Statistics 2

Assumptions

Casper Albers & Jorge Tendeiro

Lecture 7, 2019 – 2020

university of groningen

Overview

Literature for this lecture

No literature from Agresti. Please review the slides and practical exercises extra carefully on this topic.

Model assumptions

$$\boxed{y_i = \alpha + \beta x_i + \varepsilon_i} \quad \boxed{\varepsilon \sim \mathcal{N}(0, \sigma)}$$

Assumptions of the model: In the order as they appear on p. 279

1. Independent observations:
   - All observations are independent of each other.
   - True random sampling.
2. Linear relations:
   - Relation between x and $E(y)$ is a straight line.
3. Homoscedasticity:
   - The conditional standard deviation σ is constant.
4. Residuals follow a normal distribution:
   - y_i follows a normal distribution around $E(y)$.

Model assumptions

What if the assumptions are invalid?

- The analyses are no longer guaranteed the best approach or, worse, not even valid anymore.
- Tests and CI's can lead to misleading and incorrect conclusions.
- Inferences are no longer justified.
- Checks and corrections are necessary.

Thus, the validity of the model is at play here.
Today, we explain how and why.

Assumption: Independence

No relation between cases

- What happens if it doesn't hold?
  $\Rightarrow$ Biased estimation, bad inference.
- Checking independence:
  - Part of data collection protocol.
- Correcting violation:
  - Use other techniques: Paired t-test, multilevel, repeated measures models.

Assumption: Independence

Examples of designs without independent measurements:

- Study on effectivity of anxiety treatment.
  Measures: before and after treatment measurement of Hamilton Index.
- Educational study, surveying different classrooms and different schools.
  Measures: age and arithmetic proficiency test.

Assumption: Normal distributions

Residuals follow a normal distribution: $\varepsilon_i \sim \mathcal{N}(0, \sigma)$

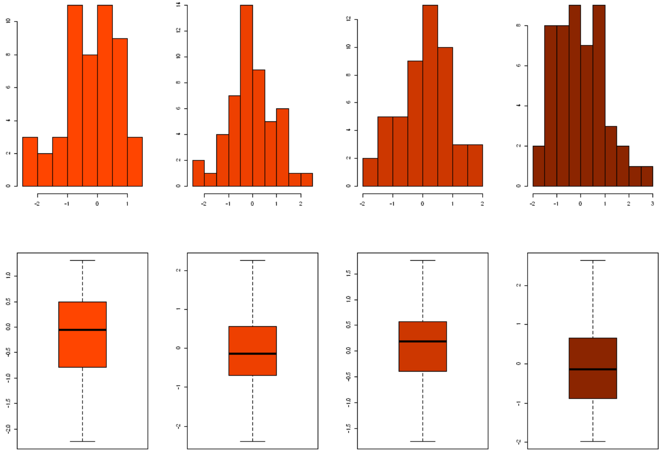
- What happens if it doesn't hold?

  $\Rightarrow$ mild consequences when other assumptions still are valid:

  Some loss of power

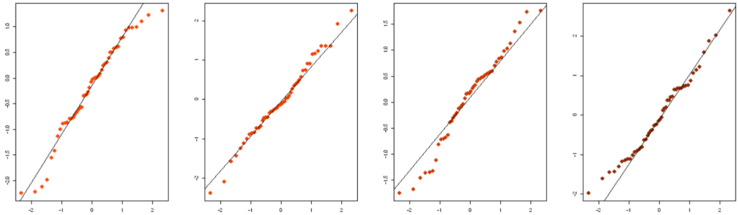
  $\Rightarrow$ when other assumptions still are invalid too: Severe consequences
- Checking normality:
  - Use sample residuals $e_i = y_i - \widehat{y_i}$.
    - Visual checks: QQ-plot (and boxplot, histogram).
    - Formal tests. (Not part of Statistics II)
  - Checking aspects of the distribution: Skewness and kurtosis.

Poor visual checks for normality



Four samples of size $n = 50$ from $\mathcal{N}(0,1)$

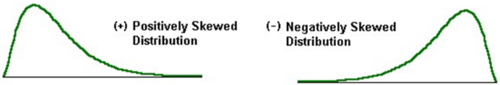
Better visual check for normality

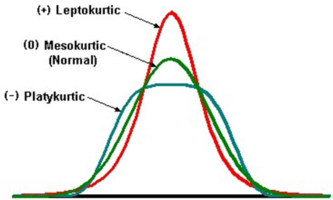


Four samples of size $n = 50$ from $\mathcal{N}(0, 1)$

- QQ-plot better check than histogram or box plot.
- Even for moderate sample sizes, fairly large deviations can occur under normality.

Skewness and kurtosis





Normal distribution has skewness = 0 and kurtosis = 0

- Strong deviation from 0 are a sign of non-normality.
- But not the other way around!

Assumption: Normal distributions

Residuals follow a normal distribution: $\varepsilon_i \sim \mathcal{N}(0, \sigma)$

- What happens if it doesn't hold?

  $\Rightarrow$ mild consequences when other assumptions still are valid:

  Some loss of power

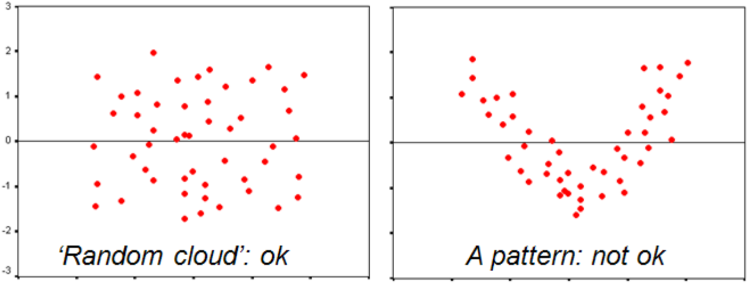
  $\Rightarrow$ when other assumptions still are invalid too: Severe consequences
- Checking normality:
  - Use sample residuals $e_i = y_i - \widehat{y_i}$.
    - Visual checks: QQ-plot (and boxplot, histogram).
    - Formal tests. (Not part of Statistics II)
  - Checking aspects of the distribution: Skewness and kurtosis.
- Correcting violation:
  - Performing data transformation.
  - Using non-parametric techniques.
  - Increasing n (CLT).
  - Removing outliers.

Assumption: Linear relations

$$\boxed{\text{Linear model: } y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i}$$

- What happens if it doesn't hold?
  $\Rightarrow$ Misspecification, bad fit, biased results.
- Checking linearity:
  - Use sample residuals $e_i = y_i - \widehat{y}_i$.
    - Residual plots: Residuals vs. other variables ($\widehat{y}$, y, x).
    - Formal tests or partial plots (not part of Statistics II).

Residual plots



Look for systematic deviations from the horizontal line

Assumption: Linear relations

$$\boxed{\text{Linear model: } y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i}$$

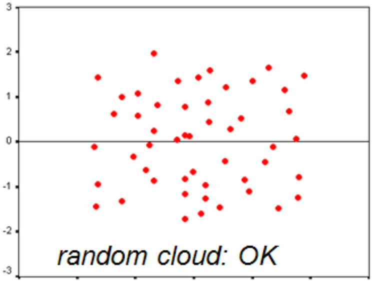
- What happens if it doesn't hold?
  $\Rightarrow$ Misspecification, bad fit, biased results.
- Checking linearity:
  - Use sample residuals $e_i = y_i - \widehat{y_i}$.
    - Residual plots: Residuals vs. other variables ($\widehat{y}$, y, x).
    - Formal tests or partial plots (not part of Statistics II).
- Correcting violation:
  - Performing data transformation.
  - Using non-linear regression (e.g., logistic, Poisson).

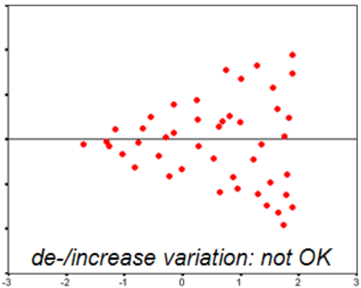
Assumption: Homoscedasticity

The residuals follow a distribution with constant variance

- What happens if it doesn't hold?
  $\Rightarrow$ Biased estimation, wrong inference.
- Checking homoscedasticity:
  - Use sample residuals.
    - Residual plots: Residuals vs. other variables ($\widehat{y}$, y, x).
    - Formal tests or partial plots (not part of Statistics II).

Residual plots





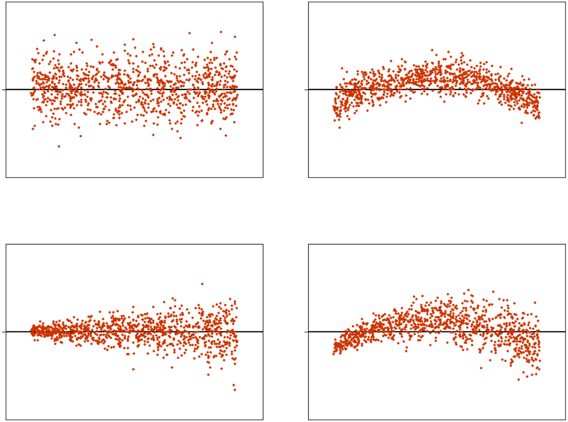
Look for systematic deviations in variation around the horizontal line

Assumption: Homoscedasticity

The residuals follow a distribution with constant variance

- What happens if it doesn't hold?
  $\Rightarrow$ Biased estimation, wrong inference.
- Checking homoscedasticity:
  - Use sample residuals.
    - Residual plots: Residuals vs. other variables ($\widehat{y}$, y, x).
    - Formal tests or partial plots (not part of Statistics II).
- Correcting violation:
  - Removing outliers.
  - Performing data transformation.
  - Using other estimation methods (not part of Statistics II).

Residual plots



Residual plots useful for detecting
violations of linearity (b,d) and homoscedasticity (c,d).

Example – Overclaiming

Atir, Rosenzweig, and Dunning (2015) studied whether experts overrate the extent of their expertise[1].

- Dependent variable
  - y: `OVCLAIM`, overclaiming based on defining 15 terms (of which 3 do not exist).
- Independent variables ($p = 2$)
  - x_1: `SPKNOW`, based on a questionnaire assessing self-perceived knowledge.
  - x_2: `ACCUR`, accuracy operationalized as the ability to distinguish between the 12 real terms and the 3 fake terms.

Sample size = 202.

[1] Atir, S., Rosenzweig, E., & Dunning, D. (2015). When knowledge knows no bounds: Self-perceived expertise predicts claims of impossible knowledge. *Psychological Science*, *26*, 1295-1303. doi: 10.1177/0956797615588195

Example – Overclaiming (correlations)

		OVCLAIM	SPKNOW	ACCUR
OVCLAIM	Pearson's r	–		
	p-value	–		
SPKNOW	Pearson's r	0.481	–	
	p-value	$< .001$	–	
ACCUR	Pearson's r	-0.672	0.033	–
	p-value	$< .001$	0.645	–

We expect that:

- OVCLAIM is linearly related to either predictor.
- The predictors SPKNOW and ACCUR are not strongly linearly related.

Example – Overclaiming (parameter estimates)

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	0.089	0.037		2.420	0.016
	SPKNOW	0.100	0.008	0.504	13.072	$<$ .001
	ACCUR	-0.754	0.042	-0.688	-17.869	$<$ .001

$$\boxed{\widehat{\text{OVCLAIM}} = 0.089 + 0.100\,\text{SPKNOW} - 0.754\,\text{ACCUR}}$$

Interpret regression coefficients:

- $a = 0.089$: The expected OVCLAIM score is equal to 0.089 when both SPKNOW and ACCUR are equal to 0.
- $b_1 = 0.100$: OVCLAIM increases by 0.100 units when SPKNOW increases by 1 unit, controlling for ACCUR (i.e., keeping ACCUR fixed).
- $b_2 = -0.754$: OVCLAIM decreases by 0.754 units when ACCUR increases by 1 unit, controlling for SPKNOW (i.e., keeping SPKNOW fixed).

These interpretations are only valid if the assumptions hold

Example – Independence

From the paper:

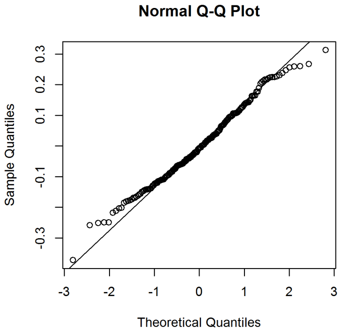
Participants

Study 1a had 100 participants (33 women, 66 men; mean age = 31 years, *SD* = 9.7; 1 participant did not report demographic information). Two additional participants failed to complete the entire study and were excluded from all analyses. Study 1b had 202 participants (85 women, 115 men, 2 whose gender was not reported; mean age = 33.5 years, *SD* = 10.1). Twelve additional participants failed to complete the entire study and were excluded from all analyses. Both samples were recruited through Amazon's Mechanical Turk and were restricted to respondents within the United States.

Perhaps not fully representative, but it is a random sample with independent observations.

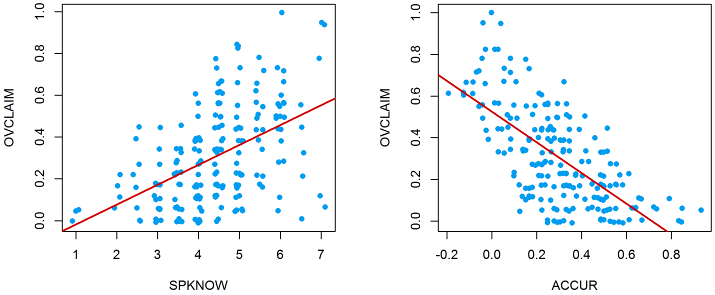
Example – normality

	Skewness (SE)	Kurtosis (SE)
Residuals	0.144 (0.171)	-0.349 (0.341)



No strong violation of the assumption.

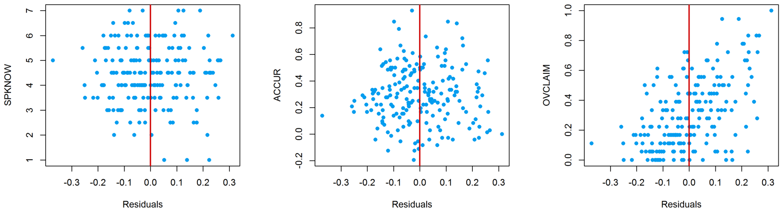
Example – linearity



Direct x vs y comparison:

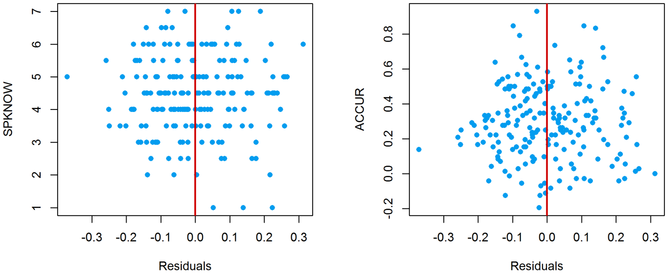
- The linearity assumption might be violated, due to a floor effect;
- Better to look at residuals plots.

Example – linearity



- The plots of residuals vs x_1 and x_2 look fine.
- The plot of residuals vs y clearly does not. Indication that there might be an interaction SPKNOW × ACCUR.
- (Note: JASP plots $x_{1,2}$ on the horizontal axis and residuals on the vertical axis. Both approaches are fine.)

Example – homoscedasticity



- No major issues w.r.t. homoscedasticity.

Next lecture – Regression with Categorical Predictors: Code Variables

Agresti, Section 12.1

Good luck with your exams the next two weeks!