Python Lab 1 SSY316

Newson Matthew, Erik Norlin

November 2023

Obtaining β using sklearn, statsmodels, and from scratch

Table 1: All three methods results in the same model parameters, indicating that the implementation of our linear regression from scratch is accurate.

model param.	sklearn	statsmodels	from sratch
β_0	0.8852	0.8852	0.8852
β_1	-2.2326	-2.2326	-2.2326
β_2	0.5587	0.5587	0.5587

Obtaining β for third degree polynomial using statsmodels, and from scratch

Table 2: Both methods results in the same model parameters, as well as similar upper and lower confidence intervals, showing that the implementation of our linear regression from scratch is accurate.

model param.	statsmodels	from sratch
β_0	-0.0105	-0.0105
β_1	10.1247	10.1247
β_2	-31.1394	-31.1394
β_3	21.1321	21.1321
$\beta_{0,lower95\%}$	-0.6270	-0.6273
$\beta_{1,lower95\%}$	4.6420	4.6418
$\beta_{2,lower95\%}$	-44.0760	-44.0761
$\beta_{3,lower95\%}$	12.6380	12.6380
$\beta_{0,upper95\%}$	0.6060	0.6064
$\beta_{1,upper95\%}$	15.6080	15.6076
$\beta_{2,upper95\%}$	-18.2030	-18.2027
$\beta_{3,upper95\%}$	29.6260	29.6261

Calculation for posterior parameters

For third degree polynomial we get that

$$\mu_n = \begin{pmatrix} 0.84 \\ -2.13 \\ 0.53 \end{pmatrix}, \Omega_n = \begin{pmatrix} 0.35 & -1.37 & 1.12 \\ -1.37 & 7.59 & -7.07 \\ 1.12 & -7.07 & 7.07 \end{pmatrix}, a_n = 10.01, b_n = 8.42$$

Model evidence of different polynomials

Table 3: The third degree polynomial results in the maximum log-evidence, which supports the visual observation that third degree polynomial fitted the sinus curve best in our case.

polynomial	log model evidence
2	-25.12
3	-18.21
4	-18.94
5	-19.25
6	-20.43

Calculation for posterior parameters of the Auto dataset

8 models were implemented with the variables as can be seen in table 4. In each model the LOSS variable represented the function value. The model that gave the largest log-modelevidence was the model with only having 'ATTORNEY' as a variable with an interception. This indicates that it is most probable that the economic loss is only dependent on the claimant was represented by an attorney or not. The other factors could be contributing to overfitting.

Table 4

model	variables and intercept
1	'INTERCEPT'
2	'INTERCEPT', 'ATTORNEY'
3	'INTERCEPT', 'CLMSEX'
4	'INTERCEPT', 'CLMAGE'
5	'INTERCEPT', 'ATTORNEY', 'CLMSEX'
6	'INTERCEPT', 'ATTORNEY', 'CLMAGE'
7	'INTERCEPT', 'CLMSEX', 'CLMAGE'
8	'INTERCEPT', 'ATTORNEY', 'CLMSEX', 'CLMAGE'

The parameters of the posterior of the most optimal model ('INTERCEPT', 'ATTORNEY') was

$$\mu_n = \begin{pmatrix} 2.5894 \\ 3.9622 \end{pmatrix}, \Omega_n = \begin{pmatrix} 0.0016 & -0.0016 \\ -0.0016 & 0.0031 \end{pmatrix}, a_n = 648.1, b_n = 34572.2382, \text{lmodevid} = -4428.3521$$