Varianzanalyse und Tukey-Test in SAS, R und JMP - die Skalierung der Erklärungsvariablen ist von enormer Bedeutung

Eckard Moll
Grashüpferweg 37
14532 Stahnsdorf
EckardMoll@web.de

Doreen Gabriel
Julius Kühn-Institut
Bundesallee 50
38116 Braunschweig
Doreen.Gabriel@julius-kuehn.de

Zusammenfassung

Mit den Daten eines Beispiels werden in SAS, R und JMP eine einfache Varianzanalyse und Tukey-Test durchgeführt. Es wird gezeigt, welche enorme Bedeutung die Skalierung der Erklärungsvariablen, d.h. des Faktors im Modell, hat.
In SAS wird im Allgemeinen mit der Wahl der Prozedur das Auswertungsverfahren festgelegt. Die Variable des Faktors im Varianzanalysemodell muss in der CLASS-Anweisung stehen. Sie kann sowohl vom Typ character als auch numerisch sein. Für die Auswertung mit R und unter Nutzung der Funktion aov muss die Erklärungsvariable als Faktor deklariert sein. Passiert das nicht und es soll eine Varianzanalyse durchgeführt werden, erfolgt bei der Durchführung des Tukey-Tests eine Fehlermitteilung. Für Varianzanalyse und Tukey-Test in JMP kann der Faktor numerisch sein, aber nur nominal oder ordinal. Ist er stetig, wird anstelle der Varianzanalyse die Regressionsanalyse gerechnet. Multiple Tests werden in diesem Fall ohne eine entsprechende Warn- oder Fehlermeldung auszugeben.

Schlüsselwörter: Skalierung, Varianzanalyse, Tukey-Test, SAS, R, JMP

1 Vorbemerkungen und Zielstellung

Es ist Basiswissen, dass die Eigenschaften der Variablen, besonders ihre Skalierung (character, numerisch: stetig, ordinal oder nominal), entscheidend für die Wahl der statistischen Analyse sind. Wird das bei der Nutzung von Software vergessen oder übersehen, kann es zu falschen Schlussfolgerungen oder Auswertungen kommen. Mit den Daten des Beispiel1 [1] sollen in SAS, R und JMP eine einfache Varianzanalyse und Tukey-Test durchgeführt werden. Besondere Bedeutung erhält dabei die Skalierung der Variablen *gruppe*. Die Daten sind:

gruppe	y			
1	15	17	19	
2	17	20	23	
3	22	25	27	30

2 Varianzanalyse und Tukey-Test in SAS, R und JMP

2.1 SAS 9.4

SAS stellt mehrere Prozeduren zur Verfügung, die in Abhängigkeit von den Modelleigenschaften und der Zielstellung Varianzanalyse und Tukey-Test genutzt werden. Grafiken werden mit Hilfe der ODS GRAPHICS Anweisung automatisch erzeugt.
Die Variable *gruppe*, der Faktor im Varianzanalysemodell, ist numerisch. Das nachfolgende Programm liefert die Ergebnisse für den Tukey-Test zum Signifikanz-niveau α = 5%.

```
DATA beispiel1;
  INPUT gruppe y @@;
CARDS;
 1 15   1 17   1 19
 2 17   2 20   2 23
 3 22   3 25   3 27   3 30
;
ODS GRAPHICS ON/ reset=all imagefmt=emf;
PROC GLM DATA = beispiel1;
  CLASS gruppe;
  MODEL y = gruppe / ss3 ;
  LSMEANS gruppe/ adjust=tukey cl;
RUN;
ODS GRAPHICS OFF;
```

Die Textausgabe der Ergebnisse wird hier nur auf die Varianztabelle, die Mittel-wertvergleiche und die Konfidenzintervalle beschränkt. Zwei Grafiken werden angelegt.

```
                          Sum of
Source           DF      Squares  Mean Square  F Value  Pr > F

Model             2  148.5000000   74.2500000     8.66  0.0128
Error             7   60.0000000    8.5714286
Corrected Total   9  208.5000000

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

                              LSMEAN
gruppe         y LSMEAN       Number

1           17.0000000            1
2           20.0000000            2
3           26.0000000            3
```

```
    Least Squares Means for effect gruppe
     Pr > |t| for H0: LSMean(i)=LSMean(j)

             Dependent Variable: y

i/j            1           2           3

   1                   0.4615      0.0122
   2       0.4615                  0.0716
   3       0.0122      0.0716

gruppe y LSMEAN 95% Confidence Limits

1     17.000000  13.003056  20.996944
2     20.000000  16.003056  23.996944
3     26.000000  22.538545  29.461455

Least Squares Means for Effect gruppe

    Difference     Simultaneous 95%
       Between  Confidence Limits for
i j     Means   LSMean(i)-LSMean(j)

1 2  -3.000000  -10.039984   4.039984
1 3  -9.000000  -15.585302  -2.414698
2 3  -6.000000  -12.585302   0.585302
```

Die erste Grafik (Abb. 1) veranschaulicht die Lage der Mittelwerte für jede Gruppe,

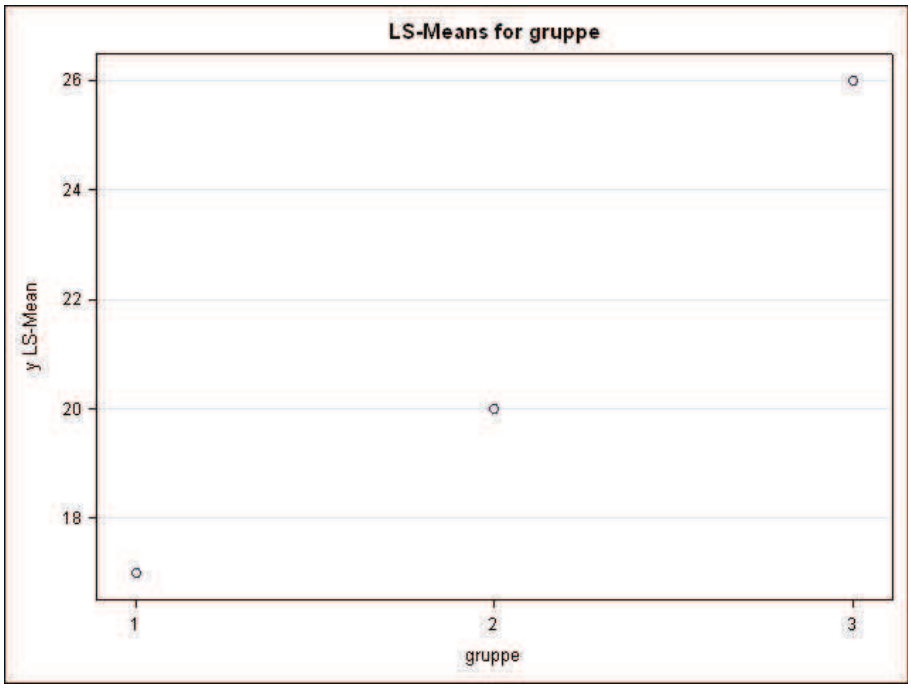

Abbildung 1: Mittelwerte jeder Gruppe

die zweite (Abb. 2) die Signifikanzentscheidungen für die Differenzen.

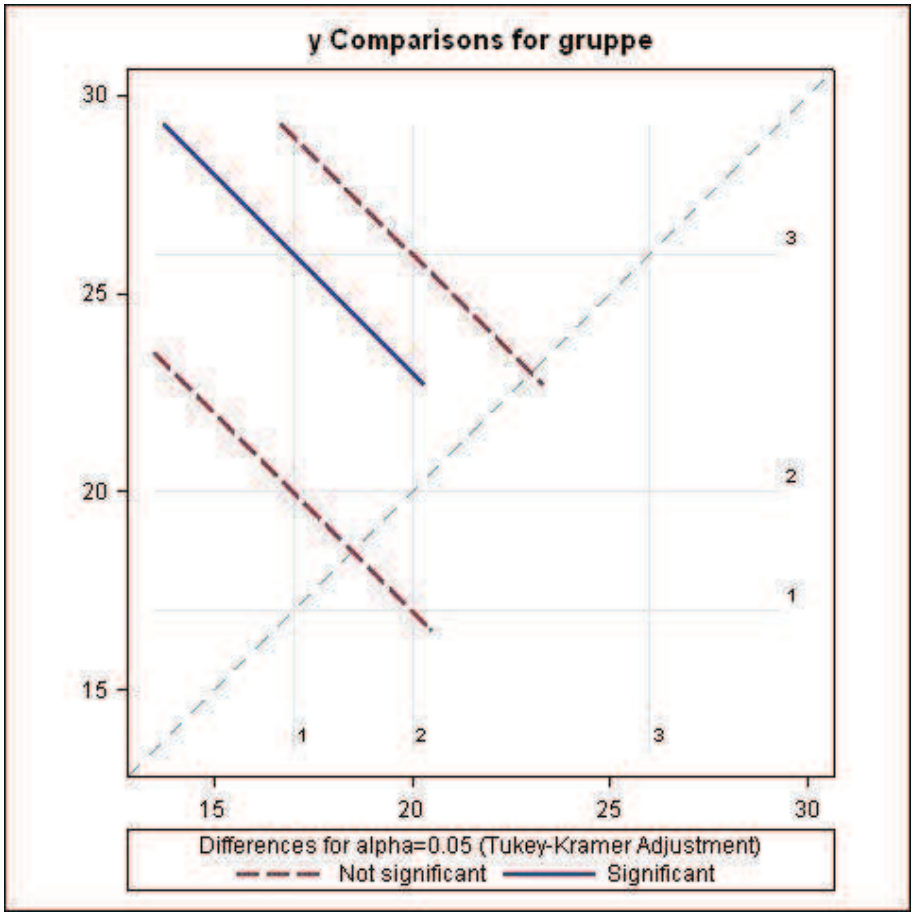

Abbildung 2: Signifikanzentscheidungen für die Differenzen

In SAS ist im Allgemeinen mit der Wahl der Prozedur das Auswertungsverfahren festgelegt. Durch die MODEL-Anweisung wird aus einer Variablen – im Beispiel die Variable *gruppe* – in Verbindung mit der CLASS-Anweisung ein Faktor im Varianzanalysemodell. Aus diesem Grund kann die Variable *gruppe* sowohl vom Typ character als auch numerisch sein.

Steht eine (numerische) Variable auf der rechten Seite der MODEL-Anweisung und fehlt in der CLASS-Anweisung, so wird sie im linearen Modell zur Kovariablen. Ist sie die alleinige Variable im Modell, wird sie zum Regressor und anstelle der Varianzanalyse wird die Regressionsanalyse gerechnet:

```
Dependent Variable: y

                              Sum of
Source                  DF       Squares  Mean Square  F Value  Pr > F
Model                    1   143.8043478  143.8043478    17.78  0.0029
Error                    8    64.6956522    8.0869565
Corrected Total          9   208.5000000
```

```
R-Square     Coeff Var      Root MSE        y Mean
0.689709      13.22678      2.843757      21.50000
Source             DF  Type III SS  Mean Square  F Value  Pr > F
gruppe              1  143.8043478  143.8043478    17.78  0.0029

                           Standard
Parameter      Estimate        Error  t Value  Pr > |t|
Intercept   11.91304348   2.44485494     4.87    0.0012
gruppe       4.56521739   1.08259997     4.22    0.0029
```

Der in der LSMEANS-Anweisung formulierte Mittelwertvergleich wird nicht ausgeführt. In der Fehlermitteilung wird darauf hingewiesen, dass die Klassifizierungsvariable fehlt.

2.2 R

2.2.1 Die Variable gruppe als Faktor (Funktion aov)

Die Aufgaben der SAS-Prozeduren übernehmen in R die speziellen Funktionsaufrufe. Für multiple Mittelwertvergleiche stehen in R mehrere Packages bereit. Für die Nutzung der Funktion **aov** wird die Variable *gruppe* des Beispiels als Faktor eingelesen.

```
beispiel1<-data.frame(gruppe=factor(c(1,1,1,2,2,2,3,3,3,3)),
y=c(15,17,19,17,20,23,22,25,27,30))

mod<-aov(y~gruppe, data=beispiel1)
summary(mod)
```

```
            Df  Sum Sq  Mean Sq  F value  Pr(>F)
gruppe       2   148.5    74.25    8.662  0.0128 *
Residuals    7    60.0     8.57
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(mod)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = y ~ gruppe, data=beispiel1)

$gruppe
    diff         lwr       upr      p adj
2-1    3  -4.0400361  10.04004  0.4614676
3-1    9   2.4146493  15.58535  0.0121868
3-2    6  -0.5853507  12.58535  0.0715599
```

```
plot(TukeyHSD(mod))
```

Diese Plot-Anweisung erzeugt eine Veranschaulichung (Abb. 3) der Signifikanzentscheidungen.

Abbildung 3: Signifikanzentscheidungen für die Differenzen

Zusätzlich kann man aus dem Package Agricolae den Tukey-Test mit der Funktion HSD.test durchführen.

```
library(agricolae)
out<-HSD.test(mod, "gruppe")
out

$statistics
  Mean       CV  MSerror      HSD r.harmonic
  21.5 13.61721 8.571429 6.740321   3.272727

$parameters
  Df ntr StudentizedRange alpha  test name.t
   7   3         4.164941  0.05 Tukey gruppe

$means
   y      std r Min Max
1 17 2.000000 3  15  19
2 20 3.000000 3  17  23
3 26 3.366502 4  22  30

$comparison
NULL
```

```
$groups
  trt means  M
1   3    26  a
2   2    20 ab
3   1    17  b
```

```
bar.group(out$groups, ylim=c(0,45), density=4, border="blue")
```

Diese Anweisung liefert ein Balkendiagramm der Mittelwerte mit der Signifikanz-entscheidung mittels Buchstaben (Abb. 4).

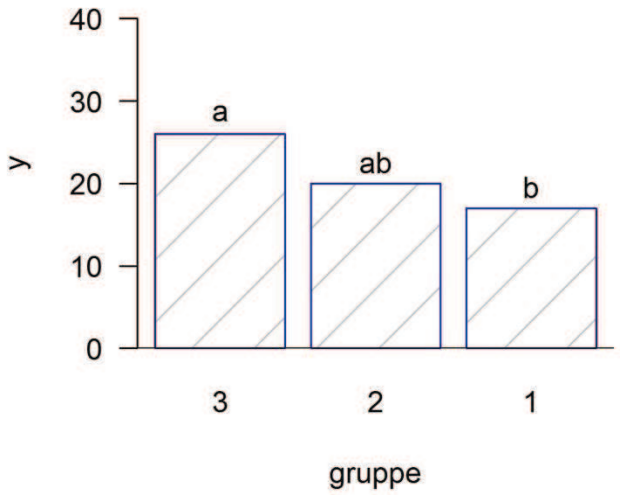

Abbildung 4: Signifikanzentscheidungen durch Buchstaben

2.2.2 Die Variable gruppe als numerische Variable (Funktion aov)

Wird der Faktor *gruppe* als *gruppe2* numerisch vereinbart und ebenfalls die Funktion aov verwendet, dann wird eine andere Varianzanalysetabelle ausgegeben und bei Wahl des Tukey-Tests erfolgt eine Fehlermitteilung.

```
beispiel1$gruppe2<-c(1,1,1,2,2,2,3,3,3,3)
mod2<-aov(y~gruppe2, data = beispiel1)
summary(mod2)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
gruppe2      1  143.8  143.80   17.78 0.00293 **
Residuals    8   64.7    8.09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(mod2)
```

```
Error in TukeyHSD.aov(mod2) : keine Faktoren im angepassten Modell
In addition: Warning message:
In replications(paste("~", xx), data = mf) :
  nicht-Faktoren ignoriert: gruppe2
```

Die Fehlerausschrift besagt, dass *gruppe2*, die numerisch vereinbarte Variable *gruppe*, nicht als Faktor für den Tukey-Test erkannt wird.

Zu klären bleibt noch, was das für eine Varianzanalyse ist, die nun gerechnet wurde. Dazu wird die Funktion lm verwendet:

```
mod2<-lm(y~gruppe2, data=beispiel1)
summary(mod2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0435 -1.3696 -0.0435  1.8152  4.3913

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.913      2.445   4.873  0.00124 **
gruppe2        4.565      1.083   4.217  0.00293 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.844 on 8 degrees of freedom
Multiple R-squared:  0.6897, Adjusted R-squared:  0.6509
F-statistic: 17.78 on 1 and 8 DF,  p-value: 0.002928

anova(mod2)

Analysis of Variance Table
Response: y
          Df  Sum Sq Mean Sq F value   Pr(>F)
gruppe2    1 143.804 143.804  17.782 0.002928 **
Residuals  8  64.696   8.087
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Diese Varianztabelle stimmt mit obiger für *gruppe2* überein. Es ist die Anpassung an ein lineares Regressionsmodell. Das Verfahren ist folglich ebenso wie mit der SAS-GLM-Prozedur ohne die CLASS-Anweisung die Regressionsanalyse. Dem entsprechend werden wie auch in SAS die Schätzwerte für die Regressionskoeffizienten der linearen Funktion y = f(*gruppe2*) berechnet.

Für die Aufgabenstellung, die Berechnung der einfachen Varianzanalyse und des Tukey-Tests, bedeutet das, dass die Erklärungsvariable *gruppe* als Faktor deklariert sein muss.

2.3 JMP 12

2.3.1 Die Variable gruppe als nominale oder ordinale Variable

Mit dem Öffnen der Datei des Beispiels oder der direkten Dateneingabe werden die Eigenschaften der Variablen automatisch festgelegt: *y*: stetig, *gruppe*: stetig. Für die Analysen Varianzanalyse und Tukey-Tests muss die numerische Variable *gruppe* nominal oder ordinal sein. D.h. die Eigenschaft der Variablen *gruppe* muss geändert werden!

Varianzanalyse und multiple Mittelwertvergleiche erreicht man über das Pull-Down-Menü Analyze → Fit Model . Die auszuwertende Variable und die Effekte im Modell sind zuzuweisen: Y: y und Add: gruppe.
Die Standardausgabe ist umfangreich und enthält mehrere Grafiken. Die Angabe hier wird nur auf die Varianztabelle begrenzt.

```
Analysis of Variance
Source       DF     Sum of Squares     Mean Square     F Ratio
Model         2          148,50000         74,2500      8,6625
Error         7           60,00000          8,5714    Prob > F
C. Total      9          208,50000                     0,0128*
```

Das zu den multiplen Mittelwertvergleichen führende rote Dreieck steht vor der Überschrift Response y. Die Wahl von ‘Estimates ▶führt zur Entscheidung ‘Multiple Comparisons‘. In einem separaten Fenster sind zu wählen der Typ der Schätzer (Least Square Means Estimates), der Effekt (voreingestellt: gruppe) und das Testverfahren ‘All Pairwise Comparisons - Tukey HSD‘.

```
Multiple Comparisons
Estimates
gruppe  Estimate Std Error DF t Ratio Prob>|t| Lower 95% Upper 95%
1      17,000000 1,6903085  7   10,06  <,0001* 13,003056 20,996944
2      20,000000 1,6903085  7   11,83  <,0001* 16,003056 23,996944
3      26,000000 1,4638501  7   17,76  <,0001* 22,538545 29,461455

Tukey HSD All Pairwise Comparisons
Quantile = 2,94498
, Adjusted DF = 7,0
, Adjustment = Tukey-Kramer

All Pairwise Differences
gruppe -gruppe Difference Std Error t Ratio Prob>|t| Lower 95% Upper 95%
1      2         -3,00000  2,390457   -1,25   0,4615  -10,0398   4,03984
1      3         -9,00000  2,236068   -4,02   0,0122* -15,5852  -2,41483
2      3         -6,00000  2,236068   -2,68   0,0716  -12,5852   0,58517
```

Die grafische Darstellung der Signifikanzentscheidungen (Abb. 5) wird mitgeliefert.

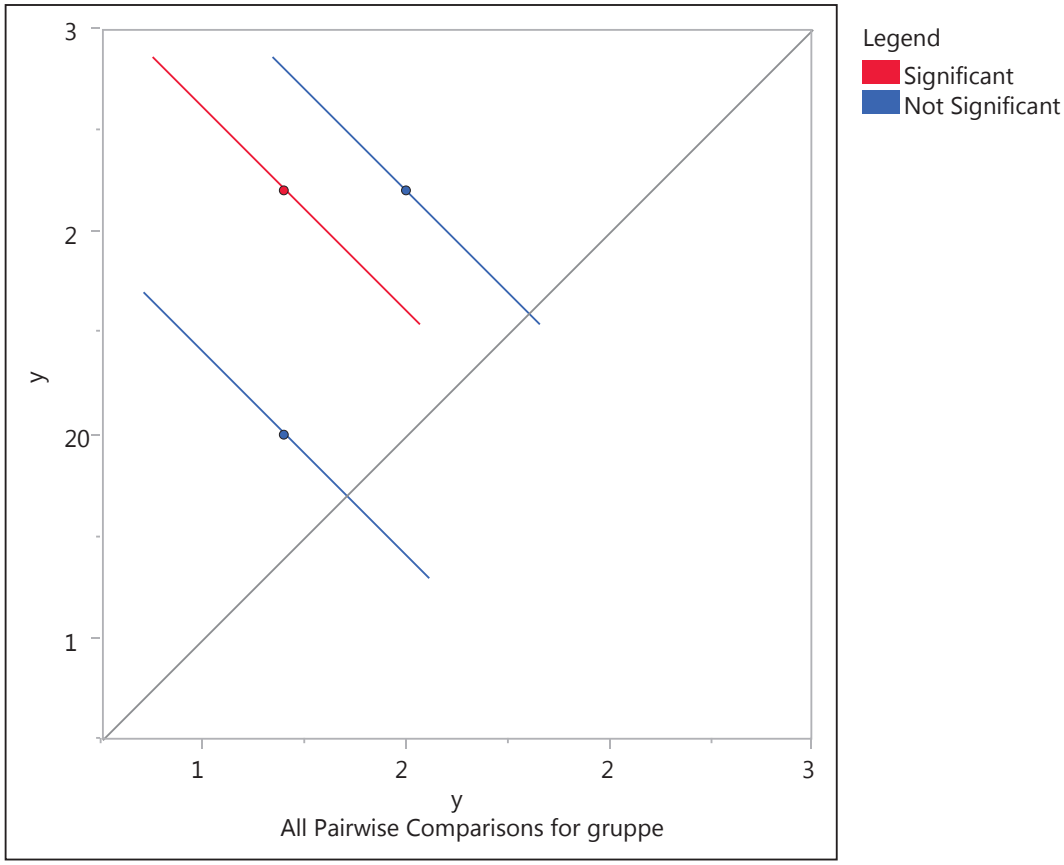

Abbildung 5: Signifikanzentscheidungen für die Paarvergleiche

2.3.2 Die Variable gruppe als stetige Variable

Wird es „vergessen“, die numerische Variable *gruppe*, den Faktor im Modell, von stetig auf nominal oder ordinal umzustellen, dann führt das zur Regressionsanalyse. Multiple Mittelwertvergleiche können durchgeführt werden; es kommt kein Warn- oder Fehlerhinweis. Diese Mittelwertvergleiche entsprechen aber nicht der Zielstellung, denn sie basieren auf dem Regressionsmodell. Um das zu veranschaulichen wird mit der stetigen Variablen *gruppe* die gleiche Analyse wie oben wiederholt. Die Varianztabelle stimmt mit der Varianztabelle der Ausgaben in SAS, wenn die Variable gruppe nicht in der CLASS-Anweisung steht, und in R für die numerische Variable *gruppe2* überein.

```
Analysis of Variance
Source      DF Sum of Squares   Mean Square   F Ratio
Model        1       143,80435      143,804   17,7823
Error        8        64,69565        8,087  Prob > F
C. Total     9       208,50000                0,0029*
```

Unter anderem werden auch die berechneten Regressionskoeffizienten ausgegeben:

```
Parameter Estimates
Term           Estimate    Std Error    t Ratio    Prob>|t|
Intercept     11,913043    2,444855       4,87      0,0012
Gruppe         4,5652174     1,0826       4,22      0,0029
```

Wenn jetzt ebenfalls über das rote Dreieck vor der Überschrift Response y ‘Estimates ▶‘ und ‘Multiple Comparisons‘ gewählt, dann kann man auch hier zu einem Tukey-Test kommen. Diese Ergebnisse unterscheiden sich vom Tukey-Test der Varianzanalyse:

```
Multiple Comparisons
Estimates
Gruppe Estimate Std Error  DF t Ratio Prob>|t| Lower 95% Upper 95%
1     16,478261 1,4922611   8   11,04   <,0001 13,037101 19,919421
2     21,043478 0,9057681   8   23,23   <,0001 18,954773 23,132183
3     25,608696 1,3259088   8   19,31   <,0001 22,551145 28,666247

Tukey HSD All Pairwise Comparisons
Quantile = 2,85742
, Adjusted DF = 8,0
, Adjustment = Tukey-Kramer

All Pairwise Differences
gruppe -gruppe Difference Std Error t Ratio Prob>|t| Lower 95% Upper 95%
1      2         -4,56522  1,082600   -4,22   0,0073   -7,6587  -1,47177
1      3         -9,13043  2,165200   -4,22   0,0073  -15,3173  -2,94354
2      3         -4,56522  1,082600   -4,22   0,0073   -7,6587  -1,47177
```

Auch die grafische Darstellung der Signifikanzentscheidungen (Abb. 6) ist eine andere als oben. Sie scheinen zur Abb. 5 parallel verschoben zu sein. Die Ursache ist schnell gefunden. Sie liegt in den Varianzen und Freiheitsgraden.

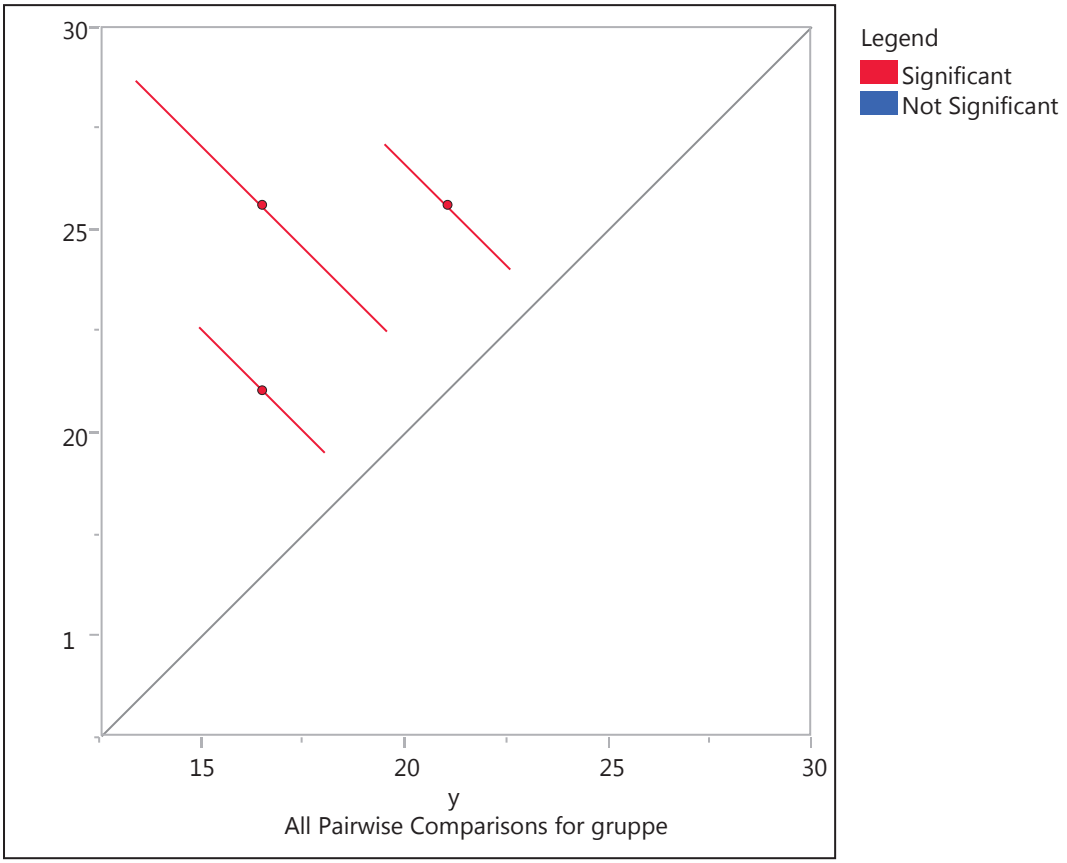

Abbildung 6: Signifikanzentscheidungen für die Paarvergleiche bei der Regressionsanalyse

Wird in JMP der Skalierung der Variablen, die Faktoren im Modell werden sollen, keine Beachtung geschenkt, dann kommt man ohne einen Hinweis bei stetiger Variable *gruppe* auch zu einem Tukey-Test. Diese Ergebnisse basieren auf der Regressionsanalyse und entsprechen nicht dem eigentlichen Ziel der Varianzanalyse mit dem Faktor *gruppe*.

3 Schlussfolgerungen

Die Skalierung der Variablen ist wesentlich für die Berechnung von Maßzahlen und der Durchführung von statistischen Analysen. Beispielhaft sollen die einfache Varianzanalyse und der Tukey-Test gerechnet werden. In SAS legt man sich dafür im Allgemeinen mit der Wahl der Prozedur, die Auflistung der Variablen in der CLASS- und MODEL-Anweisung fest, dass z.B. die Variable *gruppe* ein Faktor ist. Es macht dabei keinen Unterschied, ob dieser Faktor vom Typ character oder numerisch ist.
Für die Varianzanalyse mit R muss bei Nutzung der Funktion aov mit anschließendem Tukey-Test die Variable *gruppe* als Faktor vereinbart sein. Damit ist die Stellung dieser Variablen im Varianzanalysemodell klar.
In JMP muss der automatischen Wahl der Eigenschaften der Variablen unbedingt Beachtung geschenkt werden. Am einfachsten ist es, wenn die Variable, die Faktor werden soll, vom Typ character ist. Ist sie numerisch und wird automatisch als stetige Variable erkannt, dann muss diese Eigenschaft auf nominal oder ordinal geändert werden, wenn mit dieser Variablen als Faktor Varianzanalyse und Tukey-Test gerechnet werden sollen. Bleibt die Variable, die für die Varianzanalyse und Tukey-Test Faktor werden soll, stetig skaliert, dann wird eine Regressionsanalyse durchgeführt und die darauf basierenden Mittelwertvergleiche verfehlen die eigentliche Zielstellung der Analyse.

Literatur

[1] Schumacher, E. (2004): Vergleich von mehr als zwei Parametern
In: Moll, E., J. Gröger, M. Liesebach, P.E. Rudolph, T. Stauber und M. Ziller (Hrsg.): Einführung in die Biometrie, Heft 3,
2. Aufl., ISBN 3-930037-17-3