In [ ]:
```
from google.colab import drive
```

In [ ]:
```
drive.mount('/content/gdrive/',force_remount=True)
```

```
Mounted at /content/gdrive/
```

In [ ]:
```
import pandas as pd
import sklearn as sk
import numpy as np
import matplotlib.pyplot as plt
from sklearn import preprocessing
from sklearn.linear_model import Lasso, LassoCV
from sklearn.preprocessing import scale
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import sklearn
from sklearn import preprocessing
from patsy import dmatrix
import matplotlib.pyplot as plt
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import train_test_split
```

In [ ]:
```
df = pd.read_csv("/content/gdrive/MyDrive/stroke/insurance.csv")
```

In [ ]:
```
df
```

Out[6]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830
1334	18	female	31.920	0	no	northeast	2205.98080
1335	18	female	36.850	0	no	southeast	1629.83350
1336	21	female	25.800	0	no	southwest	2007.94500
1337	61	female	29.070	0	yes	northwest	29141.36030

1338 rows × 7 columns

In [ ]:
```
df['sex'] = df['sex'].map({'female':0, 'male':1})
df['smoker'] = df['smoker'].map({'yes':0, 'no':1})
df['region'] = df['region'].map({'southwest':0, 'southeast':1,'northeast':2,'northwest':3})
```

In [ ]:
```
df
```

Out[8]:

	age	sex	bmi	children	smoker	region	charges
0	19	0	27.900	0	0	0	16884.92400
1	18	1	33.770	1	1	1	1725.55230
2	28	1	33.000	3	1	1	4449.46200
3	33	1	22.705	0	1	3	21984.47061
4	32	1	28.880	0	1	3	3866.85520
...	...	...	...	...	...	...	...
1333	50	1	30.970	3	1	3	10600.54830
1334	18	0	31.920	0	1	2	2205.98080
1335	18	0	36.850	0	1	1	1629.83350
1336	21	0	25.800	0	1	0	2007.94500
1337	61	0	29.070	0	0	3	29141.36030

1338 rows × 7 columns

In [ ]:
```
X = df[["age","sex","bmi","children","smoker","region"]]
Y = df["charges"]
```

In [ ]:
```
X_train,X_test,Y_train,Y_test = sklearn.model_selection.train_test_split(X, Y, train_size=0.8, random_state = 0)
```

In [ ]:
```
scaler = preprocessing.StandardScaler().fit(X_train)
X_train_scale = scaler.transform(X_train)
```

In [ ]:
```
lassocv = LassoCV(alphas=None, cv=10, max_iter=10000)
lassocv.fit(scale(X_train), Y_train.values.ravel())
```

Out[12]:
```
LassoCV(cv=10, max_iter=10000)
```

In [ ]:
```
a = lassocv.alpha_
```

In [ ]:
```
a
```

Out[14]:
```
40.292167746969625
```

In [ ]:
```
lasso.set_params(alpha=a)
lasso.fit(scale(X_train), Y_train)
print('Test MSE = ',mean_squared_error(Y_test, lasso.predict(scale(X_test))))
```

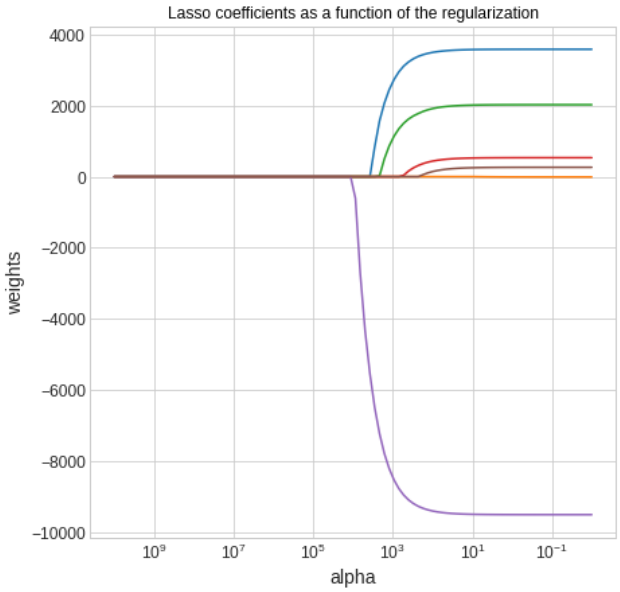
```
Test MSE =  32024426.909760844
```

In [ ]:
```
print('Train MSE = ',mean_squared_error(Y_train, lasso.predict(scale(X_train))))
```

```
Train MSE =  37761730.88241446
```

In [ ]:
```
lasso = Lasso(max_iter=10000)
coefs = []

for a in alphas*2:
    lasso.set_params(alpha=a)
    lasso.fit(scale(X_train), Y_train)
    coefs.append(lasso.coef_)
plt.figure(figsize=(7, 7))
ax = plt.gca()
ax.plot(alphas*2, coefs)
ax.set_xscale('log')
ax.set_xlim(ax.get_xlim()[::-1])
plt.axis('tight')
plt.xlabel('alpha')
plt.ylabel('weights')
plt.title('Lasso coefficients as a function of the regularization');
```



In [ ]:
```
pd.Series(lasso.coef_, index=X.columns)
```

Out[112]:
```
age         3548.561537
sex           -0.000000
bmi         1979.870977
children     494.224983
smoker     -9472.252253
region       214.264422
dtype: float64
```

As age increases by 1 year, charges increase by 3548.56. Sex is shrinked to zero, implying that it does not effect charges. As BMI increases by 1, charges increase by 1979.87. As the number of children increases by 1, charges increase by 494.22 Non-smokers' charges are less than those of smokers by 9472.