Studies of Independent Variables

11.04.2022

Bibliotheken laden, Hilfsfunktion

```
library(GGally)

debug <- T            # debug printout
debug <- F            # kein debug printout
Log <- function(string) {
  if(debug){print(string)}
}
```

Für alle MY Gruppen :

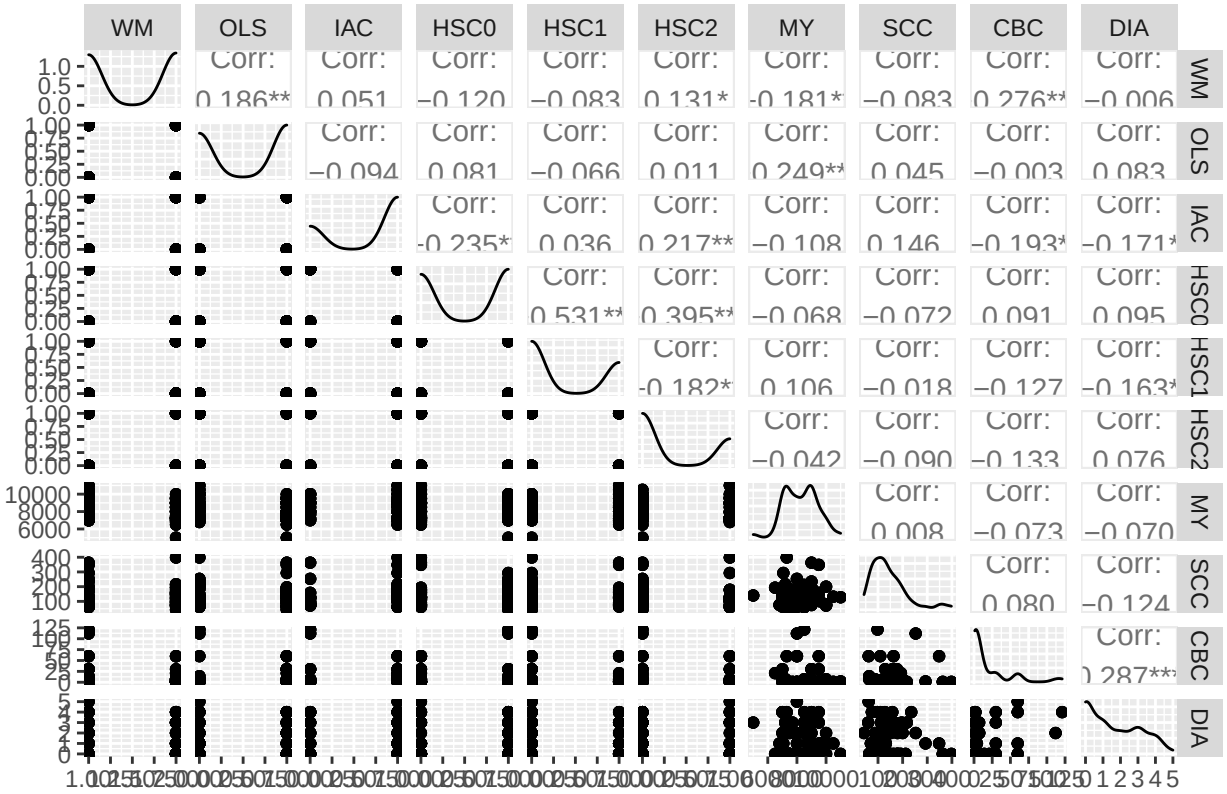
- Resistenzen.Rmd erzeugte Resistenzen[Schicht].csv, diese einlesen
- Variablen plotten und Korrelationen berechnen

"Inklusive" Variablen HSC0-HSC2

Um die Anzahl der Variablen und evtl. die Korrelationen zu reduzieren, habe ich hier wieder zurückgerechnet: Wir haben z.B. 3 = 0+1, also habe ich im Fall husbandry_system_calves = 3 die Variablen HSC0 und HSC1 auf 1 gesetzt.

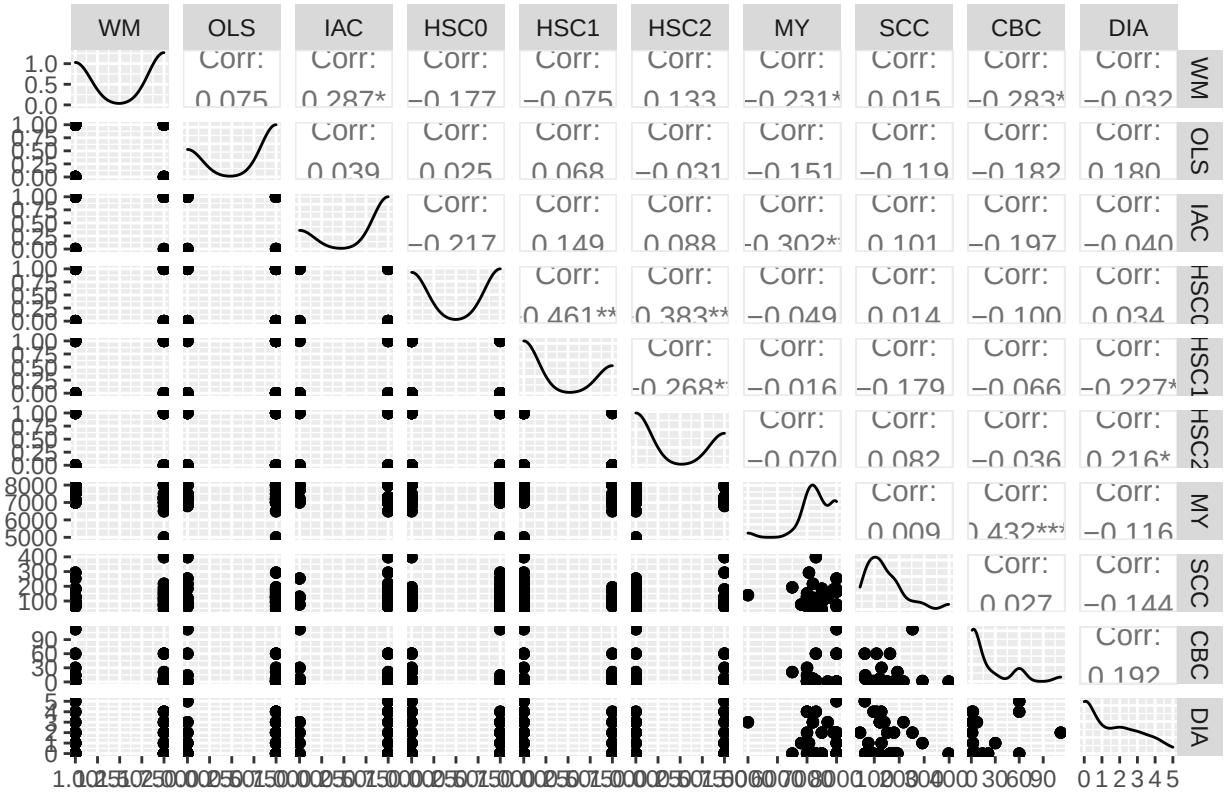
```
for( Schicht in c("U", "LE8000","GT8000") ) {     # Un-stratisfied /  Less than or Equal to 8000 / Greater Than 80
  Resistenzen <- read.csv(paste( "Resistenzen",Schicht,".csv" , sep="" ) )
  Resistenzen[,1] <- NULL         # csv schreiben fügt vorne Index-Spalte an; diese entfernen
  if(debug){View(Resistenzen)}

  df <- data.frame(WM     = Resistenzen$WM.group,    # unabhängige Variablen extrahieren
                   OLS    = Resistenzen$OLS.group,   # incl. Titel kürzen, sonst Platzprobleme ...
                   IAC    = Resistenzen$IAC.group,
                   HSC0   = Resistenzen$HSC0,
                   HSC1   = Resistenzen$HSC1,
                   HSC2   = Resistenzen$HSC2,
                   MY     = Resistenzen$MY,
                   SCC    = Resistenzen$SCC,
                   CBC    = Resistenzen$CBC,
                   DIA    = Resistenzen$DIA)
  print(ggpairs(df, title = paste("group:",Schicht), upper=list(continuos=wrap("cor",size=6))))
  print("")
}
```

group: U

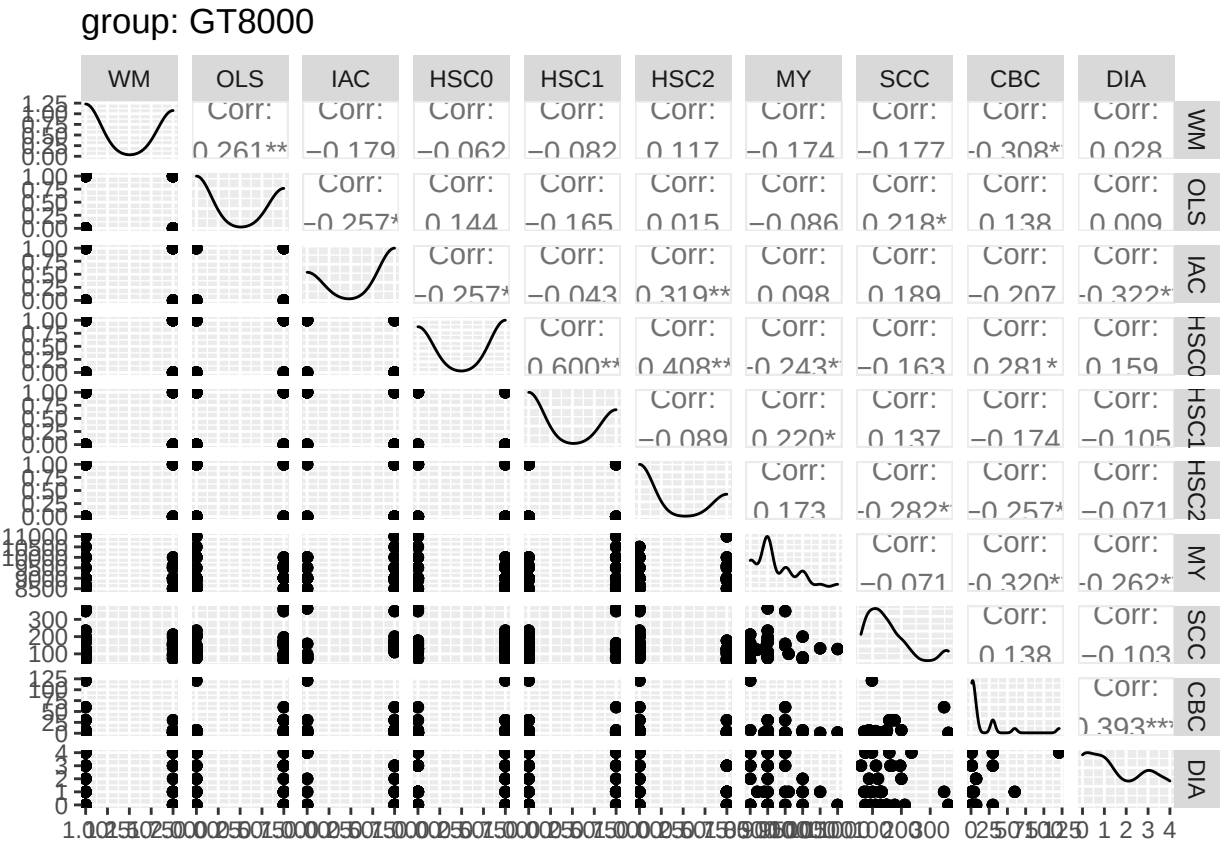



```
## [1] ""
```

group: LE8000




```
## [1] ""
```




```
## [1] ""
```

Linear Depence?

Linear Dependence implies multicollinearity, so in its presence the logistic regression would be unreliable.

The correlations with maximum magnitude are

- -53.2% (HSC0 versus HSC1) in the unstratisfied analysis
- -60.0% (HSC0 versus HSC1) in the stratisfied analysis MY > 8000

Due to collinearity problems, it might not be possible to include HSC0 and HSC1 in one multivariate logistic regression.

Outliers?

Einfach zu interpretieren sind nur die plots ohne diskrete Variablen.

In Histogrammen und Streuplots sehe ich einen Ausreisser mit MY = 5000, das ist Farm 32.

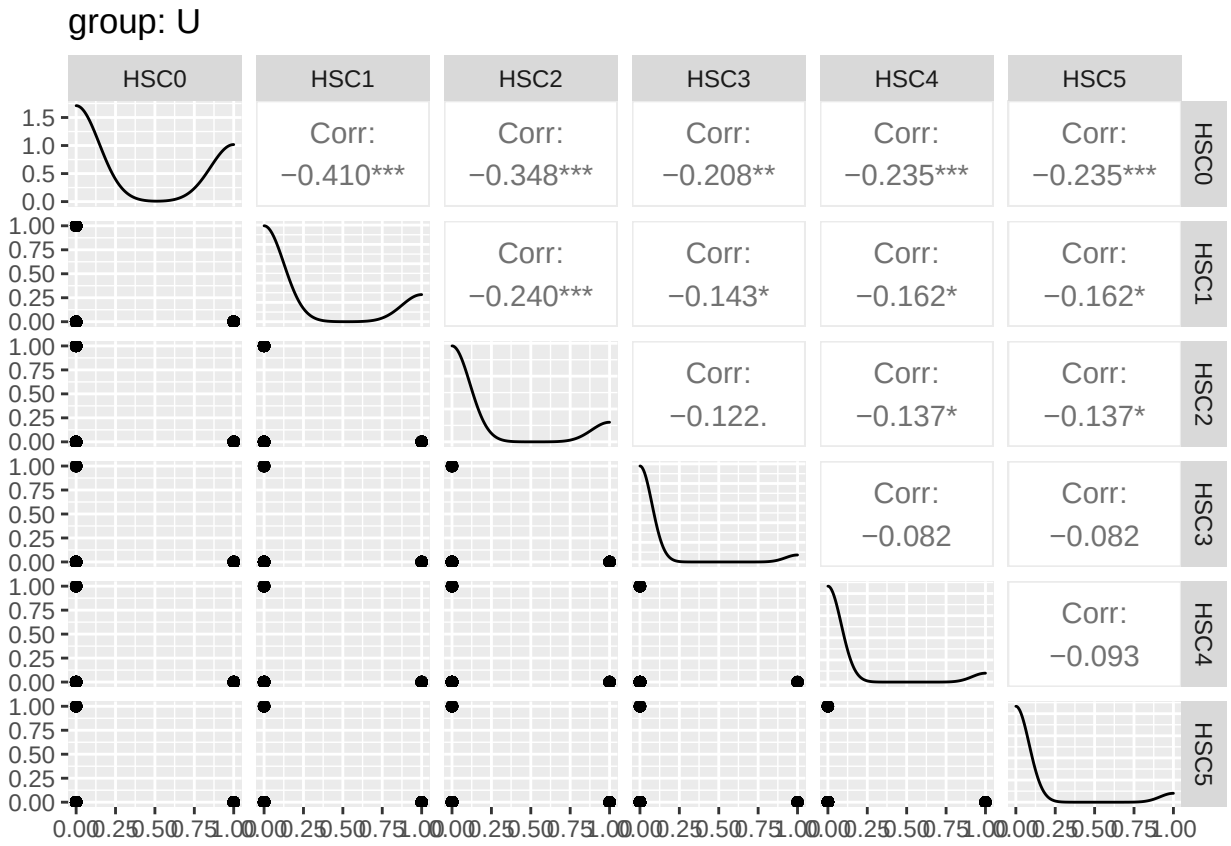
- ist sie als problematisch bekannt?
- das ist aber *kein* Problem für die Regression: MY wird zur Schichtung verwendet, nicht als unabhängige Variable

"Exklusive" Variablen HSC0-HSC5

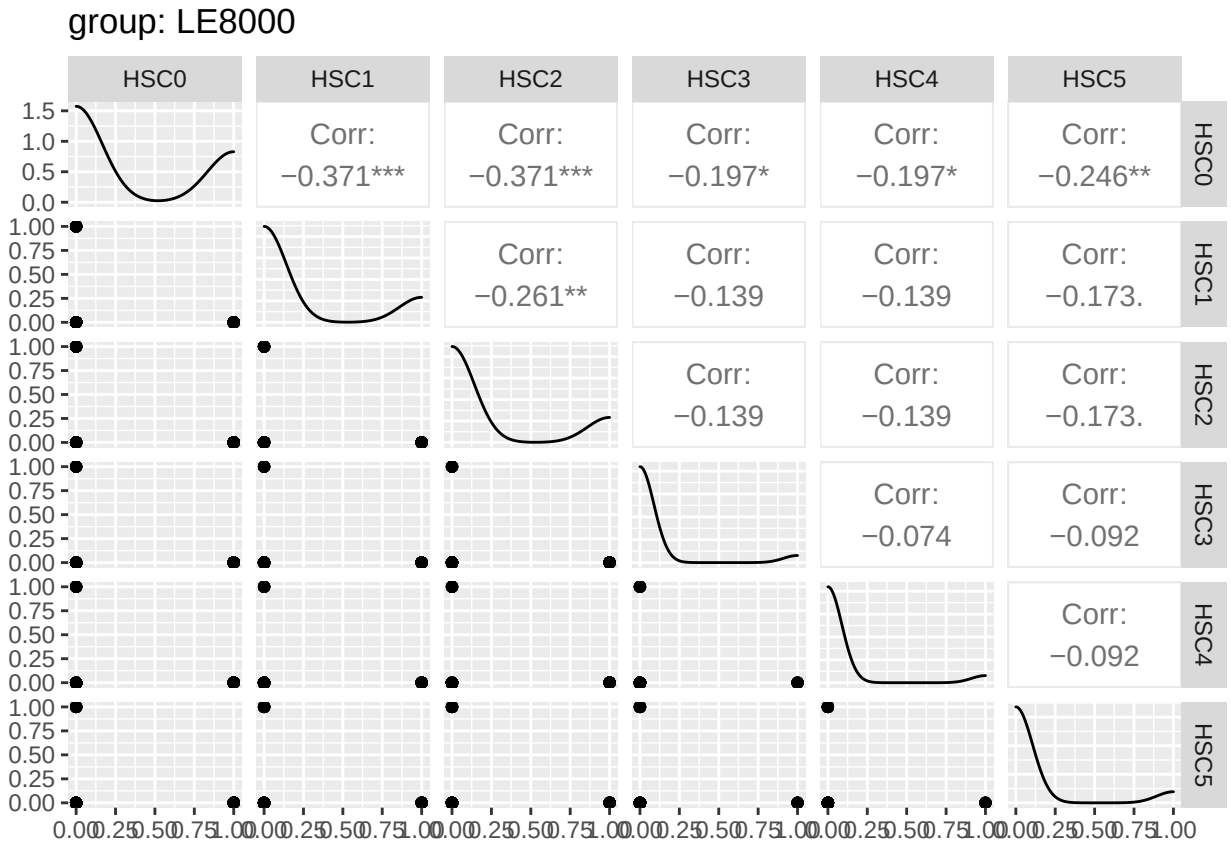
Die ursprünglichen Variablen

```
for( Schicht in c("U", "LE8000","GT8000") ) {     # Un-stratisfied /  Less than or Equal to 8000 / Greater Than 80
  Resistenzen <- read.csv(paste( "ResistenzenHSC012345/Resistenzen",Schicht,".csv" , sep="" ) )
  Resistenzen[,1] <- NULL          # csv schreiben fügt vorne Index-Spalte an; diese entfernen
  if(debug){View(Resistenzen)}

  df <- data.frame(HSC0   = Resistenzen$HSC0,
                   HSC1   = Resistenzen$HSC1,
                   HSC2   = Resistenzen$HSC2,
                   HSC3   = Resistenzen$HSC3,
                   HSC4   = Resistenzen$HSC4,
```

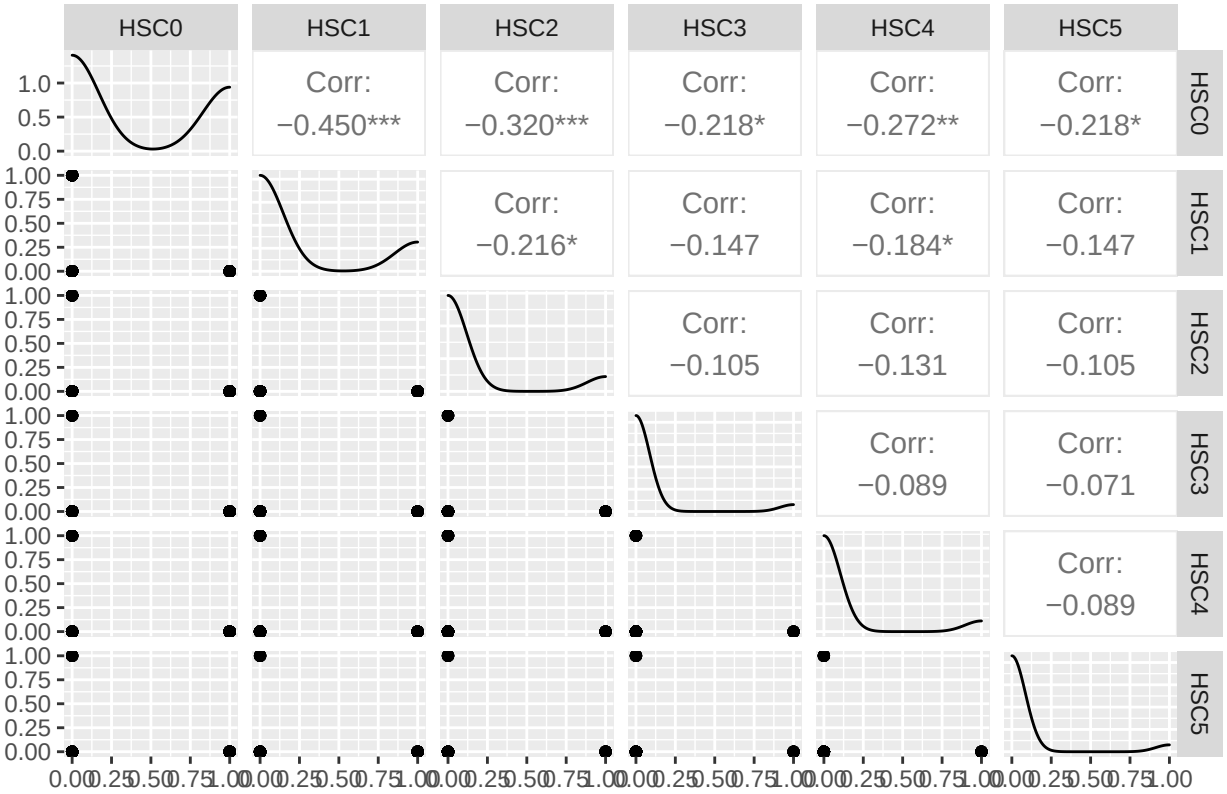
```
                  HSC5    = Resistenzen$HSC5)
  print(ggpairs(df, title = paste("group:",Schicht), upper=list(continuos=wrap("cor",size=6))))
  print("")
}
```




```
## [1] ""
```




```
## [1] ""
```




```
## [1] ""
```

Die Korrelationen sind alle negativ, das ist hier klar: nur eine der Variablen kann 1 sein, die anderen dann alle 0

Korrelationen mit maximalem Betrag:

- -0.410 HSC1 versus HSC0 in der ungeschichteten Analyse
- -0.450 HSC1 versus HSC0 in der Schicht MY > 8000

In der Tat etwas kleiner als für die “inklusiven Variablen”! Wir sollten diskutieren, welche Variablen besser zu interpretieren sind.