

Evaluation Randomisiert-Kontrollierter Studien und Experimente mit R

Auswertung Multipel Imputierter Daten: Die Rubin-Regeln

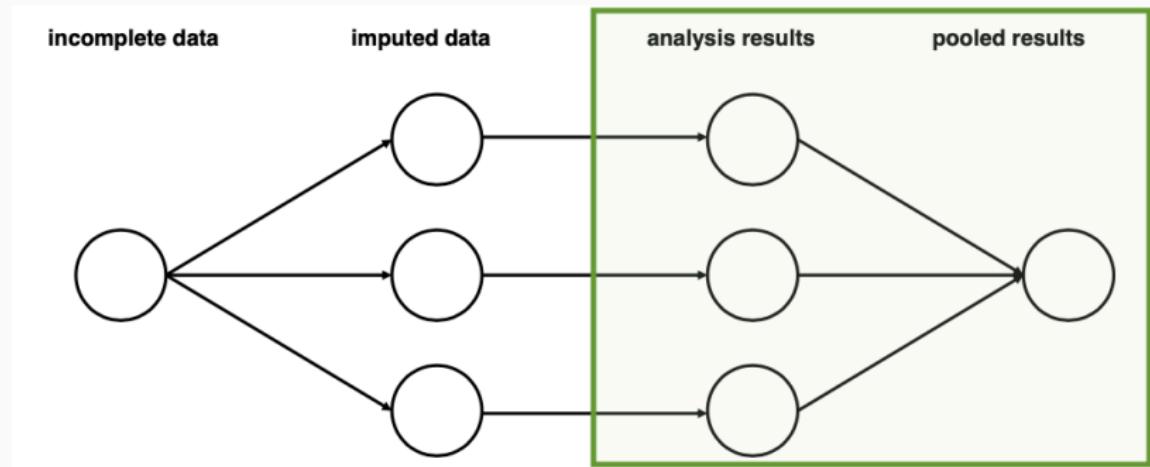
Prof. Dr. David Ebert & Mathias Harrer

Graduiertenseminar TUM-FGZ

Psychology & Digital Mental Health Care, Technische Universität München

Die 3 Variationsquellen

Wie komme ich mit m Imputationssets zu einem gemeinsamen Analyseergebnis?



adaptiert von Van Buuren & Groothuis-Oudshoorn (2011).

→ **Lösung:** Durchführung der selben Analyse in allen m imputierten Datensets parallel, dann werden alle m relevanten Parameter/Schätzwerte zu einem Wert aggregiert (“**gepoolt**”).

Dabei müssen wir miteinbeziehen, dass unsere Schätzwerte **mit Unsicherheit behaftet** sind. Diese Unsicherheit hat bei MI mindestens zwei Gründe:

1. Die Teilnehmenden der Studie stellen nur eine Stichprobe der untersuchten Studienpopulation dar, sind also mit Stichprobenfehler (“**sampling error**”) behaftet.
2. Die Daten enthalten **fehlende Werte**, deren Schätzung **unsicher** ist. Diese Unsicherheit wird dadurch reflektiert, dass multipel imputierte Werte sich zwischen Imputationssets unterscheiden (können).

Es sei Q ein zu schätzender wahrer Wert (oder ein Vektor von Werten) der Population (z.B. Populationsmittelwert, Regressionskoeffizienten, ...).

Aufgrund der zuvor genannten Gründe ist Q unbekannt und muss durch einen Schätzer \hat{Q} angenähert werden. Dies ist nur durch die beobachteten Werte Y_{obs} möglich.

Der Erwartungswert (d.h. die bestmögliche Annäherung) von Q gegeben Y_{obs} ist (Van Buuren, 2018, Kapitel 2.3.2):

$$E(Q|Y_{\text{obs}}) = E\left(E(Q|Y_{\text{obs}}, Y_{\text{mis}}) \mid Y_{\text{obs}}\right)$$

d.h. der Durchschnittswert der (imputierten) Schätzungen des Mittelwerts von Q über alle multiplen Imputationen hinweg.

Kombination von Punktschätzungen:

Es sei \hat{Q}_ℓ die Schätzung von Q im ℓ -ten von m Imputationssets. Der "gepoolte" Schätzwert von Q ist damit:

$$\bar{Q} = \frac{1}{m} \sum_{\ell=1}^m \hat{Q}_\ell$$

Aggregation von Punktschätzungen bei MI

Um Punktschätzungen (z.B. Parameter der Wahrscheinlichkeitsverteilung wie Mittelwert & Standardabweichung; Regressionsgewichte, etc.) zu poolen, wird in jedem Imputationsset der Wert des Punktschätzers berechnet, und daraufhin der Mittelwert über alle Imputationsets gebildet.

Aber wie kann die **Unsicherheit** (Varianz) von Q in MI-Daten geschätzt werden? Die Varianz von Q gegeben Y_{obs} besteht aus **zwei Komponenten**:

$$V(Q|Y_{\text{obs}}) = \underbrace{E\left(V(Q|Y_{\text{obs}}, Y_{\text{mis}}) \mid Y_{\text{obs}}\right)}_{\substack{\text{Mittelwert d. Varianzen über alle MI-Sets} \\ \rightarrow \text{Within-Variance } (\bar{U})}} + \underbrace{V\left(E(Q|Y_{\text{obs}}, Y_{\text{mis}}) \mid Y_{\text{obs}}\right)}_{\substack{\text{Varianz d. Mittelwerte über alle MI-Sets} \\ \rightarrow \text{Between-Variance } (B)}}$$

Bestimmung der gepoolten Varianz von Parametern bei MI

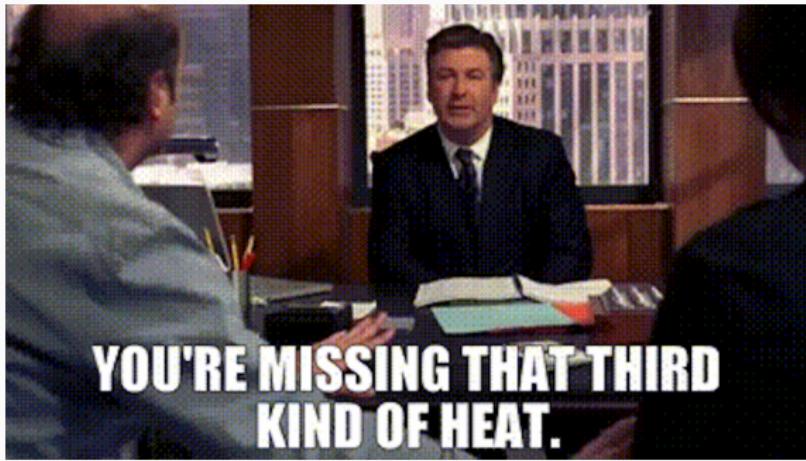
Um die Varianz eines Parameters Q zu bestimmen (z.B. für Konfidenzintervalle), muss bei MI sowohl die (gemittelte) Varianz durch den Stichprobenfehler \bar{U} , als auch die Imputationsunsicherheit B mit berücksichtigt werden. Die Berechnung der gepoolten Varianz erfolgt durch die sog. "**Rubin-Regeln**" (s. n. Folie).

"Rubin's Rules" - Die Kombinationsregeln nach Rubin (Rubin, 1987)

Die Rubin-Regeln stellen eine allgemeine Formel dar, nach der die MI-Varianz eines Punktschätzers Q im konkreten Fall berechnet werden kann:

$$\begin{aligned}\hat{V} &= \overbrace{\left(\frac{1}{m} \sum_{\ell=1}^m \bar{U}_\ell \right)}^{\bar{U}} + \left(1 + \frac{1}{m} \right) \overbrace{\left(\frac{1}{m-1} \sum_{\ell=1}^m (\hat{Q}_\ell - \bar{Q})(\hat{Q}_\ell - \bar{Q})' \right)}^B \\ &= \bar{U} + \left(1 + \frac{1}{m} \right) B \\ \Rightarrow \bar{U} + B \quad \text{as} \quad m &\rightarrow \infty\end{aligned}$$

...aber warum *drei* Variationsquellen?



Die Rubin-Formeln beziehen auch mit ein, dass immer nur eine finite Anzahl an Imputationssets generiert werden (\rightarrow Einbezug der **Simulationsvarianz**).

$$\begin{aligned}\hat{V} &= \overbrace{\left(\frac{1}{m} \sum_{\ell=1}^m \bar{U}_\ell \right)}^{\bar{U}} + \left(1 + \frac{1}{m} \right) \overbrace{\left(\frac{1}{m-1} \sum_{\ell=1}^m (\hat{Q}_\ell - \bar{Q})(\hat{Q}_\ell - \bar{Q})' \right)}^B \\ &= \bar{U} + \left(1 + \frac{1}{m} \right) B \\ \Rightarrow \bar{U} + B &\quad \text{as } m \rightarrow \infty\end{aligned}$$

→ Je größer m , desto **geringer fällt diese Komponente ins Gewicht**.

Insbesondere, wenn besonders genaue Varianzschätzungen notwendig sind, empfiehlt sich daher eine **hohe Anzahl an Imputationssets**.

Ein niedriges m führt zu **Konfidenzintervallen**, die **etwas breiter** sind als wenn $m \rightarrow \infty$ (niedrigere Effizienz). Dieser Unterschied ist in der Praxis jedoch typischerweise **überschaubar**.

Metriken & Freiheitsgrade bei MI-Analysen

Metriken für Parameterschätzungen in MI (Van Buuren, 2018, Kapitel 2.3.5)

Relative Increase in Variance Due to Nonresponse (RIV): Relativer Anstieg der Varianz aufgrund der Imputationsunsicherheit (wenn RIV > 1: Imputationsvarianz größer als “echte” Varianz in Y):

$$r_Q = \frac{B_Q/m + B_Q}{\bar{U}_Q}$$

Fraction of Missing Information Due to Nonresponse (FMI): Anteil der Information über Q , die durch die Imputationsunsicherheit “verloren geht”:

$$\gamma_Q = \frac{(r_Q + 2)/(\nu_Q + 3)}{1 + r_Q}$$

Wobei ν (“nu”) für die Freiheitsgrade bei der Schätzung von Q steht.

Freiheitsgrade bei MI-Analysen (Van Buuren, 2018, Kapitel 2.3.6)

Die Freiheitsgrade eines Modells sind definiert als **Anzahl der Beobachtungen** nach Abzug der **Modellparameter**: $\nu = n - k$ (für Mittelwert z.B. $\nu = n - 1$).

Bei MI sind manche Elemente von n nur **Schätzungen von Beobachtungen**, daher muss ν dafür **korrigiert** werden (Rubin, 1987):

$$\nu_{(MI)} = (m - 1) \left(1 + \frac{1}{r^2} \right)$$

$$\lim_{r \rightarrow 0} \nu_{(MI)} = \infty \quad \text{sowie} \quad \lim_{r \rightarrow \infty} \nu_{(MI)} = m - 1$$

Diese Formel basiert auf der Annahme, dass die **Freiheitsgrade des vollständigen Datensatzes** (den MI zu schätzen versucht) **unendlich groß** sind! Diese Approximation ist aber erst sinnvoll, wenn ein **relativ großes Sample** vorliegt.

Freiheitsgrade bei MI-Analysen (Van Buuren, 2018, Kapitel 2.3.6)

Für kleine Stichproben kann eine adaptierte Formel genutzt werden (Barnard & Rubin, 1999). Dafür müssen zuerst die **“hypothetischen” Freiheitsgrade** bestimmt werden, wenn die Daten komplett wären (“complete data degrees of freedom; ν_{com}). Es sei n die Anzahl der Beobachtungen und k die Anzahl der Parameter:

$$\nu_{\text{com}} = n - k$$

Daraus lassen sich die Freiheitsgrade der **beobachteten Werte** (ν_{obs}) bestimmen:

$$\nu_{\text{obs}} = \frac{\nu_{\text{com}} + 1}{\nu_{\text{com}} + 3} \nu_{\text{com}} \left(1 - \frac{r}{r + 1}\right)$$

Diese können wiederum zur **Korrektur** von $\nu_{(\text{MI})}$ verwendet werden:

$$\nu_{(\text{MI})}^* = \frac{\nu_{(\text{MI})} \nu_{\text{obs}}}{\nu_{(\text{MI})} + \nu_{\text{obs}}}$$

Freiheitsgrade bei MI-Analysen (Van Buuren, 2018, Kapitel 2.3.6)

Praktische Anmerkungen

- Wird die unkorrigierte Formel verwendet (default bei `{mitml}`), können die **angezeigten Freiheitsgrade** eines Modells/Parameters **größer als n sein!**
- Da die Anzahl der Freiheitsgrade bei MI anhand der obigen Formeln angenähert wird, muss ν im konkreten Fall **keine natürliche Zahl** sein (z.B. $\nu = 226.6559$)!

Konfidenzintervalle (Van Buuren, 2018, Kapitel 2.4.2)

Die berechneten Freiheitsgrade können zur Berechnung von **Konfidenzintervallen** genutzt werden. Zur Inferenz von Skalaren (= \bar{Q} ist ein einziger Wert) wird dabei häufig eine t -Verteilung angenommen.

\bar{Q} sei ein aggregierter Parameter (z.B. ein Regressionsgewicht b), und Q_0 der Referenzwert der Nullhypothese (typischerweise 0):

$$\frac{\bar{Q} - Q_0}{\sqrt{\hat{V}}} \sim t_{\nu}$$

Zusammen mit den berechneten MI-Freiheitsgraden lässt sich so ein passendes 95% Konfidenzintervall berechnen, dass die Imputationsunsicherheit berücksichtigt:

$$\bar{Q} \pm t_{\nu_{(MI)}^{(*)}, 0.975} \times \sqrt{\hat{V}}$$

Wichtige Anmerkungen zur Analyse von MI-Daten

Cave I: Aggregation von Teststatistiken

- In der Praxis ist es häufig notwendig, Teststatistiken (z.B. χ^2 oder F -Werte) in **jedem Imputationsset zu berechnen, und daraufhin zu poolen.**
- Teststatistiken können aber typischerweise nicht einfach wie bei \bar{Q} durch das **arithmetische Mittel** gepoolt werden!
- Hintergrund dafür ist, dass die **Größe der Teststatistiken von Varianz & Freiheitsgraden** abhängig ist. Diese ist/sind **bei MI größer bzw. kleiner** aufgrund der **Imputationsunsicherheit**.
- Um Teststatistiken nicht zu **überschätzen**, müssen daher **besondere Formeln** zur Aggregation eingesetzt werden.
- Implementationen in R sind beispielsweise die `micombine.chisquare` und `micombine.F` function im `{miceadds}` package.

→ Methoden zur Aggregation von Test(statistiken) sind ein **aktives Forschungsfeld**, und weitere Implementierungen in R sind zu erwarten (s. z.B. Grund et al., 2021).

Cave II: Aggregation von Korrelationen

- Der Wert einer Korrelation ρ kann die **Maximalwerte** $[-1; 1]$ **nicht überschreiten**; dies bedeutet, dass die **Varianz** von ρ **eingeschränkt** wird, je weiter $|\rho|$ gegen 1 geht.
- Dies führt dazu, dass für ρ **keine asymptotische Normalverteilung** angenommen werden kann.
- Dadurch kann zur Aggregation von Korrelationen **nicht einfach der Mittelwert** berechnet werden. Korrelationen sollten vorher einer **varianzstabilisierenden Transformation** unterzogen werden, der **Fisher z-Transformation**.
- In R können Korrelationen komfortabel mit der `micombine.cor` function im `{miceadds}` package aggregiert werden.

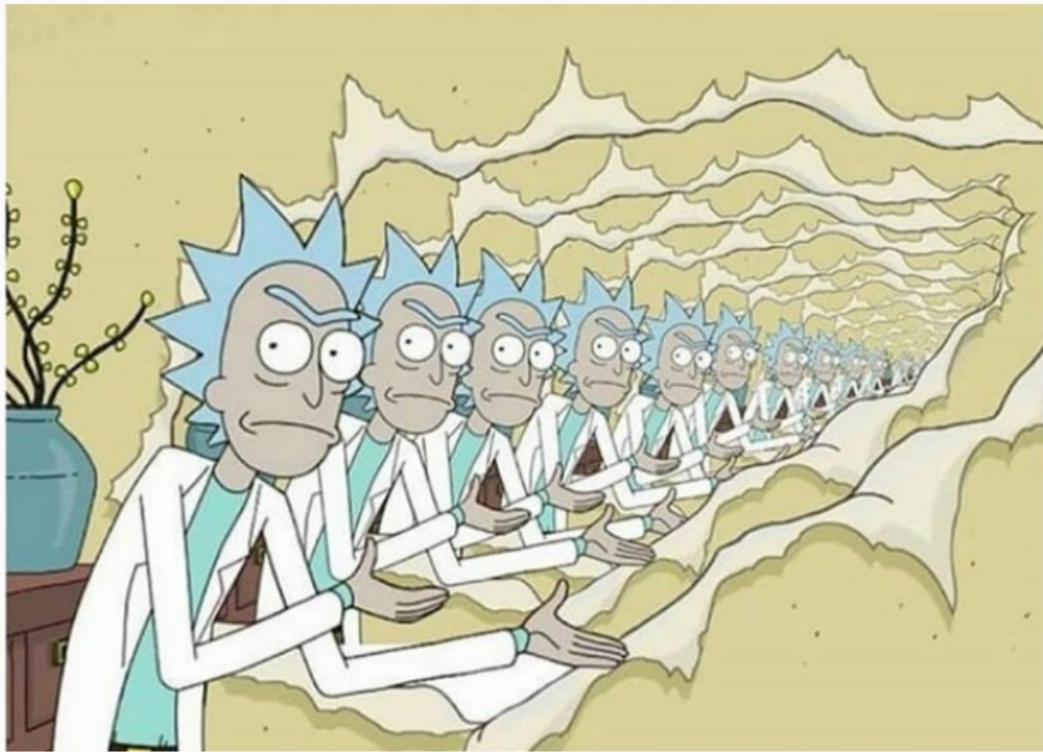
vgl. Marshall et al. (2010).

Cave III: Analyse in aggregierten Daten

- In der Praxis findet man häufig das Vorgehen, **alle MI-Sets zu einem vollständigen Datensatz** zu aggregieren, und dann Analysen in diesem (einen) Datensatz durchzuführen.
- Dies erleichtert zwar die Auswertung mit gängiger Statistiksoftware enorm, ist aber **aus statistischer Sicht unbedingt zu vermeiden!**
- Durch die Aggregation zu einem Dataset wird "**vorgegaukelt**", dass **keine Imputationsunsicherheit existiert**; dies führt zu inkorrekten, weil antikonservativen p -Werten, Konfidenzintervallen, Teststatistiken, etc.
- Eine Analyse in aggregierten Daten sollte daher höchstens dann durchgeführt werden, wenn keinerlei statistische Tests oder Quantifizierung der Parameterunsicherheit angestrebt wird; dies ist jedoch in der Praxis selten der Fall.

vgl. Van Buuren (2018), Kapitel 5.1.2.

“Rinse & Repeat”: Parallel Bearbeitung von Listenelementen



Bei der Auswertung von multipel imputierten Daten in R müssen häufig **Operationen in allen m Sets gleichzeitig durchgeführt** werden. Dies führt oft dazu, dass selbst einfache Analyseschritte **deutlich komplizierter** werden.

Eine mögliche Implementierung sind **for Loops**:

```
# Berechne aggregierten Mittelwert über alle Sets
means = vector()
for (i in 1:25){
  means[[i]] <- implist[[i]] %>% pull(pss.0) %>% mean()
}
mean <- mean(means)
```

→ **Nachteil:** trotz einfacher Operation komplexer Code, lange Rechenzeit.

Functional Programming (Wickham, 2019, Kapitel 9)

Anstatt mit dem gleichen Befehl durch alle Sets zu loopen, kann stattdessen ein **“funktionaler” Programmierstil** gewählt werden. D.h. es wird eine **Funktion** genutzt, die **wiederum selbst Funktionen** auf alle Imputationssets **anwendet**.

In Base-R sind dies Funktionen wie `apply`, `mapply`, `vapply`, `Reduce`, etc.

Besonders benutzerfreundlich und konsistent sind aber die `map`-Funktionen im package `{purrr}` (Henry & Wickham, 2020):

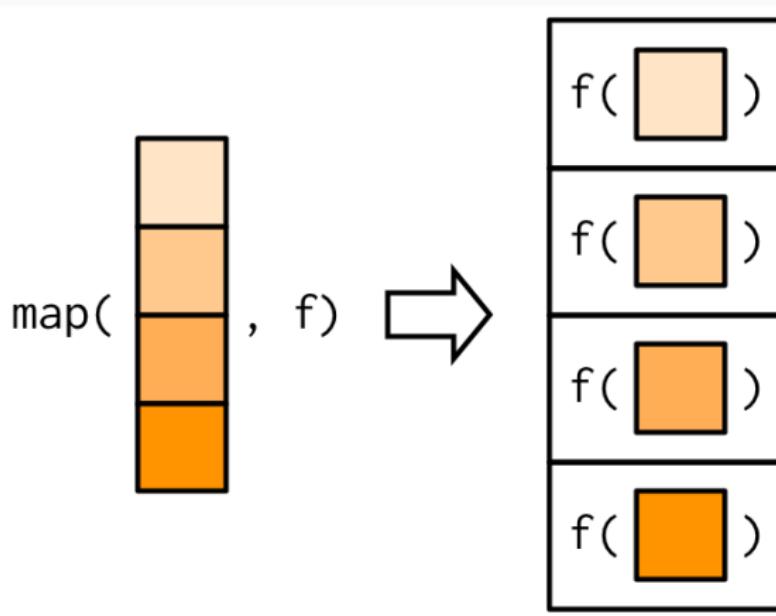
```
mean <- implist %>%  
  map_dbl(~mean(. $pss.0)) %>%  
  mean()
```

→ **Vorteil:** knapper, übersichtlicher Code; stark verkürzte Rechenzeit.



Functional Programming (Wickham, 2019, Kapitel 9)

Die Funktionsweise von map:



Functional Programming (Wickham, 2019, Kapitel 9)

Die “Geschmacksrichtungen” von map:

- `map`: Output ist ein list-Objekt.
- `map_dbl`: Output ist ein numeric-Vektor.
- `map_chr`: Output ist ein character-Vektor.
- `map_lgl`: Output ist ein logical-Vektor.
- `map_dfr`: Output ist ein `data.frame`.
- `map2(_*)`: Iteration über zwei Listen gleichzeitig.



Functional Programming (Wickham, 2019, Kapitel 9)

map akzeptiert Funktionen auf 2 Arten:

1. **“Klassisch”**: eine “voll funktionstüchtige” Funktion wird in map gesteckt.

```
# 'x' repräsentiert das individuelle Listenelement in 'list'  
list %>% map(function(x) mean(x$variable))
```

2. **“Verkürzt”**: mit “~” und “.” wird der Funktionscode abgekürzt.

```
# '.' repräsentiert das individuelle Listenelement in 'list'  
# Der Beginn einer Funktion wird durch '~' (tilde) angezeigt.  
list %>% map(~ mean(.variable))
```



Praxis-Teil



Primäre Wirksamkeitsanalyse

In der primären Wirksamkeitsanalyse wird die **Effektivität der Interventionsbedingung** evaluiert (“hatte die Interventionen einen Effekt?”).

Die Analyse fokussiert dabei auf das **primäre Outcome** (mit *a priori* definiertem Messzeitpunkt und Messinstrument), und ob darin **Unterschiede zwischen den Gruppen** bestehen:

$$\begin{aligned}|Y_{i,t}(X_1) - Y_{i,t}(X_0)| &> 0 \\ \Rightarrow |\hat{\mu}_{1,t} - \hat{\mu}_{0,t}| &> 0\end{aligned}$$

Können wir dies statistisch Nachweisen, kann geschlossen werden, dass ein Effekt der Intervention vorliegt ($|\tau| > 0$).

Das übliche Verfahren hierzu stellt die **Analysis of Covariance (ANCOVA)** dar.

Analysis of Covariance (Montgomery, 1997, Kapitel 15.3; Dunn & Smyth, 2018, Kapitel 2.9)

ANCOVAs untersuchen, ob zwei oder mehrere Gruppen sich hinsichtlich einer gemessenen Variable y unterscheiden, wenn für den **Einfluss** von (einer oder mehreren) **Kovariaten** (z.B. Baseline-Messung des primären Outcomes) kontrolliert wird.

Die Methode der AN(C)OVA geht auf R.A. Fisher zurück, und ist eng mit der experimentellen Methodik randomisierter Studien verbunden.

TABLE 38

	Degrees of Freedom.	Sum of Squares.	
Within classes . .	$n'(k - 1)$	$\sum_{j=1}^{n'} (x - \bar{x}_j)^2$	$n's^2(k - 1)(1 - r)$
Between classes . .	$n' - 1$	$k \sum_{j=1}^{n'} (\bar{x}_j - \bar{x})^2$	$n's^2[1 + (k - 1)r]$
Total . .	$n'k - 1$	$\sum_{j=1}^{n'} (x - \bar{x})^2$	$n's^2k$

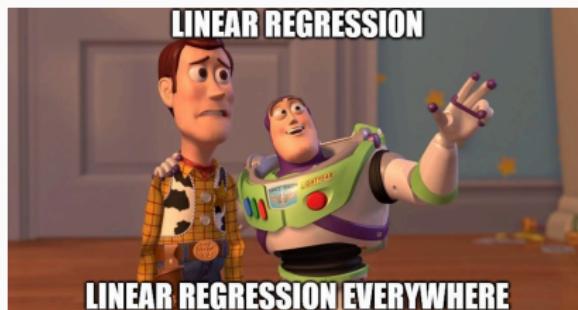
TABLE 39

	Degrees of Freedom.	Sum of Squares.	
Within classes	$n'(k - 1)$	$\sum_{j=1}^{n'} (x - \bar{x}_j)^2$	$n'(k - 1)B = n's^2(k - 1)(1 - r)$
Between classes	$n' - 1$	$k \sum_{j=1}^{n'} (\bar{x}_j - \bar{x})^2$	$(n' - 1)(k\Lambda + B) = (n' - 1)s^2[1 + (k - 1)r]$
Total	$n'k - 1$	$\sum_{j=1}^{n'} (x - \bar{x})^2$	$(n' - 1)k\Lambda + (n'k - 1)B = s^2[n'k - 1 - (k - 1)r]$

Statistical Methods for Research Workers (1925)

Analysis of Covariance (Montgomery, 1997, Kapitel 15.3; Dunn & Smyth, 2018, Kapitel 2.9)

→ Die Durchführung von ANCOVAs in RCT-Analysen ist daher auch historisch zu erklären. Konzeptuell ist das Modell hinter ANCOVAs schlicht ein **Spezialfall linearer Regression!**



Analysis of Covariance (Montgomery, 1997, Kapitel 15.3; Dunn & Smyth, 2018, Kapitel 2.9)

Es sei:

- y_{ij} der Wert des (kontinuierlichen) primären Outcomes von Person j in Gruppe i ;
- τ_i der Effekt der i -ten Behandlungsgruppe (z.B. Intervention oder Kontrolle);
- x_{ij} der Wert von ij auf einer Kovariate.

Das Modell der ANCOVA ist dann:

$$y_{ij} = \mu + \tau_i + \beta(x_{ij} - \bar{x}) + \epsilon_{ij} \quad \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, n \end{cases}$$

Die **Residuale** des Modells folgen einer Normalverteilung mit Mittelwert 0: $\epsilon \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. ANCOVAs basieren auf einer **Effekt-Kodierung** der Treatment-Variable (z.B. -1 und 1), sodass sich die τ_i -Werte zu null aufsummieren:

$$\sum_{i=1}^a \tau_i = 0$$

Analysis of Covariance (Montgomery, 1997, Kapitel 15.3; Dunn & Smyth, 2018, Kapitel 2.9)

AN(C)OVAs basieren auf dem Prinzip der **Varianzzerlegung**:

$$SS_{\text{Total}} = SS_{\text{Prädiktor}} + SS_{\text{Residual}}$$

Daten = Modellfit + Unerklärte Varianz

Variationsquelle	Quadratsumme	Freiheitsgrade	Mittlere Quadratsumme
Systematisch	$SS_{\text{Prädiktor}}$	$\nu_{\text{num}} = p - 1$	$MSS_{\text{Prädiktor}} = \frac{SS_{\text{Prädiktor}}}{p-1}$
Zufällig	SS_{Residual}	$\nu_{\text{den}} = n - p$	$SS_{\text{Residual}} = \frac{SS_{\text{Residual}}}{n-p}$
Total	SS_{Total}	$n - 1$	-

→ Bei ANCOVAs wird der Einfluss der Kovariate auf die Within-Varianz (SS_{Residual}) "herausgerechnet".

Analysis of Covariance (Montgomery, 1997, Kapitel 15.3; Dunn & Smyth, 2018, Kapitel 2.9)

Die **Signifikanz des Treatment-Effekts** wird dann über den F -Test ermittelt. Dieser vergleicht die **Variation durch den Treatmentfaktor** mit der **unerklärten Variation** in den Daten:

$$F_{\nu_{\text{num}}, \nu_{\text{den}}} = \frac{\text{SS}_{\text{Prädiktor}}/(p - 1)}{\text{SS}_{\text{Residual}}/(n - p)}$$

Reduktion von $\text{SS}_{\text{Residual}}$ bei ANCOVAs

- Durch den Einschluss prognostischer Variablen wird die unerklärte Varianz innerhalb der Gruppen verringert. Dadurch verkleinert sich $\text{SS}_{\text{Residual}}$ und der F -Wert wird größer \Rightarrow mehr Power zum Nachweis des Treatment-Effekts!
- **Cave:** Dies ist nur der Fall, wenn die Kovariate tatsächlich prognostisch relevant ist. Wenn nicht, erhöht sich nur die Anzahl der Parameter p \Rightarrow **Power sinkt!**

Analysis of Covariance (Montgomery, 1997, Kapitel 15.3; Dunn & Smyth, 2018, Kapitel 2.9)

Vorteile der Adjustierung von Baseline-Variablen:

- **Interpretierbarkeit:** *"The central question is for two patients with the same pre measurement value of x , one given treatment A and the other treatment B, will the patients tend to have different post-treatment values? This is exactly what analysis of covariance assesses."* (Harrell, 2021)
- **Power:** Adjustierung von prognostischen Variablen führt zu höherer Effizienz der Analysen (engere Konfidenzintervalle); bei dichotomen Outcomes (Odds Ratio) erhöht sich der Effekt selbst (Hernández et al., 2004).
- **Baselineunterschiede:** Kovariaten adjustieren für systematische Baselineunterschiede, sollten diese tatsächlich vorliegen.

Analysis of Covariance (Montgomery, 1997, Kapitel 15.3; Dunn & Smyth, 2018, Kapitel 2.9)

Empfehlungen

- Generell sollten nur Kovariation adjustiert werden, für die ein **starker prognostischer Zusammenhang plausibel** ist → bei der Baselinemessung des primären Outcomes voraussetzbar!
- Bei stratifizierter Randomisierung sollten die **Stratifizierungsvariablen** kontrolliert werden (Kahan & Morris, 2012).
- Alle Kovariaten sollten ***a priori*** präspezifiziert werden (Assmann et al., 2000).

Analysis of Covariance (Montgomery, 1997, Kapitel 15.3; Dunn & Smyth, 2018, Kapitel 2.9)

Anzahl der Kovariaten: Richtlinien der European Medicines Agency

“No more than **a few covariates should be included** in the primary analysis. Even though methods of adjustment, such as analysis of covariance, can theoretically adjust for a large number of covariates it is safer to pre-specify a simple model.”

“Results based on such a model are more likely to be numerically stable, the assumptions underpinning the statistical model are easier to validate and generalisability of the results may be improved.”

EMA (2015), Absatz 6.2.



Referenzen

Assmann, S. F., Pocock, S. J., Enos, L. E., & Kasten, L. E. (2000). Subgroup analysis and other (mis) uses of baseline data in clinical trials. *The Lancet*, 355(9209), 1064–1069.

Barnard, J., & Rubin, D. B. (1999). Miscellanea. Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4), 948–955.

Dunn, P. K., & Smyth, G. K. (2018). *Generalized linear models*. Springer.

EMA. (2015). *Guideline on adjustment for baseline covariates in clinical trials*.

https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-adjustment-baseline%20%20-covariates-clinical-trials_en.pdf

Grund, S., Lüdtke, O., & Robitzsch, A. (2021). *Pooling methods for likelihood ratio tests in multiply imputed data sets*.

Harrell, F. (2021). *Statistical errors in the medical literature*.

<https://www.fharrell.com/post/errmed/>

Henry, L., & Wickham, H. (2020). *Purrr: Functional programming tools*.

<https://CRAN.R-project.org/package=purrr>

- Hernández, A. V., Steyerberg, E. W., & Habbema, J. D. F. (2004). Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *Journal of Clinical Epidemiology*, 57(5), 454–460.
- Kahan, B. C., & Morris, T. P. (2012). Reporting and analysis of trials using stratified randomisation in leading medical journals: Review and reanalysis. *BMJ*, 345.
<https://doi.org/10.1136/bmj.e5840>
- Marshall, A., Altman, D. G., Royston, P., & Holder, R. L. (2010). Comparison of techniques for handling missing covariate data within prognostic modelling studies: A simulation study. *BMC Medical Research Methodology*, 10(1), 1–16.
- Montgomery, D. (1997). *Design and analysis of experiments*. John Wiley.
- Rubin, D. B. (1987). *Multiple imputation for survey nonresponse*. New York: Wiley.
- Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.
- Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(1), 1–67.
- Wickham, H. (2019). *Advanced r*. chapman; hall/CRC.