Wie funktionieren eigentlich Kontraste? Understanding contrasts.

M.Psy.205, Dozent: Dr. Peter Zezula

Johannes Brachem (johannes.brachem@stud.uni-goettingen.de)

Deutsch

Links

Extrablatt Kontraste als PDF-Datei zum Drucken

Rmd

Dummy-Variablen

Um die Funktionsweise von Kontrasten etwas näher zu beleuchten, müssen wir uns zunächst noch einmal die Arbeit mit Dummy-Variablen vor Augen führen. Was ist die Dummy-Kodierung? Sie ist eine beliebte Schreibweise für Regressionen mit kategorialen Prädiktoren. Ein kategorialer Prädiktor wiederum ist eine Variable, die zum Beispiel die Gruppenzugehörigkeit von Versuchspersonen in einem Experiment anzeigt.

Nehmen wir einmal an, wir hätten ein Experiment zur Evaluation einer Therapiemethode mit drei Gruppen durchgeführt: Eine unbehandelte Kontrollgruppe (KG), eine Experimentalgruppe, in der die VP 10 Therapieeinheiten erhalten haben (Low Dose, LD), und eine Experimentalgruppe, in der die VP 20 Therapieeinheiten erhalten haben (High Dose, HD).

Die Gruppenzugehörigkeit haben wir in der Variable `group` gespeichert. Da diese Variable aber keine Zahlen enthält, sondern inhaltliche Information, können wir `group` nicht einfach als Prädiktor nehmen. Stattdessen setzen wir zwei Dummy-Variablen ein, D_1 und D_2. Die Regressionsgleichung lautet in diesem Fall

$$\hat{y}_i = b_0 + b_1 \cdot D_1 + b_2 \cdot D_2,$$

wobei $\hat{y}_i$ unser durch das Modell geschätzter Wert in der abhängigen Variable für Versuchspersin i ist.

Die Dummy-Variablen nehmen immer bestimmte Werte an, je nachdem in welcher Gruppe eine betrachtete VP ist. Zum Beispiel könnte man sagen, D_1 soll immer den Wert 1 annehmen, wenn eine VP in der LD-Gruppe ist, ansonsten den Wert 0. Welche Werte die Dummy-Variablen in welcher Situation annehmen, ist unserer Entscheidung überlassen. Diese Entscheidung bestimmt, welche Bedeutung die Regressionskoeffizienten b_0, b_1 und b_2 haben und welche Hypothesen wir durch die so gesetzten Kontraste testen können.

Referenzkodierung

Zunächst schauen wir uns die Referenzkodierung genauer an, eine nützliche und oft verwendete Kodierung, die recht gut verständlich ist. Diese Tabelle zeigt, welche Werte die Dummy-Variablen in welchen Fällen annimmt:

Dummy-Variable	Wenn `group` = KG	Wenn `group` = LD	Wenn `group` = HD
D_1	0	1	0
D_2	0	0	1

Das sind dieselben Werte, die wir in R eingeben, wenn wir die Referenzkodierung benutzen möchten:

```
contrast1 <- c(0, 1, 0)
contrast2 <- c(0, 0, 1)

contrasts(group) <- cbind(contrast1, contrast2)
```

Die Verwendung dieser Werte führt dazu, dass b_1 aus der Gleichung oben der Unterschied zwischen der Kontrollgruppe und der Gruppe LD ist, während b_2 der Unterschied zwischen der Kontrollgruppe und der Gruppe HD ist. Warum ist das so?

Einsetzen in die Regressionsgleichung

Wir können nun Regressionsgleichungen für alle drei Gruppen aufstellen.

Kontrollgruppe Die Ursprüngliche Gleichung lautet:

$$\hat{y}_i = b_0 + b_1 \cdot D_1 + b_2 \cdot D_2$$

Nun setzen wir für D_1 und D_2 die Werte aus der Spalte für `group` = KG aus der Tabelle oben ein.

$$\hat{y}_i^{group=KG} = b_0 + b_1 \cdot 0 + b_2 \cdot 0$$

Da b_1 und b_2 mit 0 multipliziert werden, können wir sie weglassen.

$$\hat{y}_i^{group=KG} = b_0$$

Es bleibt nur noch b_0 übrig. Wir wissen, dass die Schätzung $\hat{y}_i$ bei Gruppenvergleichenden Experimenten immer der jeweilige Gruppenmittelwert ist. Das heißt, b_0 ist der Mittelwert unserer Versuchspersonen in der Kontrollgruppe.

LD-Gruppe Die Ursprüngliche Gleichung lautet:

$$\hat{y}_i = b_0 + b_1 \cdot D_1 + b_2 \cdot D_2$$

Nun setzen wir für D_1 und D_2 die Werte aus der Spalte für `group` = LD aus der Tabelle oben ein.

$$\hat{y}_i^{group=LD} = b_0 + b_1 \cdot 1 + b_2 \cdot 0$$

Da b_2 mit 0 multipliziert wird, können wir es weglassen. Auch die 1, mit der b_1 multipliziert wird, können wir weglassen.

$$\hat{y}_i^{group=LD} = b_0 + b_1$$

Hier wird also der Mittelwert in der LD-Gruppe dadurch geschätzt, dass zum Mittelwert der Kontrollgruppe (b_0) noch b_1 addiert wird. Das heißt, b_1 ist der Unterschied zwischen dem Mittelwert der KG und dem Mittelwert der LD. Wenn wir also für diesen Koeffizienten im Output der Regression einen signifikanten t-Test finden, dann heißt das, dass sich KG und LD-Gruppe signifikant unterscheiden (b_1 ist signifikant verschieden von 0).

HD-Gruppe Die Ursprüngliche Gleichung lautet:

$$\hat{y}_i = b_0 + b_1 \cdot D_1 + b_2 \cdot D_2$$

Nun setzen wir für D_1 und D_2 die Werte aus der Spalte für `group` = HD aus der Tabelle oben ein.

$$\hat{y}_i^{group=HD} = b_0 + b_1 \cdot 0 + b_2 \cdot 1$$

Da b_1 mit 0 multipliziert wird, können wir es weglassen. Auch die 1, mit der b_2 multipliziert wird, können wir weglassen.

$$\hat{y}_i^{group=HD} = b_0 + b_2$$

Hier wird also der Mittelwert in der HD-Gruppe dadurch geschätzt, dass zum Mittelwert der Kontrollgruppe (b_0) noch b_2 addiert wird. Das heißt, b_2 ist der Unterschied zwischen dem Mittelwert der KG und dem Mittelwert der HD. Wenn wir also für diesen Koeffizienten im Output der Regression einen signifikanten t-Test finden, dann heißt das, dass sich KG und HD-Gruppe signifikant unterscheiden (b_2 ist signifikant verschieden von 0).

Übersicht

Koeffizient	Bedeutung	Kontrast	Test
b_0	Mittelwert in der Kontrollgruppe	Kein Kontrast	Kein Test
b_1	Unterschied zw. Mittelwert in der Kontrollgruppe und Mittelwert in der LD-Gruppe	Kontrast 1	KG vs. LD
b_2	Unterschied zw. Mittelwert in der Kontrollgruppe und Mittelwert in der HD-Gruppe	Kontrast 2	KG vs. HD

Orthogonale Kontraste

Sehen wir uns nun orthogonale Kontraste näher an. Orthogonale Kontraste sind dadurch gekennzeichnet, dass die Werte für die Dummy-Variablen besonders sorgfältig gewählt werden, so dass die t-Tests für die einzelnen Regressionskoeffizienten nicht der Gefahr von Alpha-Fehler-Inflation unterliegen.In Kapitel 10.4.2 beschreibt Field (2012) im Detail, was orthogonale Kontraste ausmacht.

Wir sehen uns hier einen häufigen Fall von orthogonalen Kontrasten näher an. Dazu verwenden wir folgende Kodierung:

Dummy-Variable	Wenn `group` = KG	Wenn `group` = LD	Wenn `group` = HD
D_1	-2	1	1
D_2	0	-1	1

Das sind dieselben Werte, die wir in R eingeben, wenn wir hier orthognale Kontraste benutzen möchten:

```
contrast1 <- c(-2, 1, 1)
contrast2 <- c(0, -1, 1)

contrasts(group) <- cbind(contrast1, contrast2)
```

Einsetzen in die Regressionsgleichung

Wir können nun wieder Regressionsgleichungen für alle drei Gruppen aufstellen.

Kontrollgruppe Die Ursprüngliche Gleichung lautet:

$$\hat{y}_i = b_0 + b_1 \cdot D_1 + b_2 \cdot D_2$$

Nun setzen wir für D_1 und D_2 die Werte aus der Spalte für `group` = KG aus der Tabelle oben ein.

$$\hat{y}_i^{group=KG} = b_0 + b_1 \cdot -2 + b_2 \cdot 0$$

Da b_2 mit 0 multipliziert wird, können wir es weglassen. Wir schreiben auch die Multiplikation von b_1 mit -2 noch etwas schöner auf.

$$\hat{y}_i^{group=KG} = b_0 - 2b_1$$

Wir sehen nun leider: Die Koeffizienten sind nicht mehr so einfach zu interpretieren, wie bei der Referenzkodierung. Deshalb machen wir erst einmal weiter. Was wir aber hier schon sehen können: b_2 hat mit der Schätzung für die Kontrollgruppe nichts zu tun.

LD-Gruppe Die Ursprüngliche Gleichung lautet:

$$\hat{y}_i = b_0 + b_1 \cdot D_1 + b_2 \cdot D_2$$

Nun setzen wir für D_1 und D_2 die Werte aus der Spalte für `group` = LD aus der Tabelle oben ein.

$$\hat{y}_i^{group=LD} = b_0 + b_1 \cdot 1 + b_2 \cdot -1$$

Wir schreiben die Multiplikationen noch einmal um, um die Lesbarkeit zu verbessern.

$$\hat{y}_i^{group=LD} = b_0 + b_1 - b_2$$

Auch hier können wir die Bedeutung der Koeffizienten nicht mehr einfach ablesen. Machen wir erst einmal weiter.

HD-Gruppe Die Ursprüngliche Gleichung lautet:

$$\hat{y}_i = b_0 + b_1 \cdot D_1 + b_2 \cdot D_2$$

Nun setzen wir für D_1 und D_2 die Werte aus der Spalte für `group` = HD aus der Tabelle oben ein.

$$\hat{y}_i^{group=HD} = b_0 + b_1 \cdot 1 + b_2 \cdot 1$$

Wir schreiben die Multiplikationen noch einmal um, um die Lesbarkeit zu verbessern.

$$\hat{y}_i^{group=HD} = b_0 + b_1 + b_2$$

Was wir hier sehen können ist, dass b_2, also der zweite Kontrast, etwas mit dem Unterschied zwischen der LD- und der HD-Gruppe zu tun hat, da b_0 und b_1 in beiden Gruppen gleich bleiben. In der LD-Gruppe wird b_2 abgezogen, in der HD-Gruppe wird b_2 addiert. $2 \cdot b_2$ scheint daher der Unterschied zwischen der LD- und der HD-Gruppe zu sein, d.h. b_2 ist die Hälfte dieses Unterschieds.

Warum ist es ok, dass b_2 nur die Hälfte des Unterschieds zwischen der LD- und der HD-Gruppe ist? Durch diesen Umstand wird praktisch das Alpha-Fehler-Niveau kontrolliert: Der Koeffizient ist kleiner, d.h. er wird weniger einfach signifikant. In diesem Fall ist das genau das, was wir brauchen.

Bedeutung von b_1 Stellen wir noch einmal alle drei Gleichungen direkt untereinander.

$$\hat{y}_i^{group=KG} = b_0 - 2b_1$$

$$\hat{y}_i^{group=LD} = b_0 + b_1 - b_2$$

$$\hat{y}_i^{group=HD} = b_0 + b_1 + b_2$$

Hier können wir eindeutig sehen, dass der Unterschied zwischen beiden gemeinsamen Experimentalgruppen und der Kontrollgruppe im Koeffizienten b_1 steckt. Durch geschicktes Umstellen der oberen Gleichungen kann man tatsächlich zeigen, dass $b1$ exakt $\frac{1}{3}$ des Unterschieds zwischen dem Mittelwert der Kontrollgruppe und dem gemeinsamen Mittelwert der Experimentalgruppen ist.

Warum ist das ok? Wie oben bei b_2, sorgt auch hier der kleinere Koeffizient dafür, dass die Alpha-Fehler-Rate kontrolliert wird.

Übersicht

Um die Regressionskoeffizienten in diesem Fall interpretieren zu können, muss man etwas tiefer in die Mathematik der Berechnung einsteigen. Field (2012) tut das auf den Seiten 497 - 500 in anschaulicher Art und Weise. Hier geben wir Ihnen nur die Schlussfolgerungen mit auf den Weg.

Koeffizient	Bedeutung	Kontrast	Test
b_0	Gesamtmittelwert	Kein Kontrast	Kein Test
b_1	$\frac{1}{3}$ des Unterschieds zw. Mittelwert in der Kontrollgruppe und Mittelwert in beiden Experimentalgruppen gemeinsam	Kontrast 1	KG vs. EG
b_2	$\frac{1}{2}$ des Unterschieds zw. Mittelwert in der LD-Gruppe und Mittelwert in der HD-Gruppe	Kontrast 2	LD vs. HD

Orthogonale und nicht-orthogonale Kontraste

Ein Schlusswort zu orthogonalen und nicht-orthogonalen Kontrasten von Field (2012), S. 502:

> There is nothing intrinsically wrong with performing non-orthogonal contrasts. However, if you choose to perform this type of contrast you must be very careful about how you interpret the results. With non-orthogonal contrasts, the comparisons you do are related and so the resulting test statistics and p-values will be correlated to some extent. For this reason you should use a more conservative probability level to accept that a given contrast is statistically meaningful (see section 10.5).

Literatur

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. London: SAGE Publications Ltd.

English

Links

Extra sheet Contrasts in PDF format

Rmd

Dummy variables

First thing to understand contrasts is, what dummy variables are good for. So what is dummy coding? Basically it is a common way to define regressions that have categorial predictors. A categorial predictor is a variable, that codes f. e. group membership of subjects in an experiment.

Let's assume, we have made an experiment to evaluate three methods of therapy in three groups: KG: a control group that has not received any treatment, LD: an experimental group in which the subjects had 10 units of a therapy (low dose) HD: an experimental group in which the subjects had 20 therapy units (high dose).

We store group membership of our subjects in a variable called `group`. In this variable we do not store numbers but content information, we cannot simply use `group` as a predictor. Therefore we use two dummy variables, D_1 and D_2. In this case our regression equation would be:

$$\hat{y}_i = b_0 + b_1 \cdot D_1 + b_2 \cdot D_2,$$

where $\hat{y}_i$ is our value of the dependent variable for subject i, estimated by our model.

It's the combination of both dummy variables that tells us group membership for each subject. F. e. we could say, D_1 should always have a value of 1, if a subject is in group LD, and a value of 0, if not. We can decide, which value of our dummy variable codes what. But this decision defines the meaning and interpretation of our regression coefficients b_0, b_1 and b_2. Moreover this decides about the hypotheses, we test with these contrasts.

Reference coding

We now take a look to reference coding, a useful and common type of coding that is easy to understand. The table below shows the values of the dummy variables and what they code:

Dummy variable	`group` = KG	`group` = LD	`group` = HD
D_1	0	1	0
D_2	0	0	1

These are exactly the same values, we would use, to define reference coding.

```
contrast1 <- c(0, 1, 0)
contrast2 <- c(0, 0, 1)

contrasts(group) <- cbind(contrast1, contrast2)
```

These dummy values lead to the fact, that b_1 in our equation above codes the difference between the control group KG and the low dose group LD, whereas b_2 would refer to the difference between control group KG and high does group HD. But why?

Insert into rhe regression equation

We have regression equations for all three groups.

Control group The original equation is:

$$\hat{y}_i = b_0 + b_1 \cdot D_1 + b_2 \cdot D_2$$

Now we put the values of D_1 and D_2 from the column `group` = KG of the table above into the equation.

$$\hat{y}_i^{group=KG} = b_0 + b_1 \cdot 0 + b_2 \cdot 0$$

As b_1 and b_2 are multiplied by 0, they disappear

$$\hat{y}_i^{group=KG} = b_0$$

So only b_0 remains. We know, that we always estimate group means when using categorial predictors. So b_0 is the estimated mean of the observations in our control group.

Group LD The original equation is:

$$\hat{y}_i = b_0 + b_1 \cdot D_1 + b_2 \cdot D_2$$

We now put the values of column `group` = LD for D_1 and D_2 in our equation.

$$\hat{y}_i^{group=LD} = b_0 + b_1 \cdot 1 + b_2 \cdot 0$$

b_2 is multiplied by 0 and disappears. We can also leave out the 1, with which b_1 is multiplied. So we result in:

$$\hat{y}_i^{group=LD} = b_0 + b_1$$

Here we calculate the mean of group LD by adding b_1 to the mean of the control group (b_0). This means, that b_1 is the difference between the mean of the control group and the mean of the group LD. If we find a significant t-test for this coefficient in the output of our regression, we can conclude, that groups KG and LD differ significantly (b_1 is significantly different from 0).

Group HD Our original equation is:

$$\hat{y}_i = b_0 + b_1 \cdot D_1 + b_2 \cdot D_2$$

Now we insert the values of column `group` = HD from our table above for D_1 and D_2.

$$\hat{y}_i^{group=HD} = b_0 + b_1 \cdot 0 + b_2 \cdot 1$$

For the multiplication with 0 we b_1 disappears. We can also leave out the 1, with which b_2 is multiplied. We get:

$$\hat{y}_i^{group=HD} = b_0 + b_2$$

Here we estimate the mean of group HD by adding b_2 to the mean of the control group KG (b_0). Therefore b_2 is the difference between the mean of group KG and the mean of group HD. If we find a significant t-test for this coefficient in our regression output, the groups KG and HD differ significantly (b_2 is significantly different from 0).

Overview

coefficient	meaning	contrast	test
b_0	mean of control group KG	no contrast	no test
b_1	difference between mean of control group and mean of group LD	contrast 1	KG vs. LD
b_2	difference between mean of control group and mean of group HD	contrast 2	KG vs. HD

Orthogonal contrasts

Let's take a closer look at orthogonal contrasts. They are characterized by a specially selected way to set dummy variables, so that the t-tests of the regression coefficients do not inflate. Field (2012) chapter 10.4.2 describes in detail what are orthogonal contrasts.

We look at a very common type of orthogonal contrasts. To do that we use the following coding:

dummy variable	if `group` = KG	if `group` = LD	if `group` = HD
D_1	-2	1	1
D_2	0	-1	1

These are the same values, that we would enter in R to define orthogonal contrasts.

```
contrast1 <- c(-2, 1, 1)
contrast2 <- c(0, -1, 1)

contrasts(group) <- cbind(contrast1, contrast2)
```

Insert in the regression equation

Again we could note down the regression equation for all three groups.

Control group The original equation is:

$$\hat{y}_i = b_0 + b_1 \cdot D_1 + b_2 \cdot D_2$$

We now insert the values of column `group` = KG from the table above for D_1 and D_2.

$$\hat{y}_i^{group=KG} = b_0 + b_1 \cdot -2 + b_2 \cdot 0$$

For the multiplication with 0 b_2 disappears. We can write the multiplication of b_1 more beautiful.

$$\hat{y}_i^{group=KG} = b_0 - 2b_1$$

Now we see: The interpretation of the coefficients isn't as easy any more as with reference coding. But let's keep on for now. We can already see, that b_2 isn't the estimation of the control groups mean.

Group LD The original equation is:

$$\hat{y}_i = b_0 + b_1 \cdot D_1 + b_2 \cdot D_2$$

We now insert the values of column `group` = LD from the table above for D_1 and D_2.

$$\hat{y}_i^{group=LD} = b_0 + b_1 \cdot 1 + b_2 \cdot -1$$

Again we change the writing of our multiplication to increase readability.

$$\hat{y}_i^{group=LD} = b_0 + b_1 - b_2$$

Here again, we cannot easyly identify the meaning of our coefficients. Again, let's keep on for now.

Group HD The original equation is:

$$\hat{y}_i = b_0 + b_1 \cdot D_1 + b_2 \cdot D_2$$

We now insert the values of column `group` = HD from the table above for D_1 and D_2.

$$\hat{y}_i^{group=HD} = b_0 + b_1 \cdot 1 + b_2 \cdot 1$$

Again we change the writing of our multiplication to increase readability.

$$\hat{y}_i^{group=HD} = b_0 + b_1 + b_2$$

Now we can see, that b_2, i. e. the second contrast has something to do with the difference between group LD and group HD, as b_0 and b_1 are equal in both groups. In group LD b_2 is subtracted, in group HD b_2 is added. $2 \cdot b_2$ seems to be the difference between group LD and group HD or b_2 is half of this difference.

Why is it ok, that b_2 is only half of the difference between group LD and group HD? By this the level of alpha error is controlled: The coefficient is smaller i. e. it reaches significance less easy. In this cas this is, what we need.

How to interprete b_1 Let's compare all three equations directly.

$$\hat{y}_i^{group=KG} = b_0 - 2b_1$$

$$\hat{y}_i^{group=LD} = b_0 + b_1 - b_2$$

$$\hat{y}_i^{group=HD} = b_0 + b_1 + b_2$$

We can see clearly, that the difference between the two experimental groups and the control group is somehow hidden in b_1. When we rearrange the above equations, we can see, that b_1 is exactly $\frac{1}{3}$ of the difference between the mean of the control group and the mean of both experimental groups.

Why is this o.k.? Just like above with b_2, the smaller coefficient takes car of our total alpha error.

Overview

For a better interpretation of this case, we have to dive a bit deeper into the mathematics behind regression coefficients. Field (2012) does this at pages 497 - 497 in a very understandable way. We give you here only the conclusions, we draw.

coefficient	meaning	contrast	test
b_0	grand mean	no contrast	no test
b_1	$\frac{1}{3}$ of the difference between the mean of the control group and the common mean of the two experimenta, groups	contrast 1	KG vs. EG
b_2	$\frac{1}{2}$ of the difference of the mean of group LD and the mean of group HD	contrast 2	LD vs. HD

Orthogonal and non orthogonal contrasts

A final word of Field (2012, p. 502) about orhtogonal and non orthogonal contrasts:

> There is nothing intrinsically wrong with performing non-orthogonal contrasts. However, if you choose to perform this type of contrast you must be very careful about how you interpret the results. With non-orthogonal contrasts, the comparisons you do are related and so the resulting test statistics and p-values will be correlated to some extent. For this reason you should use a more conservative probability level to accept that a given contrast is statistically meaningful (see section 10.5).

Literature

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. London: SAGE Publications Ltd.

Version: 20 Mai, 2021 08:59