

# Topic 14: FH models for means with a log transformation

Tristan Pfuderer

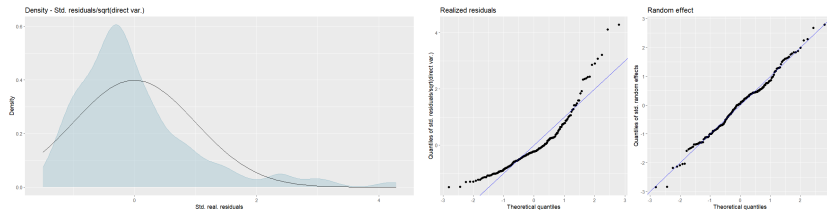
Otto-Friedrich-Universität Bamberg

15.Juli.2025

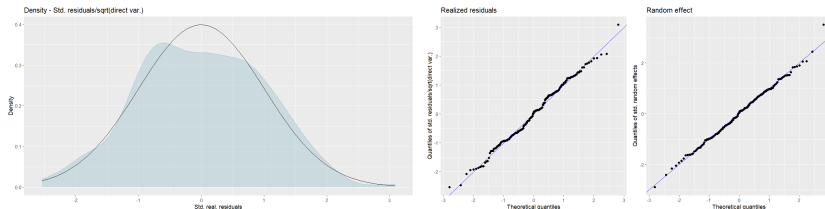
- 1 Motivation
- 2 Methodik
- 3 Simulationsstudie
- 4 Synthetic Business
- 5 Zusammenfassung und Ausblick

# Motivation

- **Grundproblem:** Entweder  $v_i$  oder  $u_i$  sind nicht normalverteilt, d. h.  $v_i, u_i \not\sim \mathcal{N}(0, \sigma^2)$ .



- **Lösung:**  $\log(\theta_i) = x_i^\top \beta + v_i + u_i$  dann  $v_i \sim \mathcal{N}(0, \sigma_v^2)$ ,  $u_i \sim \mathcal{N}(0, \sigma_u^2)$



- Ergebnisse jetzt aber auf *log* skala.
- $\log(10000)\text{€} = 9.21$  Wie soll man das Interpretieren "actionable" machen?
- Antwort: Rücktransformation von *log* auf original. Das sorgt aber für ein Großes Problem  $\Rightarrow$  Bias!
- Die korrekte Rücktransformation Analytisch zu Erklären und Lösen ist mein Ziel.

# Erklärung des Problems von nicht Normalität

- Wenn z.B. Einkommen log-normalverteilt sind:

$$\theta_i \sim \text{log-Normal}(\mu_i, \sigma^2) \Rightarrow \hat{\theta}_i^{\text{Dir}} \text{ ist nicht normalverteilt}$$

- Das klassische Fay-Herriot-Modell setzt voraus:

$$\hat{\theta}_i^{\text{Dir}} = x_i^\top \beta + u_i + e_i, \quad u_i \sim \mathcal{N}(0, \sigma_u^2), \quad e_i \sim \mathcal{N}(0, D_i)$$

- Weil: If either  $v_i$  or  $u_i \not\sim \mathcal{N}$  then “linear conditional expectation property of the multivariate normal distribution” does no longer hold.
- EBLUP nicht mehr "best" (Varianz) oder "unbiased" (auch BLUP)
- Folge: Bias und erhöhte RMSE in Originalskala trotz korrekter Umsetzung des Modells

## Unter Normalität:

$$\mathbb{E}[\theta_d \mid \hat{y}_d] = x_d^\top \beta + \gamma_d(\hat{y}_d - x_d^\top \beta)$$

(Aßmann) Theorem 4.11: "lineare bedingte Erwartung unter Normalverteilung" ermöglicht:

$$\widehat{\text{MSE}}(\hat{\theta}_d^{\text{EBLUP}}) = g_1 + g_2 + 2g_3$$

- $g_1 = \gamma_d \sigma_{e_d}^2$ : Modellvarianz
- $g_2 \approx (1 - \gamma_d)^2 \cdot \text{Var}(x_d^\top \hat{\beta})$ : Schätzfehler
- $g_3 \approx \text{Var}[(\hat{\gamma}_d - \gamma_d)(\hat{y}_d - x_d^\top \hat{\beta})]$ : Varianzschätzfehler
- $g_2, g_3$  verschwinden für  $D \rightarrow \infty$

## Ohne Normalität:

- $\mathbb{E}[\theta_d \mid \hat{y}_d]$  nicht mehr linear
- Nicht optimal  $\Rightarrow$  Verzerrung
- $g_2, g_3$  schrumpfen nicht  $\Rightarrow$  Bias und erhöhter RMSE

# Erklärung von log-Transformation

- Wirkung der **log-Transformation** auf die Daten: Große Werte (z. B. Ausreißer) werden überproportional verkleinert.

Beispiel:

$$\log(2,000 \text{ €}) \approx 7,60, \quad \log(60,000 \text{ €}) \approx 11,00$$

- Für EBLUP auch unter log-normal Daten, bietet sich eine log-transformation an.

$$\begin{aligned}\hat{\theta}_i^{\text{Dir}^*\log} &= \log\left(\hat{\theta}_i^{\text{Dir}}\right), \\ \text{var}\left(\hat{\theta}_i^{\text{Dir}^*\log}\right) &= \left(\hat{\theta}_i^{\text{Dir}}\right)^{-2} \text{var}\left(\hat{\theta}_i^{\text{Dir}}\right).\end{aligned}$$

*Quelle: Neves et al. (2013)*

- Wichtig: Auch Varianz muss korrekt Transformiert werden.

# Erklärung des Problems von naiver Rücktransformation (vom mean!)

- Jensens Inequality:

$$\mathbb{E} \left[ \exp \left( \hat{\theta}_i^B \right) \right] \neq \exp \left( \mathbb{E} \left[ \hat{\theta}_i^B \right] \right)$$

*Quelle: Molina & Rao (2006)*

- Generell:

$$\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$$

*Quelle: Aßmann Vorlesung*

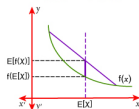
## Jensen's Inequality



States that if 'X' is an integrable random variable and  $f: \mathbb{R} \rightarrow \mathbb{R}$  is a convex or concave function, then

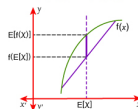
For Convex Function

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$



For Concave Function

$$\mathbb{E}[f(X)] \leq f(\mathbb{E}[X])$$



# Beweisidee von Jensens Ungleichung (Asmann)

## Beweisidee:

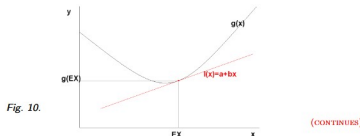
- Lege eine Tangente  $\ell(x) = a + bx$  an  $g(x)$  bei  $x = \mathbb{E}[X]$
- Da  $g$  konvex ist, gilt:

$$g(x) \geq \ell(x) = a + bx \quad \forall x$$

- Erwartungswert ergibt dann:

$$\mathbb{E}[g(X)] \geq \mathbb{E}[\ell(X)] = a + b\mathbb{E}[X] = \ell(\mathbb{E}[X]) = g(\mathbb{E}[X])$$

so that  $\mathbb{E}g(X) \geq g(\mathbb{E}X)$ , as was to be shown (for a continuous  $X$  analogous).



**Figure:** Tangente liegt unterhalb der konvexen Funktion  $g(x)$

**Theorem:** Wenn  $g$  konvex

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$$

**Beispiel:**  $g(x) = x^2$ , also konvex.

- $P(X = 1) = 0,5$  und  $P(X = 3) = 0,5$
- Erwartungswert:  $\mathbb{E}[X] = 1 \cdot 0,5 + 3 \cdot 0,5 = 2$
- Transformation:  $\mathbb{E}[X^2] = 1^2 \cdot 0,5 + 3^2 \cdot 0,5 = 0,5 + 4,5 = 5$
- Vergleich:  $\mathbb{E}[X^2] = 5 \geq \mathbb{E}[X]^2 = (2)^2 = 4$

*Interessant:*  $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \geq 0$ .

Wichtig: MSE muss ebenfalls, bias korrigiert, Rücktransformiert werden. Slud&Maiti haben hierzu eine exakte Methode entwickelt. Thema aber nicht für diese Präsentation.

- raw
- naiv
- bc\_crude
- bc\_sm

- Vergleichsgröße: Was passiert mit FH ganz ohne Transformation auf log-normal Verteilten Daten?

$$\hat{\theta}_i^{\text{Dir*log}} \xrightarrow{\text{bleibt}} \hat{\theta}_i^{\text{Dir*log}}$$

$$\hat{\theta}_i^{\text{FH-raw}} = x_i^\top \hat{\beta} + \hat{u}_i = \hat{\gamma}_i \hat{\theta}_i^{\text{Dir*log}} + (1 - \hat{\gamma}_i) x_i^\top \hat{\beta}$$

- Ergebnis ist (zum Vergleich) absichtlich Biased

- Transformiert, aber falsch bzw. Biased Rücktransformiert

$$\hat{\theta}_i^{\text{Dir*log}} = \log \left( \hat{\theta}_i^{\text{Dir}} \right),$$

$$\text{var} \left( \hat{\theta}_i^{\text{Dir*log}} \right) = \left( \hat{\theta}_i^{\text{Dir}} \right)^{-2} \cdot \text{var} \left( \hat{\theta}_i^{\text{Dir}} \right),$$

- Jetzt aber nur mit  $\exp()$  Rücktransformiert

$$\hat{\theta}_i^{\text{FH-naiv}} = \exp \left( \hat{\theta}_i^{\text{FH*log}} \right).$$

Ergebnis wieder biased: Negativer bias (Zumindest bei konkav  $\log()$ ). Korrektur ist ja + (...)

- Transformiert und korrigiert Rücktransformiert

$$\hat{\theta}_i^{\text{FH-crude}} = \exp \left( \hat{\theta}_i^{\text{FH*log}} + 0,5 \cdot \text{MSE} \left( \hat{\theta}_i^{\text{FH*log}} \right) \right)$$

$$\text{nur wenn Parameter bekannt} \stackrel{=}{=} \exp \left( \mathbb{E} \left[ \hat{\theta}_i^B \right] + \frac{\text{Var}(\hat{\theta}_i^B)}{2} \right)$$

- Explizite Nutzung von Jensens Inequality (In Molina & Rao)
- Unter **bekannten Parametern** ist der Schätzer **genau unbiased**
- Beinhaltet gesamten  $\text{MSE}(g_1 + g_2 + g_3)$  (fixed + random effects)
- Idee von  $1/2 \bullet \text{Varianz}$  ist, desto mehr Varianz desto weiter liegen sigma werte auseinander 1 ist Abstand genau 0.5 bei  $2\hat{\sigma} = 4 \bullet 0.5 = 2$  also 4 mal so großer Abstand. Deswegen muss mit Varianz korrigiert werden, nicht nur mit  $1/2$ .

# Woher kommt der Korrekturfaktor? — Beweisidee

Gegeben:  $Y \sim \mathcal{N}(\mu, \sigma^2)$ , also ist  $f_Y(y)$  = Dichte Normalverteilung.

**Ziel:**  $\mathbb{E}[\exp(Y)] = \exp\left(\mu + \frac{\sigma^2}{2}\right)$

**Beweisskizze:**

- 1 Einsetzen der Dichtefunktion in das Integral:

$$\mathbb{E}[\exp(Y)] = \int_{-\infty}^{\infty} \exp(y) \cdot f_Y(y) dy$$

- 2 Exponentialterme zusammenfassen
- 3 Konstante Terme (inkl.  $\mu + \frac{\sigma^2}{2}$ ) aus dem Integral ausklammern

4

$$= \exp\left(\mu + \frac{\sigma^2}{2}\right) \cdot \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(y - (\mu + \frac{\sigma^2}{2}))^2}{2\sigma^2}\right) dy}_{\text{Dichte einer Normalverteilung} = 1}$$

**Bleibt übrig:**

$$\mathbb{E}[\exp(Y)] = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$

- Transformiert und auch analytisch begründet/verbessert rücktransformiert:

$$\hat{\theta}_i^{\text{FH-sm}} = \exp \left( \hat{\theta}_i^{\text{FH*log}} + 0,5 \cdot \hat{\sigma}_u^2 \cdot (1 - \hat{\gamma}_i^{\text{*log}}) \right)$$

- Anders als **crude**: Nur der Bias-Anteil der Random Effects wird berücksichtigt!
- Berücksichtigt zusätzlich den Bias durch Parameterschätzung
- Theoretisch fundierter und genauer als **crude**, da Unsicherheit explizit eingerechnet wird
- Bias-Anteil aus der MSE-Zerlegung:

$$\begin{aligned} \widehat{\text{MSE}}(\hat{\theta}_d^{\text{EBLUP}}) &\approx \gamma_d \cdot \hat{\sigma}_u^2 \cdot (1 - \hat{\gamma}_d) \\ &\quad + (1 - \gamma_d)^2 \cdot \text{Var}(x_d^\top \hat{\beta}) \\ &\quad + 2 \cdot \text{Var} \left[ (\hat{\gamma}_d - \gamma_d)(\hat{y}_d - x_d^\top \hat{\beta}) \right] \end{aligned}$$

- Mithilfe einer Taylor-Reihen-Approximation zweiter Ordnung hergeleitet. Das heißt, es ist nicht mehr nur eine Tangente, sondern approximiert die Form einer Parabel durch „eine Summe von Ableitungen“ – und ist dadurch genauer. (ETH Zürich)
- Slud & Maiti haben dies gezielt auf das FH-Modell angewendet und dabei nur die stochastisch veränderlichen Komponenten korrigiert, also den Area-Random-Effect  $u_i$ .
- `bc_crude` tut das nicht: Diese Methode verwendet einen allgemeinen, nicht FH-spezifischen Ausgleich der Jensen's Inequality, indem sie den gesamten MSE inklusive Fixed Effects berücksichtigt.
- Ich habe zuvor gezeigt, wie der Crude-Korrekturfaktor über ein Integral hergeleitet werden kann, aber man kann dasselbe auch mithilfe einer Taylor-Entwicklung ableiten (übersteigt jedoch meine Kompetenz).

# Simulationsdatengenerierung (Log-Transformation)

- **Ziel:** Log-Verteilte Daten (z.B. Einkommen) simulieren

- **Design:**

- Anzahl Gebiete:  $D = 200$
- Anzahl Simulationen:  $R = 500$

- **Zufallskomponenten:**

- Area Random Effect:  $v_i \sim \mathcal{N}(0, \sigma_u^2 = 0,03)$
- Sampling error:  $e_i \sim \mathcal{N}(0, \text{var}_{\text{dir}_i})$
- Direkte Varianz:  $\text{var}_{\text{dir}_i} \sim \mathcal{U}(0,01, 0,12)$
- Kovariaten:  $x_1, x_2 \sim \mathcal{U}(0,1)$

- **Lineares Modell:**

$$y_i = 5 + 2x_{1i} - 2x_{2i} + v_i$$

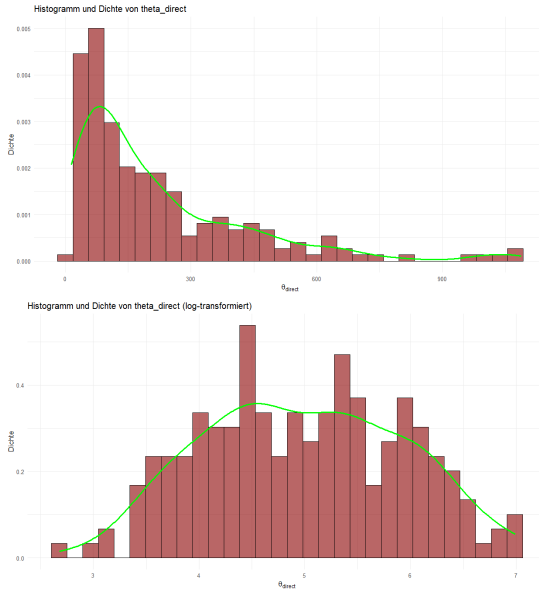
- **Zielgröße (z. B. Einkommen):**

$$\theta_i = \exp(y_i + e_i), \quad \theta_i^{\text{true}} = \exp(y_i)$$

- **Varianz nach Log-Transformation: (Neves schreibt es anders)**

$$\text{var}_{\text{dir}}^{\log} = \exp(\log(\theta))^2 \cdot \text{var}_{\text{dir}}$$

# Verteilung der SimDaten



- **Standard-FH-Schätzungen** mit `fh()`-Funktion aus `emdi`:
  - Transformation: "log", "no"
  - Rücktransformation: "bc\_crude", "bc\_sm", "no"
  - Methode: "ml"
- **Für Naiv-FH-Schätzungen** wegen `emdi`-Limitierung:
  - Manuelle Transformation:  $\log(\hat{\theta}_i^{\text{Dir}})$  und
$$\text{var}(\log(\hat{\theta}_i)) = \left(\hat{\theta}_i^{\text{Dir}}\right)^{-2} \cdot \text{var}(\hat{\theta}_i^{\text{Dir}})$$
  - Transformation: "no"
  - Rücktransformation: "no"
  - Methode: "ml"
  - Manuelle Rücktransformation:  $\hat{\theta}_i^{\text{naiv FH}} = \exp(\hat{\theta}_i^{\log \text{ FH}})$

- **RMSE:**

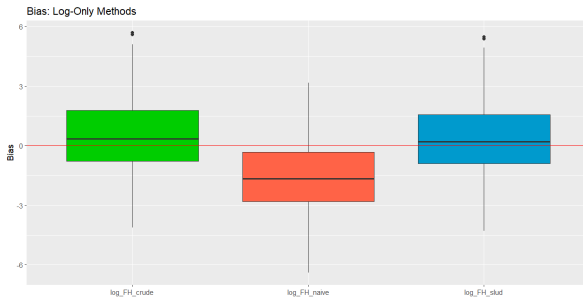
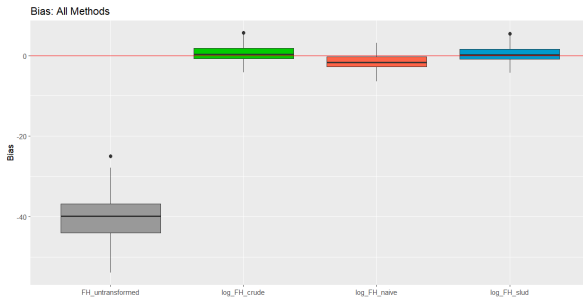
$$\text{RMSE} \left( \hat{\theta}_d^{\text{method}} \right) = \left[ \frac{1}{R} \sum_{r=1}^R \left( \hat{\theta}_d^{\text{method}(r)} - I_d^{(r)} \right)^2 \right]^{1/2}$$

- **Bias:**

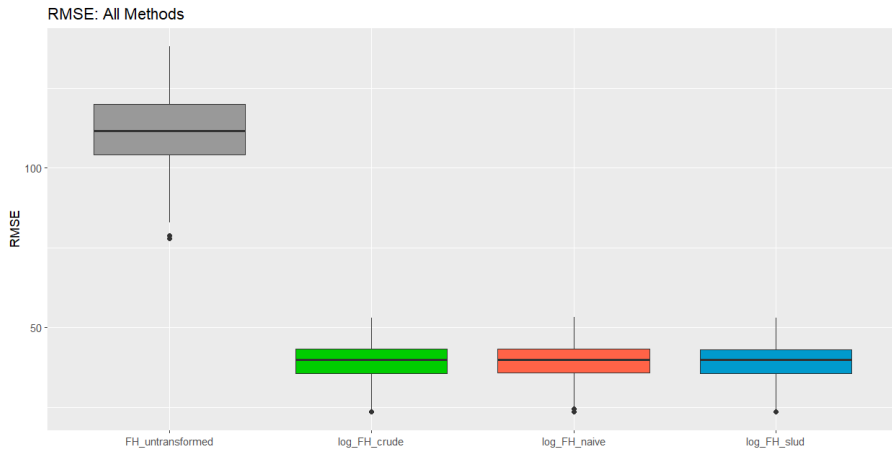
$$\text{Bias} \left( \hat{\theta}_d^{\text{method}} \right) = \frac{1}{R} \sum_{r=1}^R \left( \hat{\theta}_d^{\text{method}(r)} - I_d^{(r)} \right)$$

Beide auf  $R = 500$  Monte-Carlo-Replikationen berechnet.

# Ergebnisse (Bias)



# Ergebnisse (RMSE)



- Szenario Business data  $\Rightarrow$  Private Equity?

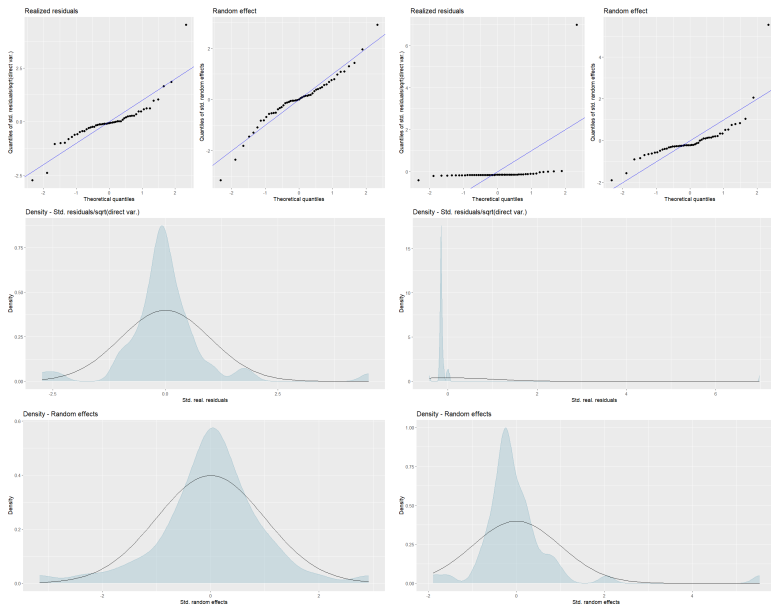
# Datensatz: INC 5000 (skaliert auf 500,000 Unternehmen)

- Basierend auf der **INC 5000 Liste** der am schnellsten wachsenden privaten US-Unternehmen (2019)
- **Skalierung:** Hochgerechnet auf eine synthetische Population mit **500,000 Unternehmen**
- Umfasst **Firmendaten:** Standort, Mitarbeiterzahl, Branche, Gründungsjahr, etc.
- Enthält 14 Variablen pro Unternehmen.

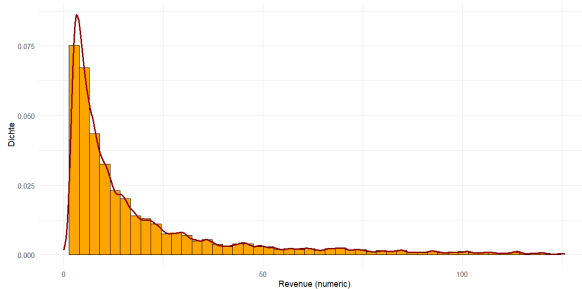
## Zielvariable: `revenue_numeric`

- Unternehmensumsatz (in Millionen USD)
- Ziel: **Schätzung von Umsatz** auf Area-Level (Bundesstaaten)
- Reale Werte mit großer Streuung und starker Rechtsschiefe
- Nur numerische Variablen in FH-Angewendet

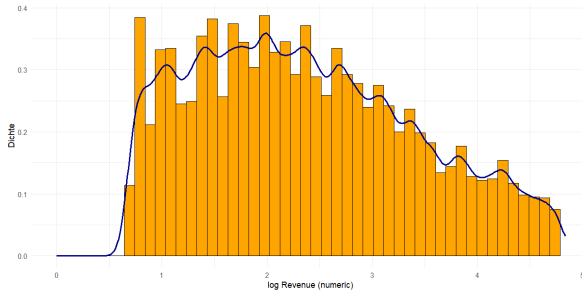
# FH Business full POP Model Plots (LOG vs RAW)

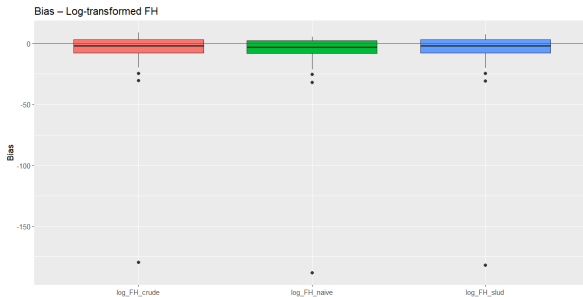
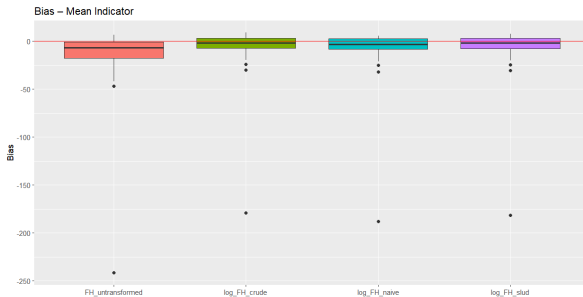


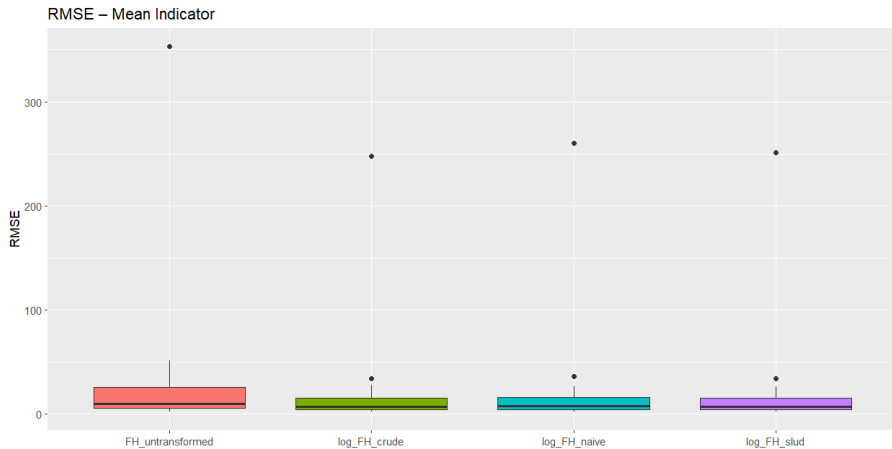
Verteilung von revenue\_numeric (bis 95. Perzentil)



Verteilung von log(revenue\_numeric) (bis 95. Perzentil)







- **Wichtig:** MSE muss ebenfalls korrekt Rücktransformiert werden. Ist aber komplexer. Eigenes Thema.
- Wie in synthetischen Daten zu sehen, log auch limitiert.  
**Alternativen:**
  - EBP bei Unit-Level mit box-cox und shift Parameter
  - MERF
- Wann sollte ich welche Backtransformation benutzen?
  - `bc_crude` wenn Rechenleistung limitiert
  - `bc_sm` wenn möglich (fast immer)

Vielen Dank!  
Fragen?

<https://mathepedia.de/JensenscheUngleichung.html>  
[https://www.researchgate.net/publication/375269760\\_AFrameworkforProducingLevelModels\\_in\\_R](https://www.researchgate.net/publication/375269760_AFrameworkforProducingLevelModels_in_R)  
DOI :  
10.1002/9781118735855Foliensatz3  
[https://wp-prd.let.ethz.ch/analysis19/chapter/taylor – approximation/](https://wp-prd.let.ethz.ch/analysis19/chapter/taylor%20-%20approximation/)