Statistical modelling with linear regression

Overview of the data

The data used is a powerlifting csv data set sourced from kaggle, and it comes from the open powerlifting project which aims to document all competitive powerlifting meets worldwide.

Powerlifters compete in meets to be officially recognised within the sport. Some meets use no extra equipment, “raw” whilst others allow the use of specialist lifting equipment. In each meet a lifter gets 3 attempts at the squat/bench/deadlift and the highest number is recorded.

At the end, the best attempts are totalled and the lifter with the highest total in their weight class wins.

Variables in the dataset

The dataset includes a variety of different variables that may be used as features to train a model.

More prominent ones include: Sex,Bodyweight,Age, Lifts, Placement, Wilks and Tested.

The best bench press recorded is the target variable for this linear regression model and all the others will be used as features.

Reading and cleaning the data

The data is read into a pandas dataframe and parsed so only useful columns are within the dataframe

```
cleaned_df =
df[['Sex','Event','Equipment','Age','BodyweightKg','Best3SquatKg','Best3BenchKg','Best3DeadliftKg','Tested']]
```

The data is further filtered so that only lifters who have completed all 3 lifts and have been tested can contribute. This is because I wished to use the other lifts as predictors.

```
cleaned_df=cleaned_df[(cleaned_df['Event'] == 'SBD') & (cleaned_df['Tested'] == 'Yes')]
```

Rows with missing values are dropped.

Expanding on the current feature choice

Wilks- This feature was dropped as it is directly calculated from the total lifted weight and the lifter’s bodyweight using the formula

Total lifted- This feature was dropped as it was a sum of the respective lifts and like the wilks would introduce multicollinearity.

After filtering, the tested and events column were dropped

Data Splitting

Using SKLearn, the data is split 70:30 so that the training data can be used to fit a linear model.

```
from sklearn.model_selection import train_test_split
```

```
X_train, X_test, y_train, y_test=train_test_split(X,y, train_size = .7)
```

Initial regression model

Use `cleaned_df.info()`

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 332269 entries, 8164 to 1423351
Data columns (total 7 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   Sex              332269 non-null  object
 1   Equipment        332269 non-null  object
 2   Age              332269 non-null  float64
 3   BodyweightKg     332269 non-null  float64
 4   Best3SquatKg     332269 non-null  float64
 5   Best3BenchKg     332269 non-null  float64
 6   Best3DeadliftKg  332269 non-null  float64
dtypes: float64(5), object(2)
memory usage: 20.3+ MB
```

Given a clean model the statsmodel is used to fit a linear regression model.

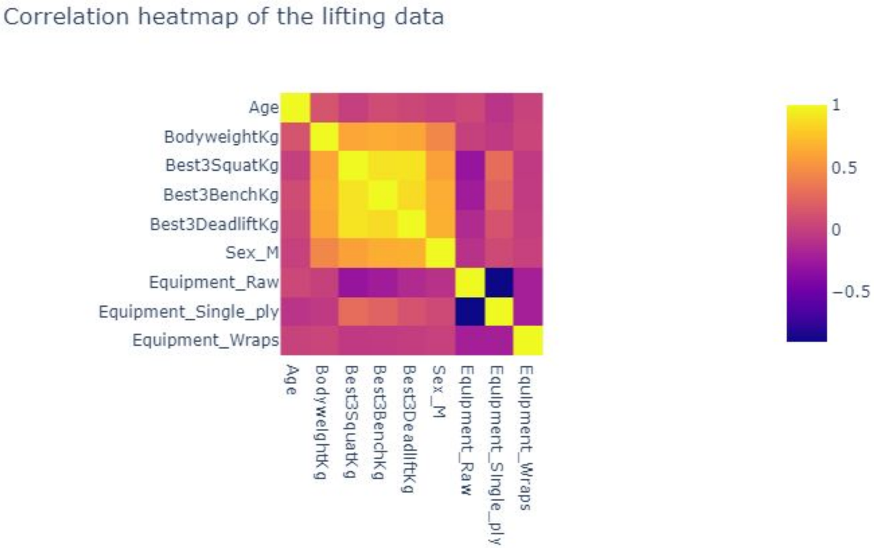
```
                 OLS Regression Results
   Dep. Variable:      Best3BenchKg          R-squared:        0.852
           Model:               OLS      Adj. R-squared:        0.852
          Method:     Least Squares         F-statistic:    2.387e+05
            Date:  Mon, 05 Apr 2021  Prob (F-statistic):        0.00
            Time:          21:32:58      Log-Likelihood:  -1.4560e+06
No. Observations:            332269                 AIC:    2.912e+06
    Df Residuals:            332260                 BIC:    2.912e+06
        Df Model:                 8
 Covariance Type:         nonrobust

                            coef  std err        t  P>|t|   [0.025   0.975]
              Intercept -19.3448    0.450  -43.005  0.000  -20.226  -18.463
              Sex[T.M]   15.4088    0.098  157.268  0.000   15.217   15.601
        Equipment[T.Raw]  -6.9240    0.428  -16.193  0.000   -7.762   -6.086
Equipment[T.Single-ply]   -3.8233    0.424   -9.027  0.000   -4.653   -2.993
      Equipment[T.Wraps]  -7.7684    0.450  -17.250  0.000   -8.651   -6.886
                     Age   0.2254    0.003   82.381  0.000    0.220    0.231
            BodyweightKg   0.2460    0.002  120.078  0.000    0.242    0.250
            Best3SquatKg   0.4463    0.001  337.451  0.000    0.444    0.449
         Best3DeadliftKg   0.1193    0.001   80.770  0.000    0.116    0.122

       Omnibus:  86468.451      Durbin-Watson:        1.696
Prob(Omnibus):      0.000   Jarque-Bera (JB):  3079921.612
          Skew:      0.572           Prob(JB):         0.00
      Kurtosis:     17.871          Cond. No.      7.54e+03
```

Multicollinearity within the data

Although the model ran well, there may be some underlying collinearity.

To test this ran both a vif test and created a correlation heatmap. Dummy variables were created to introduce this functionality.

Results indicate correlation between lifts and very high negative correlation between equipment category raw and single.




```
vif_lst

[{'Age': 1.0591169258682418},
 {'BodyweightKg': 1.8942063130630529},
 {'Best3SquatKg': 9.88962424652017},
 {'Best3BenchKg': 6.746268727284266},
 {'Best3DeadliftKg': 7.432579151013646},
 {'Sex_M': 1.9300530190329817},
 {'Equipment_Raw': 40.43518599545051},
 {'Equipment_Single_ply': 39.63785041638041},
 {'Equipment_Wraps': 8.318615645554301}]
```

Final regression model

By dropping the categorical data associated with single equipment, the final model saw a marginal increase in performance.

Calculated R--squared=0.853

OLS Regression Results

Dep. Variable:	Best3BenchKg	R-squared:	0.851
Model:	OLS	Adj. R-squared:	0.851
Method:	Least Squares	F-statistic:	1.897e+05
Date:	Tue, 06 Apr 2021	Prob (F-statistic):	0.00
Time:	17:51:18	Log-Likelihood:	-1.0200e+06
No. Observations:	232588	AIC:	2.040e+06
Df Residuals:	232580	BIC:	2.040e+06
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-23.2141	0.188	-123.463	0.000	-23.583	-22.846
Age	0.2283	0.003	69.638	0.000	0.222	0.235
BodyweightKg	0.2474	0.002	100.601	0.000	0.243	0.252
Best3SquatKg	0.4463	0.002	283.167	0.000	0.443	0.449
Best3DeadliftKg	0.1192	0.002	67.751	0.000	0.116	0.123
Sex_M	15.3281	0.117	130.687	0.000	15.098	15.558
Equipment_Raw	-3.1692	0.095	-33.448	0.000	-3.355	-2.983
Equipment_Wraps	-3.9403	0.196	-20.145	0.000	-4.324	-3.557

Omnibus:	61048.639	Durbin-Watson:	1.998
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2442203.034
Skew:	0.543	Prob(JB):	0.00
Kurtosis:	18.837	Cond. No.	1.49e+03

Potential uses

Determine if a given powerlifter has a lagging lift.

Give people a general indicator of what their strength could feasibly be at their current level