Regression

Flervariabel & Polynom

RESTORY...

Agenda

- Linjär regression med flera variabler
- Kodstruktur
- Kodexempel
- Polynomregression
- Kodexempel
- Statistik
    - p-value
    - Multivariable regression - variable selection
    - R_2
- Övningar

Linjär Regression med flera variabler

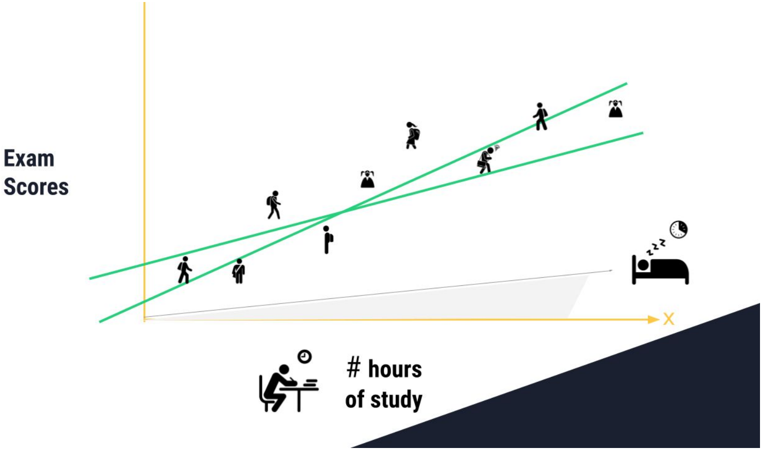
Flervariabel Linjär Regression

$$\boxed{y = b + m_0x_0 + m_1x_1 + m_nx_n}$$

y är nu beroende av två, eller fler, variabler
(years of experience AND years of education)

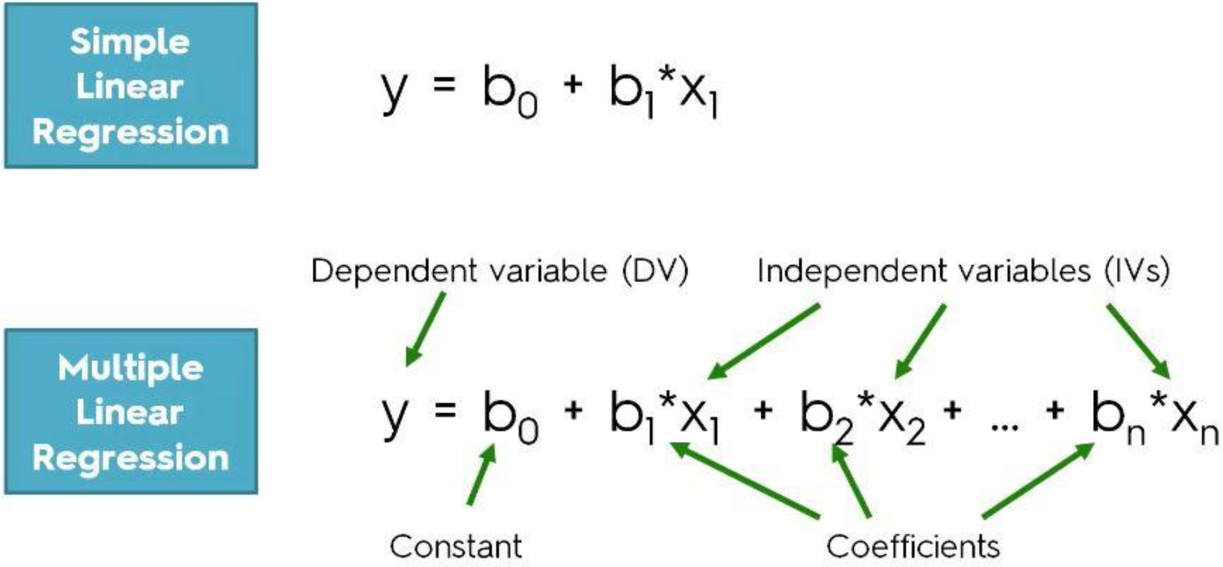
Varje ny variabel är också oberoende.

Har sin egen riktningskoefficient.



https://images.app.goo.gl/ZuAVS1F2JuiTXN1B8

Flervariabel LR



Simple Linear Regression

$$y = b_0 + b_1{}^*x_1$$

Multiple Linear Regression

Dependent variable (DV)

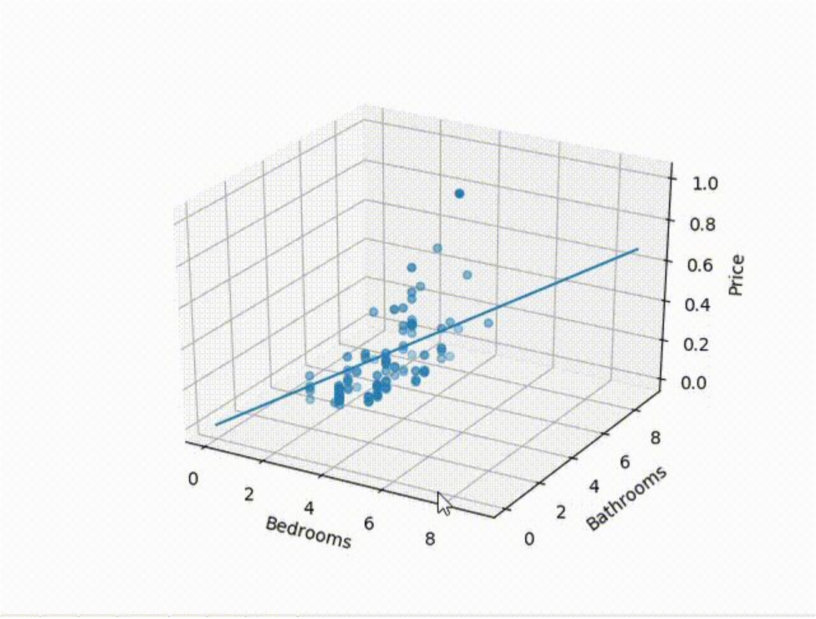
Independent variables (IVs)

$$y = b_0 + b_1{}^*x_1 + b_2{}^*x_2 + \dots + b_n{}^*x_n$$

Constant

Coefficients

Flervariabler - två oberoende variabler - 3D



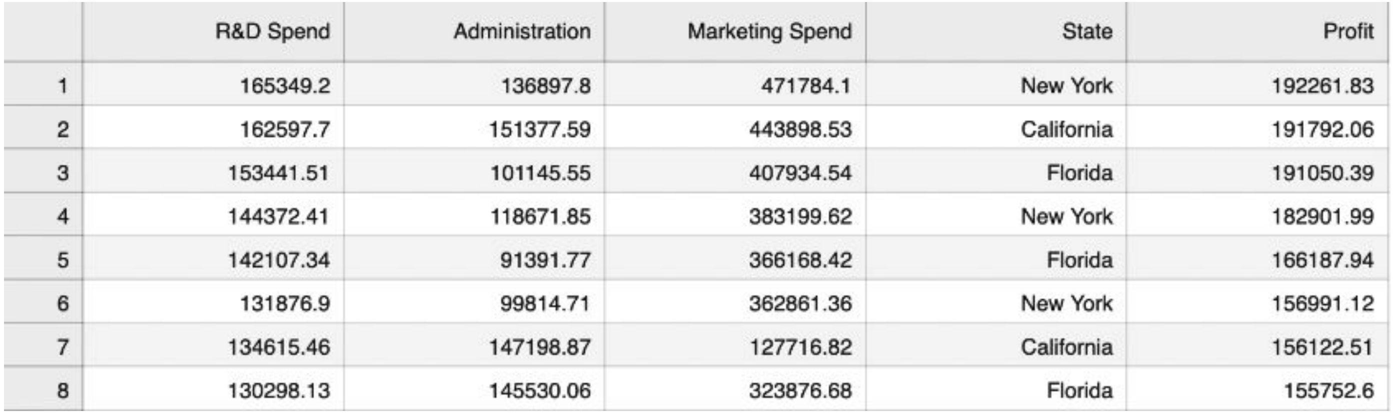
bedrooms
bathrooms

price

The dummy variable trap

- Ordinal to numerical
- Third can be inferred by the absence of the other two (included in b_0).
- sklearn model is advanced and will figure that out

$$y = b_0 + b_1{*}x_1 + b_2{*}x_2 + b_3{*}x_3 + b_4{*}D_1$$



	R&D Spend	Administration	Marketing Spend	State	Profit
1	165349.2	136897.8	471784.1	New York	192261.83
2	162597.7	151377.59	443898.53	California	191792.06
3	153441.51	101145.55	407934.54	Florida	191050.39
4	144372.41	118671.85	383199.62	New York	182901.99
5	142107.34	91391.77	366168.42	Florida	166187.94
6	131876.9	99814.71	362861.36	New York	156991.12
7	134615.46	147198.87	127716.82	California	156122.51
8	130298.13	145530.06	323876.68	Florida	155752.6

NY = D_1
CA = D_2

D_1 = 1- D_2

b_0 “includes” D_2

Kodstruktur

1. Import libraries
2. Import dataset
3. Encode categorical variables
4. Split data from “answer” column
5. Split train/tests
    a. Import “splitter”
6. Training
    a. Import model?
7. Testing

Print options

- `np.set_printoptions(precision=2)`

Kod

RESTORY...

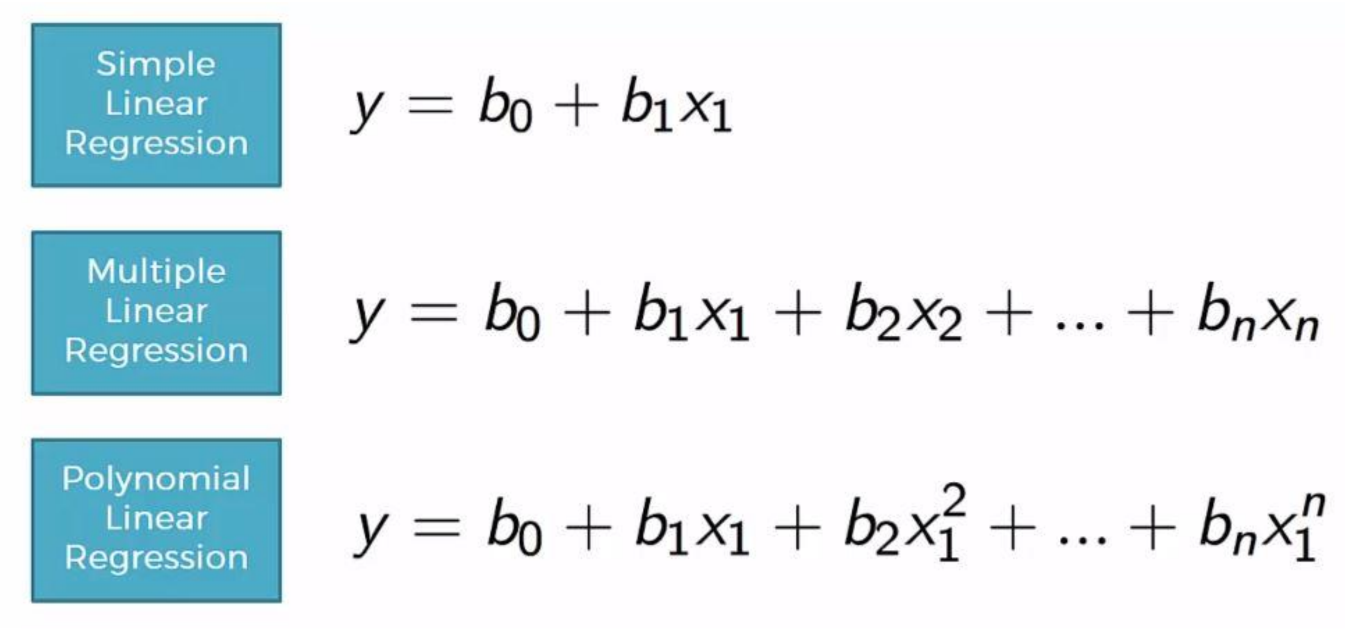
Övningar

- Ladda ned Iris data
- Omvandla nominell data till numeriska kategorier
- Gör en flervariabelregression för att förutspå Sepal längden.

- Ladda ned Pima indians diabetes data
- Gör en flervariabelregression.

Polynom

Polynom Linjär Regression



Simple Linear Regression	$y = b_0 + b_1 x_1$
Multiple Linear Regression	$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$
Polynomial Linear Regression	$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$

Polynom LR

$$y = b + m_0x_1 + m_1x_1^2 + \ldots + m_nx_1^n$$

y är beroende av x

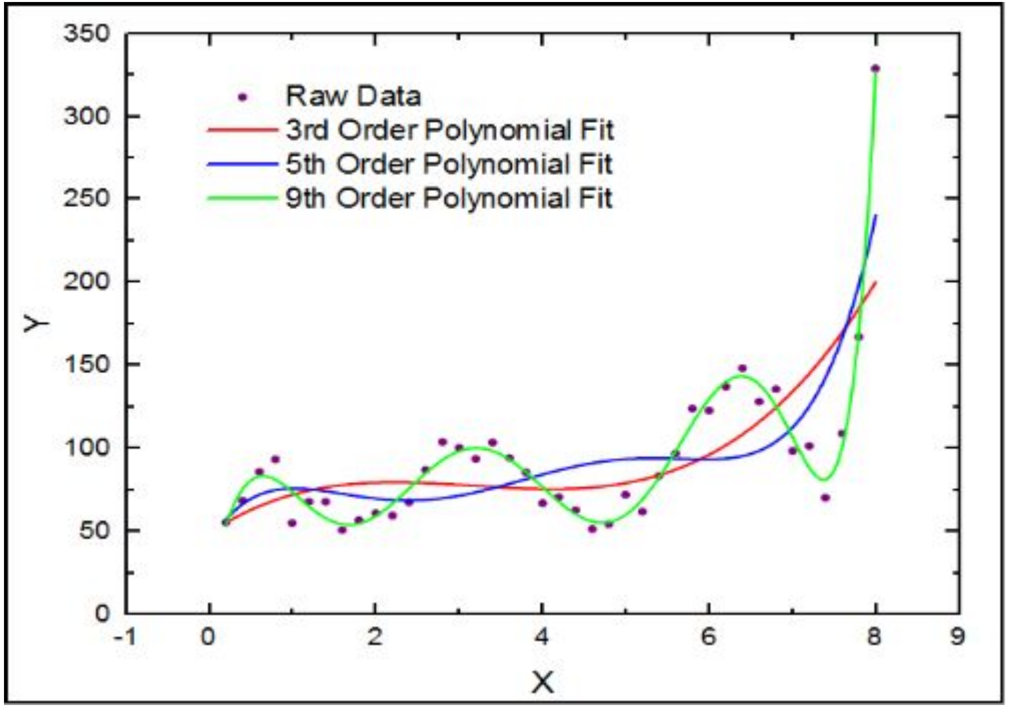
x är oberoende

förhållandet mellan dem som en n-graders polynom av x (relationship between them as a n-th degree polynomial of x)

högre grad -> mer komplex modell

Kan modellera mer komplexa förhållanden.

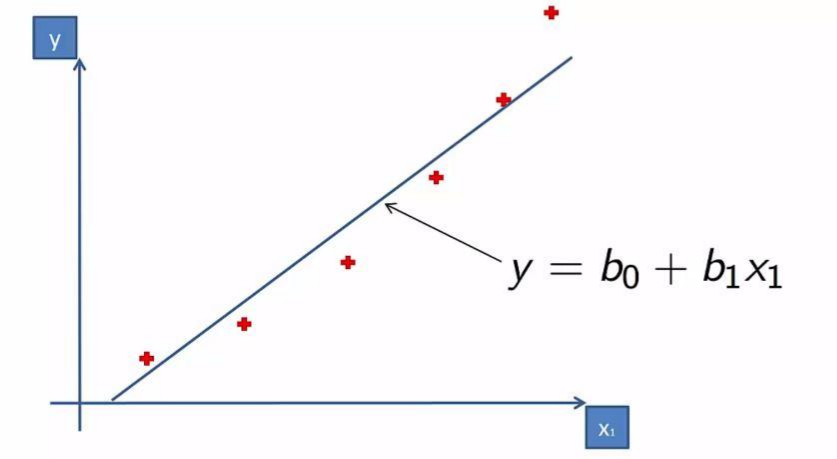
Linjär - coeffs

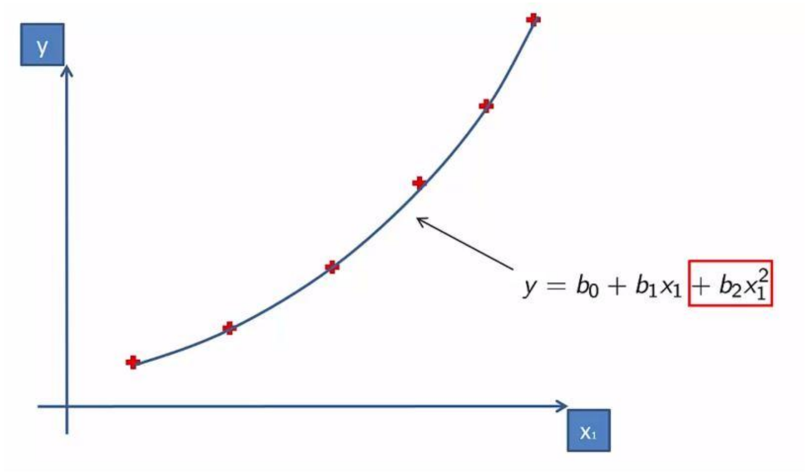


https://images.app.goo.gl/qriDKw1nLNMr2fmf9

RESTORY...

Polynom regression - när behövs det





Kod

Övning

Titta på datan i följande Kaggle exemplet och försök göra en polynomregression utan att titta på koden.

Pumpkin data

Statistik-ish

p-value

- Statistical significance
- Null hypothesis H_0 - assume true
- Is it correct? - experiment
  - 0.50
  - halveras ...
  - ***α*** = 0.05
  - domain dependent

" A p-value is the probability that random chance generated the data, or something else that is equal or rarer"

p-value ≠ probability

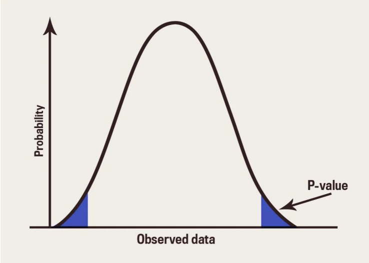
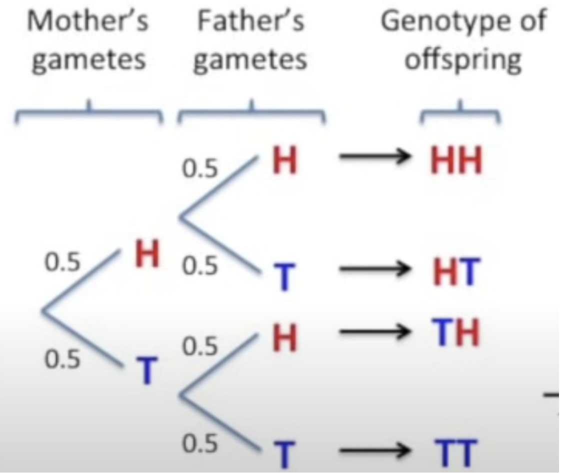
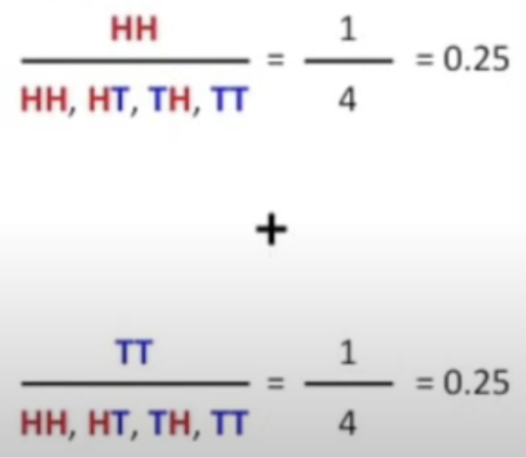


Bild ref.





$$\frac{HH}{HH, HT, TH, TT} = \frac{1}{4} = 0.25$$

$$+$$

$$\frac{TT}{HH, HT, TH, TT} = \frac{1}{4} = 0.25$$

StatQuest Youtube vid

Multivariable regression - variable selection

Why select?

- GIGO
- Noise
- Explain

1. **All** Variables
   a. Domain knowledge - predictors
   b. Forced - by company
   c. Preparing for no. 2

2. **Backward** Elimination
   a. Set significance level (e.g. sl = 0.05)
   b. Fit full model
   c. Remove predictor with highest **p**-val if **p>sl**.
   d. Refit model
   e. Repeat c & d until no p > sl.

3. **Forward** Selection
   a. Select sl
   b. Fit all simple reg. models for y and X for all x separately.
   c. Select the one with the lowest p-val ($\mathbf{x}_\mathbf{n}$).
   d. Fit $\mathbf{x}_\mathbf{n}$ with all other x variables (X-$\mathbf{x}_\mathbf{n}$) separately for every model. (y and [x_a, x_b])
   e. If the model with the lowest p-val has a p<sl, add it to the predictor, and repeat d.

Udemy kurs, Lek. 54.

Multivariable regression - variable selection

Stepwise regression: 2,3,**4**

4. ***Bidirectional*** *Elimination*

- a. Select an entry and stay sl (can be different)
- b. Add on new variable using Forward Selection (p < sl_enter).
- c. Do all steps of Backward elimination. Keep only variables where p < sl_stay is fulfilled.
- d. Repeat b and c until no new variables can enter or exit.

5. Score Comparison

- a. Set a fit score minimum
- b. Create a model for all possible variable combinations (2^n-1)
- c. Select the one with the best score

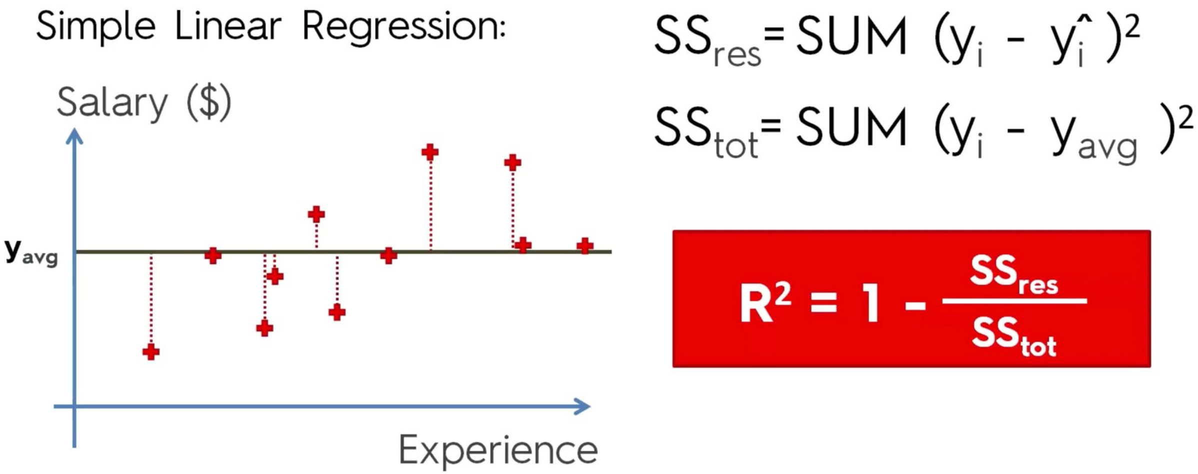
Scikit-learn does this automatically!

Udemy kurs, Lek. 54.

R^2

R^2

How much better than average?



Simple Linear Regression:

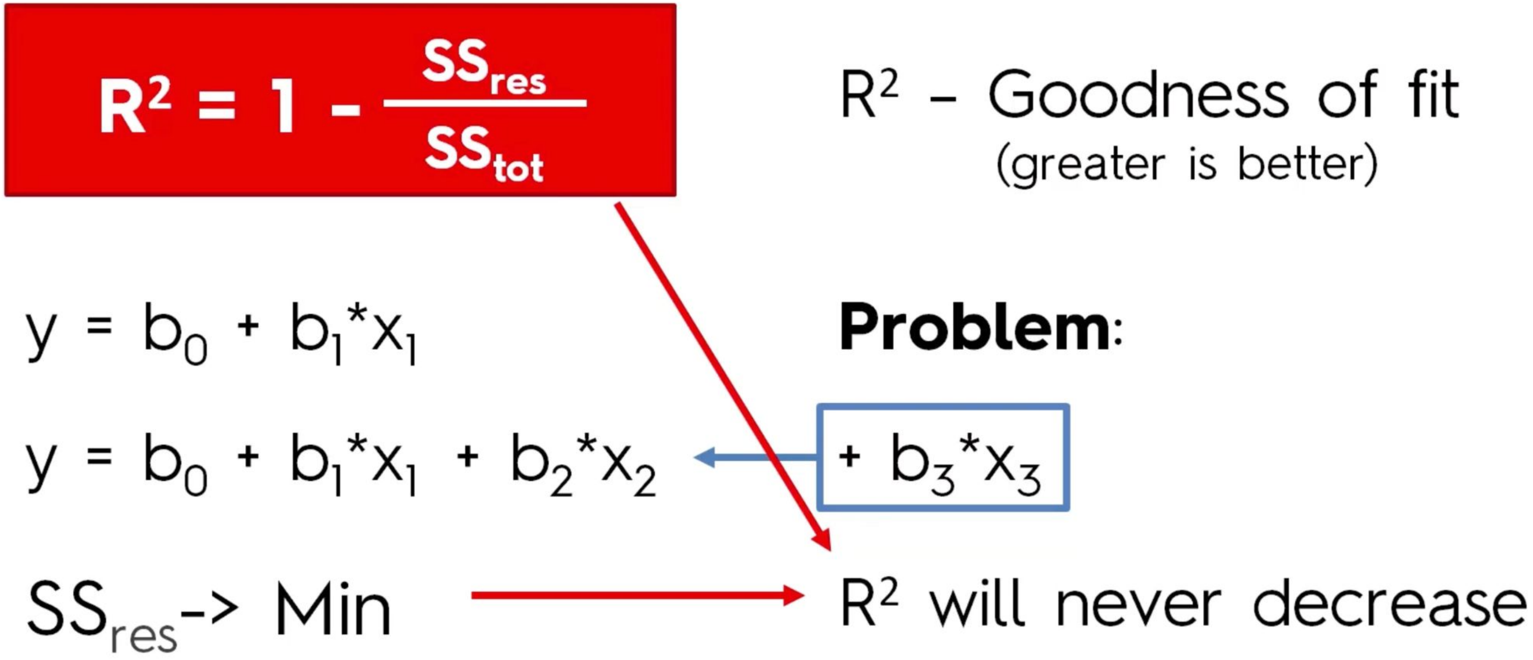
$$SS_{res} = SUM\ (y_i - \hat{y_i})^2$$

$$SS_{tot} = SUM\ (y_i - y_{avg})^2$$

$$\boxed{R^2 = 1 - \frac{SS_{res}}{SS_{tot}}}$$

Udemy kurs, Lek.99.

R^2 och flera variabler



$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

R^2 – Goodness of fit

(greater is better)

$$y = b_0 + b_1{}^*x_1$$

$$y = b_0 + b_1{}^*x_1 + b_2{}^*x_2$$

Problem:

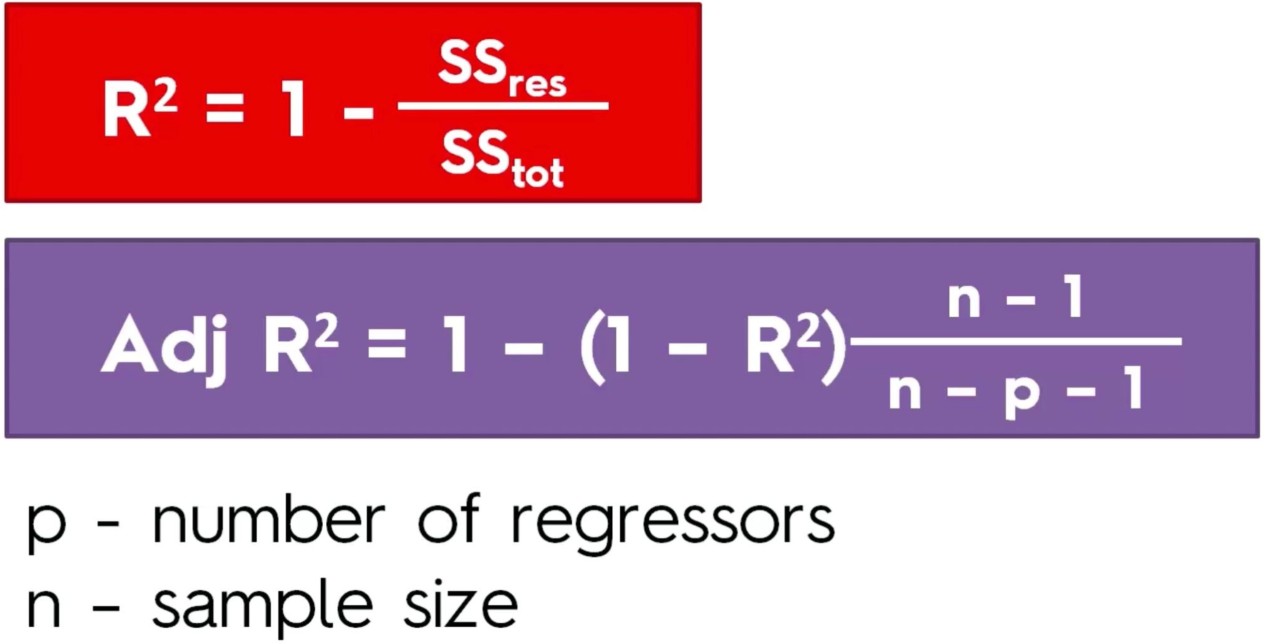
$$\boxed{+ b_3{}^*x_3}$$

SS_{res}-> Min

R^2 will never decrease

Udemy kurs, Lek. 99.

Justerad R^2



$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

$$\text{Adj } R^2 = 1 - (1 - R^2)\frac{n - 1}{n - p - 1}$$

p - number of regressors

n - sample size

Udemy kurs, Lek. 100.

Sammanfattning

- Linjär regression med flera variabler
- Kodstruktur
- Polynomregression
- Statistik
    - p-value
    - Multivariable regression - variable selection
    - R_2

Länkar

- StatQuest
- Linear regression