Paper Small Area

Tristan Pfuderer

09.September 2025

1 Einleitung

Das Thema dieser Hausarbeit ist "Topic 14: Fay–Herriot-Modelle für Mittelwerte mit einer Log-Transformation". Ziel ist es, die Probleme und Lösungsansätze im Zusammenhang mit naiven Rücktransformationen unter Log-Transformation im Kontext des Mittelwertes des FH-Modells darzustellen. Grundlage des Problems ist Jensens Inequality, welche bei einer naiven Rücktransformation zu einem systematischen Bias auf der Originalskala führt.
Die Arbeit ist wie folgt gegliedert: Zunächst wird auf die Notwendigkeit und Relevanz des Themas eingegangen. Anschließend wird erläutert, wie Log-Transformationen im FH-Modell funktionieren, wie sich zeigen lässt, dass eine naive Rücktransformation zu verzerrten (biased) Schätzungen führt und welche Korrekturansätze in der Literatur vorgeschlagen wurden. Die Applikation der verschiedenen Bias-Korrekturen wird in einer Simulationsstudie untersucht und abschließend anhand eines synthetic Business-Datensatzes demonstriert.
Es wird vorausgesetzt, dass man mit den Grundlagen des FH-Modells vertraut ist, über grundlegende statistische Kenntnisse verfügt und bereits erste Erfahrungen mit dem R-Package `emdi` gesammelt hat.

Tristan Pfuderer
Otto-Friedrich-Universität Bamberg 09.September.2025

Contents

2 Motivation

Die Small-Area-Estimation (SAE) ist ein wichtiges Forschungsfeld der angewandten Statistik, da sie die Bereitstellung verlässlicher Schätzungen für disaggregierte Regionen oder Subpopulationen ermöglicht. Grundlage ist die Idee, Informationen über Hilfsvariablen mit Modellen zu kombinieren, um die Ungenauigkeit direkter Schätzungen zu verringern. Ein häufig eingesetztes Verfahren ist das Fay–Herriot-Modell für Gebietsebene (Fay und Herriot, 1979).
In vielen praktischen Anwendungen zeigt sich, dass Variablen wie Einkommen auf der Originalskala nicht linear mit den erklärenden Variablen zusammenhängen. Daher wird das Fay–Herriot-Modell oft auf einer log-Skala geschätzt. Diese Transformation stellt die Normalität im Modell wieder her, erfordert aber eine Rücktransformation der Schätzer in die Originalskala. Dabei entsteht ein Bias.
Um diesen Bias zu reduzieren, wurden verschiedene Ansätze entwickelt. Slud und Maiti (2005) schlugen einen Korrekturterm auf Basis der Eigenschaften der Lognormalverteilung vor. Harmening et al. (2023) stellten mit dem `emdi`-Framework ein Werkzeug bereit, das unterschiedliche Varianten und Erweiterungen des Fay–Herriot-Modells implementiert. Es ermöglicht die Nutzung der von Slud und Maiti (2005) eingeführten Methode, sowie einer einfacheren Korrektur wie in Harmening et al. (2023) beschrieben. Einen umfassenden Überblick über Methoden und Anwendungen der Small-Area-Estimation bieten Rao und Molina (2015).

Um das akute Problem zu verdeutlichen, sind in Abbildung 1 zwei Diagnostic Plots aus dem Paket `emdi` dargestellt, mit denen die meisten Anwender vertraut sind.

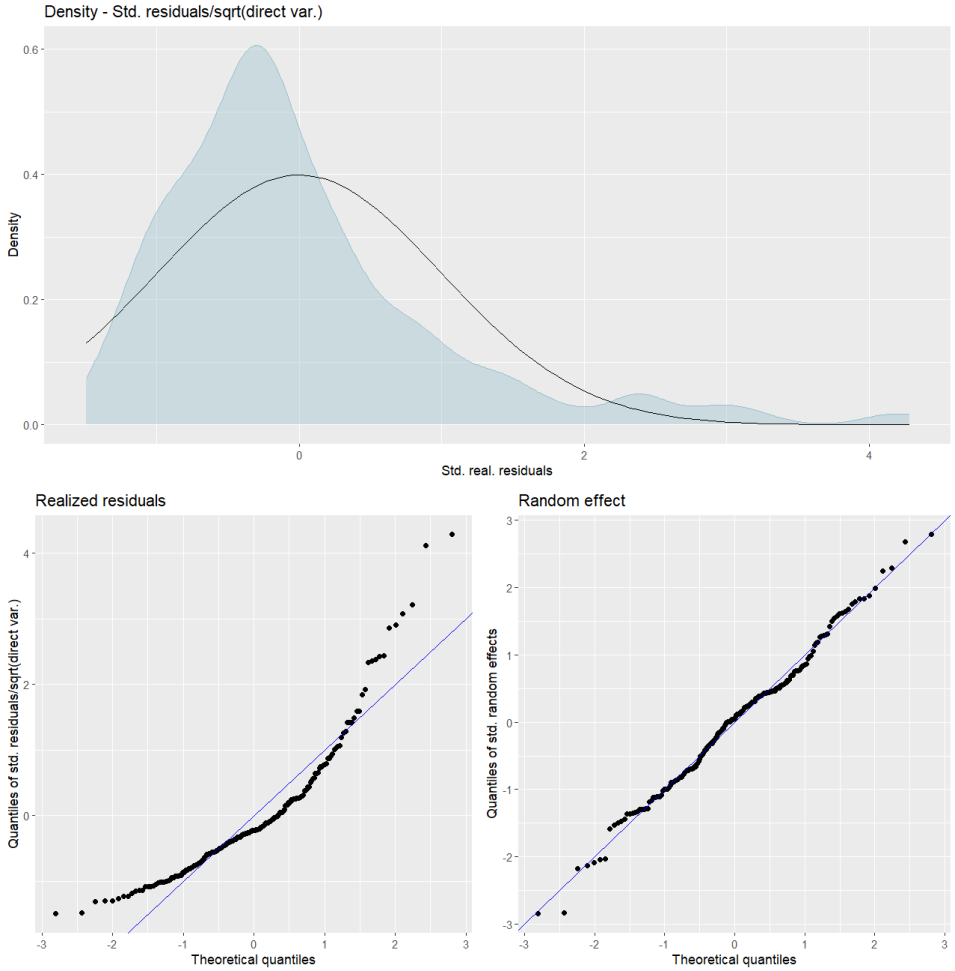


Figure 1: Verteilungen und Quantile auf Originalskala

Das Problem ist klar: der Random Area Effect (u_i) und der Sampling Error (e_i) sind nicht normalverteilt. Dies ist häufig bei einkommensbasierten Daten der Fall, bei denen viele kleine Einkommen und wenige sehr große Einkommen auftreten. Dies führt zu einer Rechtsschiefe:

$$v_i, u_i \not\sim \mathcal{N}(0, \sigma^2).$$

Die Normalverteilung wird jedoch für ein akkurates FH-Modell vorausgesetzt. Die am weitesten verbreitete Lösung ist daher eine Log-Transformation:

$$\log(\theta_i) = x_i^\top \beta + v_i + u_i, \quad v_i \sim \mathcal{N}(0, \sigma_v^2), \quad u_i \sim \mathcal{N}(0, \sigma_u^2).$$

Dies führt, wie in Abbildung 2 zu sehen, zu einer deutlichen Verbesserung im Hinblick auf die Normalität.

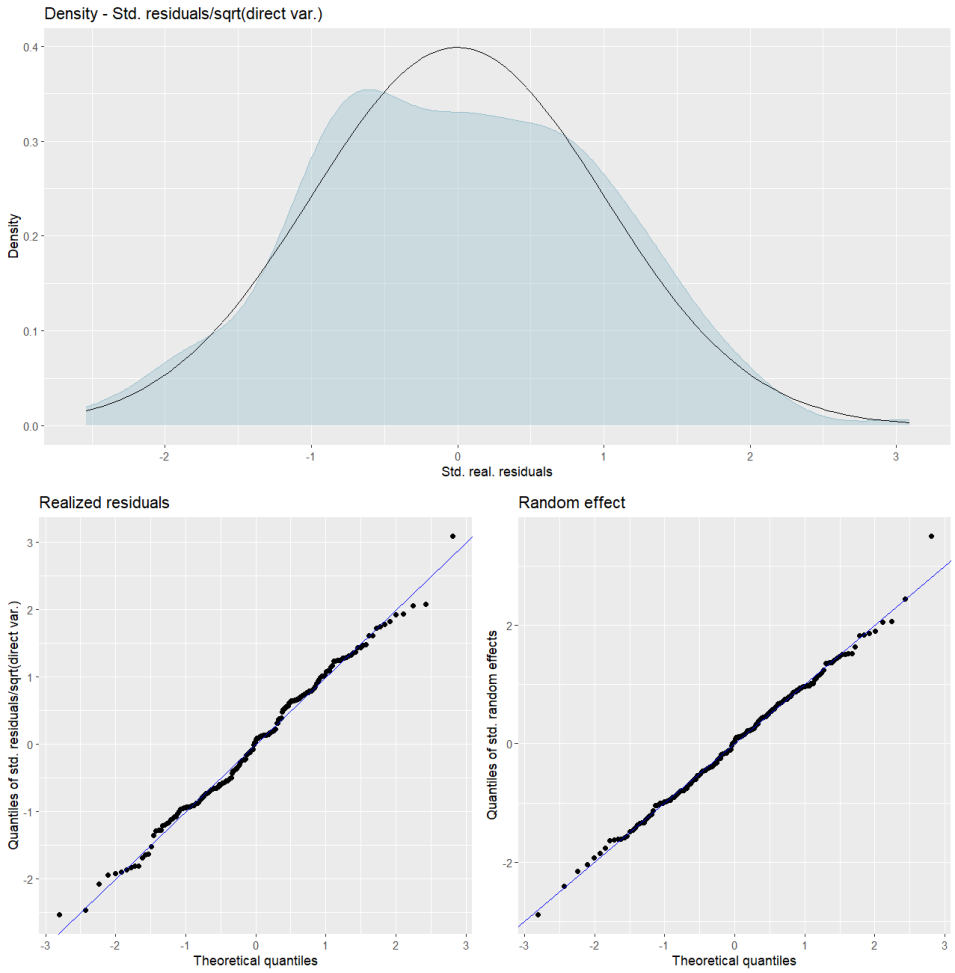


Figure 2: Verteilungen und Quantile nach Log-Transformation

Allerdings ergibt sich ein Nachteil: die Ergebnisse liegen nun auf der Log-Skala und nicht mehr auf der Originalskala vor. So sind z.B. Währungen nicht länger in Euro oder Dollar angegeben. Ein Beispiel:

$$\log(10000\,€) = 9.21.$$

Um Ergebnisse wieder auf die Originalskala (z. B. Euro) zurückzuführen, muss die Inverse der Log-Funktion angewendet werden, also die Exponentialfunktion. Da die Logarithmusfunktion jedoch nichtlinear und konkav ist, führt eine naive Rücktransformation zu einem Bias.
Die korrekten, bias-korrigierten Rücktransformationen analytisch zu erklären und praktisch anzuwenden, ist das Ziel dieser Arbeit.

3 Methodik

3.1 Problem von Nicht-Normalität im FH-Kontext

Zunächst betrachten wir das Standard Fay–Herriot (FH) Modell, um zu verstehen, warum Nicht-Normalität problematisch ist. Die Modellformel für den direkten Schätzer lautet:

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

Hierbei ist u_i der gebietsspezifische Zufallseffekt (area random effect), während e_i den Stichprobenfehler des direkten Schätzers beschreibt.
Das zentrale Resultat, das den FH-Schätzer zum EBLUP macht, ist die *lineare bedingte Erwartung unter Normalverteilung* (linear conditional expectation property of the multivariate normal distribution, vgl. Theorem 4.11, Aßmann). Dieses Theorem erlaubt es, die bedingte Erwartung linear zu schreiben als

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

mit dem Shrinkage-Faktor

$$\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \sigma_{ed}^2}.$$

Diese Linearität ist die Grundlage dafür, dass der FH-Schätzer ein EBLUP ist (d. h. aus BLUE: Beste Varianz und Unbiased). Anders als bei OLS, wo β als fester Parameter gilt und keine Normalitätsannahme notwendig ist, hängt der FH-Schätzer von Zufallseffekten ab, sodass Normalität entscheidend für die EBLUP-Eigenschaft ist.
Die Mean-Squared-Error-Zerlegung des FH-Schätzers lautet:

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

- $g_1 = \gamma_d \sigma_{ed}^2$: 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,
- $g_2, g_3 \to 0$ für $D \to \infty$ unter Normalität.

(Quelle: SAE 1 Foliensatz; Rao & Molina, 2015)

Liegt jedoch keine Normalverteilung vor, ist $E[\theta_d \mid y_d]$ nicht mehr linear darstellbar. Damit funktioniert die Zerlegung in g_1, g_2, g_3 nicht mehr und g_2, g_3 verschwinden nicht mehr asymptotisch. Die Folge sind ein systematischer Bias und ein erhöhter RMSE.

3.2 Die Log-Transformation

Um die Normalitätsvoraussetzung des FH-Modells besser zu erfüllen, kann eine Log-Transformation angewandt werden. Dabei werden große Werte stärker gestaucht als kleine, wodurch rechtsschiefe Verteilungen symmetrischer werden. Dies reduziert insbesondere den Einfluss von Ausreißern.
Die Transformation erfolgt auf den direkten Schätzer durch einfache Anwendung der Logarithmusfunktion. Ein Beispiel verdeutlicht den Effekt:

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

Da die Transformation auch die Streuung beeinflusst, muss die Varianz der direkten Schätzer entsprechend angepasst werden. Dies geschieht im `emdi`-Paket automatisch. Die zugrundeliegende Formel lautet:

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

Quelle: Neves et al. (2013)

3.3 Naive Rücktransformation vom Mittelwert

Nach der Modellierung auf der Log-Skala müssen die Schätzer wieder auf die Originalskala zurückgeführt werden. Hierbei entsteht jedoch ein mathematisch-statistisches Problem: Jensen's Inequality.
Sie besagt, dass sich das Ergebnis unterscheidet, je nachdem ob die Transformation vor oder nach der Bildung des Erwartungswerts erfolgt. Explizit wird Jensen's Inequality als Ursache für den Bias bei der Rücktransformation in Kreutzmann (2022) und Würz (2023) genannt. In den Arbeiten von Slud (2005), Harmening (2022) und Rao (2015) wird Jensen's Inequality im Kontext der Log-Transformation jedoch nicht explizit erwähnt.
Der durch Jensen's Inequality entstehende systematische Fehler äußert sich als Bias in der Originalskala. Die Richtung des Bias hängt dabei davon ab, ob die verwendete Funktion konvex oder konkav ist. Im Fall des Logarithmus handelt es sich um eine konkave Funktion.

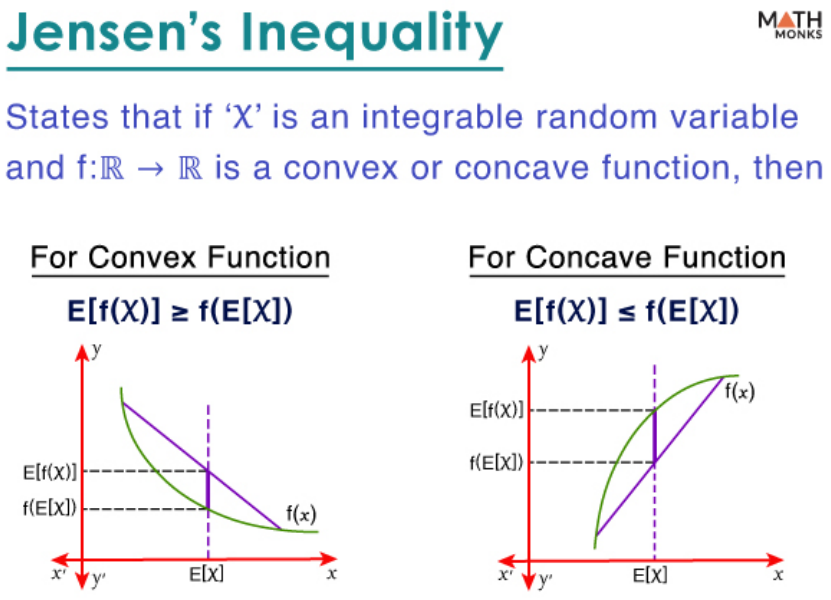
- **Jensen's Inequality im Log-Fall:**

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

- **Allgemein:**

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

(Quelle: Aßmann Vorlesung)



Quelle: Math Monks

Die obige Grafik zeigt den Effekt für konvexe und konkave Funktionen. Die Differenz auf der Y-Achse entspricht genau dem Bias, der durch die naive Rücktransformation entsteht.
Beweisidee: Die folgende Beweis-Skizze stammt aus Aßmanns Vorlesung. Die Kernaussage ist, dass die Tangente im Erwartungswert $E[X]$ die Funktion $g(x)$ bei Konvexität nach unten und bei Konkavität nach oben beschränkt. Daraus ergibt sich die Ungleichung für die Erwartungswerte.

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

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

- Für den Erwartungswert folgt:

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

so that $Eg(X) \geq g(EX)$, as was to be shown (for a continuous X analogous).

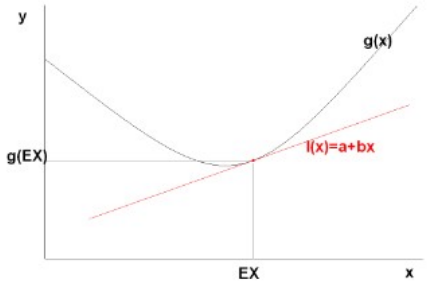


Fig. 10.

(CONTINUES)

Quelle: Aßmann Vorlesung

3.4 emdi Backtransformation Optionen

Im Folgenden widmen wir uns den Möglichkeiten für Rücktransformationen, die in der anschließenden Simulationsstudie verglichen werden. Es handelt sich um:

- raw
- naiv
- bc_crude
- bc_sm

Die ersten beiden Varianten (*raw* und *naiv*) sind bewusst **biased** und nicht im Paket `emdi` implementiert, sondern müssen manuell erstellt werden. Die Methoden bc_crude sowie bc_sm sind hingegen Teil von `emdi`. Dabei bezieht sich *sm* auf die von Slud & Maiti (2005) entwickelte Korrektur. Diese beiden Verfahren stellen bias-korrigierte Rücktransformationen dar.

3.4.1 Raw

Die erste Methode wird hier als *raw* bezeichnet und dient als Vergleichsgröße, um zu beurteilen, welchen Einfluss die Log-Transformation auf die Schätzergebnisse hat. Dabei wird **keine** Transformation angewandt, sondern das Fay–Herriot-Modell direkt auf die nicht-normalverteilten Daten geschätzt:

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

Das Ergebnis ist, wie zu erwarten, biased und soll im Vergleich zeigen, wie stark sich eine Log-Transformation positiv auf die Ergebnisgüte auswirkt.

3.4.2 Naiv

Bei der naiven Rücktransformation wird zunächst eine Logarithmierung der direkten Schätzer durchgeführt:

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

Die Varianz auf der log-Skala berechnet sich zu:

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

Bis hierhin unterscheiden sich die Methoden nicht. Der Unterschied liegt ausschließlich in der Art der Rücktransformation. Hier, im naiven Ansatz, wird die Inverse der Log-Transformation direkt angewandt:

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

Aufgrund von Jensen's Inequality entsteht hier ein systematischer Fehler (Bias), da Erwartungswertbildung und Exponentialfunktion nicht vertauschbar sind.

3.4.3 bc_crude

Die erste bias-korrigierte Rücktransformation ist `bc_crude`, die im Paket `emdi` implementiert ist. In der Literatur gibt es zwei äquivalente Schreibweisen: (1) die praktische Formulierung mit dem MSE und (2) die theoretische Variante mit der wahren Varianz.

$$\hat{\theta}_i^{\text{FH,crude}} = \exp\left(\hat{\theta}_i^{\text{FH,log}} + \tfrac{1}{2}\,\text{MSE}\left(\hat{\theta}_i^{\text{FH,log}}\right)\right), \tag{1}$$

$$\overset{\text{nur bei bekannten Parametern}}{=} \exp\left(\mathbb{E}\left[\hat{\theta}_i^{B}\right] + \tfrac{1}{2}\,\text{Var}\left(\hat{\theta}_i^{B}\right)\right). \tag{2}$$

Die zweite Formulierung ist exakt unbiased, setzt jedoch voraus, dass die wahren Parameter bekannt sind. In der Praxis nutzt man daher den gesamten MSE anstelle der reinen Varianz. Im Gegensatz zur Slud–Maiti-Korrektur basiert `bc_crude` nicht speziell auf der Struktur des FH-Modells, sondern allgemein auf der Eigenschaft des Erwartungswerts einer exponentiell transformierten normalverteilten Zufallsvariablen (Würz, 2023).

Herleitung. Die Eigenschaft lautet: für $X \sim \mathcal{N}(\mu, \sigma^2)$ gilt

$$E\left[e^X\right] = e^{\mu + \frac{1}{2}\sigma^2}.$$

Schreiben wir den Erwartungswert als Integral:

$$E\left[e^X\right] = \int_{-\infty}^{\infty} e^x f_X(x)\, dx,$$

mit der Normaldichte

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Damit ergibt sich, mit herausziehen der konstanten

$$E\big[e^X\big] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\Big(x - \tfrac{(x-\mu)^2}{2\sigma^2}\Big)\, dx.$$

Nun analysieren wir den Exponenten:

$$x - \frac{(x-\mu)^2}{2\sigma^2} = -\frac{1}{2\sigma^2}(x^2 - 2\mu x + \mu^2) + x.$$

Durch Anwenden der Exponentialregel $e^a \cdot e^b = e^{a+b}$ können wir die Terme zusammenfassen. Anschließend ergänzen wir das Quadrat (binomische Formel rückwärts), um wieder die Form einer Normaldichte zu erhalten:

$$= -\frac{(x-(\mu+\sigma^2))^2}{2\sigma^2} + \mu + \tfrac{1}{2}\sigma^2.$$

Damit schreiben wir das Integral um:

$$E\big[e^X\big] = \exp\big(\mu + \tfrac{1}{2}\sigma^2\big) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\Big(-\tfrac{(x-(\mu+\sigma^2))^2}{2\sigma^2}\Big)\, dx.$$

Nun substituieren wir

$$t = \frac{x - (\mu + \sigma^2)}{\sigma}$$

sodass das Integral zur Standardnormalverteilung wird:

$$E\big[e^X\big] = \exp\big(\mu + \tfrac{1}{2}\sigma^2\big) \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2/2}\, dt.$$

Das Integral entspricht 1, sodass schließlich folgt:

$$E\big[e^X\big] = \exp\big(\mu + \tfrac{1}{2}\sigma^2\big).$$

Übertragung auf das FH-Modell. Im Fay–Herriot-Kontext gilt:

$$\hat\theta_d^{FH} \sim \mathcal{N}(\theta_d, \sigma_d^2), \qquad \sigma_d^2 = \mathrm{MSE}(\hat\theta_d^{FH}).$$

Damit ergibt sich:

$$E\Big[\exp(\hat\theta_d^{FH})\Big] = \exp\big(\theta_d + \tfrac{1}{2}\sigma_d^2\big),$$

und die praktische Version der bias-korrigierten Rücktransformation lautet:

$$\hat\theta_d^{FH,\mathrm{crude}} = \exp\Big(\hat\theta_d^{FH} + \tfrac{1}{2}\,\mathrm{MSE}(\hat\theta_d^{FH})\Big).$$

`bc_crude` basiert direkt auf der Lognormal-Eigenschaft und liefert eine einfache, allgemein anwendbare Korrektur. Sie ist im Idealfall exakt unbiased, in der Praxis jedoch nur approximativ, da die MSE anstelle der wahren Varianz eingesetzt wird.

3.4.4 bc_sm

Die Methode `bc_sm` ist eine FH-spezifische Weiterentwicklung der Rücktransformation, die von Slud & Maiti (2005) (daher: `sm`) eingeführt wurde. Während die zuvor beschriebene `bc_crude`-Korrektur nur die Verzerrung durch Jensen's Inequality adressiert, berücksichtigt `bc_sm` zusätzlich die spezielle Struktur des Fay–Herriot-Modells, insbesondere die Varianz der Bereichseffekte u_d. Dies führt zu einer Korrektur, die stärker auf das FH-Modell zugeschnitten ist. Allerdings berücksichtigt diese Methode nicht die zusätzliche Unsicherheit durch die Parameterschätzung, weshalb sie später von Chandra (2011) erweitert wurde.
Die Formel sieht formal ähnlich aus wie bei `bc_crude`, wird jedoch um einen zusätzlichen Korrekturterm erweitert:

$$\hat{\theta}_i^{\mathrm{FH\_sm}} = \exp\Big(E[\hat{\theta}_i^{\mathrm{FH\_log}}] + \tfrac{1}{2}\hat{\sigma}_u^2 \cdot (1 - \hat{\gamma}_i^{\log})\Big).$$

Zur Herleitung betrachten Slud & Maiti (2005) den FH-Schätzer im Detail. Während `bc_crude` die Lognormal-Eigenschaft direkt auf den gesamten EBLUP anwendet, setzen sie spezifisch beim Bereichseffekt u_i an, der in `bc_crude` unberücksichtigt bleibt. Weitere Informationen zur Herleitung stammen aus Chandra (2017)
Wir starten mit der Darstellung

$$\theta_i = x_i^\top \beta + u_i,$$

wobei u_i ein zufälliger Bereichseffekt ist. Für das log-transformierte FH-Modell gilt dann

$$\theta_i^{FH\_log} = \exp(\theta_i) = \exp(x_i^\top \beta + u_i).$$

Der EBLUP für θ_i (ohne Transformation) lautet

$$\hat{\theta}_i = x_i^T \hat{\beta} + \hat{\gamma}_i\big(y_i - x_i^T \hat{\beta}\big),$$

mit dem Shrinkage-Faktor

$$\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \sigma_{ei}^2}.$$

Im Fay–Herriot-Modell gilt

$$y_i = \theta_i + e_i, \qquad e_i \sim \mathcal{N}(0, \sigma_{ei}^2), \quad u_i \sim \mathcal{N}(0, \sigma_u^2).$$

Der zentrale Schritt bei Slud & Maiti ist, den Zusammenhang zwischen dem direkten Schätzer y_i und dem Bereichseffekt u_i zu berücksichtigen. Aufgrund der Eigenschaften der bivariaten Normalverteilung gilt:

$$u_i \mid y_i \sim \mathcal{N}\big(\gamma_i(y_i - x_i^\top \beta),\ \sigma_u^2(1 - \gamma_i)\big).$$

Diese Eigenschaft besagt, dass die gemeinsame Verteilung von y_i und u_i ebenfalls normalverteilt ist, mit entsprechend angepasstem Erwartungswert und Varianz.
Für eine normalverteilte Zufallsvariable $Z \sim \mathcal{N}(\mu, \tau^2)$ gilt die Lognormal-Eigenschaft:

$$\mathbb{E}[e^Z] = \exp\big(\mu + \tfrac{1}{2}\tau^2\big)\,.$$

Wendet man dies auf $x_i^\top \beta + u_i \mid y_i$ an, ergibt sich:

$$\mathbb{E}\Big[\exp(x_i^\top \beta + u_i)\,\Big|\,y_i\Big] = \exp\Big(x_i^\top \beta + \gamma_i(y_i - x_i^\top \beta) + \tfrac{1}{2}\sigma_u^2(1-\gamma_i)\Big)\,.$$

Der zentrale zusätzliche Term ist also:

$$\tfrac{1}{2}\sigma_u^2(1-\gamma_i).$$

Slud & Maiti definieren damit den bias-korrigierten Schätzer:

$$\hat{\theta}_i^{SM} = \exp\Big(x_i^\top \hat{\beta} + \hat{\gamma}_i\,(y_i - x_i^\top \hat{\beta}) + \tfrac{1}{2}\hat{\sigma}_u^2(1-\hat{\gamma}_i)\Big)\,.$$

Diese Methode ist präziser als `bc_crude`, hat aber zwei Einschränkungen:

- Die zusätzliche Unsicherheit aus der Parameterschätzung von β und σ_u^2 wird nicht berücksichtigt (vgl. Chandra, 2011).
- Die Korrektur ist nur für *sampled domains* definiert. Für samplelose Bereiche empfehlen Harmening et al. (2023) daher eine Kombination aus `bc_crude` und `bc_sm`.

4 Simulationsstudie

Um die Methoden zu verdeutlichen, folgen nun eine Simulationsstudie sowie eine Anwendung auf reale Daten, um die bias-korrigierte Rücktransformation in Aktion zu sehen. Der Datengenerierungsprozess (DGP) für die Simulation wurde mir bereitgestellt.
Zunächst zum Aufbau der Simulationsdaten: Der Datensatz umfasst $D = 200$ Domains, die jeweils $R = 500$ Mal (Monte-Carlo) geschätzt wurden. Die Zielgröße ist rechtsschief (wie z. B. Einkommen in einer realen Anwendung). Für jedes Sample wird ein FH-Modell geschätzt, und die resultierenden Schätzer werden mit den wahren Werten θ^{true} verglichen. Dadurch können Bias und RMSE der einzelnen Methoden genau berechnet werden.

- **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: $e_i \sim \mathcal{N}(0,\ \text{vardir}_i)$
  - Direkte Varianz: $\text{vardir}_i \sim \mathcal{U}(0{,}01,\ 0{,}12)$
  - Kovariaten: $x_1,\ x_2 \sim \mathcal{U}(0, 1)$
- **Lineares Modell:**

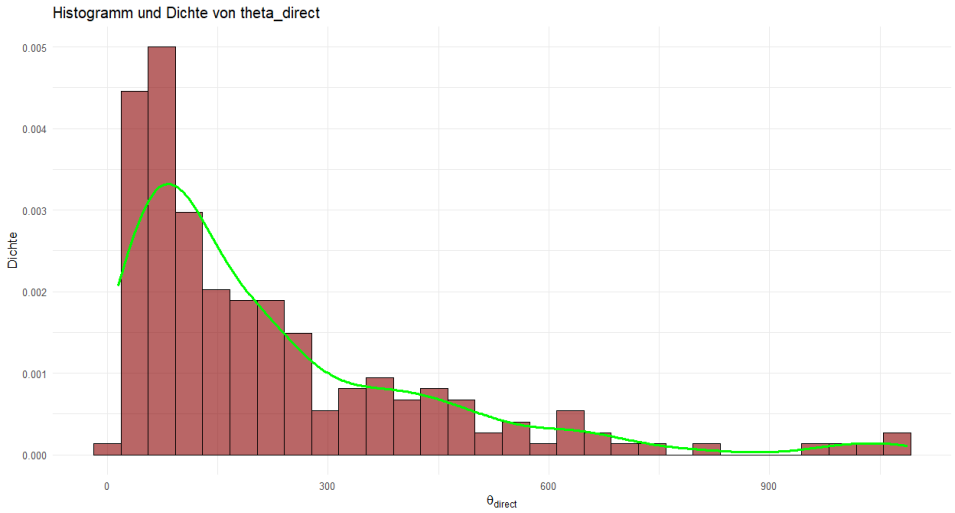
$$y_i = 5 + 2x_{1i} - 2x_{2i} + v_i$$

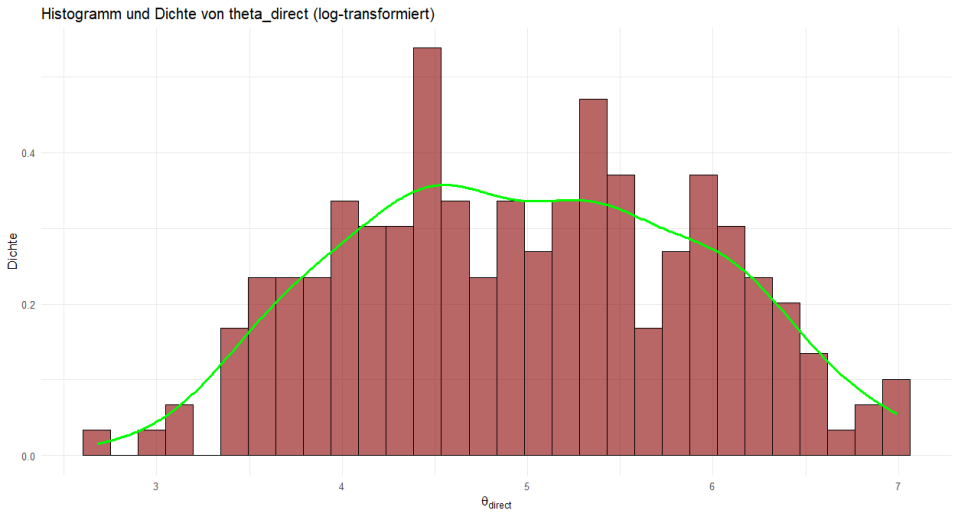
- **Zielgröße (z. B. Einkommen):**

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

- **Varianz nach Log-Transformation:**

$$\text{vardir}^{\text{log}} = \exp(\log(\theta))^2 \cdot \text{vardir}$$





Die obigen Abbildungen zeigen die Verteilung der Zielgröße im Simulationsdatensatz: einmal untransformiert und einmal log-transformiert. Im oberen Plot (untransformiert) erkennt man eine deutliche Nicht-Normalität, während im unteren Plot nach Log-Transformation eine klare Normalisierung sichtbar wird.
Die FH-Modelle wurden mit der `fh()`-Funktion aus dem R-Paket `emdi` geschätzt, mit den folgenden Einstellungen:

- **Standard-FH-Schätzungen:**
  - Transformation: `"log"`, `"no"`
  - Rücktransformation: `"bc_crude"`, `"bc_sm"`, `"no"`
  - Methode: `"ml"`

Da `emdi` keine naive Rücktransformation für Logarithmen erlaubt, musste diese manuell implementiert werden:

- **Naive FH-Schätzungen:**
  - 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 im Modell: `"no"`
  - Rücktransformation im Modell: `"no"`
  - Methode: `"ml"`
  - Manuelle Rücktransformation: $\hat{\theta}_i^{\text{naiv FH}} = \exp(\hat{\theta}_i^{\text{log FH}})$

—

4.1 Metriken

Zum Schätzen von Bias und RMSE habe ich folgende Quality Measures aus der Vorlesung benutzt wobei I(r) für den wahren Populationswert steht.

- **RMSE:**

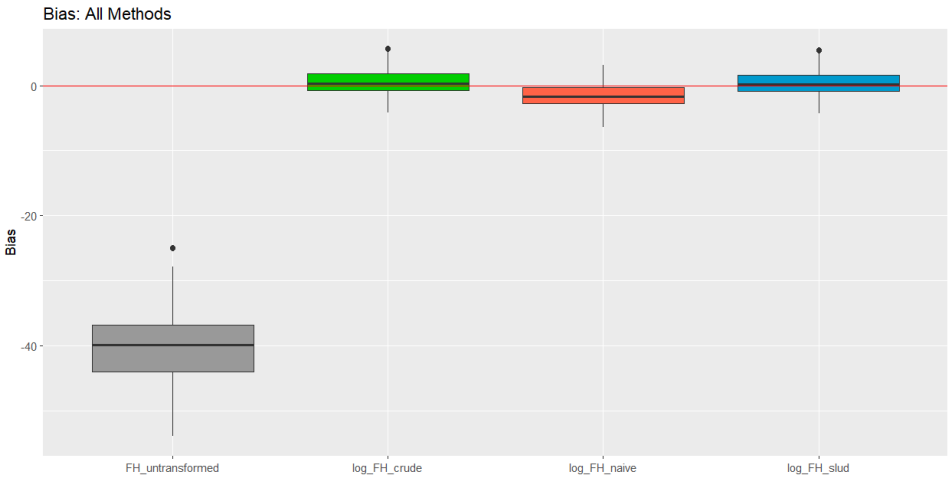
$$\text{RMSE}\left(\hat{\theta}_d^{\text{method}}\right) = \left[\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_d^{\text{method}(r)} - I_d^{(r)}\right)^2\right]^{1/2}$$

- **Bias:**

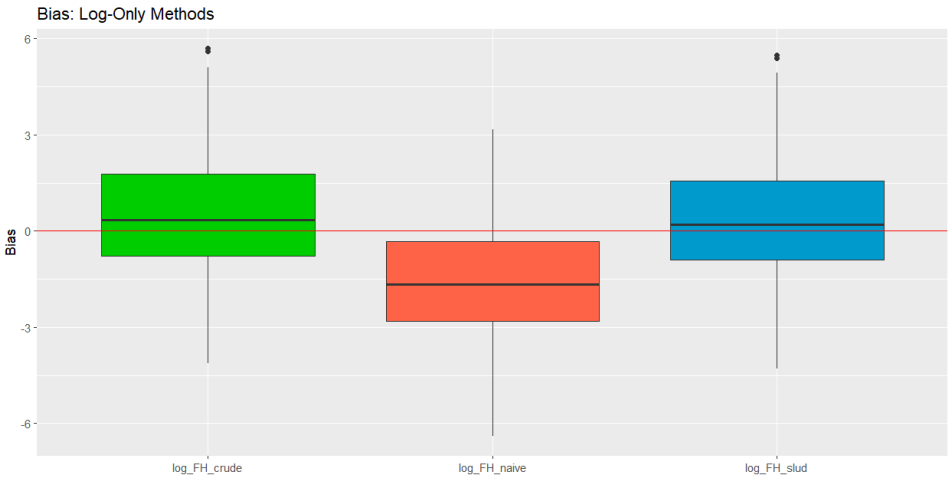
$$\text{Bias}\left(\hat{\theta}_d^{\text{method}}\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_d^{\text{method}(r)} - I_d^{(r)}\right)$$

4.2 Ergebnisse der Simulationsstudie

Zur Übersicht: **Raw** bezeichnet das FH-Modell ohne Transformation, **crude** die analytische bias-korrigierte Rücktransformation, **naiv** die manuelle Rücktransformation ohne Bias-Korrektur, und **bc_sm** die ausführliche Korrektur nach Slud & Maiti.
Wie erwartet zeigt sich: ohne Log-Transformation entsteht ein starker Bias, was die Notwendigkeit der Normalitätsannahme unterstreicht.



Ein detaillierterer Blick (Zoom) verdeutlicht zwei Punkte: (1) Die naive Rücktransformation weist etwas weniger Bias auf als zunächst erwartet. (2) Die `crude`- und `sm`-Korrekturen liefern sehr ähnliche Ergebnisse, wobei `sm` leicht überlegen ist.

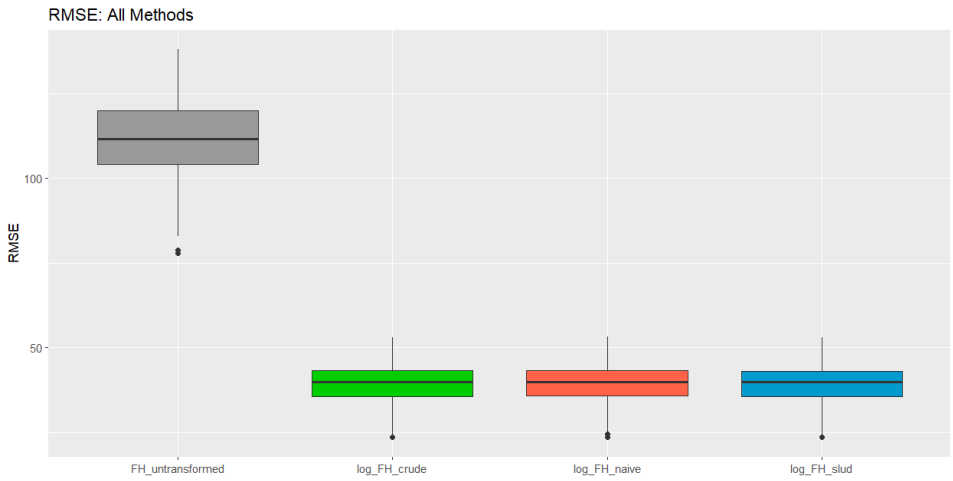


Die folgende Tabelle bestätigt dies nochmals am Mittelwert des Bias (Bias_Mean):

Table 1: RMSE and Bias (Mean and Median)

Method	RMSE_Mean	RMSE_Median	Bias_Mean	Bias_Median
FH_untransformed	111.147	111.542	-40.177	-39.822
log_FH_crude	39.017	39.992	0.479	0.345
log_FH_slud	39.010	39.993	0.308	0.199
log_FH_naive	39.020	39.889	-1.589	-1.674

Zum Abschluss der RMSE. Dieser wird durch Jensen's Inequality nicht direkt beeinflusst, sodass keine zusätzlichen Fehler durch die Rücktransformation entstehen. Dennoch wird der Nutzen der Log-Transformation auch hier sichtbar.



Die Methode `raw` zeigt – wie erwartet – einen deutlich höheren Bias und bestätigt somit die Notwendigkeit der Log-Transformation. Die naive Rücktransformation weist einen systematischen Bias auf. Die beiden bias-korrigierten Methoden `bc_crude` und `bc_sm` liefern sehr ähnliche Ergebnisse, wobei `bc_sm` aufgrund der zusätzlichen Berücksichtigung der Bereichseffekte leicht besser abschneidet.

5 Synthetic Business

5.1 Anwendung Business

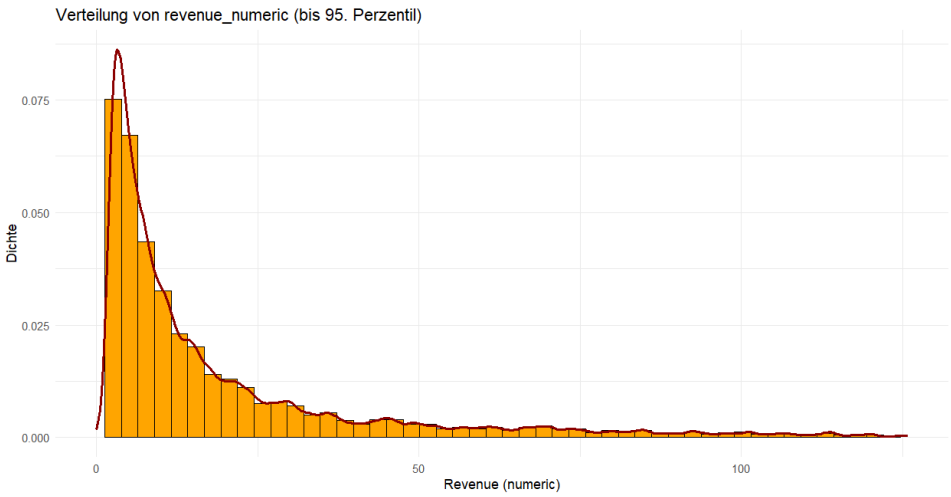
Das zweite Set-Up basiert auf einem Business-Datensatz, der aus der **INC 5000 Liste** der am schnellsten wachsenden privaten US-Unternehmen (2019)

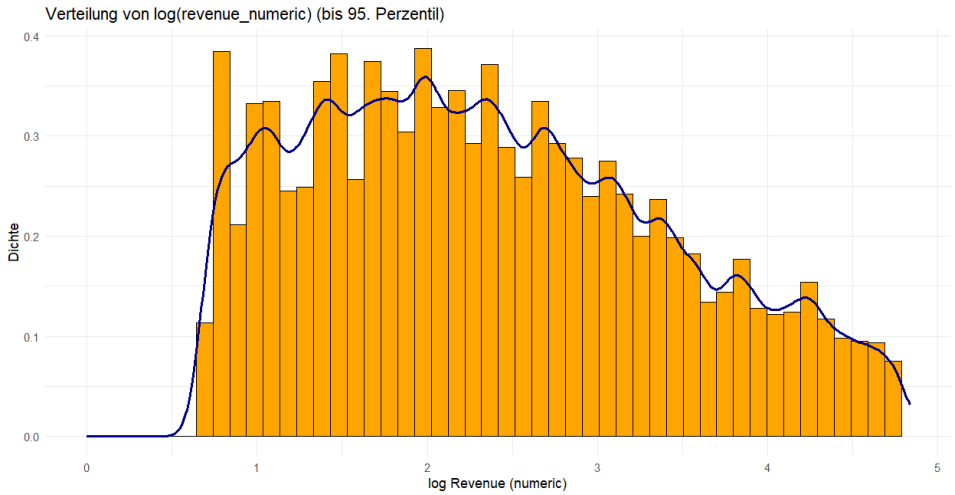
stammt. Auf Basis dieser Liste von 5000 Firmen wurde eine synthetische Population mit insgesamt **500,000 Unternehmen** erstellt.
Die Zielgröße ist der Umsatz (`revenue_numeric`), welcher stark rechtsschief verteilt ist: die meisten Unternehmen haben kleine Umsätze, während einige wenige extrem hohe Umsätze aufweisen.
Die Vorgehensweise entspricht jener der Simulationsstudie: Aus der synthetischen Population werden wiederholt Samples von Größe $n = 1000$ gezogen. Auf jedes dieser Samples wird das FH-Modell mit verschiedenen Rücktransformationsmethoden angewendet. Anschließend vergleichen wir die durchschnittlichen Schätzfehler mit den „wahren“ Werten, die aus den Populationsdaten bekannt sind.
Die "wahren" Vergleichsdaten, also θ^{true} basiert auf direct estimates. Des gesamten 500.000 Unternehmen großen Datensatz.
Eine spezifische Variablenselektion wurde nicht durchgeführt; es wurden alle nicht hoch korrelierten Variablen berücksichtigt.

Zum Datensatz:

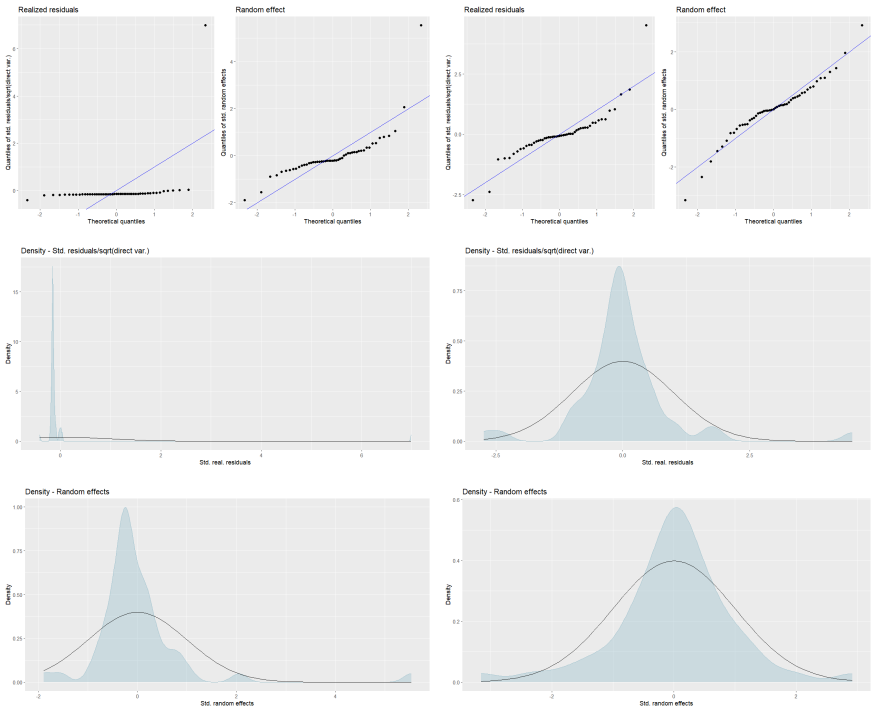
- Basierend auf der **INC 5000 Liste** (2019)
- **Skalierung:** Hochgerechnet auf **500,000 Unternehmen**
- Umfasst Standort, Mitarbeiterzahl, Branche, Gründungsjahr, etc. (14 Variablen pro Firma)
- Zielvariable: `revenue_numeric` (Unternehmensumsatz in Mio. USD)
- Ziel: Schätzung von Umsatz auf Area-Level (Bundesstaaten)

Die Abbildungen unten zeigen die Verteilung der Zielvariablen. In den Rohdaten ist die deutliche Rechtsschiefe sichtbar, während sich unter einer log-Transformation eine deutlich bessere Annäherung an die Normalität ergibt.



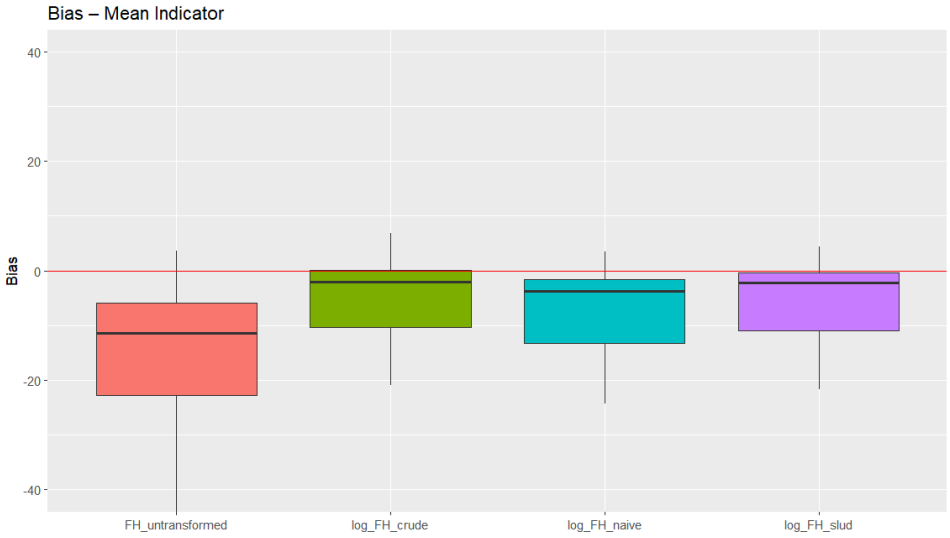


Die folgenden Standard-`emdi`-Diagnostic-Plots wurden auf den Populationsdaten erstellt. Links sind die Ergebnisse ohne Transformation dargestellt, rechts nach einer log-Transformation. Deutlich wird, dass die Normalitätsannahme durch die Transformation wesentlich besser erfüllt ist.

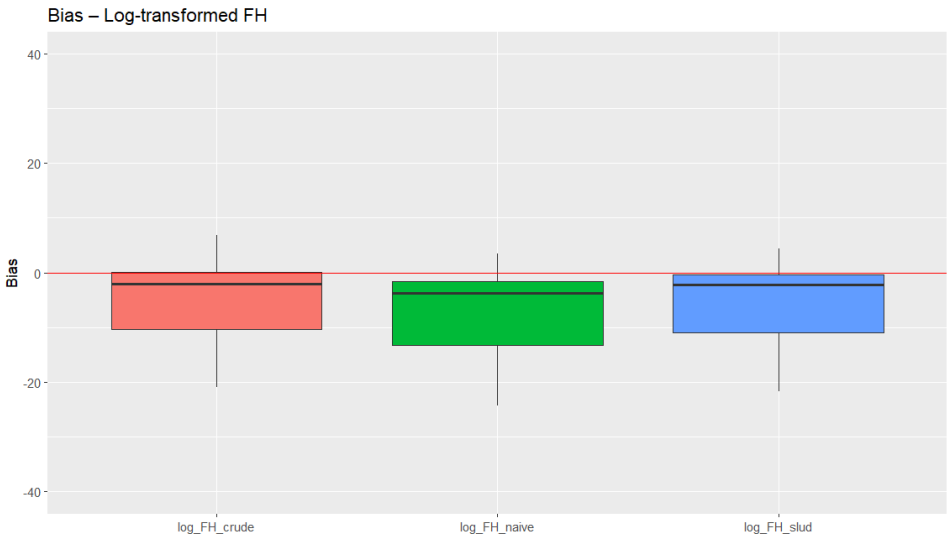


5.2 Ergebnisse Business-Datensatz

Die folgenden Boxplots zeigen die Verteilungen der Schätzungen analog zur Simulationsstudie. Überraschend ist, dass das **untransformierte FH-Modell** deutlich besser abschneidet als in den Simulationsdaten: Es ist zwar weiterhin stärker biased als die transformierten Varianten, aber die Unterschiede sind geringer als erwartet.



Bei genauerer Betrachtung der transformierten Methoden zeigt sich erneut, dass `bc_crude` und `bc_sm` sehr ähnliche Ergebnisse liefern.



Eine Übersicht bietet die folgende Tabelle:

Table 2: RMSE and Bias (Mean and Median)

Method	RMSE_Mean	RMSE_Median	Bias_Mean	Bias_Median
FH_untransformed	29.354	17.834	-23.262	-11.349
log_FH_crude	39.420	19.196	-9.058	-2.000
log_FH_slud	36.610	18.681	-10.196	-2.252
log_FH_naive	34.775	17.653	-12.355	-3.750

Auffällig ist, dass im Business-Datensatz `bc_crude` einen leicht geringeren Bias aufweist als `bc_sm`. In den Simulationsdaten war dies umgekehrt. Zudem sind die absoluten Bias-Werte insgesamt größer, was vermutlich auf die nur näherungsweise log-normale Verteilung der Zielvariablen zurückzuführen ist – ein zusätzlicher Verzerrungsfaktor.
Ein Grund für die Outperformance von `bc_crude` könnte sein, dass im realen Business-Datensatz OOS-Domains (nicht gesampelte Gebiete) vorhanden sind. Dies war in der Simulationsstudie nicht der Fall, und `bc_crude` eignet sich hier besser.

6 Zusammenfassung und Ausblick

Die vorgestellten Methoden machen deutlich, dass Rücktransformationen im log-FH-Modell entscheidend sind, um Verzerrungen zu vermeiden. Naive Rücktransformationen führen durch die Jensen'sche Ungleichung systematisch zu Bias, während `bc_crude` und `bc_sm` diesen Fehler korrigieren können. Wichtig ist dabei auch, dass nicht nur die Punktschätzer, sondern auch die MSE korrekt rücktransformiert werden müssen, was allerdings deutlich komplexer ist.
Die Simulationen haben gezeigt, dass die Log-Transformation zwar in vielen Fällen Verzerrungen reduziert, aber nicht immer ausreichend ist. In der Literatur finden sich deshalb Alternativen, zum Beispiel Empirical Best Prediction (EBP) bei Unit-Level-Modellen mit Box–Cox- und Shift-Transformation (vgl. Molina & Rao, 2015) oder moderne Ansätze wie Mixed Effects Random Forests (MERF).
Für die Wahl der Backtransformation gibt es Unterschiede: Die Methode `bc_crude` ist sinnvoll, wenn Rechenleistung begrenzt ist oder REML zur Varianzschätzung eingesetzt wird, da `bc_sm` nur mit ML funktioniert. Außerdem ist `bc_crude` im Vorteil, wenn viele nicht gesampelte Gebiete (OOS-Domains) vorliegen, da `bc_sm` dort nicht definiert ist (vgl. Harmening et al., 2023). `bc_sm` liefert dagegen im Idealfall die genaueren Ergebnisse, weil es die spezielle Struktur des FH-Modells berücksichtigt (Slud & Maiti, 2005). Eine

Weiterentwicklung ist die Erweiterung von Chandra (2011), die zusätzlich die Unsicherheit aus der Parameterschätzung (β, σ_u^2) einbezieht.
Zusammenfassend gilt: Für große Datensätze mit vielen nicht gesampelten Gebieten ist `bc_crude` die robustere Wahl, während `bc_sm` (oder die Erweiterung nach Chandra (2011) für rein gesampelte Domains die besten Resultate liefert. In der Praxis ist die Entscheidung daher eine Abwägung zwischen Genauigkeit, Rechenaufwand und der Datenstruktur.

7 Appendix

7.1 Zahlenbeispiel Jensens

Theorem: Wenn g konvex
Anhand eines kleinen Zahlenbeispiels lässt sich die Ungleichung gut erklären Ziel ist es den Mittelwert der Konvexen! Funktion x^2 zu berechnen. Wir können die Funktion entweder zuerst anwenden und dann den Mittelwert berechnen oder zuerst den Mittelwert bzw. Erwartungswert (E[X]) berechnen und dann die Funktion anwenden und die Ergebnisse vergleichen.
Beispiel: $g(x) = x^2$, also konvex.

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

- $P(X = 1) = 0{,}5$ und $P(X = 3) = 0{,}5$

Zunächst berechnen wir Anhand gegebener Wahrscheinlichkeiten den Erwartungswert ohne transformation.

- Erwartungswert: $\mathbb{E}[X] = 1 \cdot 0{,}5 + 3 \cdot 0{,}5 = 2$

Im zweiten Schritt berechnen wir den Erwartungswert also Mittelwert gleichzeitig mit der Transformation.

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

Jetzt können wir sehen, dass wenn wir Mittelwert und Transformation gleichzeitig berechnen und das Ergebnis mit der Umgekehrten Variante, vergleichen, das Ergebnis nicht gleich ist. Das ist Jensens inequality. Ebenfalls der Beweis für die Nichtnegativität der Varianz.

8 Quellen

References

[1] Chandra, H., & Chambers, R. (2011). *Small area estimation under transformation to linearity.* Survey Methodology, 37(1), 39–51.

[2] Harmening, S., Kreutzmann, A.-K., Schmidt, S., Salvati, N., & Schmid, T. (2023). A framework for producing small area estimates based on area-level models in R. *The R Journal, 15*(1), 316–341. https://doi.org/10.32614/RJ-2023-039

[3] Kreutzmann, A.-K., Marek, P., Runge, M., Salvati, N., & Schmid, T. (2022). The Fay–Herriot model for multiply imputed data with an application to regional wealth estimation in Germany. *Journal of Applied Statistics, 49*(13), 3278–3299. https://doi.org/10.1080/02664763.2021.1954223

[4] Rao, J. N. K., & Molina, I. (2015). *Small area estimation*. Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/9781118735855

[5] Würz, N., Schmid, T., & Tzavidis, N. (2022). Estimating regional income indicators under transformations and access to limited population auxiliary information. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 185*(2), 584–609. https://doi.org/10.1111/rssa.12913

[6] Aßmann, C. (). *Advanced Statistics – Foliensatz 3*. Universität Bamberg.

[7] Aßmann, C. (). *Formelsammlung Statistik*. Universität Bamberg.

[8] Jensensche Ungleichung. (o. D.). *Mathepedia*. https://mathepedia.de/Jensensche_Ungleichung.html

[9] Jensen's inequality. (o. D.). *Math Monks*. https://mathmonks.com/inequalities/jensens-inequality

[10] Log-normal distribution: Properties and proofs. (o. D.). *StatLect*. https://www.statlect.com/probability-distributions/log-normal-distribution

[11] Harmening, S., Kreutzmann, A.-K., Schmidt, S., Salvati, N., & Schmid, T. (2023). *A framework for producing small area estimates based on area-level models in R*. ResearchGate (Preprint). https://www.researchgate.net/publication/375269760_A_Framework_for_Producing_Small_Area_Estimates_Based_on_Area-Level_Models_in_R

[12] Chandra, H., Aditya, K., & Kumar, S. (2017). Small area estimation under a log transformed area level model. *Journal of Statistical Theory and Practice, 12*(1), 1–19. https://doi.org/10.1080/15598608.2017.1415174

[13] Slud, E., & Maiti, T. (2005). MSE estimation in transformed Fay–Herriot models. *Journal of the Royal Statistical Society: Series B*

(Statistical Methodology), 67(5), 83–98. https://doi.org/10.1111/j.1467-9868.2005.00533.x

[14] Fay, R. E., & Herriot, R. A. (1979). Estimates of income for small places: An application of James–Stein procedures to census data. *Journal of the American Statistical Association, 74*(366), 269–277. https://doi.org/10.2307/2286322

[15] Neves, A., Silva, D., & Correa, S. (2013). Small domain estimation for the Brazilian service sector survey. *ESTADÍSTICA, 65*(185), 13–37.

[16] OpenAI. (2025). *ChatGPT* [Large language model]. Für Codeformatierung, Rechtsschreibung, Satzklarheit, Recherche https://chat.openai.com/