Regression

Lorenz Gygax

2020-02-13

A first regression model

Here, we use an example where crop output of brussels sprouts ("Ertrag") has been measured in dependence of the distance between the plants ("Standweite") and we assume that the measurements were all independent.

Install and load the package `contrast`.

```
library (contrast)
```

Then load the data and estimate the regression curve. In addition, we are using the command `contrast` to estimate the confidence interval of the regression. We check the summary of the regression model and calculate an anova table.

```
rkohl.df <- read.table ('Rosenkohl.csv', header= TRUE, sep= ';', dec= '.')
rkohl.lin <- lm (Ertrag ~ Standweite, rkohl.df)
rkohl.contr <- contrast (rkohl.lin, list (Standweite = seq (30, 48, len= 50)))
summary (rkohl.lin)
```

```
##
## Call:
## lm(formula = Ertrag ~ Standweite, data = rkohl.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.300  -9.787   5.875  12.262  21.600
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.4500    22.6620  -0.196    0.847
## Standweite    3.0917     0.5727   5.399 9.38e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.37 on 14 degrees of freedom
## Multiple R-squared:  0.6755, Adjusted R-squared:  0.6523
## F-statistic: 29.15 on 1 and 14 DF,  p-value: 9.383e-05
```

```
drop1 (rkohl.lin, test= 'F')
```

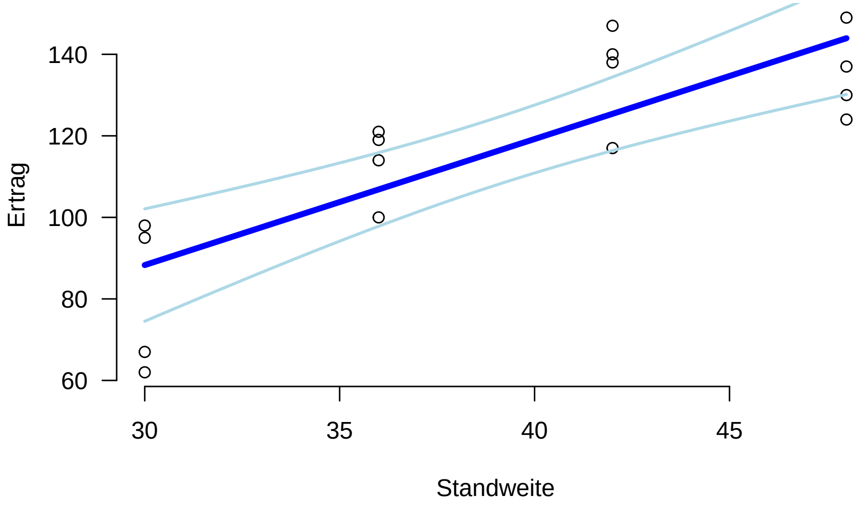
```
## Single term deletions
##
## Model:
## Ertrag ~ Standweite
##            Df Sum of Sq     RSS     AIC F value    Pr(>F)
## <none>                   3305.7  89.293
## Standweite  1      6882 10187.8 105.302  29.146 9.383e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note the formula for the regression line (the equation for the expected values.

Can a slope unequal to zero be supported statistically? What is the basis for your conclusion?

Let us have a look at the data and the estimated model:

```
par (las= 1, bty= 'n')
plot (Ertrag ~ Standweite, rkohl.df)
lines (seq (30, 48, len= 50), rkohl.contr [['Contrast']], lwd= 4, col= 'blue')
lines (seq (30, 48, len= 50), rkohl.contr [['Lower']], lwd= 2, col= 'lightblue')
lines (seq (30, 48, len= 50), rkohl.contr [['Upper']], lwd= 2, col= 'lightblue')
```



How do you judge the fit of the estimated line based on this figure?

A simple possibility for non-linearity

It could be guessed that the relationship is non-linear. We check this here using a polynomial.

```
rkohl.df [, 'SW2'] <- rkohl.df [, 'Standweite']^2
rkohl.qua <- lm (Ertrag ~ Standweite + SW2, rkohl.df)
rkohl.contr <- contrast (rkohl.qua, list (Standweite = seq (30, 48, len= 50),
                                          SW2= seq (30, 48, len= 50)^2))
pred.df <- data.frame (pred= rkohl.contr [['Contrast']] [rkohl.contr [['SW2']] ==
                                                         rkohl.contr [['Standweite']]^2],
                       low= rkohl.contr [['Lower']] [rkohl.contr [['SW2']] ==
                                                     rkohl.contr [['Standweite']]^2],
                       upp= rkohl.contr [['Upper']] [rkohl.contr [['SW2']] ==
                                                     rkohl.contr [['Standweite']]^2])
summary (rkohl.qua)
```

```
##
## Call:
## lm(formula = Ertrag ~ Standweite + SW2, data = rkohl.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.925 -11.912   2.600   7.975  18.075
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -347.8250   134.2064  -2.592  0.02235 *
## Standweite    21.2375     7.0365   3.018  0.00989 **
## SW2           -0.2326     0.0900  -2.585  0.02265 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.96 on 13 degrees of freedom
## Multiple R-squared:  0.7857, Adjusted R-squared:  0.7527
## F-statistic: 23.83 on 2 and 13 DF,  p-value: 4.487e-05
```

```
drop1 (rkohl.qua, test= 'F')
```

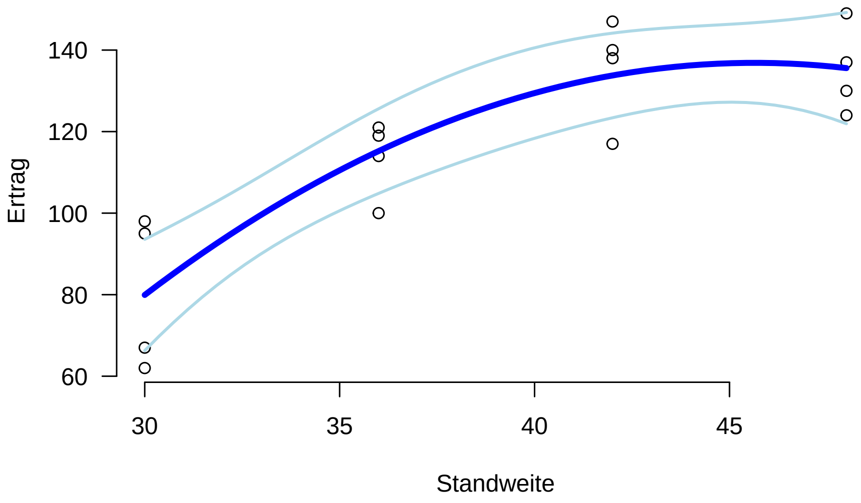
```
## Single term deletions
##
## Model:
## Ertrag ~ Standweite + SW2
##            Df Sum of Sq    RSS    AIC F value   Pr(>F)
## <none>                  2183.4 84.657
## Standweite  1    1530.0 3713.5 91.154  9.1094 0.009887 **
## SW2         1    1122.2 3305.7 89.293  6.6817 0.022645 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note the formula of the extended regression model now.

Can a non-linear component be supported statistically? What is the basid of your conclusion?

Grphically, this looks as follows:

```
par (las= 1, bty= 'n')
plot (Ertrag ~ Standweite, rkohl.df)
lines (seq (30, 48, len= 50), pred.df [['pred']], lwd= 4, col= 'blue')
lines (seq (30, 48, len= 50), pred.df [['low']], lwd= 2, col= 'lightblue')
lines (seq (30, 48, len= 50), pred.df [['upp']], lwd= 2, col= 'lightblue')
```


Ertrag
Standweite