ACTU PS5841 Data Science in Finance & Insurance - Autumn 2022 (Y. Wang)
Assignment - 8
Assigned 10/13/22, Due 10/30/22

Part A The file `Carseats.csv` records child car seat sales in 400 locations. The following linear regression model attempts to predict `Sales` in non-US locations (`US = No`):

`Sales` $\sim$ `Income + Price + ShelveLoc + Urban + Urban:Income`

where the categorical feature `ShelveLoc` is coded according to the sum-to-zero contrast, and `Urban` is coded according to the treatment contrast.

We can easily fit the regression model in python using `statsmodels` as follows.

```
from patsy.contrasts import Treatment, Sum
import statsmodels.formula.api as smf   #smf.ols
```

```
sum_contrast = Sum().code_without_intercept(['Bad', 'Good', 'Medium'])
treatment_contrast = Treatment(reference = 'No').code_without_intercept(['No', 'Yes'])
lm_smf_res = smf.ols("Sales ~ Income + Price +\
              C( ShelveLoc, sum_contrast) + C(Urban, treatment_contrast) +\
              C(Urban, treatment_contrast):Income",\
             data = df).fit()
lm_smf_res.summary()
```

[**a**] Run the above code and let your code report the fitted parameters and the residual sum of squares.

[**b**] Instead of the above, please code in python to estimate the regression coefficients using the gradient descent method with the following loss function and starting coefficient vector:

$$\text{loss} = \sum_{i}^{n} (y_i - \hat{y}_i)^2, \qquad \hat{\boldsymbol{\beta}}^{(0)} = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1)^T$$

If one uses a fixed learning rate of 0.0000003 *or* 3×10^{-7}*, the gradient descent approach achieves a loss of* 403.92700086 *at* $5,000,000$ *iterations (... 411.338296 and* 404.221267 *at* $1,000,000$ *and* $2,000,000$ *iterations ...) There is hope in navigating a seven-dimensional parameter space:) You are encouraged to explore different learning rates and investigate how they affect convergence (or non-convergence).*

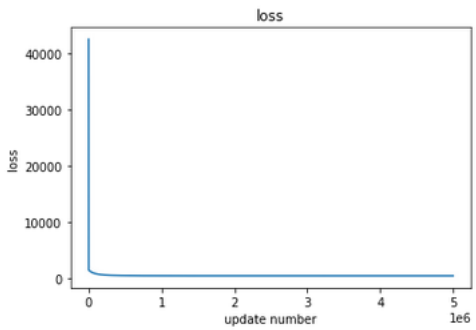
Please let your code report the following:

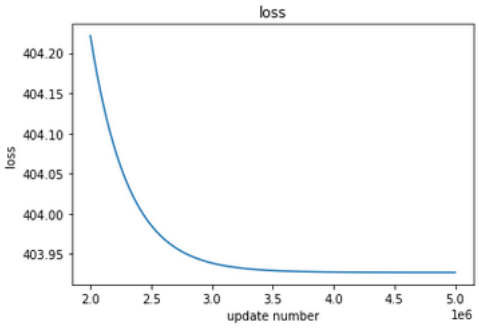
- learning rate
- the minimum loss achieved
- the number of iterations (or epochs) needed to achieve the reported minimum loss
- the estimated coefficients associated with the reported minimum loss

:

	OLS coef	GD coef
Intercept	11.757081	11.754978
C(ShelveLoc, sum_contrast)[S.Bad]	-1.956725	-1.956716
C(ShelveLoc, sum_contrast)[S.Good]	2.297505	2.297466
C(Urban, treatment_contrast)[T.Yes]	1.887842	1.889406
Income	0.026715	0.026732
C(Urban, treatment_contrast)[T.Yes]:Income	-0.024027	-0.024048
Price	-0.057761	-0.057753

[**c**] Please plot the estimated loss as a function of iterations (or epochs).





Please submit your work as hw8.ipynb and hw8.html to `Canvas`.