医疗费用线性回归预测，六个自变量和一个因变量（医疗费用），分别为：年龄，性别，体质指数，孩子个数，是否吸烟，地区，医疗费用。其中年龄，体质指数，孩子个数三个变量是定量变量，其他三个为定性变量。

一、回归分析。

假设误差服从 $N(0,\sigma^2)$ 分布，建立个人医疗费用和 3 个定量变量之间的线性回归方程并研究相应的统计推断问题。

由于我电脑上没装 sas，所以这次选择使用 python 来做这次作业。首先，第一步，我把这些数据都读进来。

```
data = pd.read_csv('../data.txt')
```

```
data.head()
```

	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

- 我们用“data.txt”中的前 1333 条数据（一共 1338 条数据）进行线性回归拟合。

然后用 sklearn 里面的函数以 age，bmi，children 为变量，charges 为应变量进行多元线性回归拟合得到结果。

```
from sklearn.linear_model import LinearRegression
X = data[['age','bmi','children']].values
y = data['charges'].values
X_train, y_train = X[:-5], y[:-5]
X_test,y_test = X[-5:],y[-5:]
```

```
lr = LinearRegression()
lr.fit(X_train,y_train)
```

- 用最后 5 条数据进行测试。请预测他的个人医疗费用，并给出置信度为 95%的置信区间。

接着用 lr.predict()函数进行预测，得到结果为

	age	sex	bmi	children	smoker	region	charges	预测值
1333	50	male	30.97	3	no	northwest	10600.5483	16989.312781
1334	18	female	31.92	0	no	northeast	2205.9808	8059.725787
1335	18	female	36.85	0	no	southeast	1629.8335	9705.113218
1336	21	female	25.80	0	no	southwest	2007.9450	6730.408091
1337	61	female	29.07	0	yes	northwest	29141.3603	17331.533438

然后用 wsl_prediction_std 求出置信度为 0.95 的置信区间[lower,upper]，结果为

	age	sex	bmi	children	smoker	region	charges	预测值	lower	upper
1333	50	male	30.97	3	no	northwest	10600.5483	16989.312781	-5959.954530	39007.358091
1334	18	female	31.92	0	no	northeast	2205.9808	8059.725787	-13538.729053	31424.971860
1335	18	female	36.85	0	no	southeast	1629.8335	9705.113218	-12737.954702	32256.152246
1336	21	female	25.80	0	no	southwest	2007.9450	6730.408091	-13924.872086	31005.225915
1337	61	female	29.07	0	yes	northwest	29141.3603	17331.533438	-5274.740633	39704.025393

二、方差分析。

根据上例子，利用同样的数据集（1338 条数据）：

- 利用方差分析知识，假设个人医疗费用服从方差分析模型，见（3.1）或（3.2）比较不同性别对个人医疗费用是否有显著（显著水平为 0.05）差异。

我们用 anova_lm()函数以性别作为变量求单变量的 p 值，结果为

	自由度	平方和	均方	F值	p值
sex	1.0	6.435902e+08	6.435902e+08	4.399702	0.036133
Residual	1336.0	1.954306e+11	1.462804e+08	NaN	NaN

可见 0.036<0.05，所以性别对个人医疗费用无显著差异。

- 利用方差分析知识（两因素等重复试验下），假设个人医疗费用服从两因素的方差分析模型,见教材（3.23）请对性别、是否吸烟两个因素，对方差进行分析（显著水平为 0.05）。

我们用两因素的方差分析模型。同样用 anova_lm()函数来求 p 值，结果为

	自由度	平方和	均方	F值	p值
sex	1.0	6.435902e+08	6.435902e+08	11.524608	7.069618e-04
smoker	1.0	1.208777e+11	1.208777e+11	2164.527244	1.190490e-281
Residual	1335.0	7.455290e+10	5.584487e+07	NaN	NaN

可见性别和是否吸烟同样无显著差异。