Statistik 2 – Tutorate
Sitzung 9: ANOVA

Fynn Siefart, Lea Elina Hofer, Samuel Rauh, Sebastian Senn,
Marco Giesselmann

Lernziele dieser Sitzung

Regression mit kategorialen unabhängigen Variablen

1. Dichotome UV
2. Multikategoriale UV

1. Dichotome UV

Geschlecht → Lebenszufriedenheit

Operationalisierung durch: **gndr** und **stflife** aus dem ESS8 (Gesamtdatensatz)

```
ess8 <- read_dta("ESS8e02_2.dta")
ess8 <- select(ess8, idno, stflife, gndr)
```

→ Variableninspektion mit **look_for()** bzw. **attributes()…**

1. Dichotome UV

```
look_for(ess8)
stflife  How satisfied with life
 [0] Extremely dissatisfied
 [1] 1
 [2] 2
 [3] 3
 [4] 4
 [5] 5
 [6] 6
 [7] 7
 [8] 8
 [9] 9
 [10] Extremely satisfied
 [NA(b)] Refusal
 [NA(c)] Don't know
 [NA(d)] No answer
gndr     Gender
 [1] Male
 [2] Female
 [NA(d)] No answer
attributes(ess8$gndr)
 $label
 [1] "Gender"

 $format.stata
 [1] "%12.0g"

 $class
 [1] "haven_labelled" "vctrs_vctr"     "double"

 $labels
     Male    Female No answer
        1         2        NA
```

Frage:
Welchem Variablentyp gehören die beiden Variablen an?

Antwort:
AV Lebenszufriedenheit: metrisch
UV Geschlecht: kategorial *(allerdings als numerische Variable im Datensatz)*

Standardverfahren zur Analyse dieser Variablenkombination*:* ***Mittelwertvergleich***

```
t.test(formula = stflife ~ gndr, var.equal = TRUE, data = ess8)

        Two Sample t-test

data:  stflife by gndr
t = 2.6563, df = 44190, p-value = 0.007902
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01383775 0.09173981
sample estimates:
  mean in group Male mean in group Female
            7.182473             7.129685
```

Interpretation:
Männer sind im Mittel geringfügig (0.05 Skalenpunkte) Lebenszufriedener als Frauen. Dieser Mittelwertunterschied ist statistisch signifikant von 0 verschieden (p=0.0079).

1. Dichotome UV

```
look_for(ess8)
stflife  How satisfied with life
 [0] Extremely dissatisfied
 [1] 1
 [2] 2
 [3] 3
 [4] 4
 [5] 5
 [6] 6
 [7] 7
 [8] 8
 [9] 9
 [10] Extremely satisfied
 [NA(b)] Refusal
 [NA(c)] Don't know
 [NA(d)] No answer
gndr      Gender
 [1] Male
 [2] Female
 [NA(d)] No answer
```

```
attributes(ess8$gndr)
$label
[1] "Gender"

$format.stata
[1] "%12.0g"

$class
[1] "haven_labelled" "vctrs_vctr"     "double"

$labels
     Male    Female No answer
        1         2        NA
```

Frage:
Welchem Variablentyp gehören die beiden Variablen an?

Antwort:
AV Lebenszufriedenheit: metrisch
UV Geschlecht: kategorial *(allerdings als numerische Variable im Datensatz)*

Standardverfahren zur Analyse dieser Variablenkombination*: **Mittelwertvergleich***

Aber: Mittelwertvergleich ist auch im Rahmen der Regressionsanalyse umsetzbar

Wie? – Ganz einfach:
- (Variableninspektion, ggf. Rekodierung – siehe auch HP)
- Umwandlung der UV in Faktorvariable mit **as_factor()**
- Stichprobenstatistik erstellen mit **table1()**
- Anschliessende Integration der faktorisierten Variable in den **lm()**-Befehl

```
ess8$gndr <- as_factor(ess8$gndr)
```

1. Dichotome UV@Reg: Variableninspektion

```
summary(ess8)
```

```
      idno              stflife              gndr
 Min.   :       1   Min.   : 0.000   Male      :21027
 1st Qu.:    1208   1st Qu.: 6.000   Female    :23351
 Median :    2589   Median : 8.000   No answer:     9
 Mean   : 31545782   Mean   : 7.155
 3rd Qu.:    11058   3rd Qu.: 9.000
 Max.   :551603139   Max.   :10.000
                     NA's   :187
```

1. Dichotome UV@Reg: Stichprobenstatistik und PostFaktor-Bereinigung

```
table1::table1(~ gndr + stflife, data = ess8)
```

	Overall (N=44387)
Gender	
Male	21027 (47.4%)
Female	23351 (52.6%)
No answer	9 (0.0%)
How satisfied with life as a whole	
Mean (SD)	7.15 (2.09)
Median [Min, Max]	8.00 [0, 10.0]
Missing	187 (0.4%)

Post-Faktorisierungs Bereinigung: Missings rekodieren

Problem: Fehlende Werte der faktorisierten Variablen **gndr** sind nicht (mehr) als NA codiert.

```
ess8$gndr <- na_if(ess8$gndr, "No answer")
```

1. Dichotome UV@Reg: Stichprobenstatistik und PostFaktor-Bereinigung

```
table1::table1(~ gndr + stflife, data = ess8)
```

	Overall (N=44387)
Gender	
Male	21027 (47.4%)
Female	23351 (52.6%)
No answer	0 (0.0%)
Missing	9 (0.0%)
How satisfied with life as a whole	
Mean (SD)	7.15 (2.09)
Median [Min, Max]	8.00 [0, 10.0]
Missing	187 (0.4%)

Post-Faktorisierungs Bereinigung: Phantomkategorien entfernen

Problem: Die vormaligen Kategorien für fehlende Werte sind als Kategorienrelikte weiterhin in den kategorialen Variablen angelegt

```
ess8$gndr <- fct_drop(ess8$gndr)
```

1. Dichotome UV@Reg: Stichprobenharmonisierung

```
table1::table1(~ gndr + stflife, data = ess8)
```

	Overall (N=44387)
Gender	
Male	21027 (47.4%)
Female	23351 (52.6%)
Missing	9 (0.0%)
How satisfied with life as a whole	
Mean (SD)	7.15 (2.09)
Median [Min, Max]	8.00 [0, 10.0]
Missing	187 (0.4%)

Stichprobenharmonisieren zur Konstanthaltung der Fallzahl:

```
ess8_noNA <- na.omit(ess8)
```

1. Dichotome UV@Reg

bivariates Regressionsmodell…

```
bi_model <- lm(stflife ~ gndr,
               data = ess8_noNA)
summary(bi_model)

Call:
lm(formula = stflife ~ gndr, data = ess8_noNA)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.18247    0.01441 498.379   <2e-16 ***
gndrFemale  -0.05279    0.01987  -2.656   0.0079 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

t-test zum Vergleich…

```
t.test(formula = stflife ~ gndr, var.equal = TRUE, data = ess8_noNA)

        Two Sample t-test

data:  stflife by gndr
t = 2.6563, df = 44190, p-value = 0.007902
alternative hypothesis:
true difference in means is not equal to 0
95 percent confidence interval:
 0.01383775 0.09173981
sample estimates:
  mean in group Male mean in group Female
            7.182473              7.129685
```

An welchen Stellen spiegeln **t-test** und bivariates **Regressionsmodell** die identischen Ergebnisse wieder?

1. Dichotome UV@Reg: Referenzmodifikation

bivariates Regressionsmodell…

```
bi_model <- lm(stflife ~ gndr,
               data = ess8_noNA)
summary(bi_model)

Call:
lm(formula = stflife ~ gndr, data = ess8_noNA)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.18247    0.01441 498.379   <2e-16 ***
gndrFemale  -0.05279    0.01987  -2.656   0.0079 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

Referenzkategorie: Die Basiskategorie des Vergleichs, aus deren Perspektive der Koeffizient interpretiert werden muss

Default:
Referenzkategorie = Kategorie mit niedrigerer ursprünglicher numerischer Codierung
oder mit niedrigerem Initial (**hier: Männer**)

Der Koeffizient der faktorisierten Geschlechtervariable zeigt an, wie sich die mittlere Lebenszufriedenheit der Frauen von der LZ der Referenzkategorie unterscheidet: **Frauen sind im Mittel etwa 0.05 Skalenpunkte weniger Lebenszufrieden als Männer**.

1. Dichotome UV@Reg: Referenzmodifikation

bivariates Regressionsmodell…

```
bi_model <- lm(stflife ~ gndr,
               data = ess8_noNA)
summary(bi_model)

Call:
lm(formula = stflife ~ gndr, data = ess8_noNA)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.18247    0.01441 498.379   <2e-16 ***
gndrFemale  -0.05279    0.01987  -2.656   0.0079 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

Frage: Was macht folgender Code?

```
ess8_noNA$gndr <- relevel(ess8_noNA$gndr, ref = "Female")
bi_model <- lm(stflife ~ gndr, data = ess8_noNA); summary(bi_model)

Call:
lm(formula = stflife ~ gndr, data = ess8_noNA)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.12968    0.01368 521.052   <2e-16 ***
gndrMale     0.05279    0.01987   2.656   0.0079 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.086 on 44190 degrees of freedom
Multiple R-squared:  0.0001597, Adjusted R-squared:  0.000137
F-statistic: 7.056 on 1 and 44190 DF,  p-value: 0.007902
```


Antwort: Der **relevel()** Befehl tausche die Referenzkategorie aus.

Der im Koeffizient vermittelte **Mittelwertunterschied** bezieht sich **neu auf die Referenzkategorie „Frauen“,** bleibt identisch hat nun aber ein positives Vorzeichen

1. Dichotome UV@Reg: Regressionstabelle

… ordentlich formatiert über Stargazer und per Hand nachgearbeitet...

```
stargazer(bi_model, type = "html",
          dep.var.caption = "",
          dep.var.labels = "",
          column.labels = c("Bivariates Modell"),
          single.row = TRUE,
          omit.stat = c("f", "ser", "adj.rsq"),
          digits = 2, digits.extra = 5,
          star.cutoffs = c(.05, .01, .001),
          notes = "Daten: ESS(2016),
          Standardfehler in Klammern",
          title = "Der Effekt vom Geschlecht
          auf die Lebenszufriedenheit",
          out = "bi_model.doc")
```

Lineares Modell: Geschlecht und Lebenszufriedenheit

	Bivariates Modell
Geschlecht	
Männlich	Referenz
Weiblich	-0.05** (0.02)
Constant	7.18*** (0.01)
Observations	44,192
R^2	0.0002
Note:	*p<0.05 **p<0.01 ***p<0.001
	Daten: ESS(2016), Standardfehler in Klammern

Interpretation:
Die Lebenszufriedenheit von Frauen liegt im Schnitt 0.05 Skalenpunkte unter der von Männern. Dieser Mittelwertunterschied ist signifikant von null verschieden, aus statistischer Perspektive wird die Hypothese also gestützt.

1. Dichotome UV@Reg: Interpretation

… ordentlich formatiert über Stargazer und per Hand nachgearbeitet...

```
stargazer(bi_model, type = "html",
          dep.var.caption = "",
          dep.var.labels = "",
          column.labels = c("Bivariates Modell"),
          single.row = TRUE,
          omit.stat = c("f", "ser", "adj.rsq"),
          digits = 2, digits.extra = 5,
          star.cutoffs = c(.05, .01, .001),
          notes = "Daten: ESS(2016),
          Standardfehler in Klammern",
          title = "Der Effekt vom Geschlecht
          auf die Lebenszufriedenheit",
          out = "bi_model.doc")
```

Lineares Modell: Geschlecht und Lebenszufriedenheit

	Bivariates Modell
Geschlecht	
Männlich	Referenz
Weiblich	-0.05** (0.02)
Constant	7.18*** (0.01)
Observations	44,192
R^2	0.0002
Note:	*p<0.05 **p<0.01 ***p<0.001
	Daten: ESS(2016), Standardfehler in Klammern

Interpretation:
Die Lebenszufriedenheit von Frauen liegt im Schnitt 0.05 Skalenpunkte unter der von Männern.
Frage: Ist dies ein grosser oder eher kleiner Unterschied?
Für diese Beurteilung werfen wir einen Blick zurück in die Stichprobenstatistik ...

1. Dichotome UV@Reg: Interpretation

	Overall (N=44387)
Gender	
Male	21027 (47.4%)
Female	23351 (52.6%)
Missing	9 (0.0%)
How satisfied with life as a whole	
Mean (SD)	7.15 (2.09)
Median [Min, Max]	8.00 [0, 10.0]
Missing	187 (0.4%)

Lineares Modell: Geschlecht und Lebenszufriedenheit

	Bivariates Modell
Geschlecht	
Männlich	Referenz
Weiblich	-0.05** (0.02)
Constant	7.18*** (0.01)
Observations	44,192
R^2	0.0002
Note:	*p<0.05 **p<0.01 ***p<0.001
	Daten: ESS(2016), Standardfehler in Klammern

Interpretation:
Die Lebenszufriedenheit von Frauen liegt im Schnitt 0.05 Skalenpunkte unter der von Männern.
Frage: Ist dies ein grosser oder eher kleiner Unterschied?
Für diese Beurteilung werfen wir einen Blick zurück in die Stichprobenstatistik ...

1. Dichotome UV@Reg: Interpretation

	Overall (N=44387)
Gender	
Male	21027 (47.4%)
Female	23351 (52.6%)
Missing	9 (0.0%)
How satisfied with life as a whole	
Mean (SD)	7.15 (2.09)
Median [Min, Max]	8.00 [0, 10.0]
Missing	187 (0.4%)

Lineares Modell: Geschlecht und Lebenszufriedenheit

	Bivariates Modell
Geschlecht	
Männlich	Referenz
Weiblich	-0.05** (0.02)
Constant	7.18*** (0.01)
Observations	44,192
R^2	0.0002
Note:	*p<0.05 **p<0.01 ***p<0.001
	Daten: ESS(2016), Standardfehler in Klammern

Interpretation:
Der mittlere geschlechtsspezifische Unterschied in der Lebenszufriedenheit entspricht nur etwa dem Vierzigstel eines typischen Unterschieds zwischen zwei zufälligen Personen. Der Mittelwertunterschied in der Lebenszufriedenheit zwischen den Geschlechtern ist also eher klein.

Abwägung t-Test/Regressionsanalyse

Der **t-Test** ist in gewissen Anwendungsbereichen mit starkem experimentellen Zugang (Psychologie, Humanmedizin) stark vertreten. In der Soziologie eher selten.

Das **zwei Gründe**:

1. Der t-Test erlaubt keine Integration von **Drittvariablen**, die Regression aber schon
2. Der t-Test ist nur bei UVs mit genau zwei Ausprägungen anwendbar, die Regression auch auch mit **multikategorialen UV**

→ Siehe folgende Analyse

2. Multikategoriale UV@Reg

Arbeitsverhältnis → Lebenszufriedenheit

ESS: «emprel»
- 1 = Angestellt
- 2 = Selbstständig
- 3 = Arbeit im Familienbetrieb

ESS: «stflife»:
- 0 bis 10
- 0 → sehr **unzufrieden**
- 10 → sehr **zufrieden**

Hypothese:
Arbeit im Familienbetrieb ist die zufriedenheitsstiftendeste aller Erwerbsformen.

Wir ***releveln*** die Arbeitsverhältnisvariable so, dass die Ausprägung „Arbeit im Familienbetrieb“ die Referenzkategorie ist.
Frage: Warum ist das sinnvoll?
Antwort: Alle Koeffizienten bilden dann Mittelwertunterschiede zu den Familienbetrieblern ab, wodurch die Logik der Hypothese direkt aufgegriffen wird.
Zusätzlicher Hinweis: Wenn die Hypothesenformulierung die Referenzkategorie nicht eindeutig vorgibt, nehmt bei ordinalen UV die niedrigste und bei nominalen UV die modale Kategorie

2. Multikategoriale UV@Reg

Arbeitsverhältnis → Lebenszufriedenheit

ESS: «emprel»
- 1 = Angestellt
- 2 = Selbstständig
- 3 = Arbeit im Familienbetrieb

ESS: «stflife»:
- 0 bis 10
- 0 → sehr **unzufrieden**
- 10 → sehr **zufrieden**

Hypothese:
Arbeit im Familienbetrieb ist die zufriedenheitsstiftendeste aller Erwerbsformen.

Datenmanagement auf der Homepage!

(Rekodieren, Faktorisieren, Relevel, Postfaktorbereinigung, Harmonisierung…)

(Achtung: Vorgehen weichen leicht von der Präsi ab, daher minimal unterschiedliche Ergebnisse)

2 Multikategoriale UV@Reg

Bivariates Regressionsmodell:

Wiederum werden die Regressionskoeffizienten als **Mittelwertdifferenzen zur Referenzkategorie** (hier: *Arbeit im Familienbetrieb*) ausgewiesen.

```
Call:
lm(formula = stflife ~ emplrel, data = ess8_noNA)

Residuals:
    Min      1Q  Median      3Q     Max
-7.4679 -1.1163  0.6411  1.6411  2.8837

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          7.46794    0.07694  97.063  < 2e-16 ***
emplrelEmployee     -0.35163    0.07775  -4.523 6.13e-06 ***
emplrelSelf-employed -0.10901    0.08266  -1.319    0.187
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.083 on 40081 degrees of freedom
Multiple R-squared:  0.001841,  Adjusted R-squared:  0.001791
F-statistic: 36.96 on 2 and 40081 DF,  p-value: < 2.2e-16
```

Interpretation:
Sowohl **Angestellte** als auch **selbstständig** arbeitende Personen weisen, entsprechend der von uns formulierten Hypothese, eine **geringere Lebenszufriedenheit** auf als **Arbeitnehmende in Familienbetrieben**.

Frage: Gibt es Störmerkmale, die kontrolliert werden sollten?

2 Multikategoriale UV@Reg

Multivariates Regressionsmodell:

Wiederum werden die Regressionskoeffizienten als Mittelwertdifferenzen zur Referenzkategorie (hier: *Arbeit im Familienbetrieb*) ausgewiesen.

…die nun aber um den Einfluss der Alters- und Bildungsvariablen bereinigt sind!

```
Call:
lm(formula = stflife ~ emplrel + agea + eduyrs, data = ess8_noNA)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3706 -1.1160  0.4046  1.4083  3.9862

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.7423204  0.0922310  73.103  < 2e-16 ***
emplrelEmployee    -0.4212887  0.0769528  -5.475 4.41e-08 ***
emplrelSelf-employed -0.1824897  0.0818481  -2.230   0.0258 *
agea               -0.0035316  0.0006071  -5.818 6.02e-09 ***
eduyrs              0.0736050  0.0027852  26.427  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Modellvergleich und Interpretation auf Basis des **stargazer()** – Outputs…

2 Multikategoriale UV@Reg

```
stargazer(bi_model, multi_model, type = "html",
          dep.var.caption = "",
          dep.var.labels = "",
          column.labels = c("Bivariates Modell",
                            "Multivariates Modell"),
          single.row = TRUE,
          omit.stat = c("f", "ser", "adj.rsq"),
          digits = 2, digits.extra = 5,
          star.cutoffs = c(.05, .01, .001),
          notes = "Daten: ESS(2016),
          Standardfehler in Klammern",
          title = "Der Effekt vom Anstellungsverhältnis
          auf die Lebenszufriedenheit",
          out = "bi_mult_model_2.doc")
```

Formatiert in MS Word

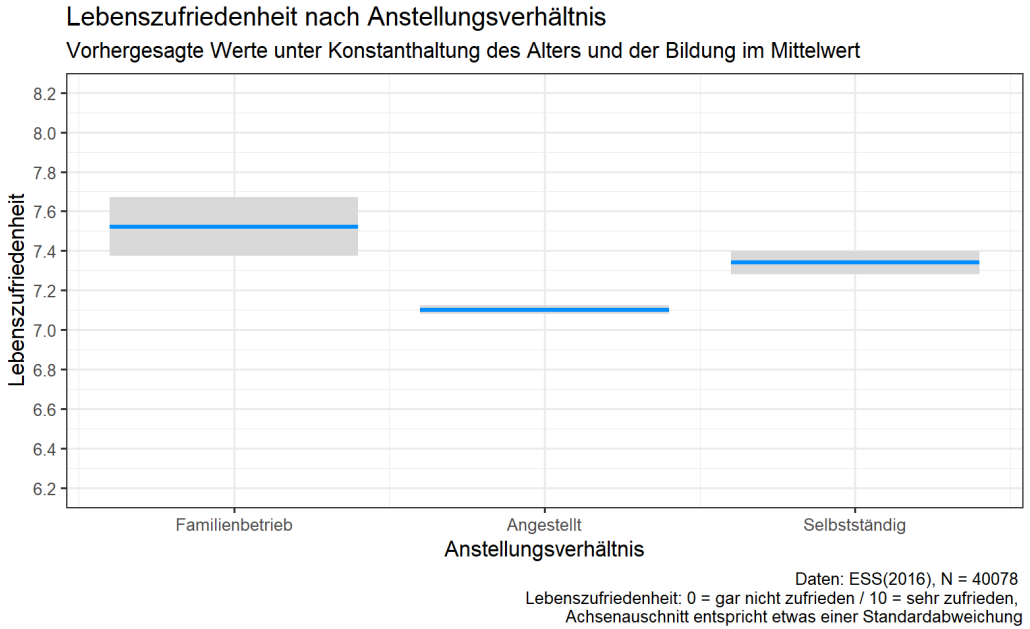
Der Effekt vom Anstellungsverhältnis auf die Lebenszufriedenheit

	Bivariates Modell	Multivariates Modell
Anstellungsverhältnis		
Familienbetrieb	Referenz	
Angestellt	-0.35*** (0.08)	-0.42*** (0.08)
Selbstständig	-0.11 (0.08)	-0.18* (0.08)
Alter in Jahre		-0.004*** (0.001)
Bildung in Jahre		0.07*** (0.003)
Konstante	7.47*** (0.08)	6.74*** (0.09)
n	40,084	40,084
R^2	0.002	0.02
Note:		*p<0.05 **p<0.01 ***p<0.001
		Daten: ESS(2016), Standardfehler in Klammern

Interpretation:

- Sowohl Angestellte als auch Selbständige sind unter Konstanthaltung der Drittmerkmale im Mittel statistisch signifikant weniger lebenszufrieden als Familienbetriebler
- Insb. der bereinigte Unterschied zwischen Angestellten und Familienbetrieblern ist inhaltlich substanziell
- Koeffizient «Selbständig» im unbereinigten Modell transportiert offenbar positiven Bildungseffekt und unterschätzt daher den negativen Einfluss von Selbständigkeit auf die Lebenszufriedenheit

2 Multikategoriale UV@Reg



Der Effekt vom Anstellungsverhältnis auf die Lebenszufriedenheit

	Bivariates Modell	Multivariates Modell
Anstellungsverhältnis		
Familienbetrieb		Referenz
Angestellt	-0.35*** (0.08)	-0.42*** (0.08)
Selbstständig	-0.11 (0.08)	-0.18* (0.08)
Alter in Jahre		-0.004*** (0.001)
Bildung in Jahre		0.07*** (0.003)
Konstante	7.47*** (0.08)	6.74*** (0.09)
n	40,084	40,084
R^2	0.002	0.02
Note:		*p<0.05 **p<0.01 ***p<0.001
		Daten: ESS(2016), Standardfehler in Klammern

Interpretation:

- Sowohl Angestellte als auch Selbständige sind unter Konstanthaltung der Drittmerkmale im Mittel statistisch signifikant weniger lebenszufrieden als Familienbetriebler
- Insb. der bereinigte Unterschied zwischen Angestellten und Familienbetrieblern ist inhaltlich substanziell
- Koeffizient «Selbständig» im unbereinigten Modell transportiert offenbar positiven Bildungseffekt und unterschätzt daher den negativen Einfluss von Selbständigkeit auf die Lebenszufriedenheit

Anmerkung zur Übung

- Auf der Tutoratswebseite findet ihr eine Übung zu dieser Einheit.

  Wir werden diese zwar nicht mehr im Plenum besprechen aber auf OLAT wurde ein Forum-Strang eingerichtet auf dem ihr Fragen oder Anmerkungen zu der Übung vermerken könnt und auf diese wir dann dort antworten können.