Statistical Computing
Week 7: 15. November 2019
Regression 2

Statistical Models: Regression part 2

This week

- R^2 statistic.
- p-values.
- Quadratic regression
- Multiple regression

Reminder from last week

- Simple linear regression: does X have an influence on Y?
- Model the true influence using $y = f(x) = b_0 + b_1 x$.
- The true underlying function is unkown, we only know the data x_i and y_i
- This function is estimated $\widehat{y}_i = \widehat{b}_0 + \widehat{b}_1 x_i$ by minimising the sum of squared residuals (method of least squares).
- Least squares method gives a formulae for $\widehat{b}_0$ and $\widehat{b}_1$.
- The fitted values are $\widehat{y}_i$.
- the residuals are $\widehat{\epsilon}_i = y_i - \widehat{y}_i$.
- Fitted in R using the function `lm()`.

Coefficient of Determination, Goodness of Fit, R^2

How well does the regression line fit the data?

It can be shown that

Variance of data= Variance of fitted values+Variance of residuals

Equivalently

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

Once we have the data, the variance of the data is fixed.
The variance of the residuals is minimised when we get the least squares estimates.

We want the 1st term to be large and the 2nd term to be small.

Dividing both sides by the variance of the data we get a measure of how good the model fits the data on a scale of 0 to 1.

$$1 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} + \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

The first term is the R^2 statistic, the ratio $\dfrac{\text{Variance of the fitted values}}{\text{Variance of Y}}$.
It is a measure of the "model fit".

Definition: The coefficient of determination or R^2 is

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{s^2_{\widehat{Y}}}{s^2_Y}$$

Properties:

$$R^2 = r_{XY}^2 \qquad \text{the correlation squared}$$

$$0 \leqslant R^2 \leqslant 1$$

When all the data points lie on a straight line then $R^2 = 1$

When X has no influence on Y then R^2 is near to zero.

Grasshopper example:

$s_{\hat{Y}}^2 = 2.018729$ $\qquad$ $s_Y^2 = 2.896952$

$$R^2 = \frac{2.018729}{2.896952} = 0.6968 \approx 0.7.$$

On a scale from 0 to 1 the model fit is *good*.

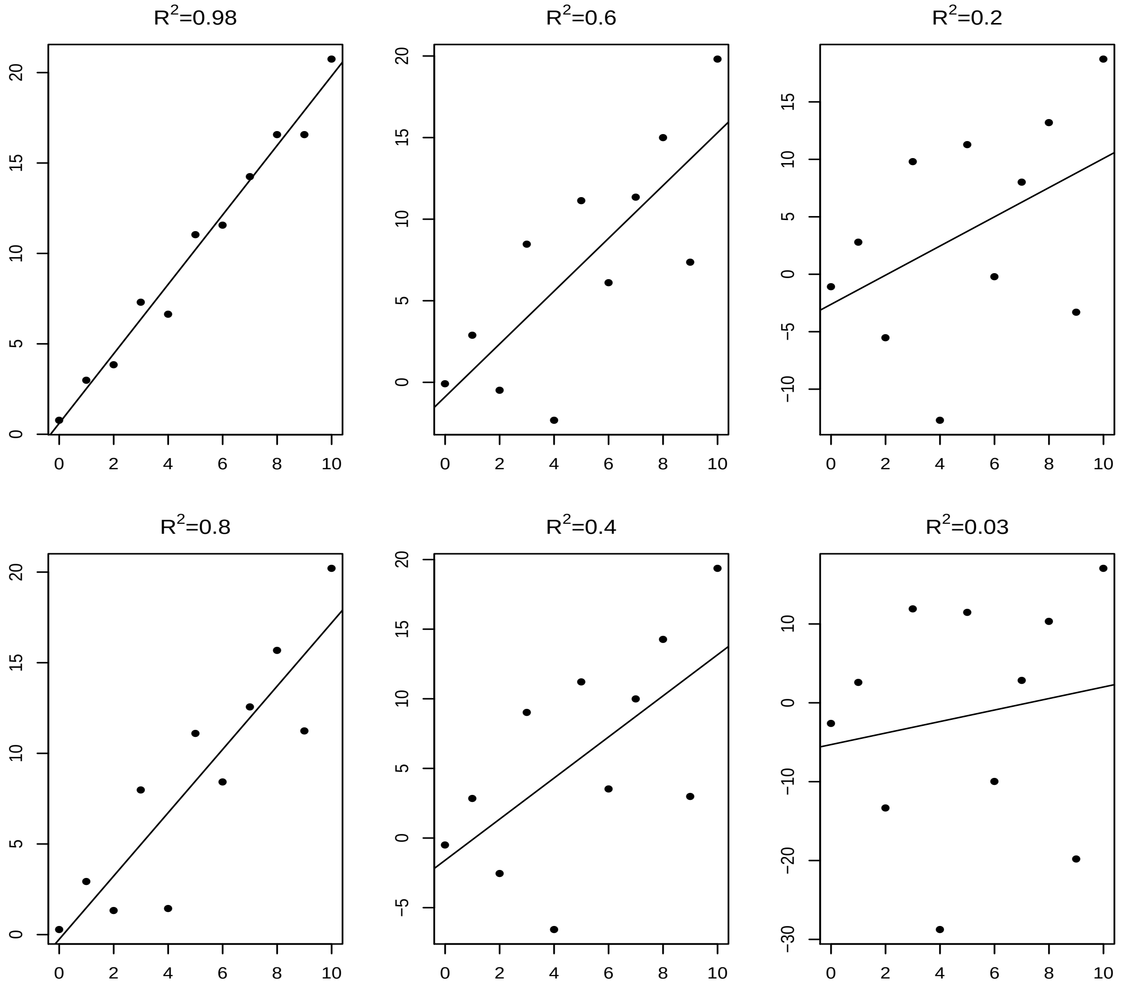
Output from the grasshopper regression model

```
> lm.obj<-lm(formula = chirp ~ temp, data = Grasshoppers)
> summary(lm.obj)
Call: lm(formula = chirp ~ temp, data = Grasshoppers)

Residuals:
     Min       1Q   Median       3Q      Max
-1.54879 -0.58426  0.01574  0.60056  1.53880

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.882  0.908     13.084 7.36e-09 ***
Temp         0.382  0.069      5.466 0.000108 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
Residual standard error: 0.9725 on 13 degrees of freedom
Multiple R-squared: 0.6968, Adjusted R-squared: 0.6735
F-statistic: 29.88 on 1 and 13 DF, p-value: 0.0001081
```

Graphical Examples of R^2



Warning!

R^2 is a popular statistic but is easy to over-use.

In a simple linear regression it is useful to know how good the model fit is. However, when choosing between several models, blindly choosing the one with the largest R^2 can quickly lead to *over fitting*. Over fitting is a common problem when analysing large datasets.

You will learn better ways of model development in the courses Regression and Machine Learning 1.

Significance of a variable

In the `lm` model output coefficient table, there is a column `Pr(>|t|)`. This column contains the so called *p*-values for each estimate.

p-values are part of hypothesis testing and help us decide whether that predictor variable has an influence over the outcome variable. You will learn about this later in the course.

For now we will use the decision rule:
If the *p*-value $\leqslant 0.05$ then we call the associated variable *statistically significant* and we conclude that the variable does have an influence over the outcome variable.

If the *p*-value > 0.05 then we call the associated variable *not statistically significant* and we conclude that the variable does not have an influence over the outcome variable.

Quadratic Regression

In quadratic regression a quadratic function of x is fitted to the data:

$$f(x) = b_0 + b_1 x + b_2 x^2$$

The best function is again defined by minimising the residual sum of squares, as with linear regression.

The model residuals are:

$$\widehat{\epsilon}_i = y_i - \widehat{f}(x_i) = y_i - (\widehat{b}_0 + \widehat{b}_1 x_i + \widehat{b}_2 x_i^2).$$

The model estimates $\widehat{b}_0$, $\widehat{b}_1$, $\widehat{b}_2$ minimise $\sum_{i=1}^{n} \widehat{\epsilon}_i^2$.

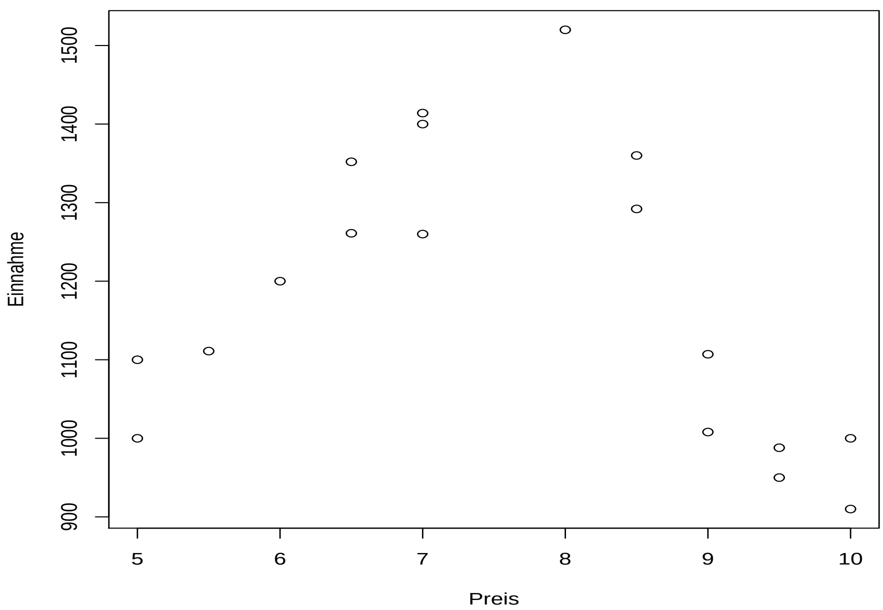
The fitted values are:

$$\widehat{y}_i = \widehat{b}_0 + \widehat{b}_1 x_i + \widehat{b}_2 x_i^2.$$

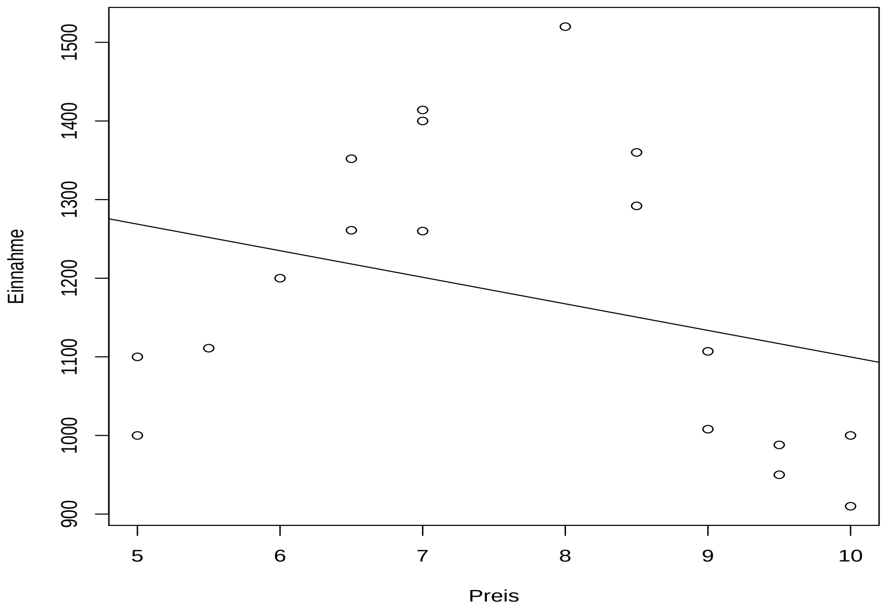
Example

The management of a museum varied the entrance price of an exhibition in order to asses the influence of price on the daily turnover. The entry price x_i in Euros is the independent variable and the daily taking y_i in Euros is the dependent variable. The data are:

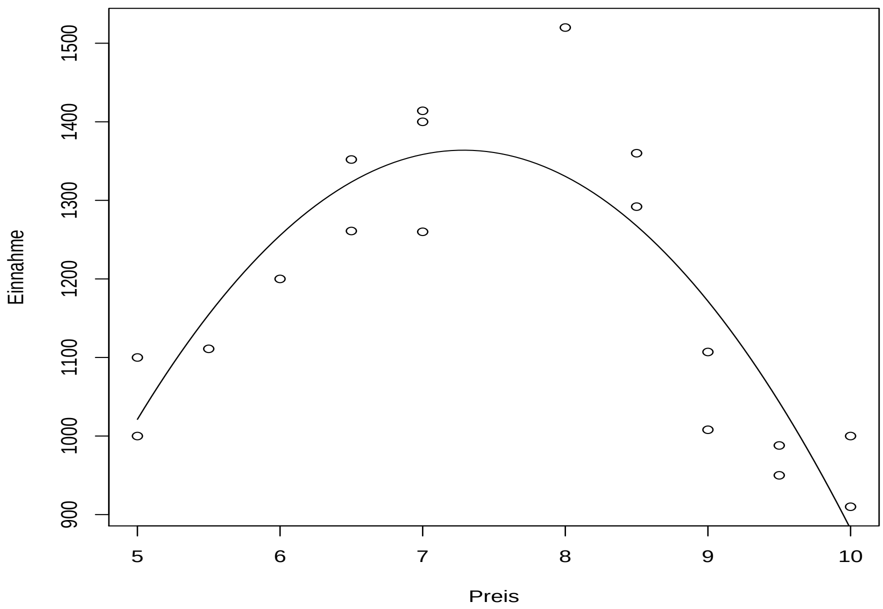
(5 ;1000)	(5 ;1100)	(5.5;1111)
(6 ;1200)	(6.5;1352)	(6.5;1261)
(7 ;1260)	(7 ;1400)	(7 ;1414)
(8 ;1520)	(8.5;1360)	(8.5;1292)
(9 ;1107)	(9 ;1008)	(9.5; 950)
(9.5; 988)	(10 ;1000)	(10 ; 910)



A straight line does not fit the data at all well

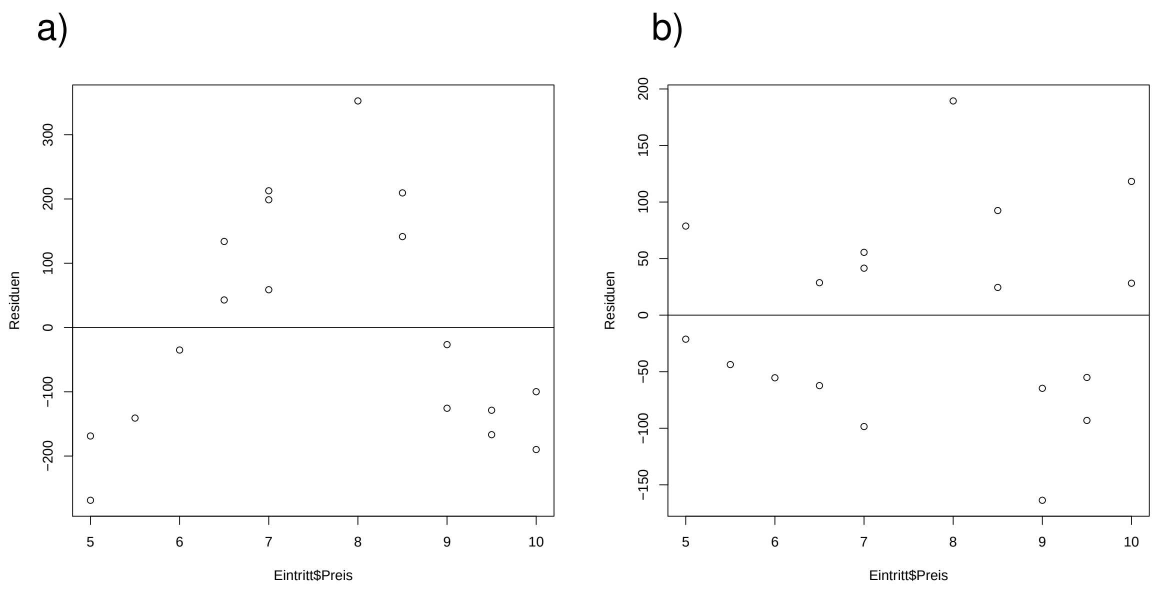


A quadratic regression fits much better.



$$F(x) = -2114 + 954.7x - 65.50x^2$$

Residual plots for a) linear and b) quadratic regression



A residual plot has the residuals on the *y* axis and either the predictor variable or the fitted values on the x axis. If the model is appropriate the the residual plot shows no *obvious pattern*.

Quadratic regression notation in R

To fit a quadratic regression in R, use a command with the general form:

```
> lm.obj2<-lm(y~x+I(x^2),data=dataframename)
```

The first argument is an R *formula*.
Note that the quadratic term has `I(x^2)`. The purpose of the `I()` function is to force the calculation to be done before the variable is added to the formula.

Also note that it is not necessary to specify the intercept term. R assumes that you want the intercept automatically

Multiple Regression

More than one variable can have an influence on the dependent variable.

The aim is to choose a model which best summarises which variables influence the outcom variable and by how much.

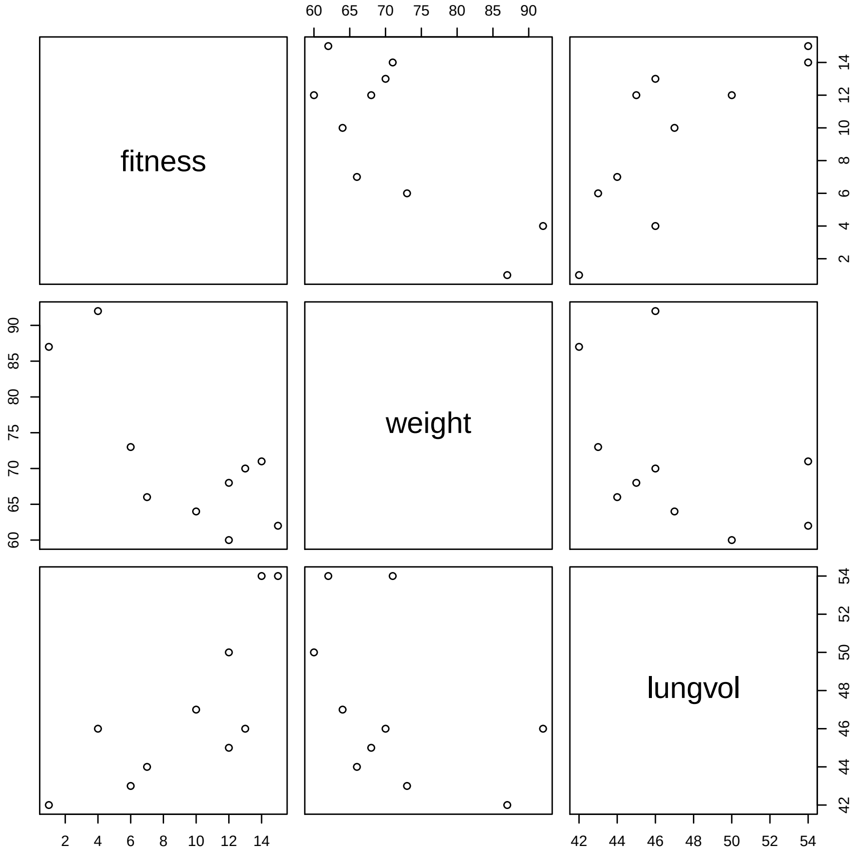
Example: two predictor variables

There are three variables relating to 10 people X_1 is weight in Kg, X_2 is lung volume in decilitre (dl) and the outcome variable Y is a fitness rating on a scale from 0 to 15.

	1	2	3	4	5
X_1 Weight	87	73	66	62	68
X_2 lung volume	42	43	44	54	45
Y Fitness	1	6	7	15	12

	6	7	8	9	10
X_1 Weight	92	60	70	71	64
X_2 lung volume	46	50	46	54	47
Y Fitness	4	12	13	14	10

Matrix plot or pairs plot



If we look at simple linear regression with Y dependent on X_1 (weight) the regression function is:

$$\widehat{f}_1(X_1) = 33.781 - 0.342X_1.$$

Simple linear regression with Y dependent on X_2 (lung volume) gives

$$\widehat{f}_2(X_1) = -30.67 + 0.851X_2.$$

If we fit the two predictor variables simultaneously we get

$$\widehat{f}_{1,2}(X_1, X_2) = -1.786 - 0.232X_1 + 0.589X_2.$$

The R command is:

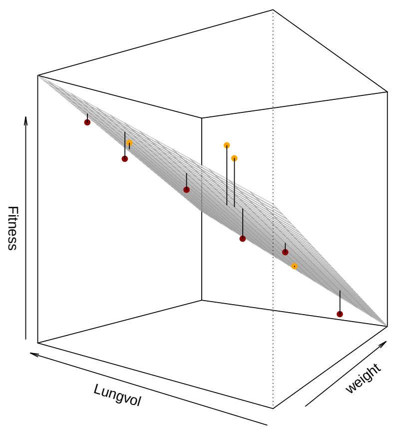
```
lm.fit<-lm(fitness~weight+lungvol,data=fit)
```

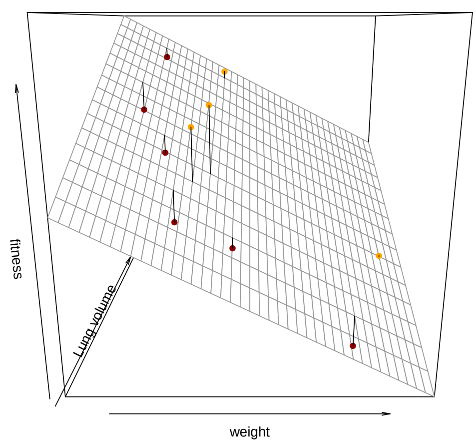
R output

```
Call:
lm(formula = fitness ~ weight + lungvol, data = fit)
Residuals:
    Min      1Q  Median      3Q     Max
-1.8049 -1.5608 -0.6082  0.3628  3.9461
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.7860    13.1958  -0.135   0.8961
weight       -0.2324     0.0816  -2.848   0.0248 *
lungvol       0.5893     0.2008   2.935   0.0219 *
--
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2.279 on 7 degrees of freedom
Multiple R-squared:  0.8149, Adjusted R-squared:  0.762
F-statistic: 15.41 on 2 and 7 DF,  p-value: 0.002728
```

The p-value for both weight and lung volume is statistically significant at the 5% level. This suggests that both variables have an influence on fitness.

Instead of a regression line the regression function is a surface (function in thwo dimensions).





$$\widehat{f}_{1,2}(X_1, X_2) = -1.786 - 0.232X_1 + 0.589X_2.$$