task

December 6, 2022

1 Set-Up

[3]:
```
import pandas as pd
import numpy as np
import patsy
import matplotlib.pyplot as plt
from sklearn.svm import SVC
from sklearn.metrics import confusion_matrix, classification_report,␣
 ↪precision_recall_fscore_support
import seaborn as sns
from sklearn import preprocessing
from sklearn.preprocessing import MinMaxScaler
import statsmodels.api as sm
import scipy.stats as stats
import warnings
from mpl_toolkits.mplot3d import Axes3D
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from sklearn import linear_model
np.random.seed(19680801)
warnings.filterwarnings('ignore')
sns.set()
sns.set_context("paper", font_scale=10)
```

2 Loading Dataframe

[20]:
```
#load csv
task = pd.read_csv('SampleTaskData[42][77][59].csv')
task = task.dropna(axis=1)
print(task)
```

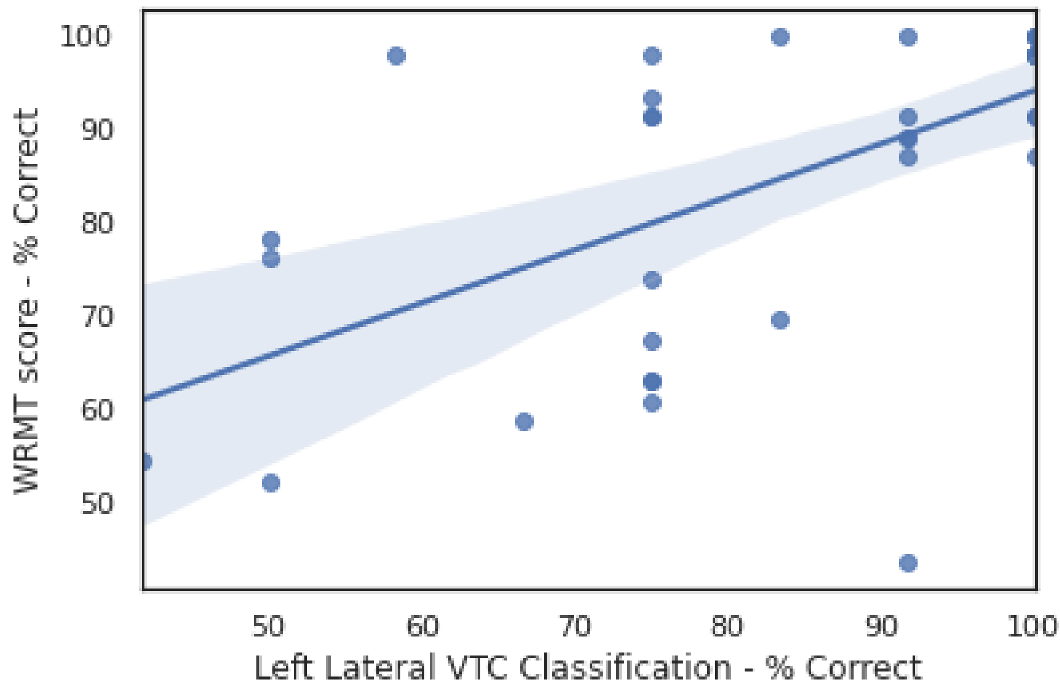
```
    Subject Number    WRMT  Left Lateral VTC Classification  Age
0                1   43.48                            91.67    6
1                2   54.35                            41.67    6
2                3   73.91                            75.00    7
```

3	4	91.30	75.00	8
4	5	52.17	50.00	7
5	6	60.87	75.00	8
6	7	78.26	50.00	9
7	8	69.57	83.33	10
8	9	63.04	75.00	10
9	10	86.96	91.67	10
10	11	67.39	75.00	10
11	12	89.13	91.67	10
12	13	76.09	50.00	10
13	14	63.04	75.00	11
14	15	58.70	66.67	11
15	16	91.30	100.00	11
16	17	91.30	100.00	11
17	18	89.13	91.67	12
18	19	100.00	91.67	22
19	20	97.83	100.00	22
20	21	97.83	100.00	22
21	22	100.00	83.33	22
22	23	97.83	100.00	23
23	24	97.83	100.00	23
24	25	97.83	75.00	23
25	26	100.00	100.00	23
26	27	97.83	100.00	23
27	28	97.83	100.00	24
28	29	86.96	100.00	24
29	30	100.00	100.00	24
30	31	98.00	58.33	24
31	32	100.00	100.00	24
32	33	91.30	75.00	24
33	34	91.30	91.67	24
34	35	100.00	100.00	25
35	36	97.83	100.00	26
36	37	93.48	75.00	28

3 Classification performances and WRMT Scores

[21]:
```
sns.set(style="white", font_scale = 1)
plot1 = sns.regplot(x ='Left Lateral VTC Classification', y ='WRMT', data =␣
 ↪task)
plot1.set_ylabel('WRMT score - % Correct')
plot1.set_xlabel('Left Lateral VTC Classification - % Correct')
```

[21]: Text(0.5, 0, 'Left Lateral VTC Classification - % Correct')



[22]:
```
from scipy.stats import linregress
linregress(task['Left Lateral VTC Classification'], task['WRMT'])
```

[22]:
```
LinregressResult(slope=0.5683353502611742, intercept=37.21850848691025,
rvalue=0.5952572867165984, pvalue=0.00010184422394926359,
stderr=0.1296793470537376, intercept_stderr=11.11905296823655)
```

[23]:
```
task.rename(columns = {'Left Lateral VTC Classification':'VTC'}, inplace = True)
outcome_1, predictors_1 = patsy.dmatrices('WRMT ~ VTC', task)
model = sm.OLS(outcome_1, predictors_1)
results = model.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   WRMT   R-squared:                       0.354
Model:                            OLS   Adj. R-squared:                  0.336
Method:                 Least Squares   F-statistic:                     19.21
Date:                Tue, 06 Dec 2022   Prob (F-statistic):           0.000102
Time:                        21:51:37   Log-Likelihood:                -147.85
No. Observations:                  37   AIC:                             299.7
Df Residuals:                      35   BIC:                             302.9
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
```

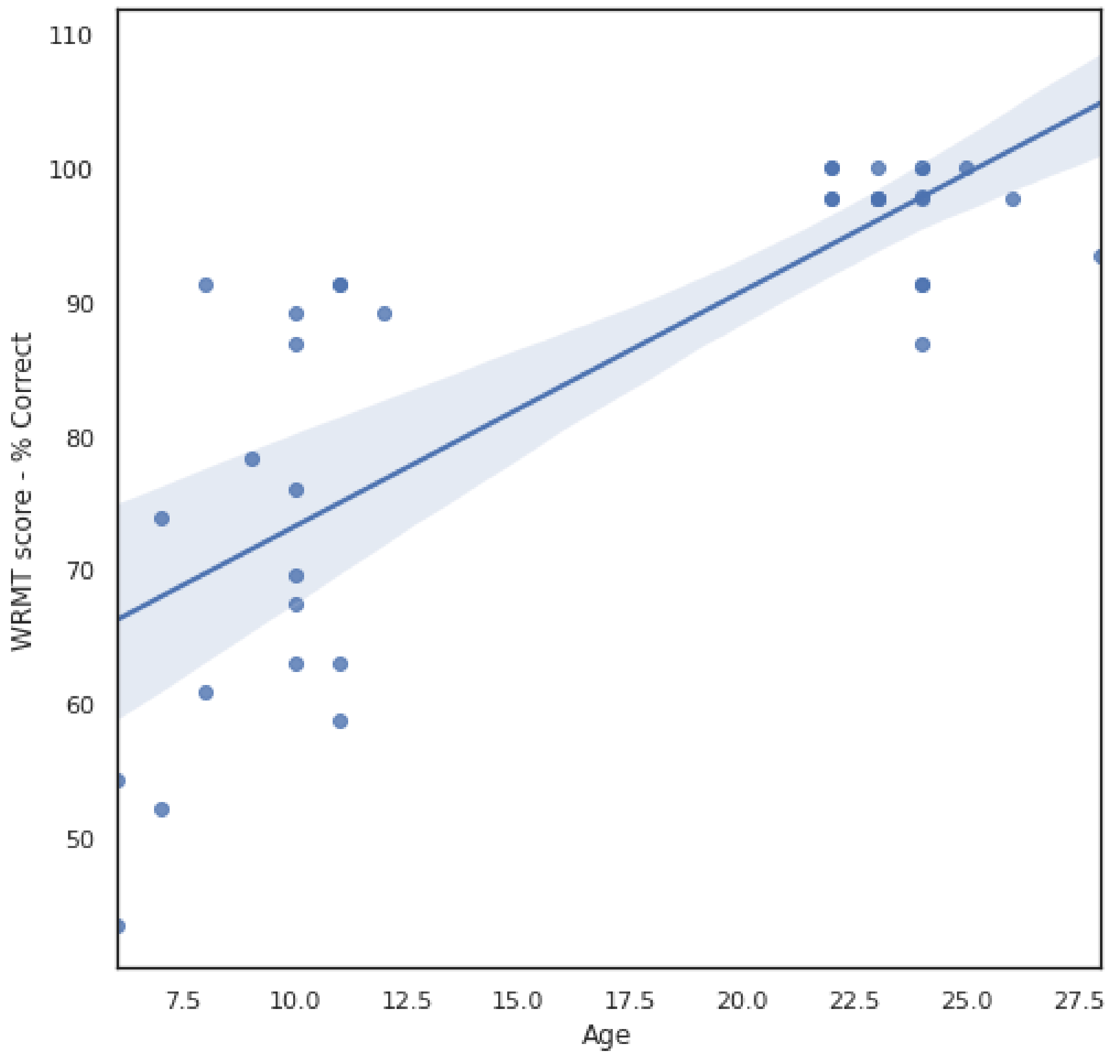
```
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     37.2185     11.119      3.347      0.002      14.646      59.791
VTC            0.5683      0.130      4.383      0.000       0.305       0.832
==============================================================================
Omnibus:                       12.093   Durbin-Watson:                   1.253
Prob(Omnibus):                  0.002   Jarque-Bera (JB):               13.517
Skew:                          -0.996   Prob(JB):                      0.00116
Kurtosis:                       5.191   Cond. No.                         429.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Interpetation: 35.4% of the variation in the WRMT score can be explained from the left lateral VTC Classification. Since the p-value for the F-test is P < .001, we can conclude that the regression model is statistically signficant. Left lateral VTC Classification has a statistically significant relationship with the WRMT score. This model has an AIC of 299.7, so we should consider this when comparing it to other models.

4 WRMT Scores and Age

[30]:
```python
sns.set(style="white", font_scale = 1)
plot2 = sns.regplot(x ='Age', y ='WRMT', data = task)
plot2.figure.set_size_inches(8,8)
plot2.set_ylabel('WRMT score - % Correct')
```

[30]:
```
Text(0, 0.5, 'WRMT score - % Correct')
```



[31]:
```
from scipy.stats import linregress
linregress(task['Age'], task['WRMT'])
```

[31]:
```
LinregressResult(slope=1.757236446994478, intercept=55.660948978497494,
rvalue=0.7920516499613185, pvalue=5.238165037003949e-09,
stderr=0.2289262941333921, intercept_stderr=4.174799520780392)
```

[32]:
```
outcome_2, predictors_2 = patsy.dmatrices('WRMT ~ Age', task)
model_2 = sm.OLS(outcome_2, predictors_2)
results = model_2.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   WRMT   R-squared:                       0.627
```

```
Model:                            OLS   Adj. R-squared:                  0.617
Method:                 Least Squares   F-statistic:                     58.92
Date:                Tue, 06 Dec 2022   Prob (F-statistic):           5.24e-09
Time:                        21:58:33   Log-Likelihood:                -137.69
No. Observations:                  37   AIC:                             279.4
Df Residuals:                      35   BIC:                             282.6
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     55.6609      4.175     13.333      0.000      47.186      64.136
Age            1.7572      0.229      7.676      0.000       1.292       2.222
==============================================================================
Omnibus:                        0.013   Durbin-Watson:                   1.626
Prob(Omnibus):                  0.993   Jarque-Bera (JB):                0.189
Skew:                          -0.011   Prob(JB):                        0.910
Kurtosis:                       2.651   Cond. No.                         45.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Interpetation: 62.7% of the variation in the WRMT score can be explained from Age. Since the p-value for the F-test is P $<$ .001, we can conclude that the regression model is statistically signficant. Age has a statistically significant relationship with the WRMT score. The AIC for this model is less than the previous model, meaning it is more likely to be the best model.

5 WRMT Scores Accounting for Multiple Variables

[27]:
```python
x = task[['VTC','Age']]
y = task['WRMT']

regr = linear_model.LinearRegression()
regr.fit(x, y)

print('Intercept: \n', regr.intercept_)
print('Coefficients: \n', regr.coef_)
```

```
Intercept:
 39.82379246929776
Coefficients:
 [0.24656801 1.46478119]
```

WRMT = (39.82379246929776) + (0.24656801)*X1 + (1.46478119)*X2

The model shown above can be used to estimate the WRMT scores considering both the subjects age (X2) and Left Lateral VTC classification (X1).

[28]:
```
outcome_3, predictors_3 = patsy.dmatrices("WRMT ~ VTC + Age", task)
mod_3 = sm.OLS(outcome_3, predictors_3)
res_3 = mod_3.fit()
print(res_3.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   WRMT   R-squared:                       0.677
Model:                            OLS   Adj. R-squared:                  0.658
Method:                 Least Squares   F-statistic:                     35.58
Date:                Tue, 06 Dec 2022   Prob (F-statistic):           4.62e-09
Time:                        21:51:58   Log-Likelihood:                -135.06
No. Observations:                  37   AIC:                             276.1
Df Residuals:                      34   BIC:                             281.0
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     39.8238      7.996      4.981      0.000      23.574      56.073
VTC            0.2466      0.108      2.277      0.029       0.027       0.467
Age            1.4648      0.252      5.822      0.000       0.953       1.976
==============================================================================
Omnibus:                        2.693   Durbin-Watson:                   1.680
Prob(Omnibus):                  0.260   Jarque-Bera (JB):                1.526
Skew:                          -0.373   Prob(JB):                        0.466
Kurtosis:                       3.658   Cond. No.                         438.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

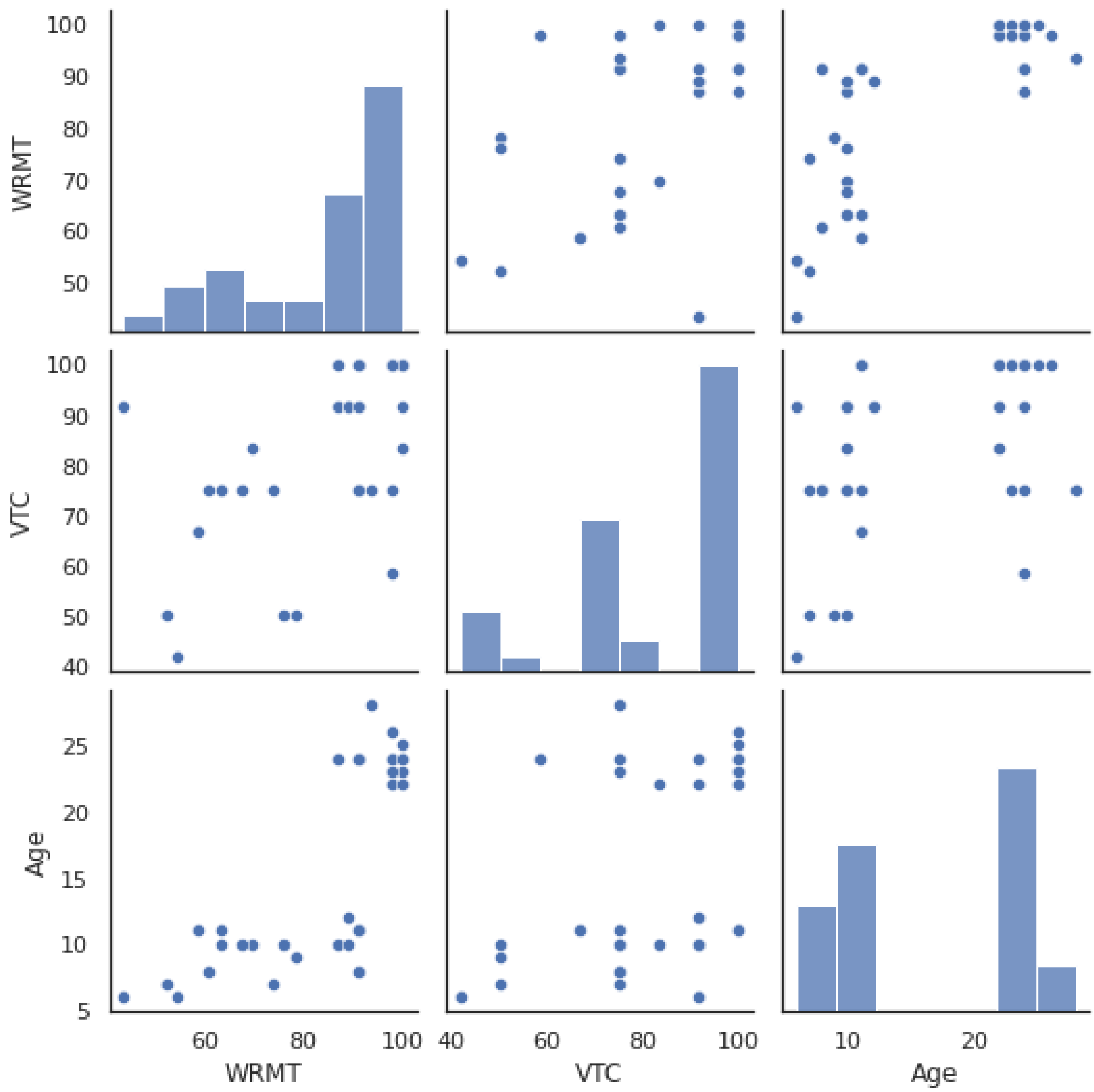
Interpetation: 65.8% (adjusted) of the variation in the WRMT score can be explained from both the left lateral VTC Classification and Age. Since the p-value for the F-test is P $<$ .001, we can conclude that the regression model is statistically signficant. This is expected because both variables have statistically significant relationships with WRMT score. Considering that this model has the lowest AIC of the three, this is most likely to be the best model.

[18]:
```
task_corr = task.drop('Subject Number', axis=1)
corr_matrix = task_corr.corr()
corr_matrix
```

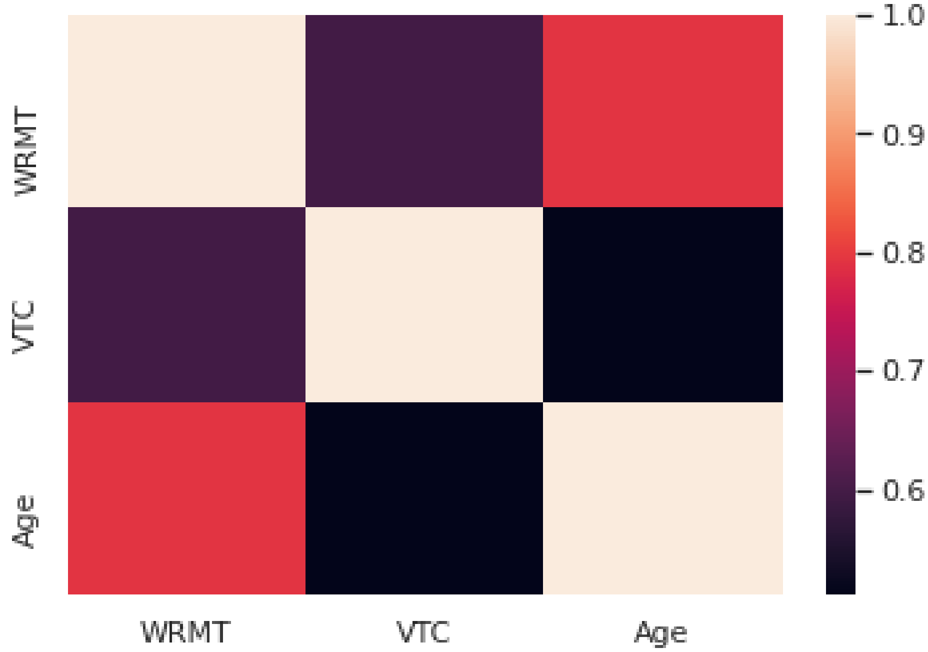
[18]:

	WRMT	VTC	Age
WRMT	1.000000	0.595257	0.792052
VTC	0.595257	1.000000	0.510441
Age	0.792052	0.510441	1.000000

[19]:
```
sns.pairplot(task_corr, kind="scatter")
plt.show()
```



[62]:
```
a = sns.heatmap(corr_matrix);
```


WRMT
VTC
Age
WRMT
VTC
Age
1.0
0.9
0.8
0.7
0.6