Assignment - 6
Assigned 10/6/22, Due 10/16/22

INTRODUCTION

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.

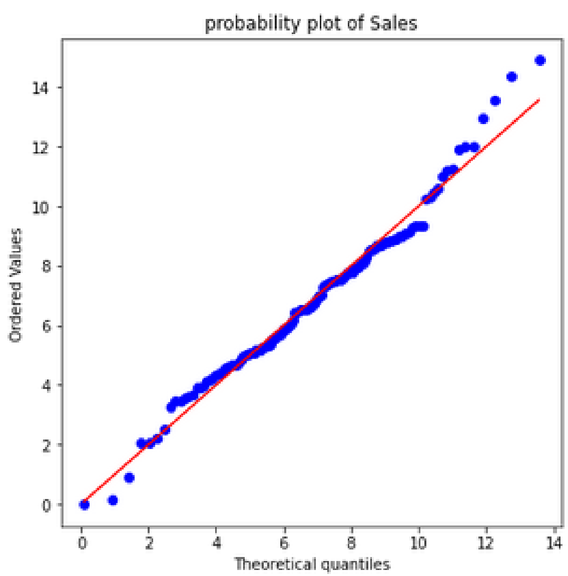
We can easily fit the regression model in python using `statsmodels` as follows.

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

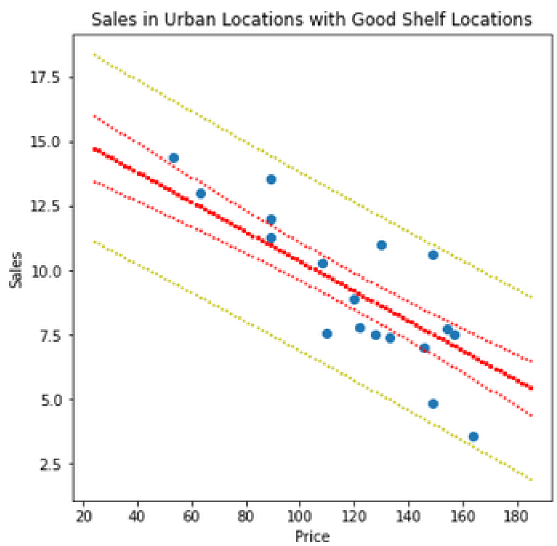
```
sum_contrast = Sum().code_without_intercept(['Bad', 'Good', 'Medium'])
treatment_contrast = Treatment(reference = 'No').code_without_intercept(['No', 'Yes'])
lm_smf_res = smf.ols("Sales ~ Income + Price +\
               C( ShelveLoc, sum_contrast) + C(Urban, treatment_contrast) +\
             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

Please code in python to complete the following tasks.

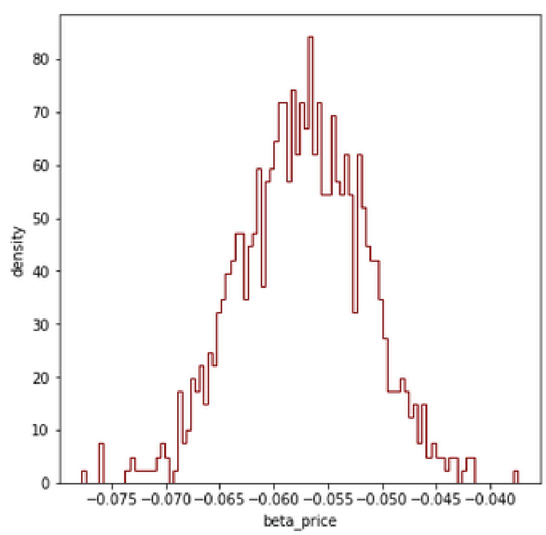
[**a**] Check if `Sales` can be assumed to be normal by producing a quantile-quantile plot with respect to the normal distribution, like the one below.
Ref: `scipy.stats.probplot()`



[**b**] The normal linear model looks applicable here. Use the **fitted model** in the **INTRODUCTION**, as well as the theoretical formulae for the confidence interval and the prediction interval, produce the following chart for `Sales` in `Urban` locations with Good `ShelveLoc`, and assuming a median income for `Income` (calculated as the median income across all non-US locations for grading purposes), but with Price varying uniformly over 24 and 185, as discussed in class.



[**c**] Without assuming normality for `Sales`, use the Bootstrap approach to estimate the distribution of the coefficient on `Price`. Produce a density plot similar to the following.
Ref: `matplotlib.pyplot.hist()`



[**d**] With the estimated coefficient distributions from the Bootstrap approach, produce the a table for inference for the regression in the **INTRODUCTION**, similar to the following.
Ref: `pandas.DataFrame.sample()`
Note: The confidence intervals shown for the estimated coefficients are ± 1.96 standard error intervals.

	Coef	std err	"t"	Lower	Upper
Intercept	11.757081	1.005850	11.688699	9.785614	13.728548
C(ShelveLoc, sum_contrast)[S.Bad]	-1.956725	0.234879	-8.330791	-2.417087	-1.496363
C(ShelveLoc, sum_contrast)[S.Good]	2.297505	0.276981	8.294824	1.754623	2.840387
C(Urban, treatment_contrast)[T.Yes]	1.887842	0.813805	2.319773	0.292785	3.482900
Income	0.026715	0.010226	2.612466	0.006672	0.046759
C(Urban, treatment_contrast)[T.Yes]:Income	-0.024027	0.011840	-2.029426	-0.047233	-0.000822
Price	-0.057761	0.005778	-9.997467	-0.069085	-0.046437

Please submit your work as hw6.ipynb and hw6.html to `Canvas`.