

Analyse zum Datensatz “Vogel”

Martin Hofbauer, Friedrich Leisch

Dieser Datensatz wurde im Zuge der Lehrveranstaltung “Vertiefung in statistische Methoden der Wildtierforschung” an der Universität für Bodenkultur Wien im Wintersemester 2018/19 von den Studierenden Hannah Ergott, Maya Hasenfuß und Theresa Schmid zur Analyse eingereicht. Der vorliegende Datensatz beschreibt die Anzahl an Beobachtungen verschiedener Vogelarten getrennt nach Wald- und Feldvögeln und die simultan gemessene Anzahl an Graslandschmetterlingen. Um Vergleichbarkeit zu gewährleisten wurden die Anzahlen skaliert, dadurch ist es möglich, dass nichtganzzahlige Anzahlen auftreten.

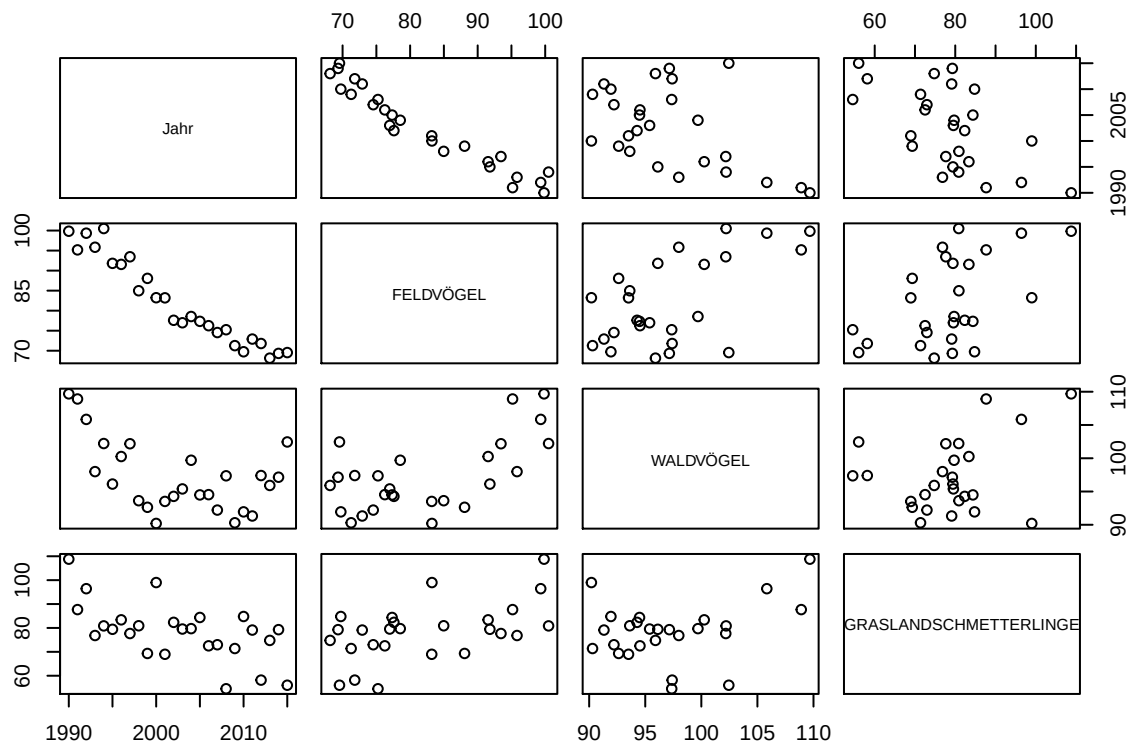
Zunächst lesen wir den vorliegenden Datensatz als csv-file ein. Mit dem hier angeführten Befehl funktioniert das nur, wenn sich die Datei direkt im aktuellen Arbeitsverzeichnis (“Working Directory”) befindet. Zum Festlegen des Arbeitsverzeichnisses kann man in R den Befehl `setwd()` verwenden. Da sich im csv-file Umlaute finden, setzen wir die Option `encoding` auf “UTF-8”, damit deutsche Umlaute in den Variablennamen korrekt dargestellt werden.

```
R> vogel <- read.csv(file="vogel.csv", encoding="UTF-8")
R> summary(vogel)
```

	Jahr	FELDVÖGEL	WALDVÖGEL	GRASLANDSCHMETTERLINGE
Min.	:1990	Min. : 68.15	Min. : 90.20	Min. : 54.54
1st Qu.	:1996	1st Qu.: 73.32	1st Qu.: 93.54	1st Qu.: 72.62
Median	:2002	Median : 78.07	Median : 96.02	Median : 79.36
Mean	:2002	Mean : 82.16	Mean : 97.21	Mean : 78.41
3rd Qu.	:2009	3rd Qu.: 91.75	3rd Qu.:100.12	3rd Qu.: 83.13
Max.	:2015	Max. :100.50	Max. :109.68	Max. :108.79

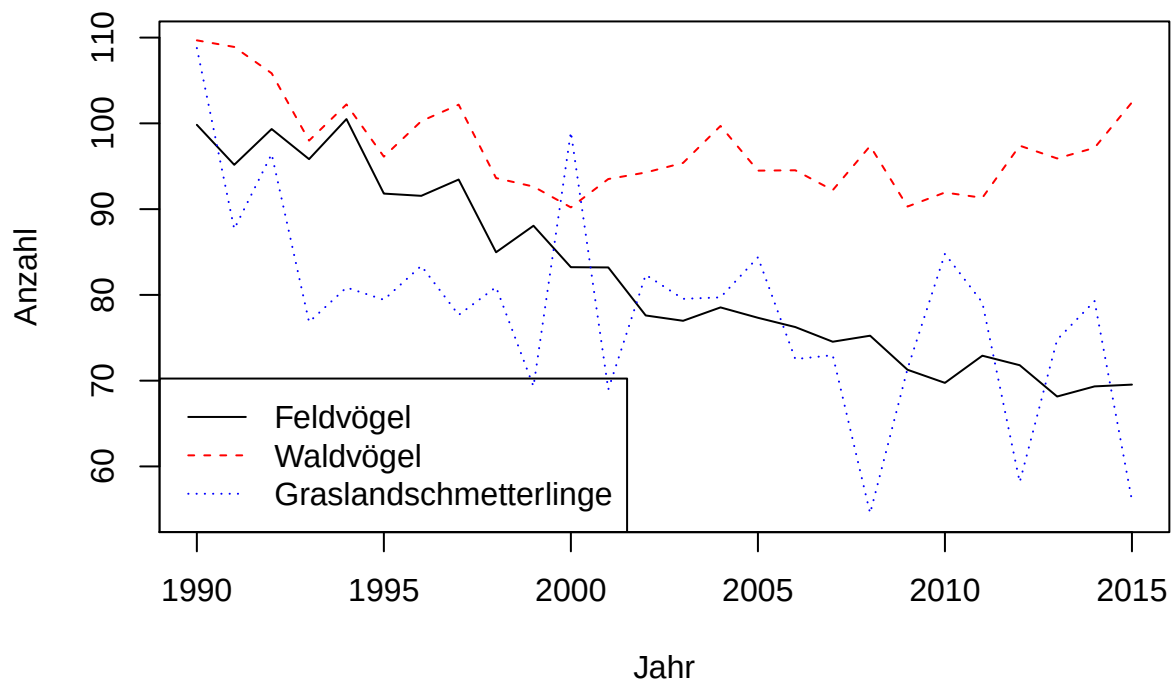
Wie aus der Zusammenfassung ersichtlich ist, handelt es sich um Beobachtungen von Graslandschmetterlingen, Waldvögeln und Feldvögeln aus den Jahren 1990 bis 2015. An der 5-Zahlen-Zusammenfassung fällt bereits auf, dass die Stichprobe der Waldvögel offensichtlich eine geringere Spannweite als die der anderen beiden Gruppen besitzt. Um aus dieser Tatsache einen Zusammenhang zum zeitlichen Verlauf herzustellen, betrachten wir zunächst die Scatterplots der Gruppen.

```
R> pairs(vogel)
```



Im Plot zwischen Jahreszahl und Feldvögeln lässt sich bereits eine deutliche Tendenz zu einer Abnahme der Feldvogelpopulation im zeitlichen Verlauf feststellen. Einen gut vergleichbaren Blick auf die zeitlichen Verläufe der Populationen erhalten wir, indem wir alle Verläufe in einem Diagramm abhängig vom zeitlichen Verlauf darstellen. Dies ist mit vielen Funktionen möglich, wir wählen hier die Funktion `matplot`.

```
R> matplot(x=vogel$Jahr, y=vogel[,2:4], type=c("l","l","l"), col=c(1,2,4),
+         xlab="Jahr", ylab="Anzahl")
R> legend(x="bottomleft", c("Feldvögel", "Waldvögel", "Graslandschmetterlinge"),
+         lty=c(1,2,3), col=c(1,2,4))
```



In diesem Plot ist unsere Vermutung einer abfallenden Feldvogelpopulation wieder gut ersichtlich. Außerdem fällt auch ein negativer Trend im Bezug auf die Anzahl der Graslandschmetterlinge auf. Die offensichtlich hohe Streuung der Anzahl der Graslandschmetterlinge (große “Zacken”) wird vermutlich durch die höhere Abhängigkeit von saisonalen Wetterbedingungen hervorgerufen.

Um die Zusammenhänge der Variablen abzuschätzen, betrachten wir zunächst die Korrelationsmatrix. Zu beachten ist hierbei, dass wir im Moment nur an den Zusammenhängen zwischen den Arten interessiert sind und deshalb die zugehörigen Jahreszahlen für diese Analyse ausschließen (die erste Spalte wird also vorübergehend nicht betrachtet).

```
R> cor(vogel[, -1])
```

	FELDVÖGEL	WALDVÖGEL	GRASLANDSCHMETTERLINGE
FELDVÖGEL	1.0000000	0.6085742	0.5274347
WALDVÖGEL	0.6085742	1.0000000	0.3037392
GRASLANDSCHMETTERLINGE	0.5274347	0.3037392	1.0000000

Es fällt auf, dass zwischen allen Arten ein positiver Zusammenhang besteht. Also scheint es auch realistisch, dass ein höherer Bestand an Schmetterlingen einen höheren Bestand an Vögeln zur Folge hat. Die Signifikanz dieser Behauptung können wir mittels Korrelationstest ermitteln:

```
R> cor.test(vogel$FELDVÖGEL, vogel$WALDVÖGEL)
```

Pearson's product-moment correlation

data: vogel\$FELDVÖGEL and vogel\$WALDVÖGEL

t = 3.7573, df = 24, p-value = 0.0009707

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.2894566 0.8059400

sample estimates:

cor

0.6085742

```
R> cor.test(vogel$FELDVÜGEL, vogel$GRASLANDSCHMETTERLINGE)
```

Pearson's product-moment correlation

data: vogel\$FELDVÜGEL and vogel\$GRASLANDSCHMETTERLINGE

t = 3.0413, df = 24, p-value = 0.005624

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.1760503 0.7595985

sample estimates:

cor

0.5274347

```
R> cor.test(vogel$WALDVÜGEL, vogel$GRASLANDSCHMETTERLINGE)
```

Pearson's product-moment correlation

data: vogel\$WALDVÜGEL and vogel\$GRASLANDSCHMETTERLINGE

t = 1.5618, df = 24, p-value = 0.1314

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.09476179 0.61834091

sample estimates:

cor

0.3037392

Lediglich für den letzten Zusammenhang erhalten wir einen p-Wert > 0.05 , daher lässt sich hier keine signifikante Aussage treffen. Um nun den zeitlichen Verlauf der Populationen abzuschätzen und zu visualisieren, bedienen wir uns linearer Regressionsmodelle. Hier ist zu beachten, dass in unserem vorliegenden Fall nur ein Modell Sinn macht, das ab 1990 startet. Daher müssen wir in der Formel des Modells den Wert 1990 von unseren Jahreszahlen subtrahieren. Damit dies möglich ist, muss die Funktion `I()` verwendet werden. Sie sorgt dafür, dass der Operator `'-'` nicht als Formeloperator, sondern als Rechenoperation gedeutet wird.

```
R> LM.1 <- lm(WALDVÜGEL~I(Jahr-1990), data=vogel)
```

```
R> LM.2 <- lm(FELDVÜGEL~I(Jahr-1990), data=vogel)
```

```
R> LM.3 <- lm(GRASLANDSCHMETTERLINGE~I(Jahr-1990), data=vogel)
```

```
R> summary(LM.1)
```

Call:

```
lm(formula = WALDVÜGEL ~ I(Jahr - 1990), data = vogel)
```

Residuals:

Min	1Q	Median	3Q	Max
-7.920	-3.312	-1.530	3.015	9.764

Coefficients:

Estimate	Std. Error	t value	Pr(> t)
----------	------------	---------	----------

```

(Intercept)    101.7422      1.7906  56.819   <2e-16 ***
I(Jahr - 1990)  -0.3622      0.1228  -2.949    0.007 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

Residual standard error: 4.698 on 24 degrees of freedom
Multiple R-squared:  0.266, Adjusted R-squared:  0.2354
F-statistic: 8.696 on 1 and 24 DF,  p-value: 0.007004

```

```
R> summary(LM.2)
```

```

Call:
lm(formula = FELDVÖGEL ~ I(Jahr - 1990), data = vogel)

```

```

Residuals:
    Min       1Q   Median       3Q      Max
-5.2348 -1.9867 -0.1555  1.9607  6.8678

```

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   99.03094    1.10111   89.94 < 2e-16 ***
I(Jahr - 1990) -1.34968    0.07554  -17.87 2.27e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

Residual standard error: 2.889 on 24 degrees of freedom
Multiple R-squared:  0.9301,    Adjusted R-squared:  0.9272
F-statistic: 319.3 on 1 and 24 DF,  p-value: 2.273e-15

```

```
R> summary(LM.3)
```

```

Call:
lm(formula = GRASLANDSCHMETTERLINGE ~ I(Jahr - 1990), data = vogel)

```

```

Residuals:
    Min       1Q   Median       3Q      Max
-18.658  -6.067  -1.183   7.623  18.519

```

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   90.2712    3.7789  23.888 < 2e-16 ***
I(Jahr - 1990) -0.9485    0.2592  -3.659  0.00124 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

Residual standard error: 9.914 on 24 degrees of freedom
Multiple R-squared:  0.3581,    Adjusted R-squared:  0.3313
F-statistic: 13.39 on 1 and 24 DF,  p-value: 0.001241

```

Anhand der Regressionskoeffizienten können wir den stark negativen Verlauf der Arten GRASLANDSCHMETTERLINGE und FELDVÖGEL erkennen. Der Regressionskoeffizient in der Regression von WALDVÖGEL ist zwar ebenfalls negativ aber bei weitem nicht so groß (betragsmäßig) wie die Koeffizienten in den anderen Regressionsanalysen.

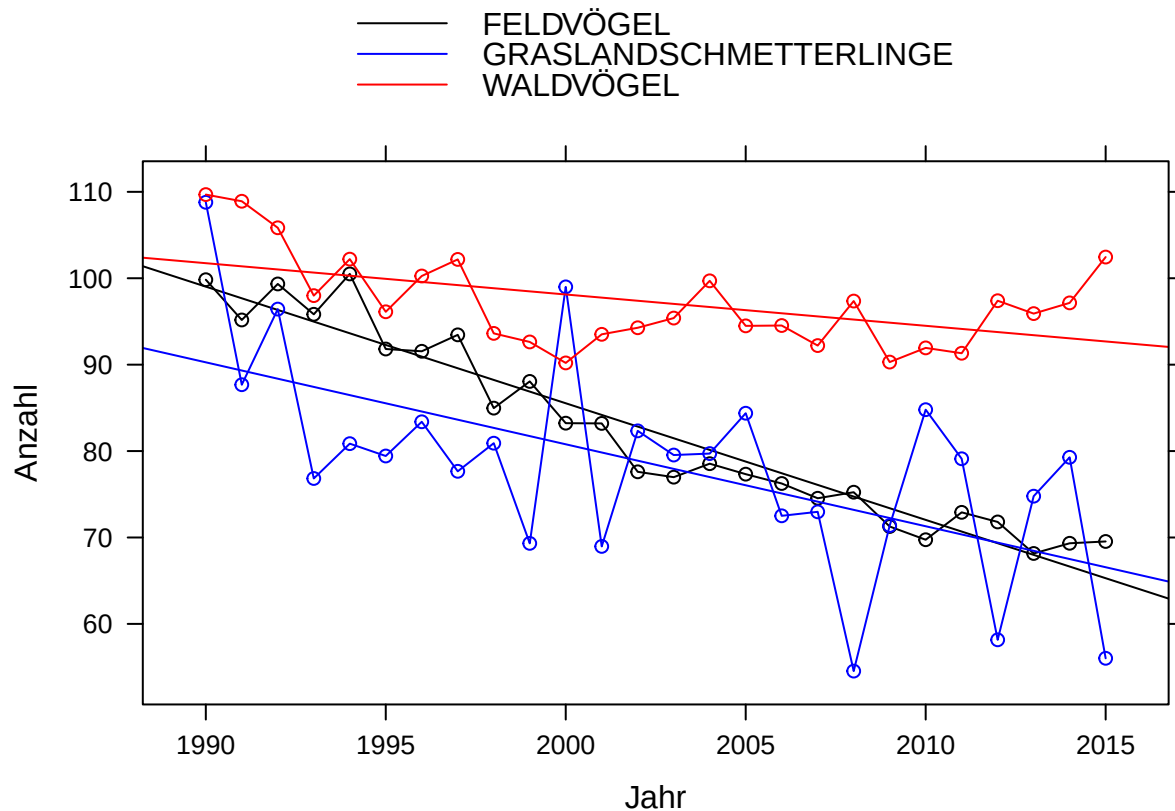
Zur weiteren Analyse und leichteren Handhabung der Daten, erstellen wir einen neuen Datensatz `vogel2`,

der jede Anzahl als eigene Beobachtung und die zugehörige Art als Gruppierungsvariable auffasst.

```
R> vogel2 <- data.frame(Jahr=vogel$Jahr,
+                       Anzahl=c(vogel$FELDVÖGEL, vogel$WALDVÖGEL,
+                               vogel$GRASLANDSCHMETTERLINGE),
+                               Typ=rep(c("Feld", "Wald", "Schmetterling"),
+                                       rep(26, 3)) )
R>                                     # Jede Gattung tritt in 26 Beobachtungen (Jahren) auf
```

Nun ist es leicht die vorliegenden Daten und die zugehörigen Regressionsgeraden in einem Diagramm darzustellen. Wir verwenden hier die Funktion `xyplot()` aus dem Package `lattice`. Mit dem Befehl `type=c("l", "r")` fügen wir zu den real vorhandenen Daten auch die zugehörigen lineare Regressionsgeraden hinzu.

```
R> library(lattice)
R> xyplot(Anzahl~Jahr, groups=Typ, data=vogel2, type=c("o", "r"), col=c(1,4,2),
+         key=list(space="top",lines=list(col=c(1,4,2), lty=c(1,1,1), lwd=1),
+         text=list(c("FELDVÖGEL", "GRASLANDSCHMETTERLINGE", "WALDVÖGEL")))) )
```



Am Vergleich der Regressionsgeraden fällt auf, dass die Variablen `GRASLANDSCHMETTERLINGE` und `FELDVÖGEL` in etwa die gleiche Steigung zu haben scheinen und `FELDVÖGEL` und `WALDVÖGEL` in etwa den gleichen Intercept. Rechnerisch können wir dies bestätigen, indem wir das zugrundeliegende Regressionsmodell betrachten. Durch die Umwandlung des Datensatzes geht dies sehr leicht mit einem einzigen linearen Regressionsmodell mit Interaktionen. Dabei kommt der Formeloperator `*` zum Einsatz. *Anmerkung: `A*B` ist für `R` gleichbedeutend mit `A + B + A:B`; es werden also beide Variablen in das Modell aufgenommen und `A` erhält neben dem normalen Regressionskoeffizienten noch eine Anpassung für die verschiedenen Faktorstufen von `B` (Interaktionen)*

```
R> LM <- lm(Anzahl~I(Jahr-1990)*Typ, data=vogel2)
R> levels(vogel2$Typ)
```

```
[1] "Feld"          "Schmetterling" "Wald"
```

```
R> summary(LM)
```

Call:

```
lm(formula = Anzahl ~ I(Jahr - 1990) * Typ, data = vogel2)
```

Residuals:

Min	1Q	Median	3Q	Max
-18.658	-3.048	-1.080	3.000	18.519

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	99.0309	2.4966	39.666	< 2e-16 ***
I(Jahr - 1990)	-1.3497	0.1713	-7.881	2.56e-11 ***
TypSchmetterling	-8.7597	3.5307	-2.481	0.015438 *
TypWald	2.7113	3.5307	0.768	0.445053
I(Jahr - 1990):TypSchmetterling	0.4011	0.2422	1.656	0.102028
I(Jahr - 1990):TypWald	0.9874	0.2422	4.077	0.000116 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.55 on 72 degrees of freedom

Multiple R-squared: 0.7511, Adjusted R-squared: 0.7338

F-statistic: 43.45 on 5 and 72 DF, p-value: < 2.2e-16

Anhand der Regressionskoeffizienten bestätigt sich unsere Vermutung. Als Referenzkategorie der Gruppierungsvariable `Typ` wählt R, wie anhand der Reihenfolge der levels von `Typ` ersichtlich ist, den `Typ Feld`. Daher gehört in diesem Fall der Intercept zur Regression des Typs `Feld`. Die anderen Intercepts ergeben sich durch Addition der Regressionskoeffizienten `TypSchmetterling` und `TypWald`. Unter Einbezug des Standardfehlers ist es also möglich, dass der Intercept von `Wald` und `Feld` gleich ist. Analog erkennen wir, dass der Regressionskoeffizient der Steigung der Regressionsgerade für `Feld` und `Schmetterling` zwar ähnlicher als der von `Wald` ist, selbst unter Einbezug des Standardfehlers aber keine Gleichheit erzielt wird.

Zum besseren Verständnis der Bedeutung und der Interpretation der Regressionskoeffizienten mit Interaktionen, seien im Folgenden die drei Regressionsmodelle separat aufgelistet:

Für `Typ=Feld`: $Anzahl = 99.0309 - 1.3497 \cdot Jahr$

Für `Typ=Wald`: $Anzahl = (99.0309 + 2.7113) + (-1.3497 + 0.9874) \cdot Jahr = 101.74 - 0.36 \cdot Jahr$

Für `Typ=Schmetterling`: $Anzahl = (99.0309 - 8.7597) + (-1.3497 + 0.4011) \cdot Jahr = 90.27 - 0.95 \cdot Jahr$