Predicting Crab Pre-Molt Size from Post-Molt Size Using Bootstrap Methods

A Punchline report – Durga Prasad N

The Issues:

The dataset used in this lab was collected as part of a study on female Dungeness crabs. The study was conducted by researchers Hankin, Diamond, Mohr, and Ianelli with help from the California Department of Fish and Game and commercial crab fishers from northern California and southern Oregon. The dataset includes two types of data. The first is the pre-molt and post-molt widths of the carapaces of 472 female Dungeness crabs, which were collected by scientists and commercial fisheries over three fishing seasons in 1981, 1982, and 1992.

We address the questions:

- Exploration of the relationship between post-molt size and pre-molt size using a scatterplot and simple linear regression analysis.
- Estimates the precision and reliability of the coefficients and their corresponding predictions.

Findings:

The EDA of the dataset showed some interesting insights in the datasets. There is a considerable difference in sizes of molts noticed when plotted smoothed histograms.

Overall, the results of our analysis showed that there is a strong positive relationship between post-molt and pre-molt sizes of crabs, and that post-molt size is a significant predictor of pre-molt size. Specifically, our model estimates that a one-unit increase in post-molt size corresponds to an increase of 0.80 units in pre-molt size, on average.

Discussion:

The findings of our analysis have important implications for understanding

the relationship between post-molt and pre-molt sizes in crabs. Our results show that there is a strong linear relationship between these two variables, with larger post-molt sizes predicting larger pre-molt sizes. Additionally, we also found that the standard error of the intercept (beta-0) was 2.66 and the standard error of the slope (beta-1) was 0.02, indicating that our model is precise in its predictions.

Appendix A: Methods

Data collection: The crab molt data was collected by measuring the size of crabs in the laboratory both before and after molting.

Variable creation: The dataset contains only two variables Post-molt and Pre-molt which represent the size of the crabs before and after molting.

Analytic method:

The Excel sheet containing the data was imported into a dataframe. Only two variables, Pre-Molt and Post-Molt, were extracted and their descriptive statistics (minimum, maximum, median, mean, standard deviation, skewness, and kurtosis) were calculated. Probability density function plot is created for both factors overlaid on top of each other. A scatter plot was also created to show the relationship between PreMolt and PostMolt size.

Appendix B: Results

The dataset contains 472 records and 2 columns. And the descriptive stats are as follows.

skimpy summary

Data Summary

dataframe	Values
Number of rows	472
Number of columns	2

Data Types

Column Type	Count
float64	2

number

column_name	NA	NA %	mean	sd	p0	p25	p75	p100	hist
postsize	0	0	140	15	39	140	150	170	
presize	0	0	130	16	31	120	140	160	

End

The kurtosis and skewness are calculated using standard dataframe methods and the values are as follows.

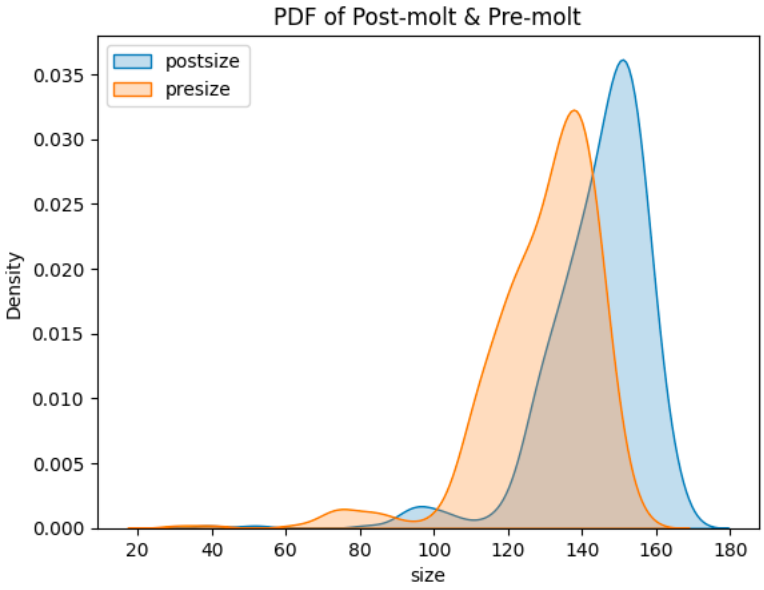
Skewness → postsize -2.354391

presize -2.009880

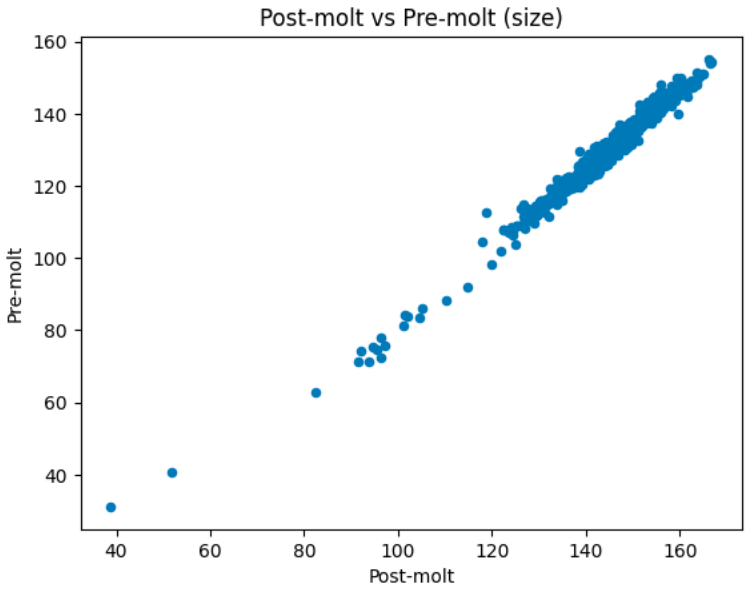
Kurtosis → postsize 10.236847

presize 6.851370

The smooth histogram approximations to the size of disparity shows a visible difference between Pre & Post molt sizes.



The scatter plot of post size vs pre size is shows a linear relationship between the independent and dependent variables.



Bootstrap method (sampling with replacement) is performed to compute the standard errors of the coefficients and the values were relatively small, indicating that our model is a good fit for the data.

Standard Errors:

beta-0: 2.66

beta-1: 0.02

Appendix C: Code

The statistical analysis is performed using following code and a linear model is trained using sklearn package of python.

a. Loading the dataset into pandas dataframe

```
df = pd.read_excel("./data/crab_molt.xls")
```

b. Dataframe describe method is used to generate descriptive stats of all the numerical columns.

skimpy summary

Data Summary

dataframe	Values
Number of rows	472
Number of columns	2

Data Types

Column Type	Count
float64	2

number

column_name	NA	NA %	mean	sd	p0	p25	p75	p100	hist
postsize	0	0	140	15	39	140	150	170	
presize	0	0	130	16	31	120	140	160	

End

c. To computed skewness & kurtosis pandas offers skew, kurtosis methods.

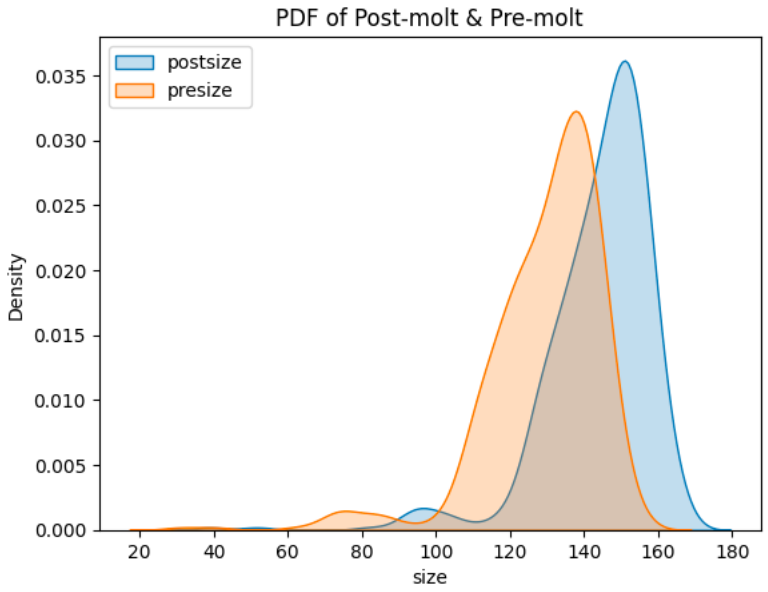
```
df.skew()
```

```
postsize   -2.354391
presize    -2.009880
dtype: float64
```

```
df.kurtosis()
```

```
postsize    10.236847
presize      6.851370
dtype: float64
```

d. Here is the code to compute overlayed smooth histogram (PDF)'s.



e. A scatter plot of Pre-molt vs Post-molt is generated using below code.

```
ax = df.plot.scatter(x='postsize', y='presize')
ax.set_xlabel("Post-molt")
ax.set_ylabel("Pre-molt")
ax.set_title("Post-molt vs Pre-molt (size)")
```

f. Finally, the standard error for the coefficients is calculated using the code shown below.

```
model = LinearRegression()

samples = []
for _ in range(1000):
    sample = df.sample(df.shape[0], replace=True)
    clone_model = clone(model)
    clone_model.fit(sample[['postsize']], sample['presize'])
    samples.append({"b0": clone_model.intercept_, "b1": clone_model.coef_[0]})
```

```
print(f"Standard Errors:\n\tbeta-0: {np.std([i['b0'] for i in samples]):.2f}\n\tbeta-1: {np.std([i['b1'] for i in samples]):.2f}")
```