The Wing Length of Blackbirds Analysis Based on OLS Model

35427962

Abstract—**The purpose of this paper is to study the related factors of weight change of blackbirds. Since the weight of a blackbird varies greatly, its wing length is generally used to represent its weight instead of its real weight.**

GLM model is used to analyse information on blackbirds from the same garden in the East Midlands between 1988 and 2015. Missing values are filled in using the mean of the column data. Since the same bird is recorded many times in the dataset, it is necessary to filter the duplicate records in the data cleaning. The relevant data show that the time factor should be ignored in the research, so the earliest records of the same bird should be extracted after grouping. The classification variables are used as the independent variable and the wing length as the dependent variable for OLS fitting. As a result, the overall and single coefficients of the model are analyzed by significance. However, the goodness of fit of the model is only 44.8%.

The results show that there is a positive correlation between age and wing length, as well as the wing length of the male blackbirds is longer than that of the female.

Key Words: Blackbirds; OLS; Linear Regression

I. Introduction

According to Feu (n.d.) published in Mathematics Education Innovation, blackbirds have unstable body weight, sometimes less than 90g, sometimes more than 130g. They can change their weight according to conditions and gain weight on cold nights to burn approximately 5% of their body mass for warmth. At the end of cold conditions, they lose weight and fly faster to avoid predators. The way to assess the standard size of a blackbird is to measure the length of its wings which is the distance from the carpal joint to the wingtip.

The dataset used in this analysis is 25 years of blackbirds captured in the same garden in East Midland, including foreign birds with foot rings and a small number of native British birds. When the birds are first caught, they are put on foot rings so that they can clearly trace back to the date of the first capture.

Some birds have been repeatedly trapped, according to the data requirements, this part of the data needs to be ignored. There are some missing values in the dataset, including age, gender, wing length and weight.

II. Methods

In this paper, NumPy, Pandas, and Statsmodels libraries in python are used to clean, calculate and analyze the dataset.

For one thing, data needs to be filled and cleaned up. Firstly, based on the information given by the dataset, vacancy value need to be filled at first. The gender of qualitative variables will be filled with the data of the same bird, while the wings and weight of quantitative variables will be filled with the average value of the whole row.

Secondly, genders and ages are recoded using dummy variables. There are only two values, namely 0 and 1. When the variables meet the conditions, the value of the matrix is 1, vice versa.

Finally, according to the suggestions in the analysis of Feu (n.d.) in the data, it can be seen that different data of the same bird species need to be ignored. I grouped the *Ring Number* variable and extracted only the first observation of the grouping results, and then reconstructed the data frame to be calculated.

For another, the adjusted dataset is fitted by the ordinary least square (OLS) model. At the outset, the 0-1 matrix composed of dummy variables is used as the independent variable, while the wing length is the dependent variable. In order to prevent multicollinearity in the model, only n-1 dummy variables are selected for calculation when the type of dummy variable is n.

Therefore, the final OLS model was given as

$$y = \beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 x_5+\beta_6 x_6+\epsilon \quad (1)$$

where y is the length of blackbird's wing, x_1 is female, x_2 is male, x_3 is adult, x_4 is the first-year, and x_5 is juvenile.

III. Results

The following results are calculated using Statsmodels (Seabold, S. and Perktold, J., 2020) library of python as follow.

TABLE I
OLS Regression Results

Var:	Wing	R-sq:	0.448
Model:	OLS	Ad. R-sq:	0.447
Method:	Least Squares	F:	346.4
P:	3.12e-272	Log-Likelihood:	-5441.6
No. Obs:	2141	AIC:	1.090e+04
Df Res:	2135	BIC:	1.093e+04
Df Model:	5		
Cov Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	130.45	0.64	204.07	0.00	129.20	131.71
Sex_F	-2.09	0.37	-5.66	0.00	-2.82	-1.37
Sex_M	2.23	0.36	6.24	0.00	1.52	2.93
Age_A	1.41	0.54	2.60	0.01	0.35	2.47
Age_F	-2.23	0.54	-4.15	0.00	-3.29	-1.18
Age_J	-3.96	0.57	-6.95	0.00	-5.07	-2.84

According to the table I, it can be seen that the p-value of F-test of overall model is 3.12×10^{-272}, which is far less than 0.001, and the p-value of all coefficients' T-test is less than 0.05. This shows that the model is much significant. As well as, the value of R^2 is 0.448, which means the goodness of fit of the model was only 44.8%, independent variables do not explain dependent variables well. The reason is that the model ignores the effect of time on wing length, which makes the model produce endogenous problems.

It can be found in the actual interpretation of the model, the wing length of female is lower than that of male significantly. Similarly, The length of wings of adult blackbirds is longer than that of juveniles. Moreover, the older the age group, the more obvious the increase of wing length.

IV. Discussion

In this paper, the relationship between wing length and sex and age of blackbirds was quantitatively analyzed by the OLS model. The results showed that there was a positive correlation between age and wing length and males tend to have longer wings than females. The disadvantage of this paper is that it does not consider the change of time, which means that the influence of time factor on the length of blackbird wing is reflected in ϵ. Since the length of blackbirds' wing has periodic time accumulation effect, the residual term must be related to its lag term, which violates the assumption of OLS estimation that the residual term is independent and identically distributed, and the result of model estimation does not conform to the excellent property of blue in Gaus-Markov theorem. In addition, the processing of the missing value is also relatively simple, there is no hierarchical interpolation fitting.

References

[BTO(2020)] BTO, "*Common blackbird guide: species facts, how to identify males, females and juveniles - Discover Wildlife*," 2020. [Online]. Available: https://www.discoverwildlife.com/animal-facts/birds/facts-about-blackbirds

[Feu(n.d.)] C. D. Feu. (n.d.) *Blackbird Data Information*. [Online]. Available: https://mei.org.uk/files/pdf/Blackbird-Data-Information.pdf/

[Mosimann and James(1979)] J. E. Mosimann and F. C. James, "NEW STATISTICAL METHODS FOR ALLOMETRY WITH APPLICATION TO FLORIDA RED-WINGED BLACKBIRDS," vol. 33, no. 1, pp. 444–459, 1979.

[Rossum(1995)] G. V. Rossum, "Python," Centrum voor Wiskunde en Informatica (CWI), Amsterdam, Tech. Rep. CS-R9526, May 1995.

[Harris and Millman(2020)] C. R. Harris and K. J. Millman, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[Seabold and Perktold(2010)] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with python," *Proceedings of the 9th Python in Science Conference*, pp. 57–61, 2010.

Appendix

A. Responding to Peer Review

The use of keywords is a useful way to aid the readers understanding. A glossary in the appendix to explain the key words would go even further to help allow non- mathematically minded people understand the paper. When using acronyms, it is useful to filly define them once before using them for the rest of the paper (What is OLS?).

Mentioning the omission of repeated measurements of a blackbird is useful in making the results reproducable. It is worth discussing how you did this and how it may effect the results. Cleaning the data according to species is a good decision, but a reference to how ring numbers are linked to species would be useful for the reader.

The model and conclusions are explained well and the IMRAD structure is well followed. To aid reader understanding in variable correlation, plots could be used. The model is well shown, but the use of the error term should be explained.

- Thank you for your suggestions on the shortcomings of my article. For the OLS abbreviation, I have revised the introduction section. For the explanation of the error item, I put it in the discussion section. Limited by the length of the article, I put the visualization results and data cleaning code into the appendix.

Thank you for submitting your report; I really enjoyed reading it! Also, I like how you analysed the data. Below, I wrote some comments that I hope will help you improve your report.

I think that in the abstract there is slightly too much emphasis on the method of analysis. You may want to consider shortening the method section (of the abstract) and expanding a little on the background and what you found, and what the results mean. You can put the details concerning the dataset you used in the method section. To make space for this, I would recommend removing the last subsection of Table 1 (the omnibus, skew, and kurtosis part).

In the introduction, you could mention briefly some literature concerning blackbirds' size in relation to their sex and age. It may also be useful, if you have space, to mention what you expect the results to be given the literature (or state hypotheses). Here are links to some references that may be useful: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0030-1299.2006.14183.xhttps://doi.org/10.1080/00063656309476036http://www.jstor.org/stable/2407788

You mentioned that the independent variables do not explain variation in the dependent variable well, but, in the context of the blackbirds dataset (where the measurements were taken in a natural setting with no control over extraneous variables), I would consider the variation in the independent variables to explain the variation in the dependent variable relatively well. Specifically, in this case, I think that if the predictor variables ac- count for 45% of the variance in the outcome variable, the predictors explain a substantial and meaningful proportion of the variance in the outcome. Overall, well-done.

- Thank you for your advice and encouragement. For the results in the table, I will delete it. As well as, thank you for your supplementary references. I checked it and it really helped a lot. About the final question, I cannot point out this in the report due to the space limitation. I suppose that if the missing important explanatory variable has nothing to do with other explanatory variables, that is to say, the explanatory variable is orthogonal to the other explanatory variables, then the variance of unbiased estimator of parameter estimator is also unbiased, but it will cause biased estimation of constant term.

B. Python Code

```
[1]: %matplotlib inline
     import numpy as np
     import pandas as pd
     import matplotlib.pyplot as plt
     import statsmodels.api as sm
     import statsmodels.formula.api as smf
     from statsmodels.formula.api import ols
     from statsmodels.compat import lzip

     plt.rc("figure", figsize=(16,8))
     plt.rc("font", size=14)
```

```
[2]: # Load data
     df = pd.read_csv("blackbird.csv", header = 0)
```

```
[3]: # Fill NaN of Sex
     for i in range(df.shape[0]):
         if df.iloc[i, :]["Sex"] == np.nan:
             df.iloc[i, :] = df[df["Ring number"] == df.iloc[1, :]["Ring␣
      ↪number"]]["Sex"]
         else:
             pass

     # Fill NaN of Wing & Weight
     df = df.fillna(df.mean(numeric_only = True))
```

```
[4]: # Data clean
     df['Day'] = df['Day'].astype("str")
     df['Month'] = df['Month'].astype("str")
     df['Year'] = df['Year'].astype("str")
     df['Time'] = df['Time'].astype("str")
     df["Sex"] = df["Sex"].astype("category")
     df["Age"] = df["Age"].astype("category")
     df['Time_index'] = df['Year'] + "-" + df['Month'] + "-" + df['Day'] + " " +␣
      ↪df["Time"] + ":00:00"
     df['Time_index'] = pd.DatetimeIndex(df["Time_index"])
     dummies_Sex = pd.get_dummies(df["Sex"], prefix="Sex")
     dummies_Age = pd.get_dummies(df["Age"], prefix="Age")
     df = pd.concat([df, dummies_Sex, dummies_Age],axis=1)

     # Ignore the same bird using Ring number
     df_grouped = df.groupby("Ring number").head(1)
```

```
[5]: # OLS
     X = df_grouped.loc[:, ["Sex_F", "Sex_M", "Age_A", "Age_F", "Age_J"]]
```

```python
y = df_grouped["Wing"]
X_model = sm.add_constant(X)
model = sm.OLS(y, X_model)
results = model.fit()
```

[6]:

```python
results.summary()
```

[6]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                   Wing   R-squared:                       0.448
Model:                            OLS   Adj. R-squared:                  0.447
Method:                 Least Squares   F-statistic:                     346.4
Date:                Tue, 17 Nov 2020   Prob (F-statistic):          3.12e-272
Time:                        21:51:09   Log-Likelihood:                -5441.6
No. Observations:                2141   AIC:                         1.090e+04
Df Residuals:                    2135   BIC:                         1.093e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        130.4525      0.639    204.069      0.000     129.199     131.706
Sex_F         -2.0915      0.369     -5.664      0.000      -2.816      -1.367
Sex_M          2.2263      0.357      6.244      0.000       1.527       2.926
Age_A          1.4096      0.542      2.602      0.009       0.347       2.472
Age_F         -2.2311      0.538     -4.146      0.000      -3.286      -1.176
Age_J         -3.9560      0.569     -6.950      0.000      -5.072      -2.840
==============================================================================
Omnibus:                        9.569   Durbin-Watson:                   1.864
Prob(Omnibus):                  0.008   Jarque-Bera (JB):               10.200
Skew:                          -0.120   Prob(JB):                      0.00610
Kurtosis:                       3.238   Cond. No.                         22.7
==============================================================================

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

[7]:

```python
df_grouped.groupby("Sex")["Age"].count()
```

[7]:

```
Sex
F     835
M    1198
U     108
```

```
Name: Age, dtype: int64
```

[8]:

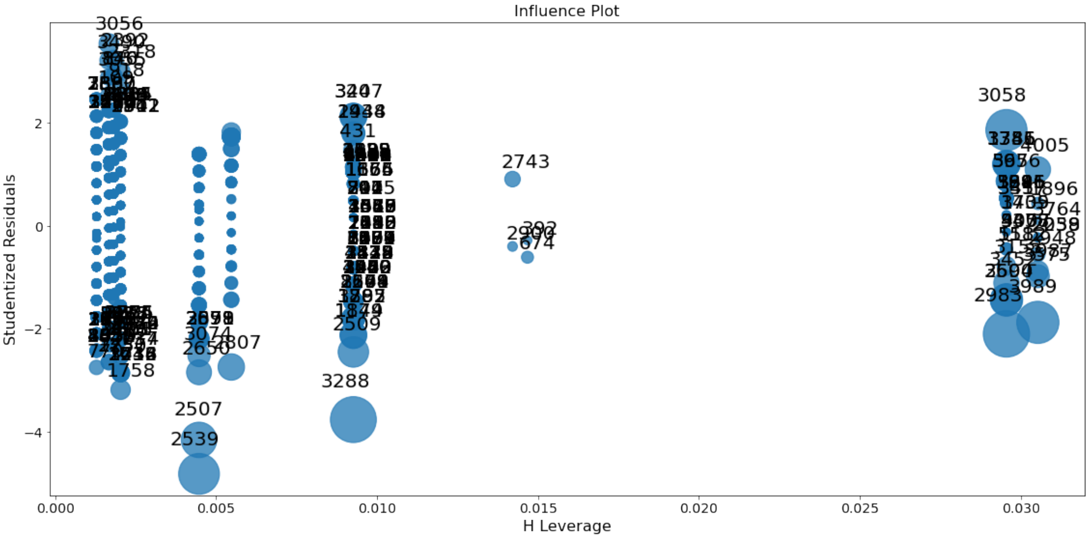
```
df_grouped.groupby("Age")["Sex"].count()
```

[8]:

```
Age
A      713
F     1063
J      331
U       34
Name: Sex, dtype: int64
```

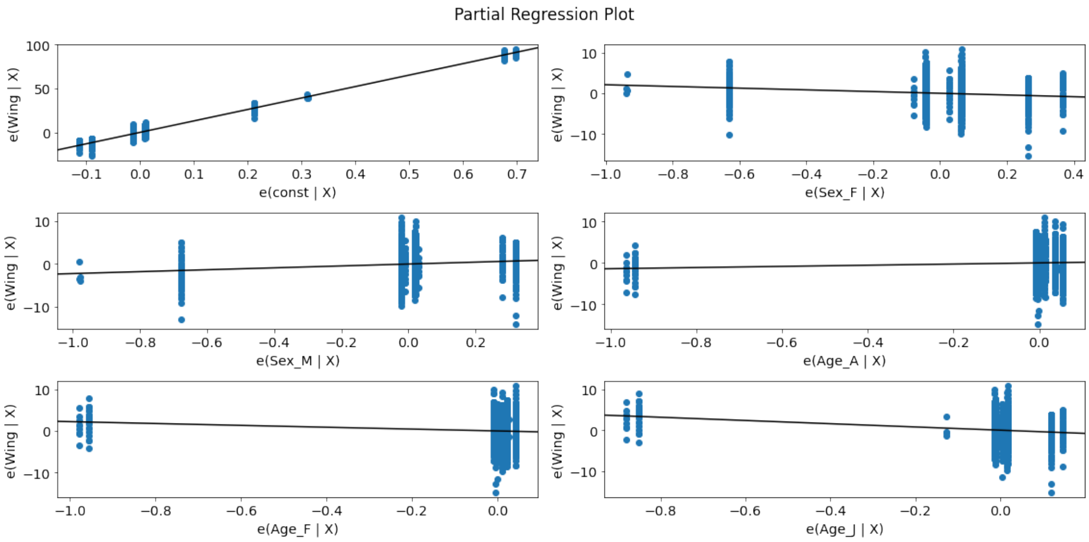
[9]:

```
fig = sm.graphics.influence_plot(results, criterion="cooks")
fig.tight_layout(pad=1.0)
```



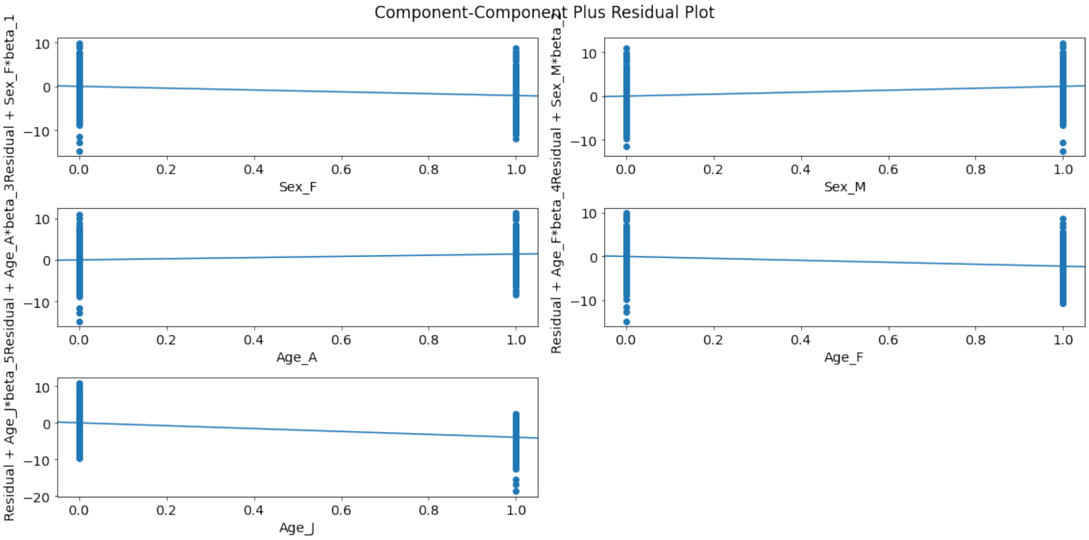
[10]:

```
fig = sm.graphics.plot_partregress_grid(results)
fig.tight_layout(pad=1.0)
```



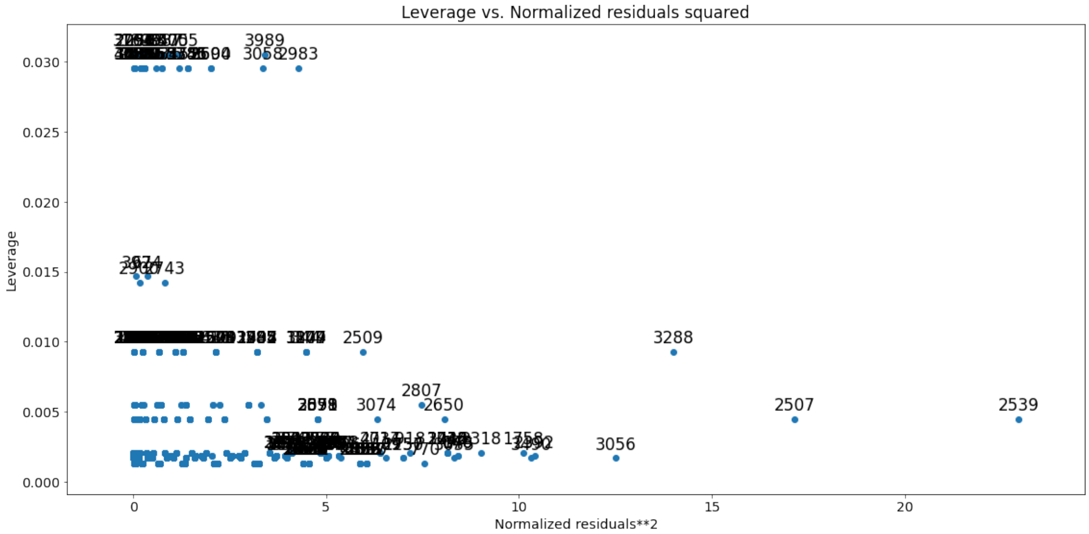
[11]:

```
fig = sm.graphics.plot_ccpr_grid(results)
fig.tight_layout(pad=1.0)
```



[12]:

```
fig = sm.graphics.plot_leverage_resid2(results)
fig.tight_layout(pad=1.0)
```


Leverage vs. Normalized residuals squared
Leverage
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Normalized residuals**2
0
5
10
15
20
3989
3058
2983
2509
3288
2807
3074
2650
2507
2539
3056
1758

[ ]: