CSCI 6360

Project 1

February 18, 2026

Transformed Regression: Sqrt, Log1p, Box-Cox

In this section, we applied three different transformation techniques to the response variables (mpg, house price, and insurance charge) across the Auto MPG, Boston House Price, and Medical Cost datasets. The distributions for mpg, house price, and medical charge exhibit right-skewed patterns. Therefore, these transformations could significantly improved model accuracy. To measure and compare the quality of fit for the different transformation models, we extracted 15 metrics, including R^2, adjusted R^2, MSE, and MAE. Each model was evaluated using both in-sample validation and a validation set with an 80-20% split. Finally, we applied feature selection to identify the optimal set of features that best describe the response variable. For box-cox transformation we reqired a optimal lamda parameter which was 0.19, 0.85, and 0.04 for mpg, house price, and insurance cost respectively.

Transformed Regression using Scala

We used TranRegression function form the modeling to perfrom Sqrt, Log1p, Box-Cox on the Auto MPG, Boston House Price, and Medical Cost datasets. The function takes predictor variables, a response variable, feature names, a factorization method, and transformation and inverse transformation methods as input, and fits the model based on the specified transformation. An example of the code used to fit the square-root (sqrt) transformation model is provided below.

```
f = ("sqrt, sq, "sqrt")
val mod = new modeling.TranRegression (ox, y, ox_fname,
                                       modeling.Regression.hp,
                                       f._1, f._2, f._3)
```

Transformed Regression on the Auto MPG Dataset using Scala: In-Sample and Validation Results

Tables 1 and 2 present the quality of fit for the in-sample and validation (80-20% split) evaluations using the Auto MPG data. For the in-sample case, the models utilized all available data. The results indicate that the log1p and Box-Cox transformations outperform the sqrt transformation, which is further confirmed by the adjusted R^2 values. The Mean Square Error (MSE) for the sqrt, log1p, and Box-Cox transformations reached 10.02, 9.17, and 9.40, respectively, with the log1p transformation achieving the lowest error. Similarly, in the validation accuracy results, where 80% of the data was randomly used for model training and 20% for testing, the log1p transformation achieved the highest R^2 and adjusted R^2 values, as well as the lowest MSE.

Table 1: Auto MPG (In-Sample): Sqrt, Log1p, and Box–Cox($\lambda = 0.19$)

Metric	sqrt	log1p	box-cox(λ=0.19)
rSq	0.835138	0.849102	0.845366
rSqBar	0.832569	0.846751	0.842956
sst	23819.0	23819.0	23819.0
sse	3926.85	3594.23	3683.22
sde	3.16757	3.02514	3.06455
mse0	10.0175	9.16895	9.39596
rmse	3.16504	3.02803	3.06528
mae	2.34675	2.18422	2.22945
smape	10.2046	9.32001	9.54480
m	392.000	392.000	392.000
dfr	6.00000	6.00000	6.00000
df	385.000	385.000	385.000
fStat	325.048	361.067	350.793
aic	-993.873	-976.527	-981.320
bic	-966.074	-948.728	-953.521

Table 2: Auto MPG (Validation): Sqrt, Log1p, and Box–Cox($\lambda = 0.19$)

Metric	sqrt	log1p	box-cox(λ=0.19)
rSq	0.846480	0.852864	0.851819
rSqBar	0.844088	0.850571	0.849510
sst	4731.23	4731.23	4731.23
sse	726.337	696.133	701.080
sde	3.05724	2.97969	2.99552
mse0	9.31202	8.92478	8.98820
rmse	3.05156	2.98744	2.99803
mae	2.11846	1.98665	2.01582
smape	9.00068	8.28831	8.41880
m	78.0000	78.0000	78.0000
dfr	6.00000	6.00000	6.00000
df	385.000	385.000	385.000
fStat	353.804	371.939	368.862
aic	-183.700	-182.048	-182.322
bic	-167.203	-165.551	-165.825

Transformed Regression on the Boston House Price Dataset: In-Sample and Validation Results

For the house price prediction, the log1p transformed model performed significantly worse than the sqrt and Box-Cox transformations. The Box-Cox and sqrt models performed similarly, although Box-Cox achieved highest R^2 and adjusted R^2 of 0.998 in the in-sample evaluation. A

similar trend was observed in the validation, where Box-Cox and sqrt outperformed the log1p transformation, with Box-Cox achieving the highest R^2 and adjusted R^2.

Table 3: House Price (In-Sample): Sqrt, Log1p, and Box–Cox($\lambda = 0.85$)

Metric	sqrt	log1p	box-cox(λ=0.85)
rSq	0.986164	0.922376	0.997549
rSqBar	0.986067	0.921828	0.997531
sst	6.42325e+13	6.42325e+13	6.42325e+13
sse	8.88711e+11	4.98598e+12	1.57458e+11
sde	29823.1	70573.3	12554.0
mse0	8.88711e+08	4.98598e+09	1.57458e+08
rmse	29811.3	70611.5	12548.2
mae	24296.5	52989.3	10078.9
smape	4.85788	9.21218	2.12365
m	1000.00	1000.00	1000.00
dfr	7.00000	7.00000	7.00000
df	992.000	992.000	992.000
fStat	10100.8	1683.94	57668.5
aic	-11705.6	-12567.9	-10840.3
bic	-11666.3	-12528.6	-10801.0

Table 4: House Price (Validation): Sqrt, Log1p, and Box–Cox($\lambda = 0.85$)

Metric	sqrt	log1p	box-cox(λ=0.85)
rSq	0.984017	0.906459	0.997398
rSqBar	0.983904	0.905799	0.997380
sst	1.33700e+13	1.33700e+13	1.33700e+13
sse	2.13694e+11	1.25064e+12	3.47839e+10
sde	32755.5	78812.3	13217.8
mse0	1.06847e+09	6.25321e+09	1.73920e+08
rmse	32687.5	79077.2	13187.9
mae	26140.1	58024.9	10655.9
smape	5.18595	9.88665	2.22264
m	200.000	200.000	200.000
dfr	7.00000	7.00000	7.00000
df	992.000	992.000	992.000
fStat	8724.77	1373.28	54329.4
aic	-2346.74	-2523.43	-2165.20
bic	-2320.35	-2497.04	-2138.81

Transformed Regression on the Medical Cost Dataset: In-Sample and Validation Results

In the Medical Cost dataset, the sqrt transformation performed significantly better than the log1p and Box-Cox transformations. The sqrt transformation achieved an R^2 of 0.753 and an adjusted R^2 of 0.751 in the in-sample evaluation, compared to 0.730 and 0.729, respectively, for the validation set. The adjusted R^2 also shows a similar trend, where sqrt, log1p, and Box-Cox achieved 0.751, 0.520, and 0.574 for in-sample evaluation and 0.729, 0.569, and 0.607 for validation, respectively.

Table 5: Medical Cost (In-Sample): Sqrt, Log1p, and Box–Cox($\lambda = 0.04$)

Metric	sqrt	log1p	box-cox(λ=0.04)
rSq	0.752657	0.522782	0.576265
rSqBar	0.751168	0.519909	0.573715
sst	1.96074e+11	1.96074e+11	1.96074e+11
sse	4.84975e+10	9.35702e+10	8.30835e+10
sde	6001.71	8358.44	7870.39
mse0	3.62463e+07	6.99329e+07	6.20953e+07
rmse	6020.49	8362.59	7880.06
mae	3613.90	4219.51	4052.90
smape	27.6903	26.2889	26.0851
m	1338.00	1338.00	1338.00
dfr	8.00000	8.00000	8.00000
df	1329.00	1329.00	1329.00
fStat	505.514	181.986	225.924
aic	-13525.1	-13964.7	-13885.2
bic	-13478.3	-13917.9	-13838.4

Table 6: Medical Cost (Validation): Sqrt, Log1p, and Box–Cox($\lambda = 0.04$)

Metric	sqrt	log1p	box-cox(λ=0.04)
rSq	0.730154	0.571048	0.609124
rSqBar	0.728530	0.568466	0.606771
sst	4.06432e+10	4.06432e+10	4.06432e+10
sse	1.09674e+10	1.74340e+10	1.58865e+10
sde	6405.50	8088.13	7716.02
mse0	4.10764e+07	6.52957e+07	5.94998e+07
rmse	6409.08	8080.58	7713.61
mae	3839.79	4116.83	3998.83
smape	30.1151	27.3599	27.2984
m	267.000	267.000	267.000
dfr	8.00000	8.00000	8.00000
df	1329.00	1329.00	1329.00
fStat	449.505	221.156	258.882
aic	-2701.24	-2763.11	-2750.71
bic	-2668.95	-2730.83	-2718.42

Transformed Regression using statsmodels

Here, we used Python statsmodels library to reproduce all results generated by the TranRegression package.

Transformed Regression with Sqrt, Log1p, and Box–Cox Transformations on the Auto MPG Dataset: In-Sample, Validation, Forward, and Backward Results

The following three tables present the results for the sqrt, log1p, and Box-Cox transformations, including in-sample evaluation, validation, and forward and backward feature selection. From these tables, we can see that, similar to the Scala results, the log1p and Box-Cox transformations perform similarly, with log1p being slightly better than Box-Cox. For the in-sample case, the sqrt, log1p, and Box-Cox transformations achieved R^2 values of 0.835, 0.849, and 0.845, respectively, with adjusted R^2 values of 0.833, 0.847, and 0.843. Similarly, for the validation case, the R^2 values were 0.822, 0.832, and 0.830, with adjusted R^2 values of 0.807, 0.818, and 0.816.

Table 7: Auto MPG Regression with Square-Root Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.835	0.822	0.823	0.822
rSqBar	0.833	0.807	0.818	0.814
sst	23818.993	5372.801	5372.801	5372.801
sse	3926.849	955.390	950.390	958.806
mse	4.017	6.094	10.030	9.137
rmse	3.165	3.478	3.468	3.484
mae	2.347	2.376	2.375	2.350
m	392.000	79.000	79.000	79.000
dfr	6.000	6.000	2.000	3.000
df	385.000	72.000	76.000	75.000
fStat	825.235	120.822	220.454	161.034

Table 8: Auto MPG Regression with Log1p Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.849	0.832	0.806	0.804
rSqBar	0.847	0.818	0.801	0.794
sst	23818.993	5372.801	5372.801	5372.801
sse	3594.229	900.453	1044.122	1052.182
mse	3.169	5.398	11.217	9.319
rmse	3.028	3.376	3.635	3.649
mae	2.184	2.221	2.710	2.713
m	392.000	79.000	79.000	79.000
dfr	6.000	6.000	2.000	4.000
df	385.000	72.000	76.000	74.000
fStat	1063.694	138.083	192.956	115.912

Table 9: Auto MPG Regression with Box-Cox Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.845	0.830	0.829	0.829
rSqBar	0.843	0.816	0.825	0.823
sst	23818.993	5372.801	5372.801	5372.801
sse	3683.218	914.477	917.735	916.540
mse	3.396	5.576	9.617	8.602
rmse	3.065	3.402	3.408	3.406
mae	2.229	2.262	2.356	2.234
m	392.000	79.000	79.000	79.000
dfr	6.000	6.000	2.000	3.000
df	385.000	72.000	76.000	75.000
fStat	988.221	133.268	231.627	172.688

Transformed Regression with Sqrt, Log1p, and Box–Cox Transformations on the Boston House Price Dataset: In-Sample, Validation, Forward, and Backward Results

For the house price prediction dataset, the ScalaTion and statsmodels summaries are consistent, as the Box–Cox transformation was identified as the best-performing model under both approaches.

Table 10: Boston House Price Regression with Square-Root Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.986	0.987	0.281	0.981
rSqBar	0.986	0.986	0.277	0.981
sst	64232463468052.547	11927062037792.469	11927062037792.469	11927062037792.469
sse	888710772890.930	155868297872.986	8579357295076.382	223122191452.058
mse	888710765.891	779341482.365	42896786474.382	1115610953.260
rmse	29811.252	27916.688	207115.394	33400.763
mae	24296.468	22632.355	177100.990	27022.336
m	1000.000	200.000	200.000	200.000
dfr	7.000	7.000	1.000	4.000
df	992.000	192.000	198.000	195.000
fStat	10182.286	2157.718	78.041	2622.765

Table 11: Boston House Price Regression with Log1p Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.922	0.934	-81398.531	0.877
rSqBar	0.922	0.931	-81809.639	0.875
sst	64232463468052.547	11927062037792.469	11927062037792.469	11927062037792.469
sse	4985983480190.929	789371506083.466	970857251871500928.000	1461933128041.826
mse	4985983473.191	3946857523.417	4854286259357504.000	7309665636.209
rmse	70611.497	62824.020	69672708.139	85496.583
mae	52989.293	48059.016	20766229.670	61057.028
m	1000.000	200.000	200.000	200.000
dfr	7.000	7.000	1.000	4.000
df	992.000	192.000	198.000	195.000
fStat	1697.515	403.131	-199.998	357.921

Table 12: Boston House Price Regression with Box-Cox Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.998	0.997	0.924	0.990
rSqBar	0.998	0.997	0.924	0.989
sst	64232463468052.531	11927062037792.469	11927062037792.469	11927062037792.469
sse	157457720227.919	32215998366.736	903559871079.215	124824257960.295
mse	157457713.228	161079984.834	4517799354.396	624121285.801
rmse	12548.216	12691.729	67214.577	24982.420
mae	10078.898	10289.124	56408.842	20247.645
m	1000.000	200.000	200.000	200.000
dfr	7.000	7.000	1.000	4.000
df	992.000	192.000	198.000	195.000
fStat	58133.527	10549.192	2440.016	4727.542

Transformed Regression with Sqrt, Log1p, and Box–Cox Transformations on the Medical Cost Dataset: In-Sample, Validation, Forward, and Backward Results

Finally, for the medical cost dataset, the sqrt transformation explained the insurance costs better than the other methods, despite the relatively lower R^2 values. For the in-sample case, the R^2 values were 0.753, 0.523, and 0.576 for the sqrt, log1p, and Box-Cox transformations, respectively, with adjusted R^2 values of 0.751, 0.520, and 0.574. These results were verified by the validation set, where the sqrt transformation achieved an R^2 of 0.706 and an adjusted R^2 of 0.697, outperforming the log1p (R^2 of 0.505, Adjusted R^2 of 0.490) and Box-Cox ((R^2 of 0.546, Adjusted R^2 of 0.532) models.

Table 13: Medical Cost Regression with Square-Root Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.753	0.706	0.702	0.704
rSqBar	0.751	0.697	0.699	0.699
sst	196074221568.367	35797001122.000	35797001122.000	35797001122.000
sse	48497528289.829	10509483431.295	10649651869.890	10582381967.672
mse	36246276.223	39214482.415	39737503.977	39486494.879
rmse	6020.489	6262.147	6303.769	6283.828
mae	3613.896	3611.073	3687.410	3606.738
m	1338.000	268.000	268.000	268.000
dfr	8.000	8.000	3.000	5.000
df	1329.000	259.000	264.000	262.000
fStat	508.937	80.606	210.946	127.713

Table 14: Medical Cost Regression with Log1p Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.523	0.505	-415.056	-53.303
rSqBar	0.520	0.490	-419.784	-54.340
sst	196074221568.367	35797001122.000	35797001122.000	35797001122.000
sse	93570168847.798	17725546380.471	14893568669580.479	1943894999500.171
mse	69932853.620	66140090.435	55573017420.808	7253339545.374
rmse	8362.587	8132.656	235739.300	85166.540
mae	4219.512	4231.425	66672.183	25327.152
m	1338.000	268.000	268.000	268.000
dfr	8.000	8.000	3.000	5.000
df	1329.000	259.000	264.000	262.000
fStat	183.219	34.154	-89.119	-52.613

Table 15: Medical Cost Regression with Box-Cox Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.576	0.546	-61.221	-10.125
rSqBar	0.574	0.532	-61.928	-10.337
sst	196074221568.367	35797001122.000	35797001122.000	35797001122.000
sse	83083467025.754	16248889297.412	2227326392657.103	398225509782.019
mse	62095258.835	60630175.946	8310919372.586	1485916076.276
rmse	7880.055	7786.539	91164.244	38547.582
mae	4052.895	4083.531	31806.787	14527.915
m	1338.000	268.000	268.000	268.000
dfr	8.000	8.000	3.000	5.000
df	1329.000	259.000	264.000	262.000
fStat	227.454	40.302	-87.898	-48.782