A07 Hoermann

Aufgabe 07

a)

- n ... index
- p, e ... regressors (independent variables)
- s, v, d ... dependent variables

```
df = read.csv("regr.csv")
cor(df)
```

```
##              n           p          e           s           v           d
## n  1.000000000 0.005942531 0.00559404 -0.01367667 -0.03169528 0.008111184
## p  0.005942531 1.000000000 0.07811189  0.89399345  0.88275842 0.699126301
## e  0.005594040 0.078111892 1.00000000  0.50275362  0.48441650 0.761721139
## s -0.013676667 0.893993454 0.50275362  1.00000000  0.99398317 0.930266958
## v -0.031695282 0.882758416 0.48441650  0.99398317  1.00000000 0.907741943
## d  0.008111184 0.699126301 0.76172114  0.93026696  0.90774194 1.000000000
```

Interpretation

There is a strong correlation between v & s, d & s, v & d and d & e. The low correlation between e and p makes sense, as those are the two independent variables. The low linear correlation of e towards s and v could mean that they are somehow other connected. Maybe quadratic.

```
Regr = lm(df$s~df$p+df$e+I(df$p^2)+I(df$e^2)+I(df$p*df$e)+I(df$p^2*df$e)+I(df$p*df$e^2)-1)
summary(Regr)
```

```
##
## Call:
## lm(formula = df$s ~ df$p + df$e + I(df$p^2) + I(df$e^2) + I(df$p *
##     df$e) + I(df$p^2 * df$e) + I(df$p * df$e^2) - 1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55295 -0.22161  0.01591  0.24363  0.48227
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## df$p              -0.476346   0.408135  -1.167    0.246
## df$e               0.190604   0.222278   0.858    0.393
## I(df$p^2)          6.358917   0.082726  76.867   <2e-16 ***
## I(df$e^2)         -0.022992   0.019333  -1.189    0.237
## I(df$p * df$e)     6.308444   0.034458 183.074   <2e-16 ***
## I(df$p^2 * df$e) -0.008824   0.007946  -1.110    0.270
## I(df$p * df$e^2)  0.003813   0.004287   0.889    0.376
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2932 on 93 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.702e+07 on 7 and 93 DF,  p-value: < 2.2e-16
```

```
Regr2 = lm(df$s~df$p+I(df$p*df$e)+I(df$p^2*df$e)+I(df$p*df$e^2)+I(df$p*df$d))
summary(Regr2)
```

```
##
## Call:
## lm(formula = df$s ~ df$p + I(df$p * df$e) + I(df$p^2 * df$e) +
##     I(df$p * df$e^2) + I(df$p * df$d))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.75011 -0.35874  0.01729  0.40443  1.50324
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -22.510626   1.049864  -21.44   <2e-16 ***
## df$p              13.168769   0.496660   26.52   <2e-16 ***
## I(df$p * df$e)     4.557513   0.045202  100.83   <2e-16 ***
## I(df$p^2 * df$e)   0.173373   0.004686   37.00   <2e-16 ***
## I(df$p * df$e^2)  -0.045616   0.002156  -21.16   <2e-16 ***
## I(df$p * df$d)     2.540141   0.034626   73.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6108 on 94 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.013e+06 on 5 and 94 DF,  p-value: < 2.2e-16
```

regr2.csv

```
df = read.csv("regr2.csv")
cor(df)
```

```
##              n           p          e           s           v           d
## n  1.000000000 0.005942531 -0.0477928 -0.03476709 -0.05442709 -0.02247575
## p  0.005942531 1.000000000  0.8743364  0.97508578  0.94781179  0.96519940
## e -0.047792799 0.874336427  1.0000000  0.93820670  0.91350344  0.97062132
## s -0.034767093 0.975085781  0.9382067  1.00000000  0.99071976  0.98707027
## v -0.054427091 0.947811794  0.9135034  0.99071976  1.00000000  0.96015193
## d -0.022475749 0.965199396  0.9706213  0.98707027  0.96015193  1.00000000
```

Interpretation

There is a strong linear correlation between all the variables (except the index), as the coefficients are nearly all over 0.9. The correlation between e and p is still high, which is surprising as those are the regressors. The analysis leads to the guess that those variables can be described by a linear function.

```
summary(Regr)
```

```
##
## Call:
## lm(formula = df$s ~ df$p + df$e + I(df$p^2) + I(df$e^2) + I(df$p *
##     df$e) + I(df$p^2 * df$e) + I(df$p * df$e^2) - 1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.55295 -0.22161  0.01591  0.24363  0.48227
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## df$p             -0.476346   0.408135  -1.167    0.246
## df$e              0.190604   0.222278   0.858    0.393
## I(df$p^2)         6.358917   0.082726  76.867   <2e-16 ***
## I(df$e^2)        -0.022992   0.019333  -1.189    0.237
## I(df$p * df$e)    6.308444   0.034458 183.074   <2e-16 ***
## I(df$p^2 * df$e) -0.008824   0.007946  -1.110    0.270
## I(df$p * df$e^2)  0.003813   0.004287   0.889    0.376
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2932 on 93 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.702e+07 on 7 and 93 DF,  p-value: < 2.2e-16
```

The P values are still pretty high, and the t value close to zero which means no good, but the residual standard error is pretty low, thus maybe some intersection is to the wrong degree.

```
Regr2 = lm(df$s~I(df$p*df$d)+I(df$e*df$v)+df$e)
summary(Regr2)
```

```
##
## Call:
## lm(formula = df$s ~ I(df$p * df$d) + I(df$e * df$v) + df$e)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7444 -0.7243  0.1572  1.1389  4.0995
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -3.385e+01  1.620e+00  -20.90   <2e-16 ***
## I(df$p * df$d)  5.689e+00  2.719e-02  209.21   <2e-16 ***
## I(df$e * df$v)  1.322e-03  8.273e-05   15.98   <2e-16 ***
## df$e            9.190e+00  2.438e-01   37.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.89 on 96 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 3.019e+05 on 3 and 96 DF,  p-value: < 2.2e-16
```

The output above shows that the values considered vor the estimation fit rather well, but the residual standard error is still a bit high, which could mean that some intersection is missing or to the wrong degree.