A2. Warum sind die Standardfehler alle gleich?

Warum modellbasierte Mittelwerte denselben Standardfehler teilen

Dr. Paul Schmidt

Um alle in diesem Kapitel verwendeten Pakete zu installieren und zu laden, kann man folgenden Code ausführen:

```
for (pkg in c("tidyverse", "emmeans", "nlme")) {
  if (!require(pkg, character.only = TRUE)) install.packages(pkg)
}

library(tidyverse)
library(emmeans)
library(nlme)
```

Eine Frage, die in Beratung und Lehre immer wieder auftaucht, klingt ungefähr so:

> Ich habe adjustierte Mittelwerte aus meinem linearen Modell berechnet und festgestellt, dass die Standardfehler über alle Gruppen hinweg identisch sind. Ist das ein Fehler?

Es ist kein Fehler. Es ist eine direkte Folge dessen, was ein gewöhnliches lineares Modell über die Daten annimmt. Dieses Kapitel erklärt, woher die identischen Standardfehler kommen, warum sie sich von den gruppenweisen Standardfehlern unterscheiden, die man aus der deskriptiven Statistik erhält, und wie man zu einem Modell übergeht, das eine separate Varianz pro Gruppe schätzt, wenn die Daten dies erfordern.

Glossar: Modellbasierte Mittelwerte

In diesem Kapitel wird der Begriff *adjustierte Mittelwerte* als Kurzform für Mittelwerte verwendet, die aus einem gefitteten linearen Modell geschätzt und nicht direkt aus den Rohbeobachtungen berechnet werden. Je nach Software oder Lehrbuch werden sie auch wie folgt genannt:

- **geschätzte marginale Mittelwerte** (der Name, der vom Paket `{emmeans}` verwendet wird),
- **Least-Squares-Mittelwerte** (historisch `lsmeans` in SAS und älteren R-Paketen),
- **modellbasierte Mittelwerte** oder einfach **vorhergesagte Mittelwerte**.

Alle diese Begriffe bezeichnen dasselbe Konzept: einen Mittelwert, der aus den Koeffizienten des Modells berechnet wird und daher die Annahmen des Modells erbt - einschließlich der Annahme einer gemeinsamen Fehlervarianz.

Die Kurzversion

Ein gewöhnliches lineares Modell nimmt eine einzige, gemeinsame Fehlervarianz für alle Beobachtungen an (Homoskedastizität). Adjustierte Mittelwerte werden aus den

Koeffizienten des Modells abgeleitet, sodass ihre Standardfehler aus dieser einen Varianzschätzung gebildet werden. Bei einem balancierten Design ergibt sich dadurch für jeden Gruppenmittelwert exakt derselbe Standardfehler. Gruppenweise deskriptive Standardfehler unterscheiden sich, weil sie eine separate Varianz pro Gruppe verwenden.

Wenn die identischen Standardfehler falsch erscheinen, lautet die zu stellende Frage nicht “Warum sind sie gleich?”, sondern vielmehr “Rechtfertigen meine Daten die Annahme einer gemeinsamen Varianz?” - und das ist eine Frage der Modelldiagnostik (siehe Anhang A1).

Ein konkretes Beispiel

Mit dem integrierten Datensatz `PlantGrowth` liefern ein einfaktorielles lineares Modell und `emmeans()` einen Standardfehler pro Gruppenmittelwert:

```
mod1 <- lm(weight ~ group, data = PlantGrowth)

emmeans(mod1, specs = ~ group)
```

```
 group emmean    SE df lower.CL upper.CL
 ctrl    5.03 0.197 27     4.63     5.44
 trt1    4.66 0.197 27     4.26     5.07
 trt2    5.53 0.197 27     5.12     5.93

Confidence level used: 0.95
```

Jeder Wert in der `SE`-Spalte ist `0.197`. Vergleicht man dies nun mit derselben Zusammenfassung, die direkt aus den Daten gruppenweise berechnet wird:

```
PlantGrowth %>%
  group_by(group) %>%
  summarise(
    mean = mean(weight),
    stddev = sd(weight),
    n = n(),
    stderr = sd(weight) / sqrt(n())
  )
```

```
# A tibble: 3 × 5
  group  mean stddev     n stderr
  <fct> <dbl>  <dbl> <int>  <dbl>
1 ctrl   5.03  0.583    10  0.184
2 trt1   4.66  0.794    10  0.251
3 trt2   5.53  0.443    10  0.140
```

Hier unterscheiden sich die Standardfehler deutlich zwischen den Gruppen. Beide Tabellen beschreiben dieselben Daten, warum stimmen sie also nicht überein?

Warum sich die deskriptiven Standardfehler unterscheiden

Wenn Mittelwert, Standardabweichung und Standardfehler getrennt pro Gruppe berechnet werden, wird jede Gruppe faktisch als eigene Stichprobe behandelt. Jede Stichprobe hat ihre eigene Varianzschätzung, und der Standardfehler des Stichprobenmittelwerts folgt der Lehrbuchformel

$$\mathrm{SE} = \frac{s}{\sqrt{n}},$$

wobei s die Standardabweichung der Gruppe und n ihre Größe ist. Verschiedene Gruppen erzeugen unterschiedliche s, sodass sich die Standardfehler unterscheiden.

Warum die modellbasierten Standardfehler identisch sind

Ein gewöhnliches lineares Modell nimmt an, dass die Fehlervarianz für jede Beobachtung gleich ist, unabhängig von der Gruppe. Das Fitten von `lm()` liefert eine einzige gepoolte Varianzschätzung $\hat{\sigma}^2$, die aus den Residuen über alle Gruppen hinweg berechnet wird. Jeder adjustierte Gruppenmittelwert ist dann eine Funktion der gefitteten Koeffizienten, und sein Standardfehler wird aus demselben $\hat{\sigma}^2$ gebildet:

$$\mathrm{SE}\ (\bar{y}_g) = \sqrt{\hat{\sigma}^2 / n_g}.$$

In einem balancierten Design, in dem n_g für jede Gruppe gleich ist, ergibt sich ein identischer Standardfehler für alle adjustierten Mittelwerte. Dies ist keine Eigenheit von `emmeans()` - es ist die direkte mathematische Folge der Homoskedastizitätsannahme, die in `lm()` fest verankert ist. Die zugrunde liegende Annahme wird ausführlicher in Anhang A1 - Modelldiagnostik behandelt.

> **i** Zusammenfassung
>
> - **Deskriptive Mittelwerte** verwenden eine separate Varianz pro Gruppe, sodass sich ihre Standardfehler unterscheiden.
> - **Adjustierte Mittelwerte aus** `lm()` teilen eine einzige gepoolte Varianzschätzung, sodass ihre Standardfehler identisch sind (in balancierten Designs).
>
> Ob man unterschiedliche oder identische Standardfehler möchte, hängt davon ab, ob eine gemeinsame Varianz eine sinnvolle Annahme für die eigenen Daten ist.

Welcher Ansatz ist besser?

Es gibt keine allgemeingültige Antwort: Die beiden Ansätze beruhen auf unterschiedlichen Annahmen, und die richtige Wahl hängt davon ab, ob diese Annahmen zu den Daten passen. Eine nützlichere Frage ist, ob die Homoskedastizitätsannahme des gewöhnlichen Modells vertretbar ist. Ist sie es nicht, ist ein lineares Modell, das heterogene Varianzen pro Gruppe zulässt, die natürliche Alternative.

Eine Varianz pro Gruppe zulassen

Die Funktion `gls()` aus `{nlme}` fittet ein Modell der verallgemeinerten Kleinste-Quadrate-Schätzung (GLS), das eine benutzerdefinierte Varianzstruktur über das Argument `weights` akzeptiert. Mit `varIdent(form = ~ 1 | group)` weist man das Modell an, eine separate Residualvarianz für jede Stufe von `group` zu schätzen:

```
mod2 <- gls(weight ~ group,
            weights = varIdent(form = ~ 1 | group),
            data = PlantGrowth)

emmeans(mod2, specs = ~ group)
```

```
 group emmean    SE   df lower.CL upper.CL
 ctrl    5.03 0.184 9.01     4.61     5.45
 trt1    4.66 0.251 9.00     4.09     5.23
 trt2    5.53 0.140 9.00     5.21     5.84

Degrees-of-freedom method: satterthwaite
Confidence level used: 0.95
```

Die adjustierten Mittelwerte aus `mod2` haben nun drei verschiedene Standardfehler - und sie stimmen mit den oben berechneten deskriptiven Werten überein. Das ist zu erwarten: Mit einer separaten Varianz pro Gruppe reduziert sich der modellbasierte SE auf die gruppenweise Formel $s_g/\sqrt{n_g}$.

Die beiden Modelle über AIC vergleichen

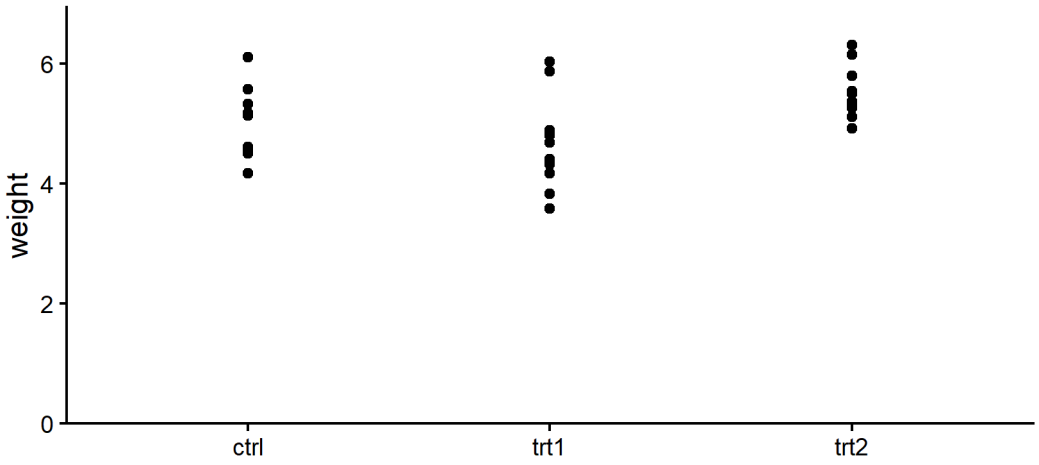
Beide Modelle beschreiben dieselben Daten, treffen aber unterschiedliche Annahmen. Da `mod2` das Modell `mod1` als Spezialfall enthält (die Varianzen könnten zufällig übereinstimmen), kann man sie mit einem Informationskriterium wie AIC vergleichen. Ein niedrigerer AIC zeigt den besseren Kompromiss zwischen Anpassung und Komplexität an:

```
AIC(mod1, mod2)
```

```
Warning in AIC.default(mod1, mod2): Modelle sind nicht alle mit der gleichen
Datensatzgröße angepasst worden
```

```
     df      AIC
mod1  4 61.61904
mod2  6 66.98890
```

Für `PlantGrowth` gewinnt das einfachere homoskedastische Modell. Die zusätzliche Flexibilität, drei Varianzen statt einer zu schätzen, bringt keine ausreichende Verbesserung der Anpassung, um die zusätzlichen Parameter zu rechtfertigen. Ein kurzer Blick auf die Rohdaten macht dies plausibel:



Die Streuung von `weight` innerhalb jeder Gruppe wirkt weitgehend vergleichbar. In einem Szenario, in dem etwa `ctrl` eine sichtbar breitere Streuung als `trt1` und `trt2` zeigen würde, würde AIC vermutlich das heteroskedastische Modell bevorzugen.

> **Wann modellbasierte Mittelwerte vorzuziehen sind**
>
> Über die Varianzfrage hinaus haben adjustierte Mittelwerte einen praktischen Vorteil, den gruppenweise deskriptive Statistiken nicht bieten können: Sie stammen aus einem Modell, das Blockeffekte, Kovariaten oder zufällige Effekte einbeziehen kann. Ein deskriptiver Mittelwert für `ctrl` kennt nur die `ctrl`-Beobachtungen; ein modellbasierter Mittelwert für `ctrl` zieht Information aus dem gesamten Design und adjustiert für die übrige Struktur im Experiment. Für jede Analyse, die über das einfachste einfaktorielle Layout hinausgeht, sind modellbasierte Mittelwerte in der Regel das richtige Ziel der Inferenz.

Weiterführende Literatur

> **Zusätzliche Ressourcen**
>
> - Kozak & Piepho (2019): Analyzing designed experiments: Should we report standard deviations or standard errors of the mean or standard errors of the difference or what? - eine ausführliche Diskussion darüber, welches Unsicherheitsmaß für geplante Experimente berichtet werden sollte, einschließlich der Unterscheidung zwischen SE und SED.
> - Stack Exchange: Standard error in estimated marginal means are all the same - mit einer maßgeblichen Antwort von Russell Lenth, dem Autor von `{emmeans}`.
> - Stack Exchange: Interpreting the standard error from emmeans.
> - IBM: Estimated Marginal Means all have the same standard error in SPSS - dasselbe Phänomen, betrachtet aus der SPSS-Perspektive.
> - Anhang A1 - Modelldiagnostik - wie man prüft, ob die Homoskedastizitätsannahme hinter den identischen SE für die eigenen Daten sinnvoll ist.

Bibliography