Assignment 4

October 21, 2022

1 Assignment 4

Matthias Rathbun, Mir Khan, Jay Nagabhairu, Jason Ng, Phoebe Collins
10/21/2022

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
import statsmodels.formula.api as smf
import scipy.stats
```

[2]:
```
df = pd.read_csv("ACT_04_Data.csv")
```

[3]:
```
df.head()
```

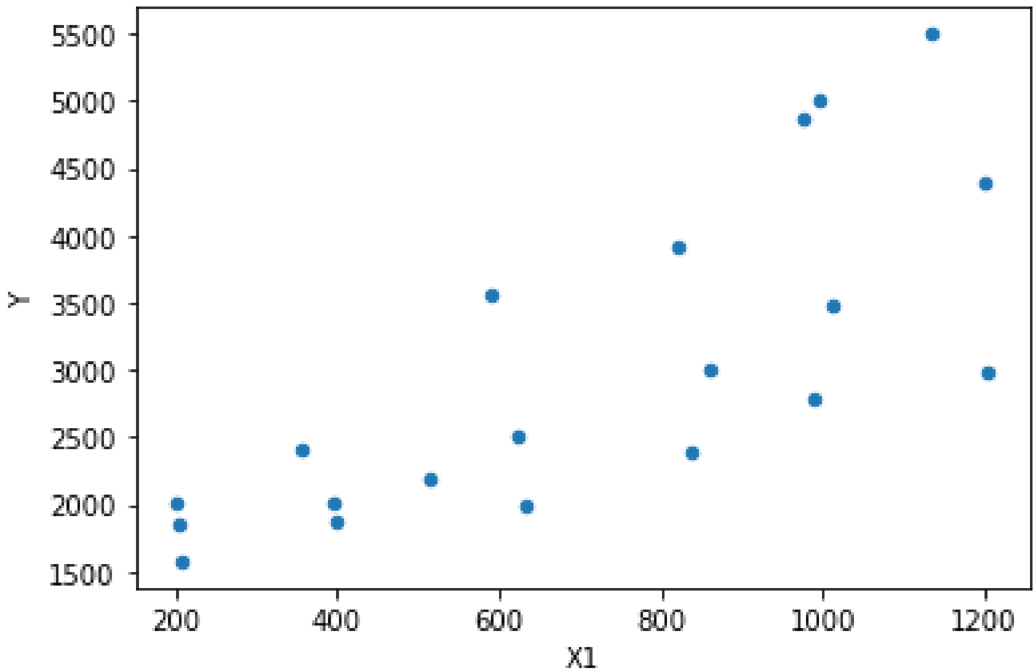
[3]:

	Y	X1	X2	X12	X1SQ	X2SQ
0	2010	201	75	15075	40401	5625
1	1850	205	50	10250	42025	2500
2	2400	355	75	26625	126025	5625
3	1575	208	30	6240	43264	900
4	3550	590	75	44250	348100	5625

[4]:
```
df = df.rename(columns = {"  Y     ":"Y"})
```

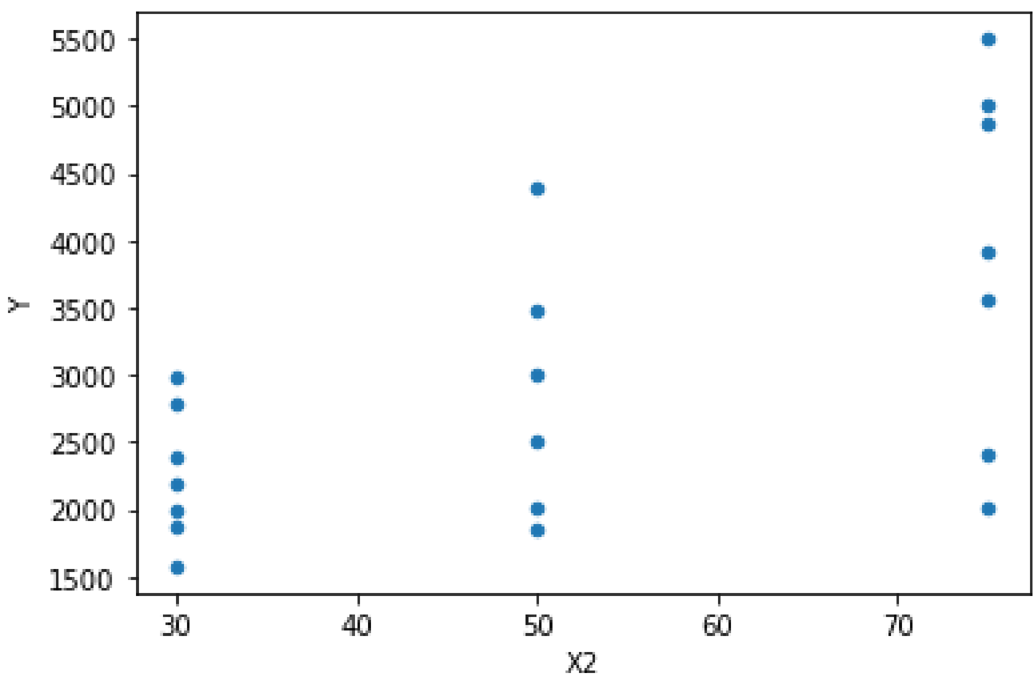
[5]:
```
df.plot.scatter(x = "X1", y = "Y")
```

[5]:
```
<AxesSubplot:xlabel='X1', ylabel='Y'>
```



[6]:
```
df.plot.scatter(x = "X2", y = "Y")
```

[6]: <AxesSubplot:xlabel='X2', ylabel='Y'>



[7]:
```
lm1 = smf.ols(formula = "Y~X1+X2", data = df).fit()
lm1.summary()
```

[7]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.900
Model:                            OLS   Adj. R-squared:                  0.888
Method:                 Least Squares   F-statistic:                     76.28
Date:                Fri, 21 Oct 2022   Prob (F-statistic):           3.23e-09
Time:                        13:21:12   Log-Likelihood:                -146.29
No. Observations:                  20   AIC:                             298.6
Df Residuals:                      17   BIC:                             301.6
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   -566.4182    312.265     -1.814      0.087   -1225.240      92.404
X1             2.5509      0.267      9.564      0.000       1.988       3.114
X2            34.2847      4.680      7.326      0.000      24.410      44.159
==============================================================================
Omnibus:                        3.525   Durbin-Watson:                   1.870
Prob(Omnibus):                  0.172   Jarque-Bera (JB):                1.410
Skew:                           0.191   Prob(JB):                        0.494
Kurtosis:                       1.756   Cond. No.                     2.77e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 2.77e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[8]:
```
sm.stats.anova_lm(lm1, typ=1)
```

[8]:

	df	sum_sq	mean_sq	F	PR(>F)
X1	1.0	1.537060e+07	1.537060e+07	98.894347	1.679621e-08
X2	1.0	8.340715e+06	8.340715e+06	53.664106	1.185020e-06
Residual	17.0	2.642216e+06	1.554245e+05	NaN	NaN

[9]:
```
lm2 = smf.ols(formula = "Y~X1+X2+X12", data = df).fit()
lm2.summary()
```

[9]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.978
Model:                            OLS   Adj. R-squared:                  0.974
Method:                 Least Squares   F-statistic:                     239.4
Date:                Fri, 21 Oct 2022   Prob (F-statistic):           1.68e-13
Time:                        13:21:12   Log-Likelihood:                -131.03
No. Observations:                  20   AIC:                             270.1
Df Residuals:                      16   BIC:                             274.0
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   1340.1506    292.598      4.580      0.000     719.870    1960.431
X1            -0.1735      0.381     -0.455      0.655      -0.981       0.634
X2            -2.8062      5.379     -0.522      0.609     -14.210       8.598
X12            0.0525      0.007      7.590      0.000       0.038       0.067
==============================================================================
Omnibus:                        2.452   Durbin-Watson:                   1.397
Prob(Omnibus):                  0.294   Jarque-Bera (JB):                1.582
Skew:                          -0.457   Prob(JB):                        0.453
Kurtosis:                       1.969   Cond. No.                     3.01e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 3.01e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[10]:
```
sm.stats.anova_lm(lm2, typ=1)
```

[10]:

	df	sum_sq	mean_sq	F	PR(>F)
X1	1.0	1.537060e+07	1.537060e+07	428.222921	5.653519e-13
X2	1.0	8.340715e+06	8.340715e+06	232.371222	5.998742e-11
X12	1.0	2.067913e+06	2.067913e+06	57.611787	1.089461e-06
Residual	16.0	5.743028e+05	3.589392e+04	NaN	NaN

[11]:
```
lm3 = smf.ols(formula = "Y~X1+X2+X12+X1SQ+X2SQ", data = df).fit()
lm3.summary()
```

[11]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.986
Model:                            OLS   Adj. R-squared:                  0.981
Method:                 Least Squares   F-statistic:                     199.4
Date:                Fri, 21 Oct 2022   Prob (F-statistic):           1.71e-12
Time:                        13:21:12   Log-Likelihood:                -126.50
No. Observations:                  20   AIC:                             265.0
Df Residuals:                      14   BIC:                             271.0
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   2405.7779    495.491      4.855      0.000    1343.055    3468.501
X1            -1.5734      0.680     -2.315      0.036      -3.031      -0.115
X2           -32.4259     17.270     -1.878      0.081     -69.466       4.614
X12            0.0542      0.006      9.092      0.000       0.041       0.067
X1SQ           0.0010      0.000      2.379      0.032    9.37e-05       0.002
X2SQ           0.2684      0.158      1.695      0.112      -0.071       0.608
==============================================================================
Omnibus:                        0.504   Durbin-Watson:                   1.835
Prob(Omnibus):                  0.777   Jarque-Bera (JB):                0.589
Skew:                           0.295   Prob(JB):                        0.745
Kurtosis:                       2.402   Cond. No.                     1.06e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.06e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[12]:
```
sm.stats.anova_lm(lm3, typ=1)
```

[12]:

	df	sum_sq	mean_sq	F	PR(>F)
X1	1.0	1.537060e+07	1.537060e+07	589.561805	7.645837e-13
X2	1.0	8.340715e+06	8.340715e+06	319.920281	4.860083e-11
X12	1.0	2.067913e+06	2.067913e+06	79.317822	3.844920e-07
X1SQ	1.0	1.343881e+05	1.343881e+05	5.154650	3.950402e-02
X2SQ	1.0	7.491751e+04	7.491751e+04	2.873570	1.121613e-01
Residual	14.0	3.649972e+05	2.607123e+04	NaN	NaN

[13]:
```
lm4 = smf.ols(formula = "Y~X1+X2+X12+X1SQ", data = df).fit()
lm4.summary()
```

[13]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.983
Model:                            OLS   Adj. R-squared:                  0.979
Method:                 Least Squares   F-statistic:                     220.9
Date:                Fri, 21 Oct 2022   Prob (F-statistic):           3.91e-13
Time:                        13:21:12   Log-Likelihood:                -128.36
No. Observations:                  20   AIC:                             266.7
Df Residuals:                      15   BIC:                             271.7
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   1746.3777    325.522      5.365      0.000    1052.544    2440.211
X1            -1.5279      0.720     -2.121      0.051      -3.063       0.008
X2            -4.2211      4.907     -0.860      0.403     -14.681       6.238
X12            0.0545      0.006      8.616      0.000       0.041       0.068
X1SQ           0.0009      0.000      2.141      0.049    3.89e-06       0.002
==============================================================================
Omnibus:                        0.481   Durbin-Watson:                   1.617
Prob(Omnibus):                  0.786   Jarque-Bera (JB):                0.590
Skew:                           0.255   Prob(JB):                        0.745
Kurtosis:                       2.330   Cond. No.                     6.54e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 6.54e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[14]:
```python
sm.stats.anova_lm(lm4, typ=1)
```

[14]:

	df	sum_sq	mean_sq	F	PR(>F)
X1	1.0	1.537060e+07	1.537060e+07	524.099340	4.418034e-13
X2	1.0	8.340715e+06	8.340715e+06	284.397677	3.675408e-11
X12	1.0	2.067913e+06	2.067913e+06	70.510704	4.718263e-07
X1SQ	1.0	1.343881e+05	1.343881e+05	4.582299	4.913919e-02
Residual	15.0	4.399147e+05	2.932765e+04	NaN	NaN

[15]:
```python
data = [[lm1.rsquared, lm1.mse_resid],[lm2.rsquared, lm2.mse_resid],[lm3.
↪rsquared, lm3.mse_resid],[lm4.rsquared, lm4.mse_resid]]
results = pd.DataFrame(data, columns=['R Squared', 'MSE'])
```

```
results.index.name = 'Model'
results.index += 1
```

[16]:
```
results
```

[16]:
```
       R Squared            MSE
Model
1       0.899740  155424.460740
2       0.978208   35893.923075
3       0.986150   26071.228596
4       0.983307   29327.647321
```

[17]:
```
q1 = scipy.stats.norm.ppf(0.0125)
```

[18]:
```
q2 = scipy.stats.norm.ppf(0.99)
```

[19]:
```
q3 = scipy.stats.t.ppf(0.0125, 333)
```

[20]:
```
q4 = scipy.stats.t.ppf(0.99, 345)
```

[21]:
```
q5 = scipy.stats.chi2.ppf(0.025, 125)
```

[22]:
```
q6 = scipy.stats.t.ppf(0.975, 245)
```

[23]:
```
q7 = scipy.stats.f.ppf(0.01, 12, 250)
```

[24]:
```
q8 = scipy.stats.f.ppf(0.99, 24, 500)
```

[25]:
```
data = [["Normal",None,None, 0.0125, q1],["Normal",None,None, 0.99,␣
↪q2],["Student t",333,None, 0.0125, q3],["Student t",345,None, 0.99, q4],
       ["Chi-Square",125,None, 0.025, q5],["Chi-Square",245,None, 0.975,␣
↪q6],["F",12,250, 0.01, q7],["F",24,500, 0.99, q8]]
```

[26]:
```
quantiles = pd.DataFrame(data, columns=['Distribution', 'Degrees Freedom I',␣
↪'Degrees Freedom II', 'Probability', 'Quantile'])
quantiles = quantiles.set_index("Distribution")
```

[27]:
```
quantiles
```

[27]:
```
              Degrees Freedom I  Degrees Freedom II  Probability   Quantile
Distribution
Normal                      NaN                 NaN       0.0125  -2.241403
Normal                      NaN                 NaN       0.9900   2.326348
Student t                 333.0                 NaN       0.0125  -2.251584
Student t                 345.0                 NaN       0.9900   2.337205
Chi-Square                125.0                 NaN       0.0250  95.945725
Chi-Square                245.0                 NaN       0.9750   1.969694
F                          12.0               250.0       0.0100   0.293798
```

F	24.0	500.0	0.9900	1.828539