**使用多元线性回归根据多个因素预测医疗费用**

**1. 实验目的**

* 熟练掌握使用多元线性回归构建回归模型
* 熟练掌握通过多元线性回归的表达式得到影响因变量的主要因素

**2. 实验内容**

预测个人未来医疗费用支出，用于辅助医疗保险作出收取保费的决策，对保险公司有重要意义。

本实验使用多元线性回归模型，通过年龄、性别、bmi 等因素预测个人未来的医疗费用支出，并分析哪个特征对因变量影响较大。

**3. 实验数据**

**数据集位置**

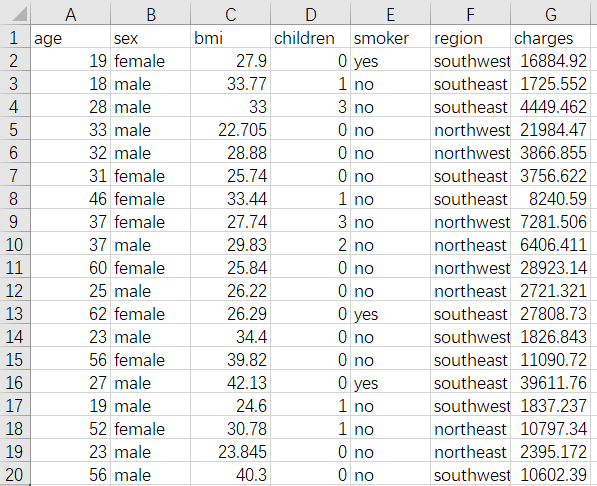
数据集存放在/home/dataset目录下。数据集名字是insurance.csv

**数据集格式**

数据集关于医疗费用，给保险公司用。

数据集包含 1338 个样本数据（行）和 7 个字段（列）。

7个列分别是年龄、性别、bmi、孩子数量、是否抽烟、所在区域、过去医疗费用支出。



**数据字段解释**

* age：数值型；年龄
* sex：字符串类型；性别；有2个值，分别是female（女性）、male（男性）
* bmi：数值型；身体bmi指数
* children：数值型；有几个孩子
* smoker：字符串类型；是否是吸烟者；有2个值，分别是yes（是吸烟者）、no（不是吸烟者）
* region：字符串类型；所在区域；有4个值，分别是southwest（西南部）、southeast（东南部）、northwest（西北部）、northeast（东北部）
* charges：数值型；客户的医疗花费

**4. 实验知识点**

* 简单线性回归模型
* MSE
* 小提琴图

**5. 实验时长**

1学时

**6. 实验环境**

* Linux Ubuntu 操作系统
* Jupyter 代码编辑器
* Python 3.6.9
* numpy 1.18.5
* pandas 1.1.5
* matplotlib 3.3.3
* seaborn 0.11.1
* scikit-learn 0.24.2

**7. 实验分析**

1. 导入包、导入数据、做数据预处理
2. 构建多元线性回归模型
3. 得到模型数学表达式
4. 找到影响因变量的最大因素

**8. 实验过程**

**8.1 导入包**

In [1]:

*# 导入包*

**import** **numpy** **as** **np**

**import** **pandas** **as** **pd**

**import** **matplotlib.pyplot** **as** **plt**

**import** **seaborn** **as** **sns**

**8.2 导入数据集**

In [2]:

*# 导入数据集*

data = pd.read\_csv('/home/dataset/insurance.csv')

data.head(5)

Out[2]:

|  | **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 |

**8.3 数据预处理**

**8.3.1 检测缺失值**

In [3]:

*# 检测缺失值*

null\_df = data.isnull().sum()

null\_df

Out[3]:

age 0

sex 0

bmi 0

children 0

smoker 0

region 0

charges 0

dtype: int64

**8.3.2 标签编码&独热编码**

In [4]:

*# 标签编码&独热编码*

data = pd.get\_dummies(data, drop\_first = **True**)

**8.3.3 得到自变量和因变量**

In [5]:

*# 得到自变量和因变量*

y = data['charges'].values

data = data.drop(['charges'], axis = 1)

x = data.values

**8.3.4 拆分训练集和测试集**

In [6]:

*# 拆分训练集和测试集*

**from** **sklearn.model\_selection** **import** train\_test\_split

x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size = 0.2, random\_state = 1)

print(x\_train.shape)

print(x\_test.shape)

print(y\_train.shape)

print(y\_test.shape)

(1070, 8)

(268, 8)

(1070,)

(268,)

**8.4 构建多元线性回归模型**

In [7]:

*# 构建多元线性回归模型*

**from** **sklearn.linear\_model** **import** LinearRegression

regressor = LinearRegression(normalize = **True**, fit\_intercept = **True**)

regressor.fit(x\_train, y\_train)

Out[7]:

LinearRegression(normalize=True)

**8.5 得到模型表达式**

In [8]:

*# 得到模型表达式*

print('数学表达式是：**\n** Charges = ', end='')

columns = data.columns

coefs = regressor.coef\_

**for** i **in** range(len(columns)):

print('**%s** \* **%.2f** + ' %(columns[i], coefs[i]), end='')

print(regressor.intercept\_)

数学表达式是：

Charges = age \* 257.49 + bmi \* 321.62 + children \* 408.06 + sex\_male \* -242.15 + smoker\_yes \* 23786.49 + region\_northwest \* -396.10 + region\_southeast \* -1038.38 + region\_southwest \* -903.03 + -11297.610008539421

由上述数学表达式可见，smoker\_yes变量对因变量影响较大

**8.6 画出吸烟与医疗费用的小提琴图**

由上述数学表达式可以得到，smoker\_yes变量对因变量影响较大。下面通过小提琴图验证一下。

In [9]:

*# 画出吸烟与医疗费用的小提琴图*

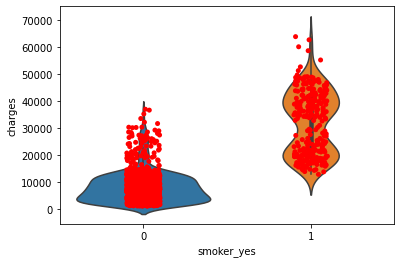
data['charges'] = y

sns.violinplot(x='smoker\_yes', y='charges', data=data)

sns.stripplot(x='smoker\_yes', y='charges', jitter=**True**, color='red', data=data)

Out[9]:

<AxesSubplot:xlabel='smoker\_yes', ylabel='charges'>



由小提琴图可见：

1. 不吸烟者（左图）大多数集中在中位数以下，中位数以上的点占少数；
2. 吸烟者（右图）小提琴图上下比较对称分布较均匀，且最小值都达到不吸烟者医疗费用的中位数；
3. 2个小提琴图对比说明吸烟者的平均医疗费用远远高于不吸烟者的平均医疗费用；

**8.7 预测测试集**

In [10]:

*# 预测测试集*

y\_pred = regressor.predict(x\_test)

In [11]:

compare\_df = pd.DataFrame(y\_test, columns=['truth'])

compare\_df['pred'] = y\_pred

compare\_df.head(10)

Out[11]:

|  | **truth** | **pred** |
| --- | --- | --- |
| **0** | 1646.42970 | 4383.680900 |
| **1** | 11353.22760 | 12885.038922 |
| **2** | 8798.59300 | 12589.216532 |
| **3** | 10381.47870 | 13286.229192 |
| **4** | 2103.08000 | 544.728328 |
| **5** | 38746.35510 | 32117.584008 |
| **6** | 9304.70190 | 12919.042372 |
| **7** | 11658.11505 | 12318.621830 |
| **8** | 3070.80870 | 3784.291456 |
| **9** | 19539.24300 | 29468.457254 |

由预测值与真实值的比较结果可知，相差不是很大，表明预测得相对准确。

**9. 实验结果（结论）**

1. 当多个自变量和因变量存在线性关系时，可以通过构建多元线性回归模型对因变量做出预测。
2. 通过多元线性回归的数学表达式，可以找到对因变量影响最大的那个自变量。