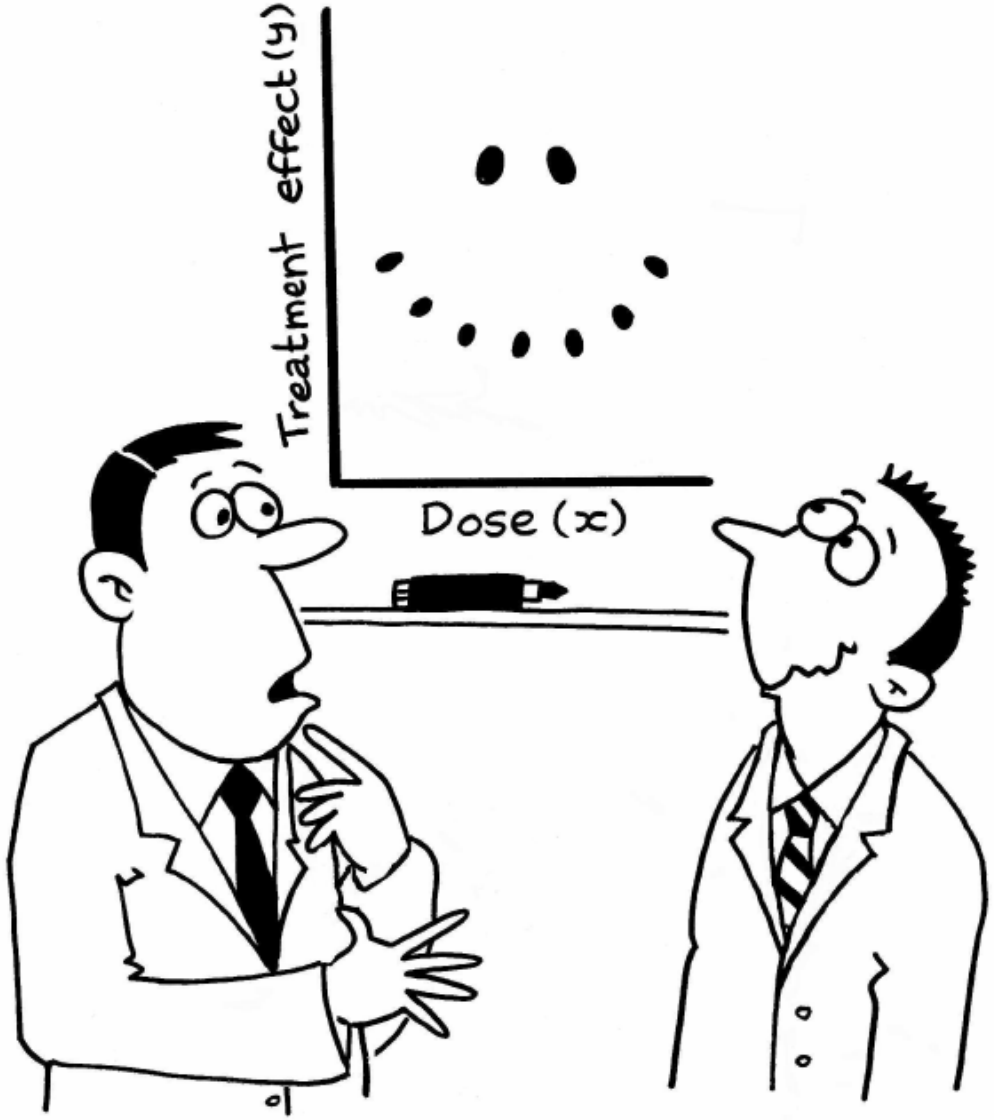
- Simple Cox model fitted to the North Central Cancer Treatment Group lung cancer data set*
- If proportionality is valid, then we should not see any association between the residuals and time
- Can formally test the correlation for each covariate
- Can also formally test the “global” proportionality

*Loprinzi CL et al. *Journal of Clinical Oncology*. 12(3) :601-7, 1994.

Conclusions

- Residuals are incredibly powerful for diagnosing issues in regression models
- If a model doesn't satisfy the required assumptions, don't expect subsequent inferences to be correct
- Assumptions can usually be assessed using methods other than (or in combination with) residuals

- Always report in manuscript
  - What diagnostics were used, even if they are absent from the Results section
  - Any corrections or adjustments made as a result of diagnostics



"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

Thanks for listening

Any questions?

Statistical Primer article to be published soon!

Slides available (shortly) from: www.glhickey.com