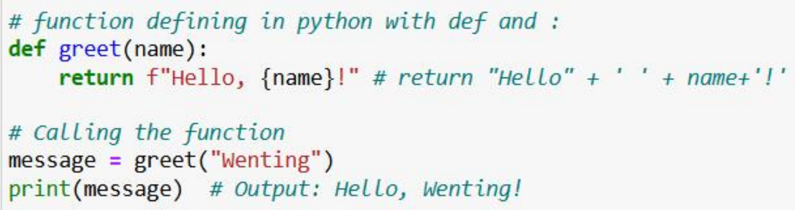
MM5425 HomeworkPlus: Linear Regression

Objectives:

- Gently get started with how to use Pandas in python
- Check the attributes' relevance using the correlation
- Predict accurately the insurance cost with multivariate linear regression
- Check the performance metrics of the linear regression model

Part I: Python Basics ~ function



```
# function defining in python with def and :
def greet(name):
    return f"Hello, {name}!" # return "Hello" + ' ' + name+'!'

# Calling the function
message = greet("Wenting")
print(message)  # Output: Hello, Wenting!
```

Exercise 1.1: Sales Tax Calculator

Write a function called calculate_sales_tax that takes two arguments: the price of an item and the sales tax rate (as a percentage). The function should return the total price including sales tax. Call the function and display the total for $170 with 12.5% tax rate.

Exercise 1.2: Employee Bonus Calculator

Task: Write a function called calculate_bonus that takes two arguments: the employee's salary and the performance rating (an integer from 1 to 5). The function should return the bonus amount based on the following criteria:

Rating 5: 20% of salary
Rating 4: 15% of salary
Rating 3: 10% of salary
Rating 2: 5% of salary
Rating 1: 0% of salary

Try the function to calculate the bonus for an employee with salary 28000 and performance rating is 3.

Part II: Linear Regression with Excel

Download the dataset insurance.csv from Blackboard in WK4. Open the file and check the contents.

This dataset has 1 target variable and 6 independent variables:

Age: age of primary beneficiary.

Gender: insurance contractor gender, female, male.

BMI: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg/m^2) using the ratio of height to weight, ideally 18.5 to 24.9.

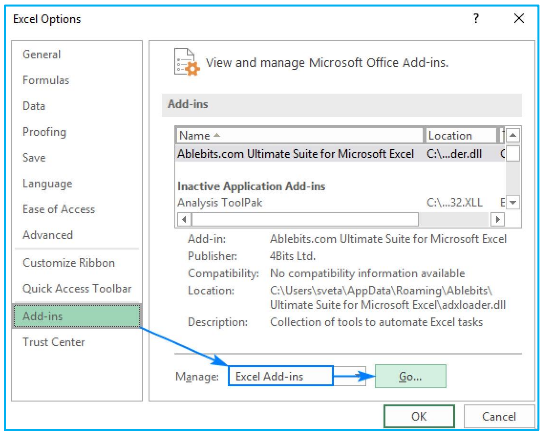
Children: Number of children covered by health insurance/Number of dependents.

Smoker: Is the person a smoker or not.

Region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.

Charges: Individual medical costs billed by health insurance.

1. Open file insurance.csv
2. Select "Data" from the toolbar. The "Data" menu displays.
3. Select "Data Analysis". The Data Analysis - Analysis Tools dialog box displays.
   - **If** you don't see 'Data Analysis',
     for Windows system, you need:
     1. In your Excel, click File > Options.
     2. In the Excel Options dialog box, select Add-ins on the left sidebar, make sure Excel Add-ins is selected in the Manage box, and click Go.



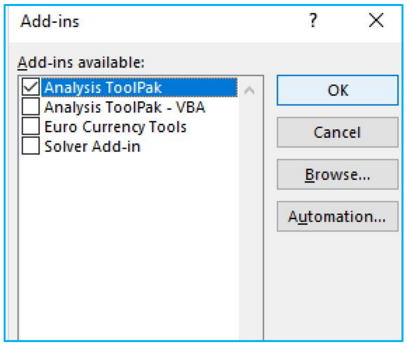
Ablebits.com Ultimate Suite for Microsoft Excel C:\...der.dll

Inactive Application Add-ins
Analysis ToolPak C:\...32.XLL

Add-in: Ablebits.com Ultimate Suite for Microsoft Excel
Publisher: 4Bits Ltd.
Compatibility: No compatibility information available
Location: C:\Users\sveta\AppData\Roaming\Ablebits\Ultimate Suite for Microsoft Excel\adxloader.dll
Description: Collection of tools to automate Excel tasks

Manage: Excel Add-ins

     3. *In the Add-ins* dialog box, tick Analysis Toolpak, and click *OK. That will add Data Analysis tools to the Data tab to your excel*

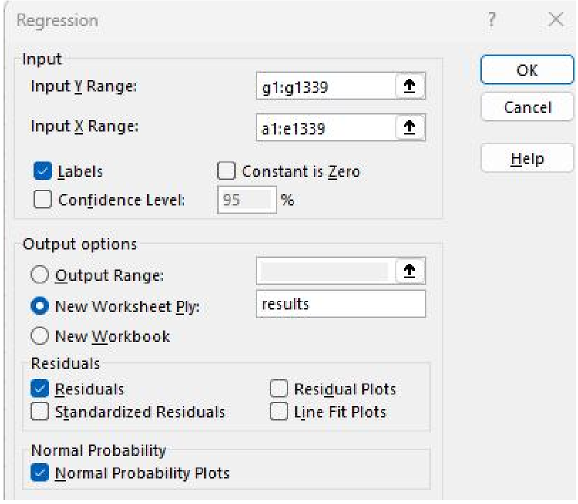


Add-ins available:
☑ Analysis ToolPak
□ Analysis ToolPak - VBA
□ Euro Currency Tools
□ Solver Add-in

For iOS:

➔ Start Excel for Mac.
➔ Click Tools, and then click Add-Ins.
➔ Click the Data Analysis ToolPak or Solver option to enable it. Then, click OK.
➔ Locate Data Analysis ToolPak or Solver on the Data tab.

4. From the menu, select "Regression" and click "OK".
5. In the Regression dialog box, click the "Input Y Range" box and select the dependent variable data. Try the following options:



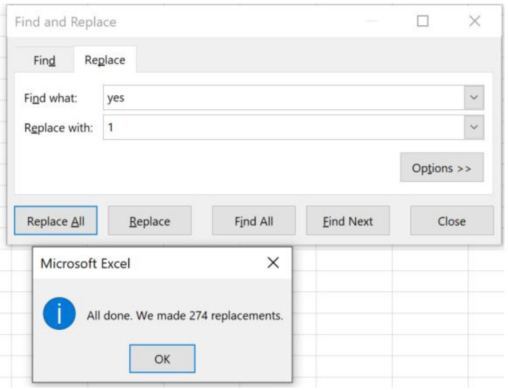
Input
Input Y Range: g1:g1339
Input X Range: a1:e1339
☑ Labels ☐ Constant is Zero
☐ Confidence Level: 95 %

Output options
○ Output Range:
● New Worksheet Ply: results
○ New Workbook

Residuals
☑ Residuals ☐ Residual Plots
☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability
☑ Normal Probability Plots

6. There's an error message showing that only numerical data can be used in the linear regression. Therefore, pressing ctrl+f keys, we replace 'yes' with 1, 'no' with 0, 'female' with 0, 'male' with 1



Find what: yes
Replace with: 1

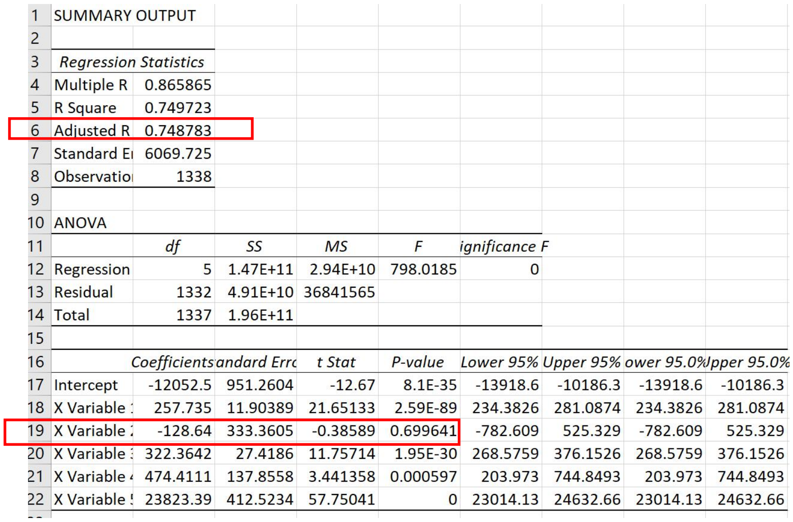
All done. We made 274 replacements.

(Question: what do we do with column region?)

7. Try the regression dialog box again.

8. Check the results in excel. Look for the following statistics:
Adjusted R square: ____________________
Coefficient of the intercept: _______________ and its p value: _________________
Coefficient of age: _________________________ and its p value: _________________
Coefficient of gender: _______________________ and its p value: _________________
Coefficient of bmi: _________________________ and its p value: _________________
Coefficient of children: _____________________ and its p value: _________________
Coefficient of smoker: _____________________ and its p value: _________________

With the calculated p-value, which attributes can be excluded in the model?



1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.865865							
5	R Square	0.749723							
6	Adjusted R	0.748783							
7	Standard E	6069.725							
8	Observatio	1338							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*ignificance F*			
12	Regression	5	1.47E+11	2.94E+10	798.0185	0			
13	Residual	1332	4.91E+10	36841565					
14	Total	1337	1.96E+11						
15									
16		*Coefficients*	*andard Errc*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Jpper 95.0%*
17	Intercept	-12052.5	951.2604	-12.67	8.1E-35	-13918.6	-10186.3	-13918.6	-10186.3
18	X Variable 1	257.735	11.90389	21.65133	2.59E-89	234.3826	281.0874	234.3826	281.0874
19	X Variable 2	-128.64	333.3605	-0.38589	0.699641	-782.609	525.329	-782.609	525.329
20	X Variable 3	322.3642	27.4186	11.75714	1.95E-30	268.5759	376.1526	268.5759	376.1526
21	X Variable 4	474.4111	137.8558	3.441358	0.000597	203.973	744.8493	203.973	744.8493
22	X Variable 5	23823.39	412.5234	57.75041	0	23014.13	24632.66	23014.13	24632.66

In the first part, the Adjusted R square is the factor used in indicating how good the model fits.
In the second part of ANOVA (analysis of variance), the df is the degree of freedom, F is the F statistic.
The most useful component in this section is Coefficients. We can see that gender (p = 0.6996) is not statistically significant at α = 0.05.

Part III: Linear regression with Python:

Introduction to Pandas

- Python Pandas provides high level, flexible and fast sets of tools to manipulate data into the right format. Pandas Focuses on loading/reading data, exploring data, basic data operations, data selection, data transformation, handling missing values, data visualization and etc..

- DataFrame provides rectangular table of data containing an ordered collection of columns of different value types (numeric, string, boolean…)

- For each column in the dataframe, we may refer it with the column name: df.age or df[‘age’]

- The data statistics of different columns can use: df.age.mean()

- We can check the correlation between the numerical variables: df.corr()

- We will try some Pandas built in plotting functions: plot(), hist(), boxplot()

- Other common operations:

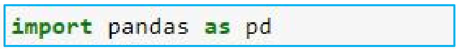
    - Replace values: df.gender.replace([‘M’, ‘F’], [1,0])

    - drop_duplicates(): df.drop_duplicates()

    - check the missing values: df.isnull().sum()

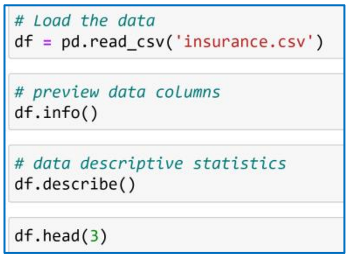
    - drop the missing values: df.dropna(inplace = True)

1. Start your jupyter notebook or google colab
2. Import the following packages: pandas, seaborn and pyplot from matplotlib



```
import pandas as pd
```

3. Read the data set and preview the data



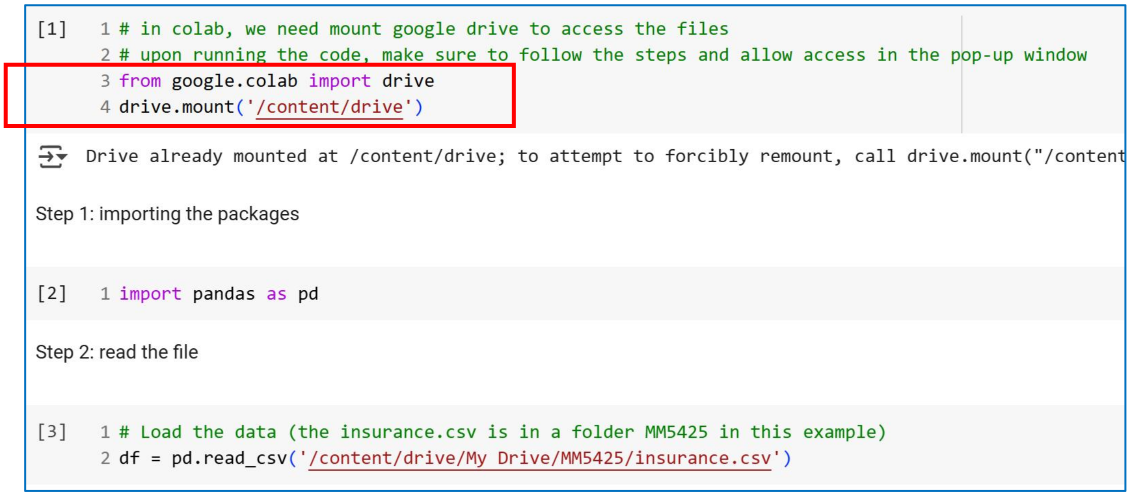
```
# Load the data
df = pd.read_csv('insurance.csv')

# preview data columns
df.info()

# data descriptive statistics
df.describe()

df.head(3)
```

(**For students who use colab,** you need save the insurance.csv to your google drive, and add the
drive to colab. Please use the following as reference:



```
[1]  # in colab, we need mount google drive to access the files
     # upon running the code, make sure to follow the steps and allow access in the pop-up window
     from google.colab import drive
     drive.mount('/content/drive')

     Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content
```

Step 1: importing the packages

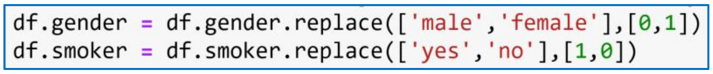
```
[2]  import pandas as pd
```

Step 2: read the file

```
[3]  # Load the data (the insurance.csv is in a folder MM5425 in this example)
     df = pd.read_csv('/content/drive/My Drive/MM5425/insurance.csv')
```

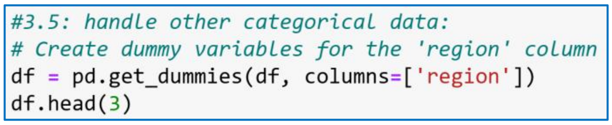
)

4. In the data preparation, we replace the categorical values of Gender and Smoker to be values of 1 or 0



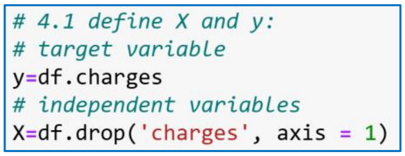
```
df.gender = df.gender.replace(['male','female'],[0,1])
df.smoker = df.smoker.replace(['yes','no'],[1,0])
```

5. For another categorical values of Region, we need use dummy variables



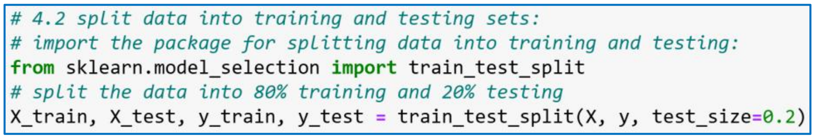
```
#3.5: handle other categorical data:
# Create dummy variables for the 'region' column
df = pd.get_dummies(df, columns=['region'])
df.head(3)
```

6. To use the linear regression model, we need find the target variable and the dependent variableshttps://www.timeanddate.com/worldclock/hong-kong/hong-kong



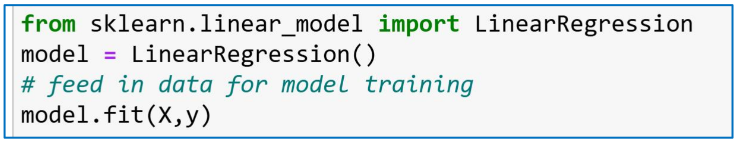
```
# 4.1 define X and y:
# target variable
y=df.charges
# independent variables
X=df.drop('charges', axis = 1)
```

7. (optional) We will split the data for training and testing in our model training:



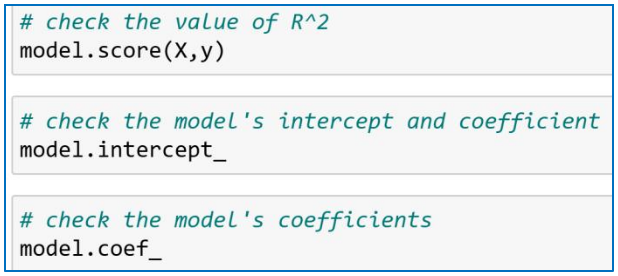
```
# 4.2 split data into training and testing sets:
# import the package for splitting data into training and testing:
from sklearn.model_selection import train_test_split
# split the data into 80% training and 20% testing
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

8. Train the model



```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
# feed in data for model training
model.fit(X,y)
```

9. At this step, we simply check the intercept and slopes in Step 5:

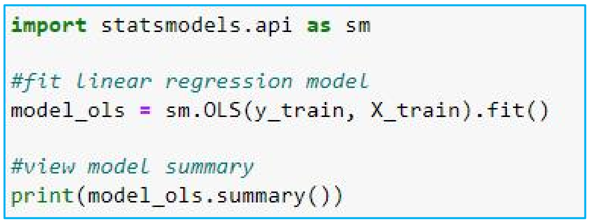


```
# check the value of R^2
model.score(X,y)
```

```
# check the model's intercept and coefficient
model.intercept_
```

```
# check the model's coefficients
model.coef_
```

10. To have a summary of the model:



```
import statsmodels.api as sm

#fit linear regression model
model_ols = sm.OLS(y_train, X_train).fit()

#view model summary
print(model_ols.summary())
```

11. Is the result consistent with that from Excel? What is the charges equation as predicted? Charges =

12. If you have a new customer who is 32 years old male with bmi=28.9, two children, non-smoking, from southeast district. How much is the predicted insurance charge for him?

```
import numpy as np
new_customer = np.array([32,1,28.9,2,0,0,0,1,0]).reshape(1,-1)
print(model.predict(new_customer))
```

Submission: please submit your Jupyter notebook file with results