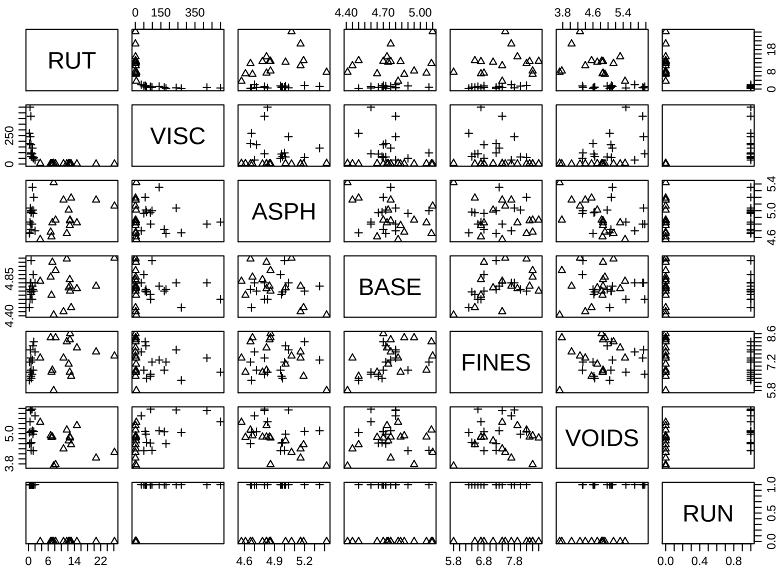
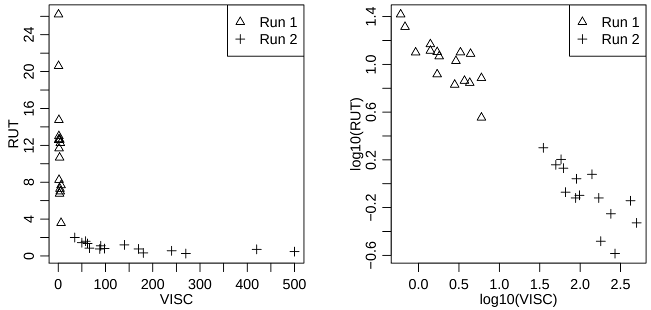
Weiterbildungs-Lehrgang in Angewandter Statistik ETHZ 2009/2011 12.10.07

Angewandte Regression — Musterlösungen zur Serie 4

1. a) In der Scatterplot-Matrix fällt auf, dass `RUT` nicht linear von `VISC` abhängt. Ausserdem sind beide Variablen sehr rechtsschief verteilt (Histogramm zeichnen!) Werden diese Variablen logarithmiert, so ist der Zusammenhang linear. Markiert man den `RUN` mit den Werten 0 mit ◦, so sieht man in den andern Streudiagrammen, dass die `RUT`-Werte höher und die `VISC`-Werte tiefer sind als beim zweiten `RUN`. Es gibt also Unterschiede zwischen den beiden Versuchsreihen.






```
> pairs(d.asp,col=d.asp$RUN+1,pch=d.asp$RUN+2)
> par(mfrow=c(2,2)); hist(d.asp$RUT); hist(d.asp$VISC)
> plot(d.asp$RUT, d.asp$VISC); plot(log(d.asp$VISC),log(d.asp$RUT))
```

b) Regression mit dem vollen Modell. Die Erkenntnis aus a) bestätigt sich: Der Faktor Run ist signifikant. Der Einfluss der Versuchsreihe ist wichtig. Man könnte somit auch getrennt für jeden Run Modelle untersuchen.

```
> d.asp$RUN <- factor(d.asp$RUN)
> r.asp <- regr(log10(RUT) ~ log10(VISC) + ASPH + BASE + FINES + VOIDS + RUN,
                data=d.asp)
> summary(r.asp)

Call:
regr(formula = log10(RUT) ~ log10(VISC) + ASPH + BASE + FINES +
    VOIDS + RUN, data = d.asp)
Terms:
                coef  stcoef signif  R2.x df p.value
(Intercept) -2.5108  0.0000 -1.139    NA  1   0.027
log10(VISC) -0.5133 -0.7989 -3.404 0.701  1   0.000
ASPH         0.4981  0.1783  2.092 0.176  1   0.000
BASE         0.1011  0.0294  0.345 0.175  1   0.483
FINES        0.0189  0.0226  0.267 0.172  1   0.587
VOIDS        0.1375  0.1309  1.391 0.254  1   0.008
RUN         -0.2688 -0.2249 -1.019 0.682  1   0.046
St.dev.error:  0.113   on 24 degrees of freedom
Multiple R^2:  0.972  Adjusted R-squared: 0.965
F-statistic:    140   on 6 and 24 d.f.,  p.value:    0
```

Bemerkung: Eigentlich sollte man die First-Aid-Transformationen $\arcsin(\sqrt(x))$ (in R: `asin(sqrt(x))`) für die Variablen `ASPH`, `BASE`, `FINES` und `VOIDS` anwenden. Da aber die Anteile in % der Variablen sich nicht allzu stark streuen (siehe dazu die Daten), kann man direkt das Modell ohne die Arcussinus-Transformationen anwenden. Dies bestätigt auch der folgende R-Output:

```
>
Call:
regr(formula = log10(RUT) ~ log10(VISC) + asin(sqrt(ASPH/100)) +
    asin(sqrt(BASE/100)) + asin(sqrt(FINES/100)) + asin(sqrt(VOIDS/100)) +
    RUN, data = d.asp)

Terms:
                           coef      stcoef     signif      R2.x df p.value
(Intercept)          -6.1902077  0.00000000 -1.4230051        NA  1  0.0072
log10(VISC)          -0.5099809 -0.79371662 -3.3901134 0.7008491  1  0.0000
asin(sqrt(ASPH/100))  21.8646074  0.18020807  2.1064803 0.1813002  1  0.0002
asin(sqrt(BASE/100))   4.3084738  0.02938097  0.3439493 0.1800831  1  0.4846
```

```
asin(sqrt(FINES/100))  0.9845241  0.02275171  0.2668891 0.1784040  1  0.5868
asin(sqrt(VOIDS/100))  5.9059017  0.13092926  1.3868225 0.2581337  1  0.0086
RUN                   -0.2749975 -0.23009145 -1.0424004 0.6826960  1  0.0417

St.dev.error:  0.1129   on 24 degrees of freedom
Multiple R^2:  0.9724  Adjusted R-squared: 0.9655
F-statistic:   140.7   on 6 and 24 d.f.,  p.value:     0
```

c) Wir betrachten das volle Modell (M0)

$$\log(\texttt{RUT}) = \beta_0 + \beta_1 \log(\texttt{VISC}) + \beta_2\ \texttt{ASPH} + \beta_3\ \texttt{BASE} + \beta_4\ \texttt{FINES} + \beta_5\ \texttt{VOIDS} + \beta_6\ \texttt{RUN}$$

und eliminieren schrittweise die am wenigsten signifikanten Variable. Das Weglassen von mehreren nicht signifikanten Variablen kann zu falschen Ergebnissen führen, da Variablen untereinander abhängig sind: dh, sobald eine nicht signifikante Variable weggelassen wird, kann eine andere nicht signifikante Variable, die mit der weggelassen Variablen korreliert, signifikant werden.
Schrittweise Elimination der Variablen `FINES` (Modell M1) und `BASE` (Modell M2)

```
#Modell M1
r.asp <- regr(log10(RUT) ~ log10(VISC) + ASPH + BASE + VOIDS + RUN,
                data=d.asp)
summary(r.asp)
>
Call:
regr(formula = log10(RUT) ~ log10(VISC) + ASPH + BASE + VOIDS +
    RUN, data = d.asp)

Terms:
                  coef      stcoef    signif      R2.x df p.value
(Intercept) -2.5608978  0.00000000 -1.185124        NA  1  0.0221
log10(VISC) -0.5219729 -0.81238062 -3.602621 0.6937442  1  0.0000
ASPH         0.5023895  0.17979617  2.149975 0.1741936  1  0.0002
BASE         0.1293445  0.03759239  0.481717 0.1150528  1  0.3306
VOIDS        0.1456384  0.13863675  1.573039 0.2164143  1  0.0034
RUN         -0.2626051 -0.21972263 -1.015417 0.6808500  1  0.0468

St.dev.error:  0.1115   on 25 degrees of freedom
Multiple R^2:  0.9719  Adjusted R-squared: 0.9663
F-statistic:   172.9   on 5 and 25 d.f.,  p.value:     0

#Modell M2
r.asp <- regr(log10(RUT) ~ log10(VISC) + ASPH + VOIDS + RUN,
                data=d.asp)
summary(r.asp)

 >
Call:
regr(formula = log10(RUT) ~ log10(VISC) + ASPH + VOIDS + RUN,
    data = d.asp)

Terms:
                  coef     stcoef     signif      R2.x df p.value
(Intercept) -1.7417923  0.0000000 -1.3091929        NA  1  0.0123
log10(VISC) -0.5474748 -0.8520708 -4.0684523 0.6709958  1  0.0000
ASPH         0.4649615  0.1664013  2.1146027 0.1243717  1  0.0002
VOIDS        0.1437032  0.1367946  1.5571013 0.2156756  1  0.0036
RUN         -0.2223544 -0.1860448 -0.9105507 0.6627635  1  0.0725

St.dev.error:  0.1115   on 26 degrees of freedom
Multiple R^2:  0.9708  Adjusted R-squared: 0.9663
F-statistic:     216   on 4 and 26 d.f.,  p.value:     0
```


Obwohl die Variable `RUN` nicht mehr signifikant ist, behalten wir sie in unserem Endmodell:

$$\log(\texttt{RUT}) = \beta_0 + \beta_1 \log(\texttt{VISC}) + \beta_2\ \texttt{ASPH} + \beta_3\ \texttt{VOIDS} + \beta_4\ \texttt{RUN}\,.$$

Ohne die Variable `RUN` erhalten wir das Modell M3:

```
#Modell M3
r.asp <- regr(log10(RUT) ~ log10(VISC) + ASPH + VOIDS,
                data=d.asp)
summary(r.asp)

>
Call:
regr(formula = log10(RUT) ~ log10(VISC) + ASPH + VOIDS, data = d.asp)

Terms:
                  coef     stcoef     signif      R2.x df p.value
(Intercept) -1.5705972  0.0000000  -1.142833        NA  1  0.0266
log10(VISC) -0.6608715 -1.0285575 -12.424347 0.1315141  1  0.0000
ASPH         0.4330037  0.1549642   1.911779 0.1129992  1  0.0005
VOIDS        0.1462267  0.1391968   1.519158 0.2153218  1  0.0043

St.dev.error:  0.1165   on 27 degrees of freedom
Multiple R^2:  0.9668  Adjusted R-squared: 0.9632
F-statistic:   262.5   on 3 and 27 d.f.,  p.value:     0
```

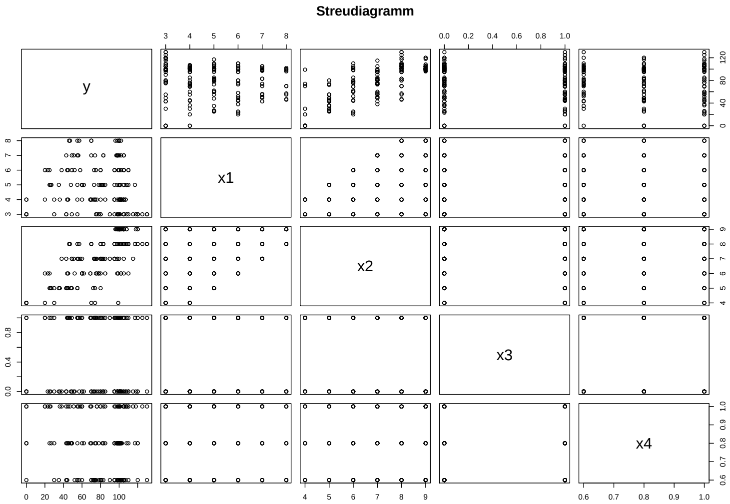
Die Tabelle mit den geschätzxten Standardabweichungen der Fehler $\hat{\sigma}$ und R^2:

Modell	$\hat{\sigma}$	R^2
M0	0.113	0.972
M1	0.1115	0.9719
M2	0.1115	0.9708
M3	0.1165	0.9668

Bemerkung: Falls man alle nicht signifikanten Variablen auf einmal eliminiert, so erkommt direkt auf das Modell M2.

2. a) R-Code:

```
t.url <- "http://stat.ethz.ch/Teaching/Datasets/WBL/cricket.dat"
cricket <- read.table(t.url, header=TRUE)
pairs(cricket)
```



b) R-Code:

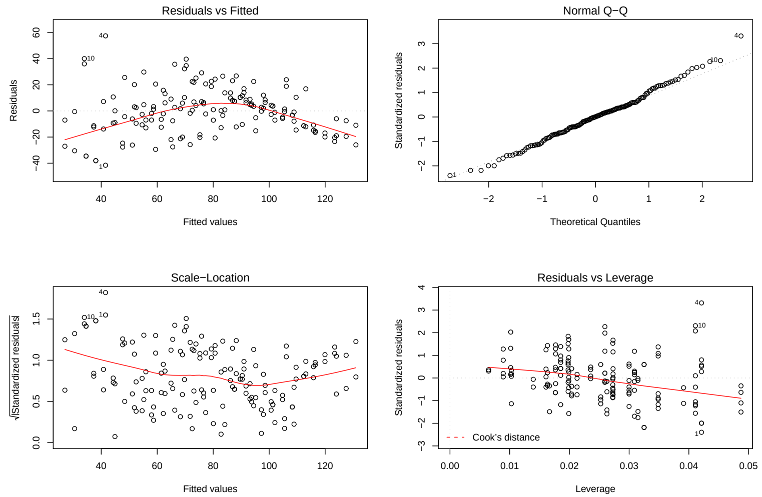
```
r.mod1 <- regr(y~x1 + x2 + x4, data=cricket)
summary(r.mod1)
```

R-Output:

```
Call:
regr(formula = y ~ x1 + x2 + x4, data = cricket)

Terms:
                  coef       stcoef     signif       R2.x df p.value
(Intercept)   3.158791  0.00000000  0.1637925          NA  1  0.7467
x1           -7.567347 -0.38126474 -3.8618856 0.08775209  1  0.0000
x2           17.884014  0.90104813  9.1268467 0.08775209  1  0.0000
x4          -17.403846 -0.09137595 -1.0145928 0.00000000  1  0.0468

St.dev.error:  17.71   on 152 degrees of freedom
Multiple R^2:  0.6841  Adjusted R-squared: 0.6779
F-statistic:   109.7   on 3 and 152 d.f.,  p.value:     0
```

Kommentar:

- Tukey-Ascombe-Plot: der Plot zeigt einen sich in x-Richtung schliessenden Trichter, mit Krümmung nach unten. Dies sieht man schon im Streudiagram für die Variablenpaare (x_1, y) und (x_2, y). Die Glättung selbst sieht aus wie eine nach unten offene Parabel. Eine Transformation der Variablen ist als erstes zu untersuchen.
- QQ-Plot: der Plot sieht ok aus.
- Residual Plots einzelner Variablen: besonders auffällig ist der Residual Plot bezüglich x_2.

c) Da der Residual Plot ein sich schliessender Trichter ist, benützen wir die Quadrat-Funktion als Transformation. Eine Quadrat-Transformation der erklärenden Variablen ist nicht nötig, da die Quadrat-Terme in der nächsten Aufgabe betrachtet werden.

- Zielvariable: der Residual Plot ist nicht mehr so trichter-förmig. Die Variable $x4$ ist nicht mehr signifikant.

Wir fahren mit zwei Modellen weiter. Das Modell 1 ist das Modell mit der quadrierten Zielvariable und das Modell 2 das untransformierte Modell.

Modell 1: Wir wählen also nur die Quadrat-Funktion für die Zielvariable.

```
Call:
regr(formula = I(y^2) ~ x1 + x2 + x4, data = cricket)

Terms:
                  coef      stcoef     signif       R2.x df p.value
(Intercept) -2997.638  0.00000000 -1.1845655         NA  1  0.0206
x1          -1259.258 -0.46938631 -4.8975325 0.08775209  1  0.0000
x2           2444.419  0.91115311  9.5068857 0.08775209  1  0.0000
x4          -1206.538 -0.04686628 -0.5360367 0.00000000  1  0.2913
```

```
St.dev.error:   2324   on 152 degrees of freedom
Multiple R^2:  0.7023  Adjusted R-squared: 0.6965
F-statistic:   119.5   on 3 and 152 d.f.,  p.value:     0
```

d) Modell 1: Keine der quadratischen Terme ist signifikant.

Modell 2: Der $x2^2$-Term ist signifikant.

```
>
Call:
regr(formula = y ~ x1 + x2 + x4 + I(x1^2) + I(x2^2) + I(x4^2),
    data = cricket)

Terms:
                    coef     stcoef     signif      R2.x df p.value
(Intercept)  -47.8936921  0.0000000 -0.4519772        NA  1  0.3732
x1            -5.8794601 -0.2962241 -0.4661358 0.8620025  1  0.3585
x2            43.7583117  2.2046698  2.7017983 0.8925297  1  0.0000
x4          -108.9423077 -0.5719832 -0.4694889 0.9280184  1  0.3551
I(x1^2)       -0.1874833 -0.1003208 -0.1576083 0.8622259  1  0.7559
I(x2^2)       -1.9085718 -1.3045225 -1.6044457 0.8921419  1  0.0018
I(x4^2)       57.2115385  0.4818572  0.3955126 0.9280184  1  0.4357

St.dev.error:  17.24   on 149 degrees of freedom
Multiple R^2:  0.7065  Adjusted R-squared: 0.6947
F-statistic:   59.79   on 6 and 149 d.f.,  p.value:     0
```

e) Modell 1: die Interaktion $x_1 * x_2$ ist signifikant. (Man lässt zuerst die Interaktion $x2 * x4$ weg: eine neue Regression zeigt, dass $x1 * x4$ die Signifikanz verliert. Am Schluss bleibt nur noch $x_1 * x_2$.)

```
>
Call:
regr(formula = I(y^2) ~ x1 + x2 + I(x1 * x2) + I(x1 * x4) + I(x2 *
    x4), data = cricket)

Terms:
                  coef     stcoef     signif      R2.x df p.value
(Intercept)  3071.1672  0.0000000  0.4898528        NA  1  0.3347
x1          -1946.4463 -0.7255346 -1.0348930 0.8773440  1  0.0426
x2            904.1758  0.3370300  0.7634241 0.8052180  1  0.1335
I(x1 * x2)    217.0999  0.8355993  1.1682988 0.8797715  1  0.0223
I(x1 * x4)  -1242.2314 -0.4465065 -0.8389071 0.8384386  1  0.0995
I(x2 * x4)    770.0226  0.3157973  0.7295591 0.8013431  1  0.1515

St.dev.error:   2285   on 150 degrees of freedom
Multiple R^2:  0.7159  Adjusted R-squared: 0.7064
F-statistic:    75.6   on 5 and 150 d.f.,  p.value:     0
```

Modell 2: Die Interaktionen $x_1 * x_2$ und $x_2 * x_4$ sind signifikant.

```
Call:
regr(formula = y ~ x1 + x2 + x4 + I(x2^2) + I(x1 * x2) + I(x1 *
    x4) + I(x2 * x4), data = cricket)
```



```
Terms:
                   coef     stcoef     signif      R2.x df p.value
(Intercept)   57.859929  0.0000000  0.7303935        NA  1  0.1510
x1           -24.316864 -1.2251536 -1.6489711 0.8910723  1  0.0014
x2            31.365073  1.5802627  1.8140596 0.9070952  1  0.0005
x4          -143.321429 -0.7524849 -1.9220919 0.7932754  1  0.0002
I(x2^2)       -3.091552 -2.1130980 -2.4366430 0.9066771  1  0.0000
I(x1 * x2)     2.644049  1.3755464  1.7446119 0.8973546  1  0.0007
I(x1 * x4)    -5.205782 -0.2529173 -0.4827814 0.8455145  1  0.3416
I(x2 * x4)    21.414116  1.1870592  1.9859333 0.8646033  1  0.0001

St.dev.error:  15.91   on 148 degrees of freedom
Multiple R^2:  0.7518  Adjusted R-squared:  0.74
F-statistic:   64.03   on 7 and 148 d.f.,  p.value:     0
```

f) Es stellt sich heraus, dass das folgende Modell brauchbar ist (keine Interaktionen)
R-Code:

```
r.mod1 <- regr(I(y^2)~x1+I(x2-x1)+ I((x2 - x1)^2), data=cricket)
summary(r.mod1)
plot(r.mod1)
```

R-Output:

```
Call:
regr(formula = I(y^2) ~ x1 + I(x2 - x1) + I((x2 - x1)^2), data = cricket)

Terms:
                     coef     stcoef    signif      R2.x df p.value
(Intercept)    -4721.1438  0.0000000 -2.667378        NA  1   0e+00
x1              1185.1611  0.4417668  4.392449 0.1604648  1   0e+00
I(x2 - x1)      3533.6868  1.4312573  5.315595 0.6864120  1   0e+00
I((x2 - x1)^2)  -210.6319 -0.4657168 -1.766397 0.6797481  1   6e-04

St.dev.error:   2244   on 152 degrees of freedom
Multiple R^2:  0.7224  Adjusted R-squared: 0.7169
F-statistic:   131.8   on 3 and 152 d.f.,  p.value:     0
```