Lineare Regression Mietpreis-Beispiel

Institut für Angewandte Statistik und EDV, Universität für Bodenkultur Wien
http://www.rali.boku.ac.at/statedv.html

Die aktuelle Version dieses Dokuments finden Sie unter:
http://www.rali.boku.ac.at/statistik-r-beispiele.html

Letzte Änderung: 2013-11-18
CY, RW

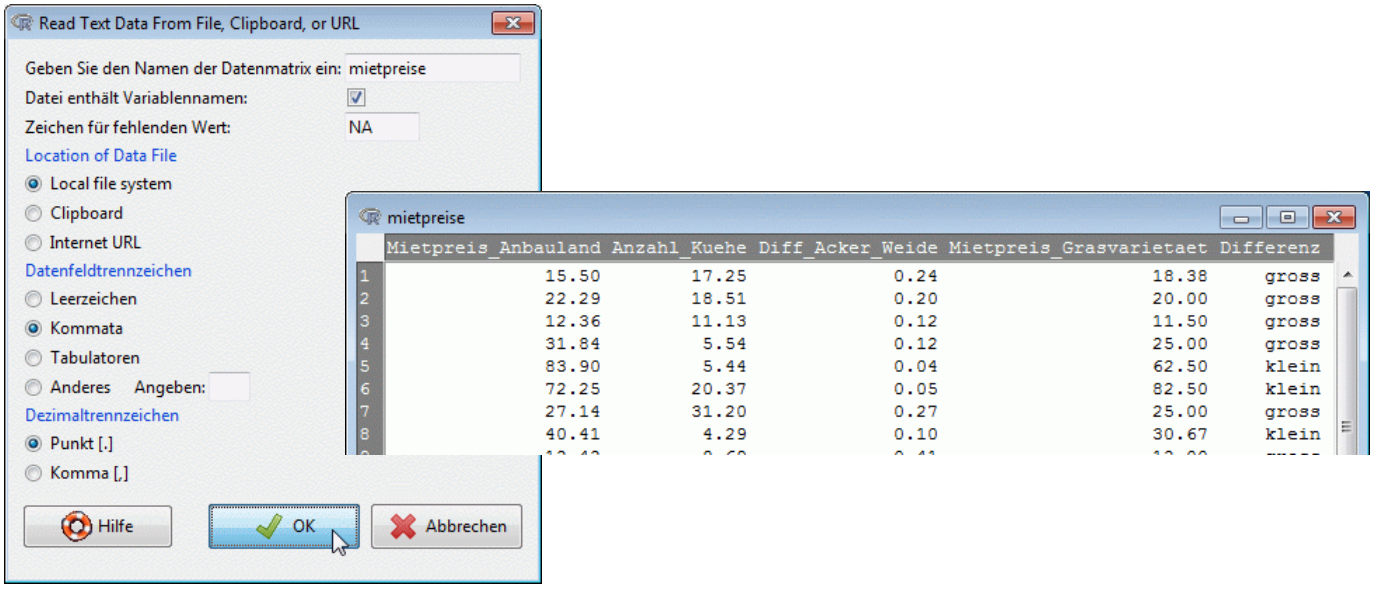
Ausführliche Informationen zur Installation von "R", zur Installation von "R Commander", und zu anderen statistischen Verfahren finden Sie auf der Seite: **http://www.rali.boku.ac.at/statistik-r.html**

Inhalt

Einlesen einer CSV-Datei

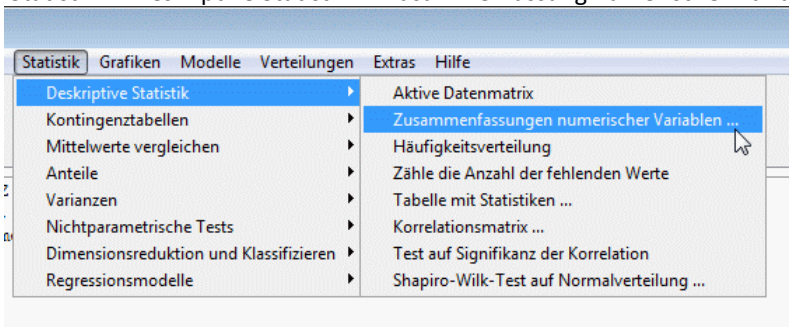
(Hinweis: Ausführliche Erklärungen zum Einlesen von CSV-Dateien finden Sie unter: http://www.rali.boku.ac.at/statistik-r-einlesen-csv.html)

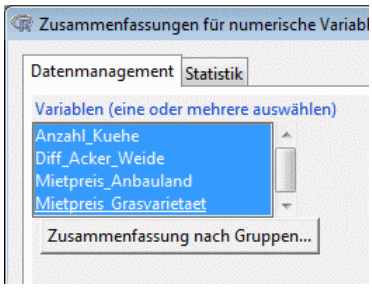
In der Datei '`Mietpreise.csv`' wird der Mietpreis pro Acre Weidefläche in Minnesota, die Anzahl der Kühe pro Quadratmeile, die Differenz zwischen Acker-und Weidefläche so wie der Mietpreis pro Acre Ackerland aufgelistet.
Die Variable Differenz ist eine Rekodierung der Differenz zwischen Acker-und Weidefläche mit den Bezeichnungen „klein“ falls `Diff_Acker_Weide < 0.17` und „gross“ falls `Diff_Acker_Weide >= 0.17` .



Zusammenfassung numerischer Variablen

Zur Analyse ist es hilfreich, eine Zusammenfassung der numerischen Variablen zu betrachten:
'Statistik' > 'Deskriptive Statistik' > 'Zusammenfassung numerischer Variablen ...'



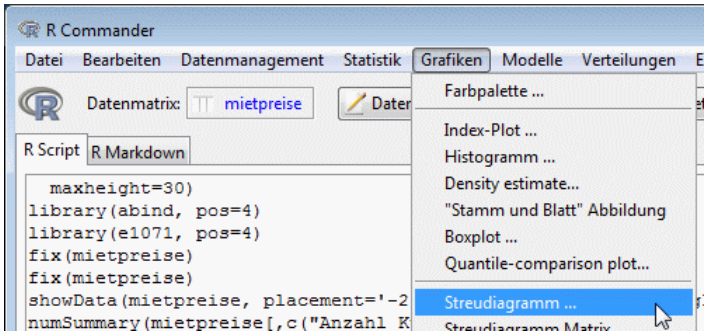



```
> numSummary(mietpreise[,c("Anzahl_Kuehe", "Diff_Acker_Weide", "Mietpreis_Anbauland",
+   "Mietpreis_Grasvarietaet")], statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,
+   1))
                            mean         sd    IQR   0%    25%   50%    75%  100%  n
Anzahl_Kuehe           20.5632836 15.3361474 24.105 1.53  7.110 16.12 31.215 58.60 67
Diff_Acker_Weide        0.1697015  0.1444946  0.170 0.02  0.065  0.12  0.235  0.72 67
Mietpreis_Anbauland    43.7891045 21.2739797 36.780 6.17 24.420 44.56 61.200 83.90 67
Mietpreis_Grasvarietaet 42.4662687 22.8206764 35.880 5.00 23.500 39.17 59.380 99.17 67
```

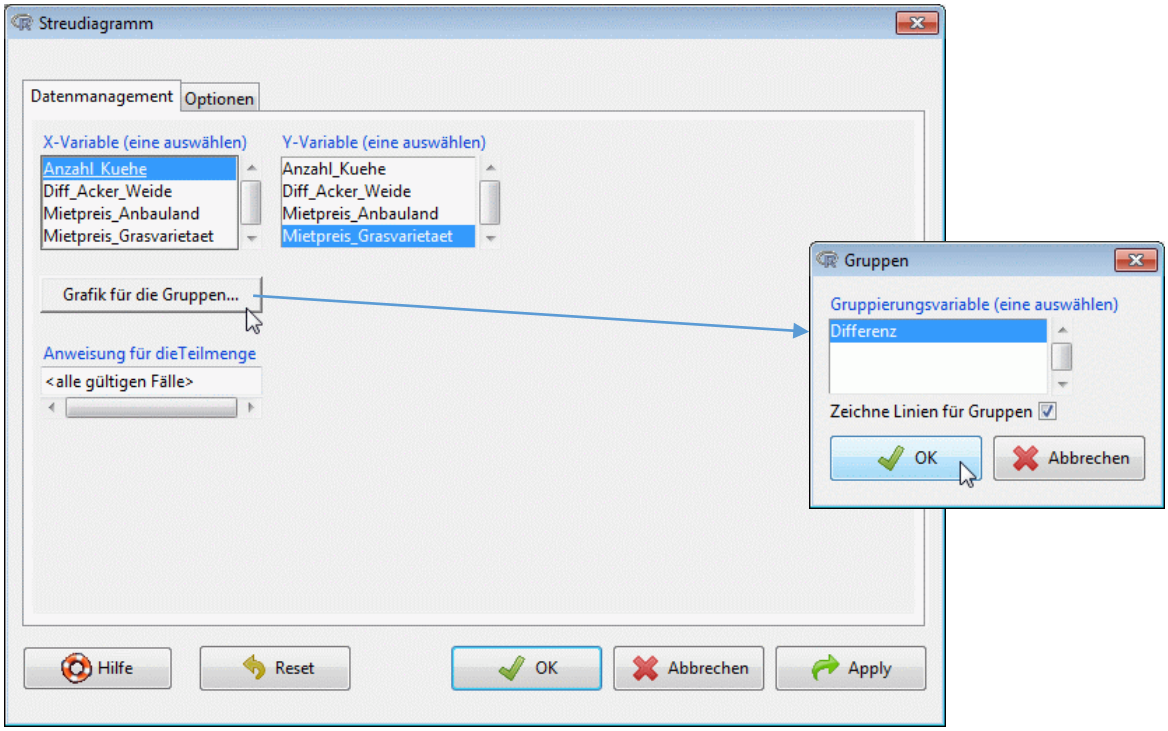
Streudiagramme

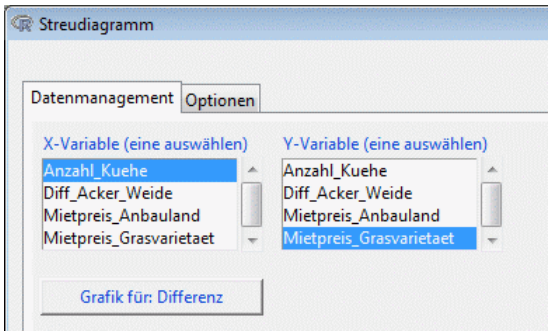
Streudiagramme inklusive Regressionsgeraden können für einzelne Variablen erzeugt werden. Möchte man nach Gruppen getrennte Regressionsgeraden haben geht man folgendermaßen vor:

'Grafiken' > 'Streudiagramm ...'

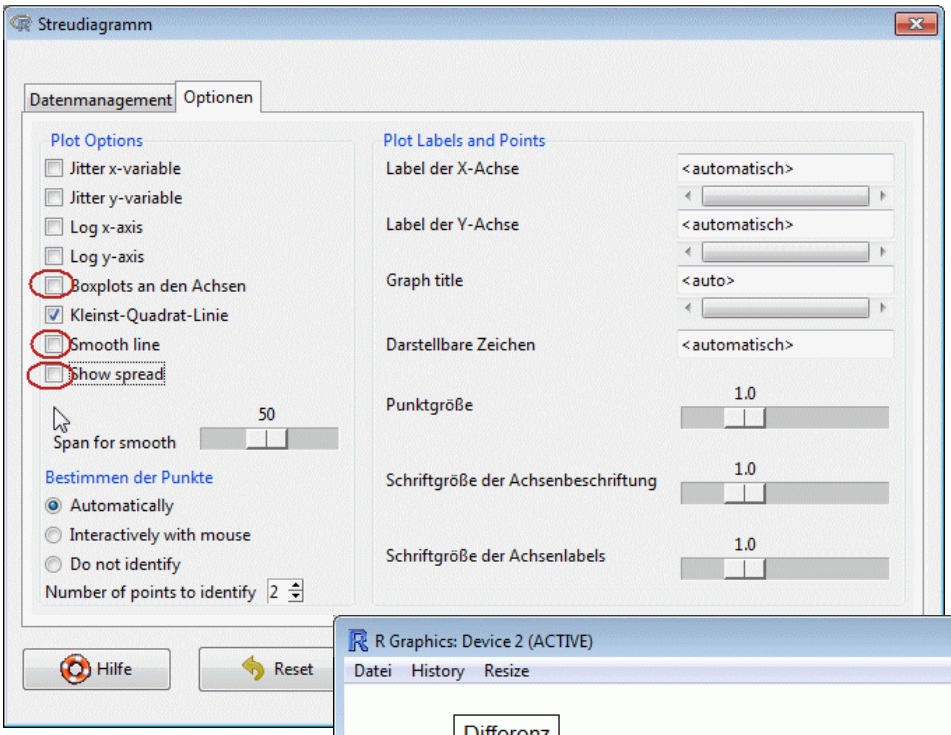


X-Variable, Y-Variable, Gruppierungsvariable auswählen:





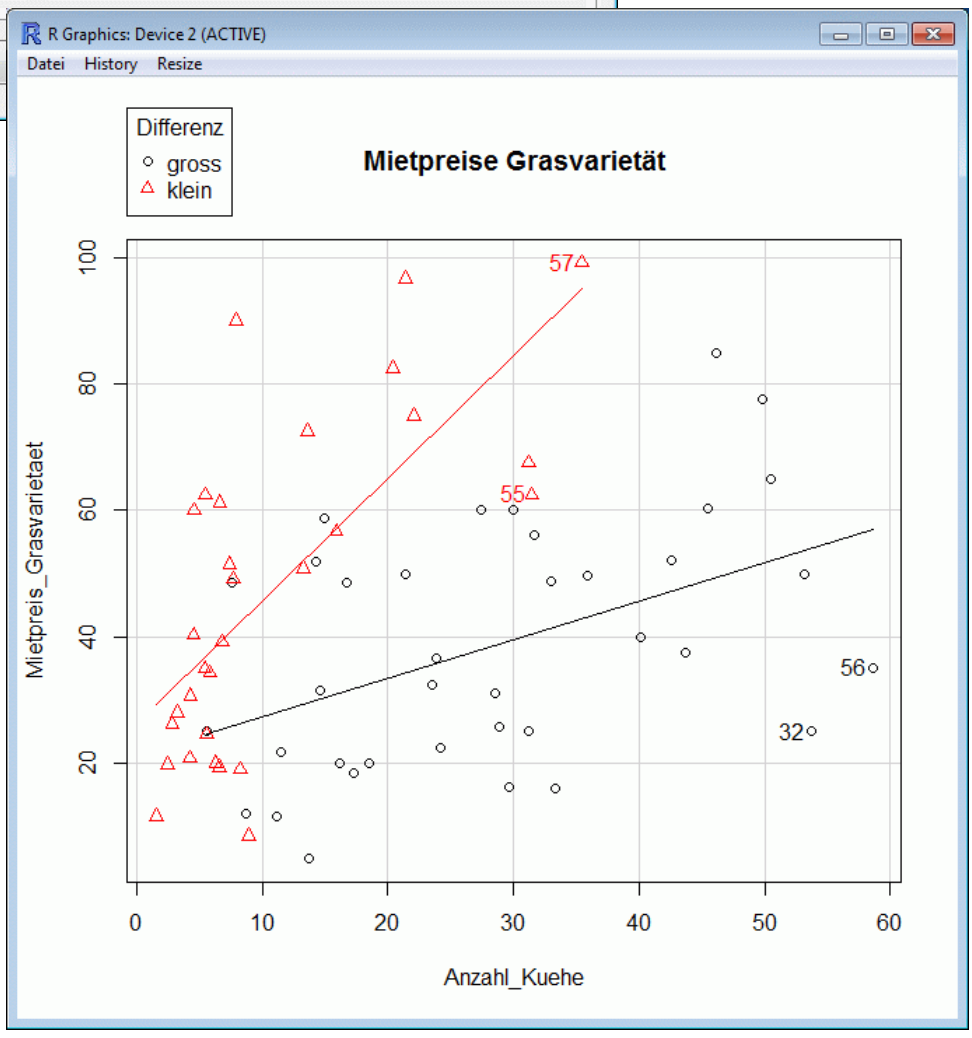
Reiter 'Optionen':



Deaktivieren Sie im Reiter 'Optionen' im Abschnitt Plot Options' folgende Optionen:

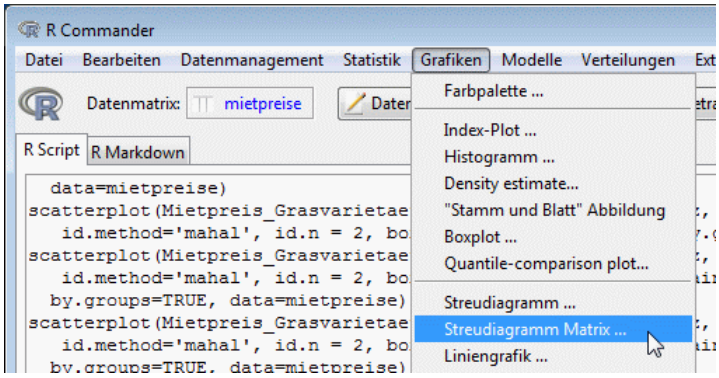
Boxplots an den Achsen
Smooth line
Show spread

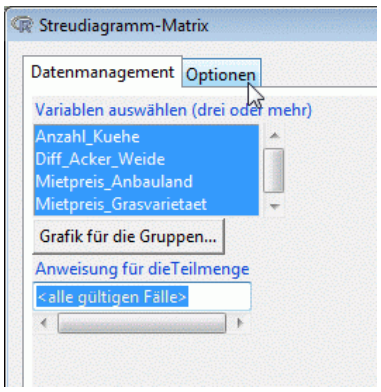
'Apply' für Vorschau der Grafik

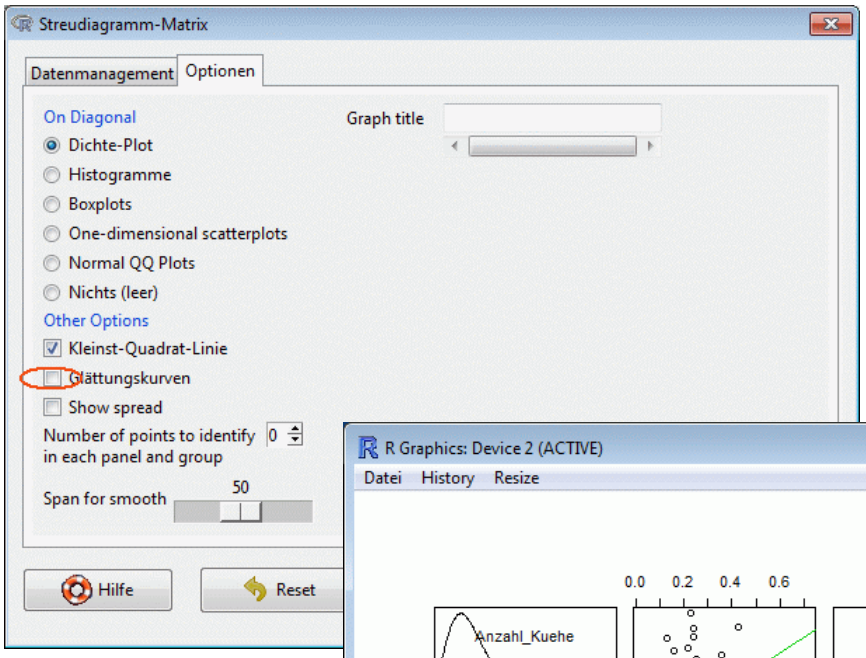


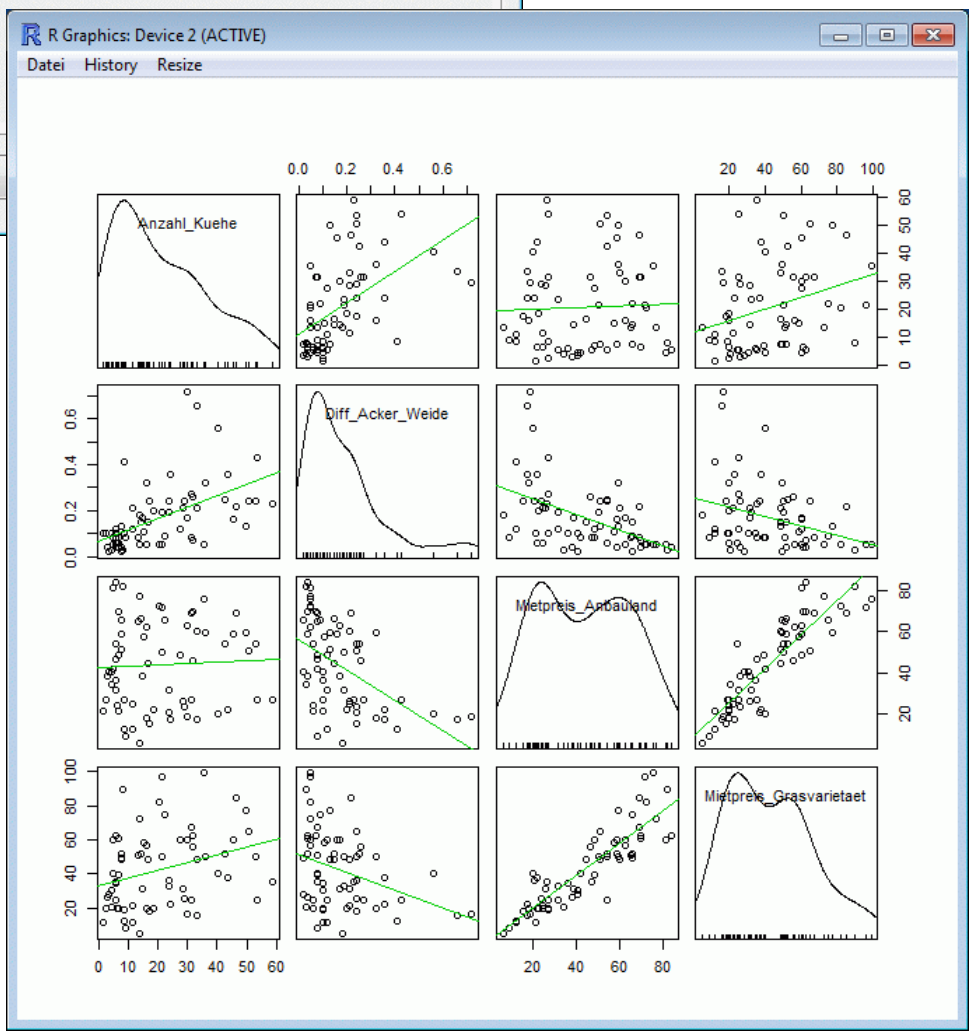
Streudiagramm-Matrix

Um alle möglichen Variablenkombinationen zu plotten, kann man eine Streudiagramm-Matrix bilden.
'Grafiken' > 'Streudiagramm Matrix ...':



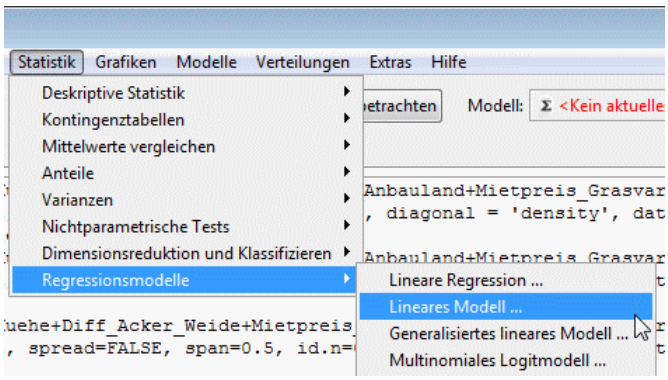


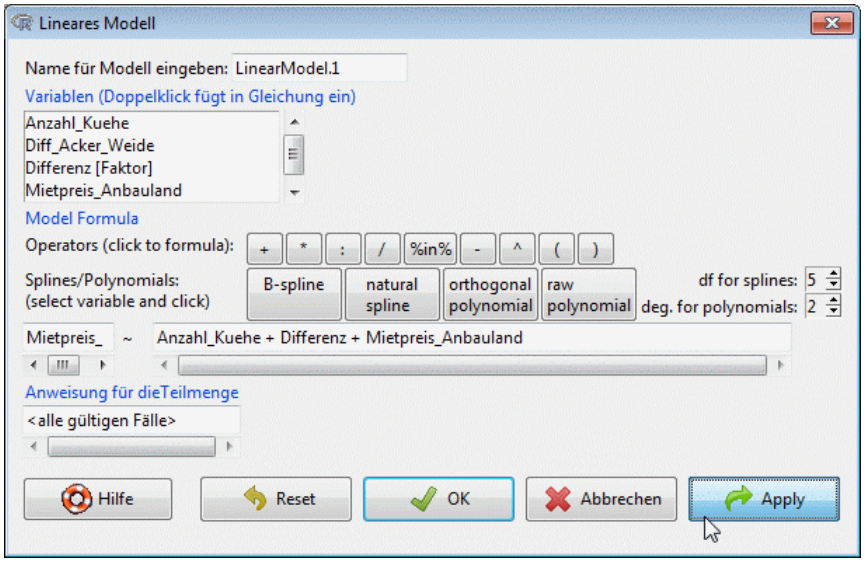




Eine Streudiagrammmatrix kann aber keinen Zusammenhang, der durch eine Linearkombination 2er oder mehr Variablen entsteht aufdecken.
Dazu wird ein lineares Modell gebildet.

Lineares Regressionsmodell





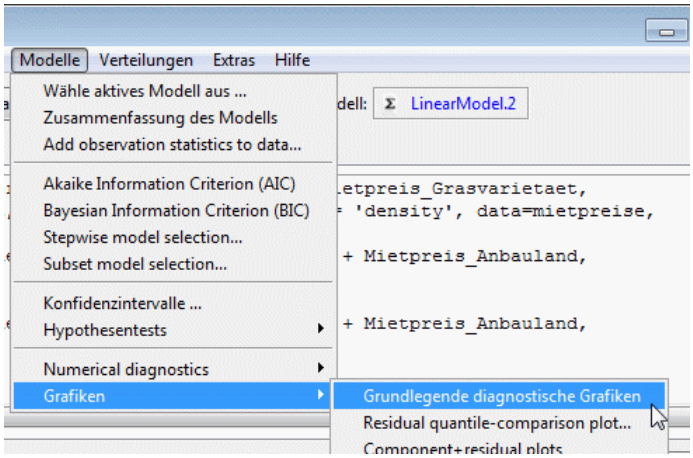

```
Call:
lm(formula = Mietpreis_Grasvarietaet ~ Anzahl_Kuehe + Differenz +
    Mietpreis_Anbauland, data = mietpreise)

Residuals:
     Min       1Q   Median       3Q      Max
-21.9244  -5.9507   0.1114   4.1996  27.7917

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -6.93592    3.17347  -2.186   0.0326 *
Anzahl_Kuehe       0.40708    0.09473   4.297 6.11e-05 ***
Differenz[T.klein] 0.74964    3.13267   0.239   0.8117
Mietpreis_Anbauland 0.92936   0.05999  15.491  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.959 on 63 degrees of freedom
Multiple R-squared:  0.8529, Adjusted R-squared:  0.8459
F-statistic: 121.7 on 3 and 63 DF,  p-value: < 2.2e-16
```

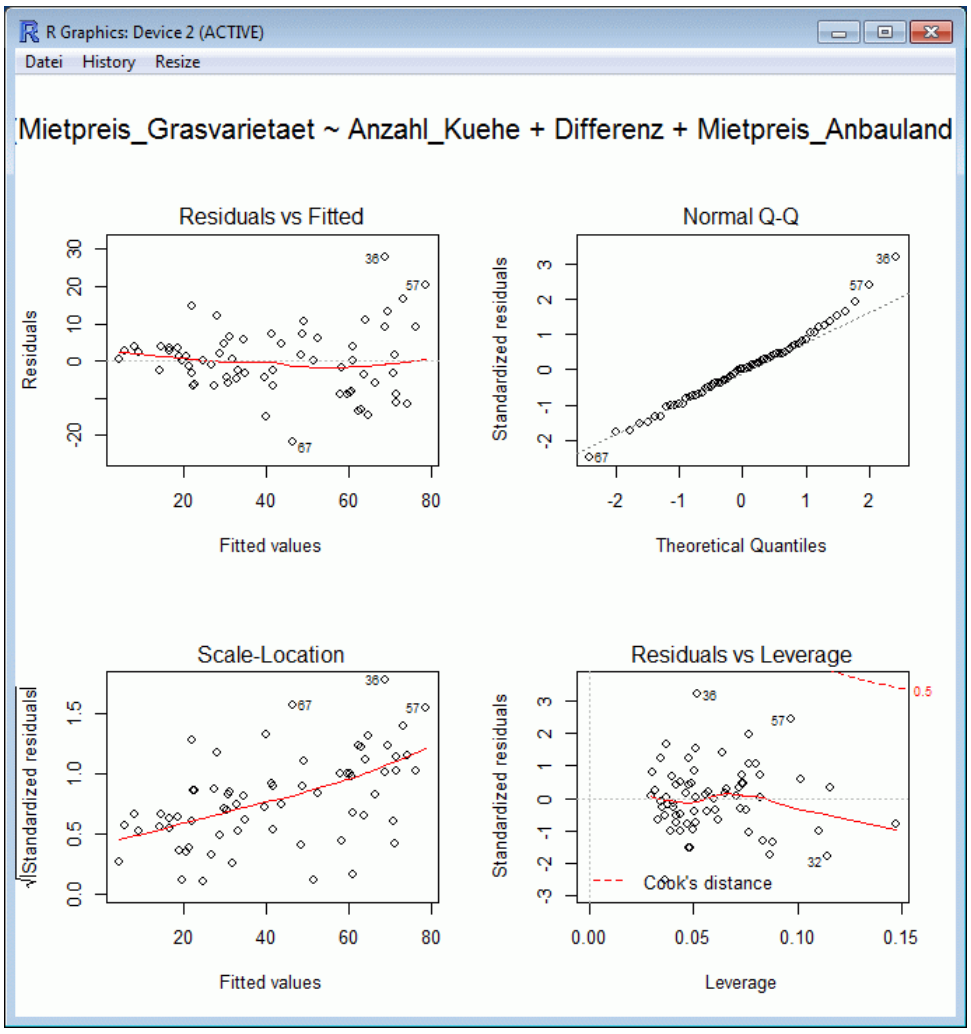
Aus der Tabelle 'Coefficients' können die Werte für die Regressionsparameter abgelesen werden.
alpha (Intercept = Achsenabschnitt) = -6.93592
beta1 = 0.40708, beta2=0.74064, beta3=0.92936
Das Modell lautet also:

```
Mietpreis_Grasvarietaet = -6.94 + 0.40*Anzahl_Kuehe + 0.74*Differenz + 0.93*Mietpreis_Anbauland
```

Diagnostische Plots

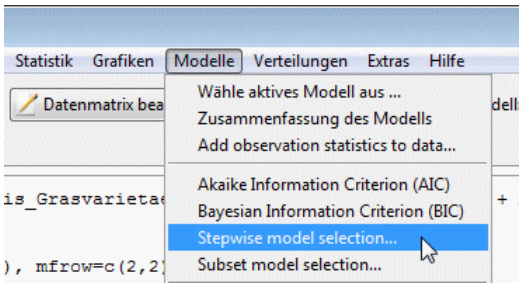
Diagnostische Plots dienen dazu, die Modellvoraussetzungen zu überprüfen.

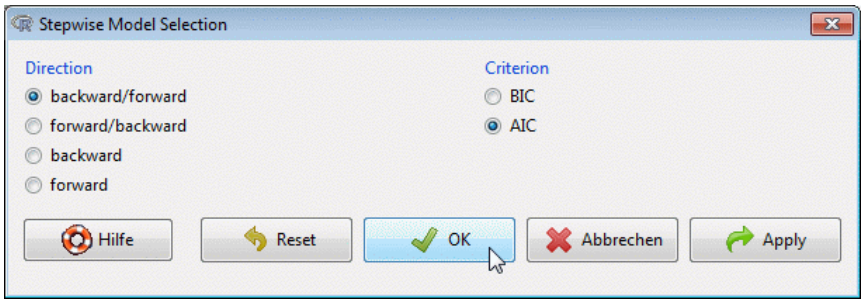




Variablenselektion

Um redundante Variablen aus dem Modell zu entfernen kann eine rückwärts/vorwärts oder vorwärts/rückwärts Selektion auf Basis des AIC durchgeführt werden.



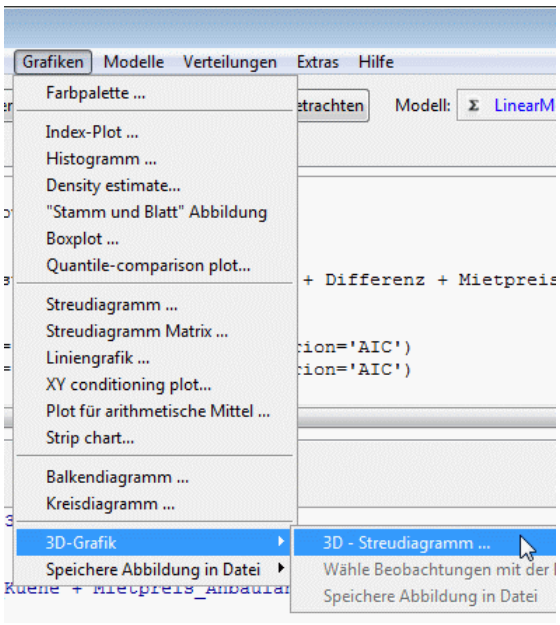


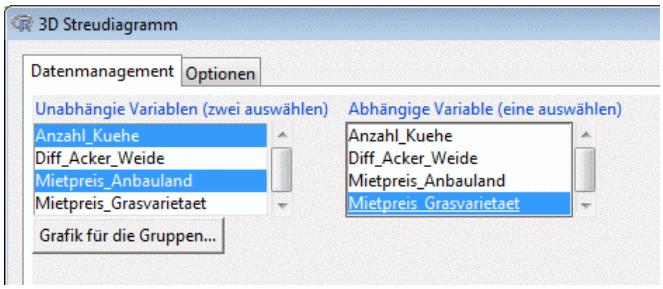

```
Step:  AIC=295.76
Mietpreis_Grasvarietaet ~ Anzahl_Kuehe + Mietpreis_Anbauland

                     Df Sum of Sq     RSS    AIC
<none>                             5061.4 295.76
+ Differenz           1       4.6  5056.8 297.69
- Anzahl_Kuehe        1    2384.5  7445.9 319.62
- Mietpreis_Anbauland 1   26146.5 31207.9 415.63

Call:
lm(formula = Mietpreis_Grasvarietaet ~ Anzahl_Kuehe + Mietpreis_Anbauland,
    data = mietpreise)

Coefficients:
        (Intercept)         Anzahl_Kuehe  Mietpreis_Anbauland
            -6.6127               0.3923               0.9366
```

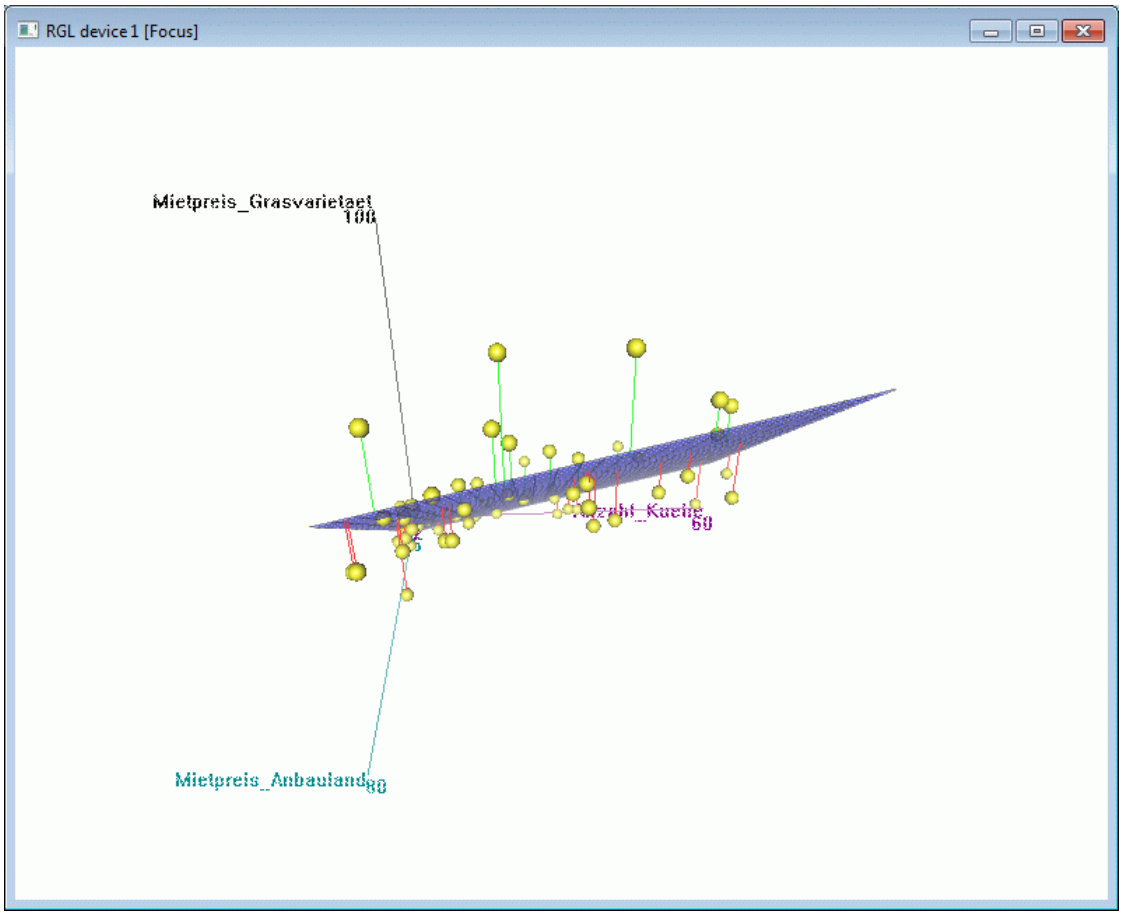
Übrig bleibt ein reduziertest Modell, das die Variablen `Mietpreis_Grasvarietaet` und die erklärenden Variablen `Anzahl_Kuehe` und `Mietpreis_Anbauland` enthält.

3D-Plots für 2 erklärende Variable

Den linearen Zusammenhang der 3 Variablen kann man mit einem 3D Plot betrachten.



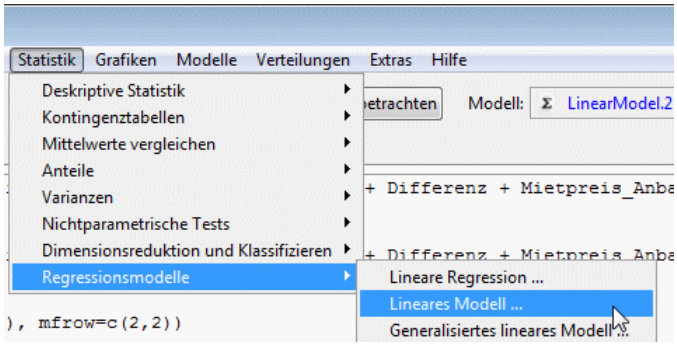




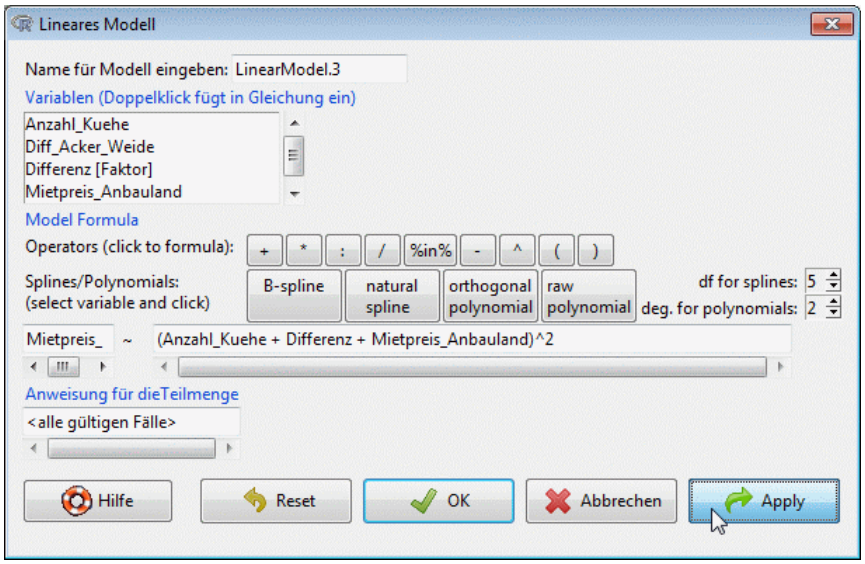
Die Abweichungen von der Fläche sind die Residuen des Modells.

Lineares Modell mit Wechselwirkungen

Manchmal ist es hilfreich auch Wechselwirkungen miteinzubeziehen.



Um alle möglichen Wechselwirkungen von 2 Variablen zu betrachten reicht es aus, die rechte Seite der Modellgleichung zu “quadrieren”.




```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       2.362995   6.518946   0.362   0.7183
Anzahl_Kuehe                      0.120689   0.228927   0.527   0.6000
Differenz[T.klein]               -9.985625   6.430571  -1.553   0.1257
Mietpreis_Anbauland               0.771136   0.171849   4.487 3.33e-05 ***
Anzahl_Kuehe:Differenz[T.klein]   0.498164   0.278209   1.791   0.0784 .
Anzahl_Kuehe:Mietpreis_Anbauland  0.004157   0.005416   0.768   0.4457
Differenz[T.klein]:Mietpreis_Anbauland  0.090378   0.166646   0.542   0.5896
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.479 on 60 degrees of freedom
Multiple R-squared:  0.8745, Adjusted R-squared:  0.8619
F-statistic: 69.68 on 6 and 60 DF,  p-value: < 2.2e-16
```