Analysis of Pre-Molt and Post-Molt Sizes in Crabs: Modeling Growth and Evaluating Linear Regression Assumptions

Issues

The data describes the pre-molt sizes and the post-molt sizes in Dungeness crabs. It has been taken from the California Department of Fish and Game and commercial crab fishers from northern California and southern Oregon. Main issue is to find out if the pre molt value is different from the post molt value depending on situations such as weather, water etc.
Here we answer about the following questions,

1. To find out if there is any overlapping of the histogram data between the two variables.

2. To find out the simple linear regression with post molt size as the predictor variable and the pre molt size as the predicted variable.

3. To find Pearson's r^2 regression value.

4. To find out descriptive statistics of the residuals.

5. To test the distribution of residuals for normality using a quantile plot and the Shapiro-Wilks test.

6. To visually check for heteroscedasticity

Findings

The data set consists of crab-molt data in pairs: (post-molt size, pre-molt size) as an Excel (xls) file.

From the data obtained, we can find out that the predicted variable depends on the predictor variable. We can also find out that the average size of the variables increases by almost 15 mm. For the Post-molt size, most of the data resides in that 145-160 mm range. And for the pre-molt size, that data resides around 120-135 mm range.

Another finding is that all the data points are very near to the regression line with a skewness of -2.46, -2.044 and a kurtosis value of 12.55, 8.21 which is extreme. The residual data summary is very different from the model summary.

Discussions

From the data, we can conclude that there is a difference of an increment of around 15 mm. The Regression analysis can be often used to address questions involving biological populations that exhibit natural variability. We can find out that the pre-molt data is dependent on the post-molt variable. So we have two values for the skewness and kurtosis which means each of the value is so pre-molt and post-molt respectively. The size measurements were made on the external carapace along the widest part of the shell, excluding spines. All measurements are in millimetres. Similarly, a crab's premolt size varies about theregression line. The residual is a name for the difference between a crab's actual premolt size and the regression line prediction of it.

Appendix A: Method

The data is an excel sheet and imported into python with the help of importing libraries. And the data is checked for any nan values. A summary is printed to find out the median, mean and standard deviation.

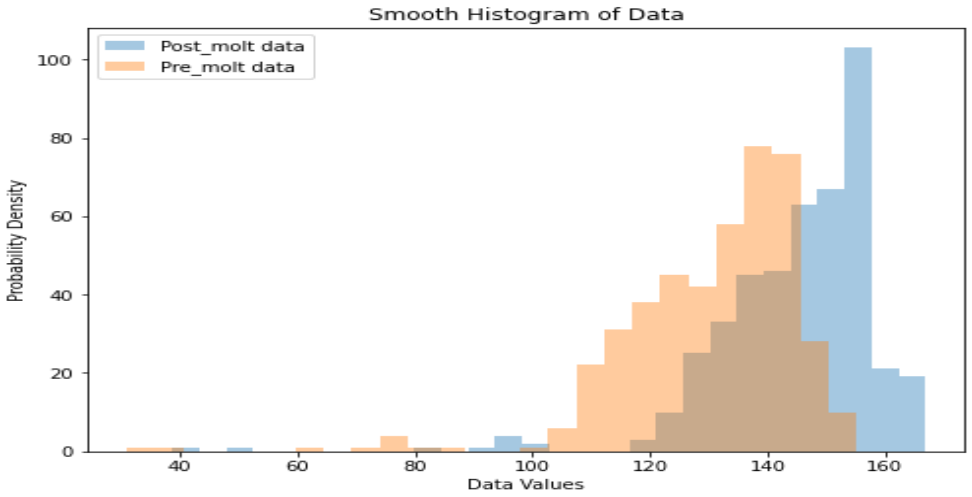
The data is then split using the 'iloc' function to create a separate variable of each. After splitting, we then create a histogram of each variable. Then we create a smooth histogram with seaborn library. This histogram shows a figure of the overlooked histograms of both the variables.

We also create a visual presentation of the scatter plot of the predictor variable and the predicted variable. Most of the data resides on the top right of the figure. We then find the regression line and find out the pearson's r squared value.

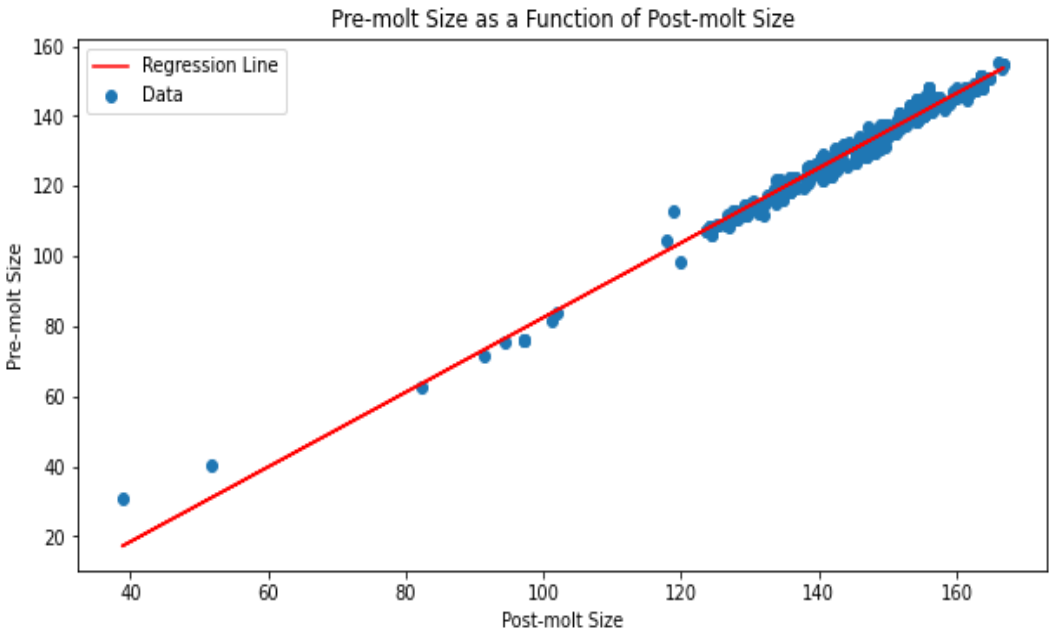
The descriptive statistics of the residuals is done and then we plot a scatter plot to test the normality of it. It is between the Residuals and the Quantities of standard normal. Shapiro wilks test is also performed. Then we plot the residual against the dependent variable and heteroscedasticity is found out.

Appendix B: Results

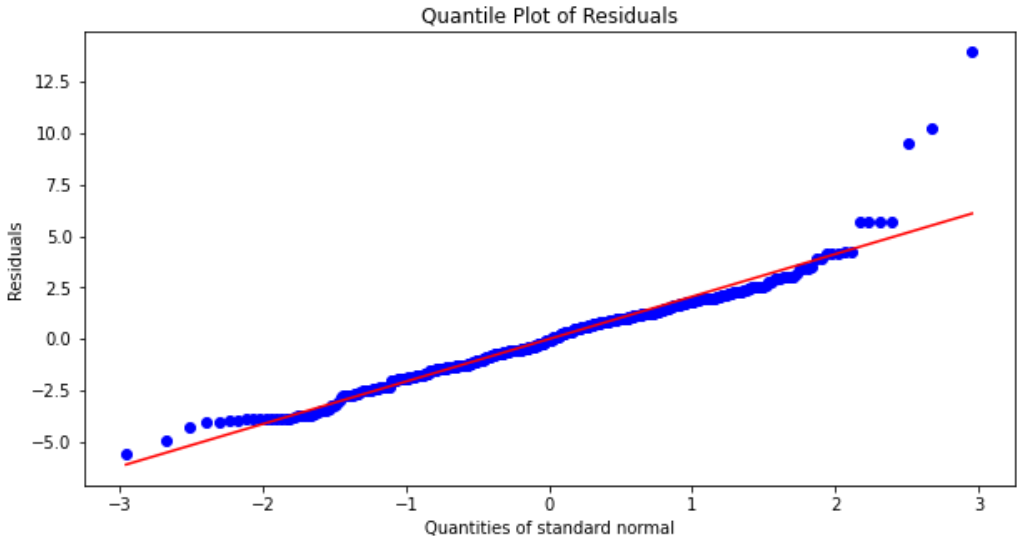
From the data set, there are around 445 values of pre molt and post molt. We create a smooth histogram overlooking both the figures of those variables. With the help of this, we can easily compare the data values and probability densities.



A simple linear regression model is created. We can find out that all the data points are in the top right of the figure and very few other data points are scattered here and there. The linear regression line goes through the data points which means the data is normally distributed. It means that the line has been fitted to the data points in a way that minimises the overall distance between the line and the point

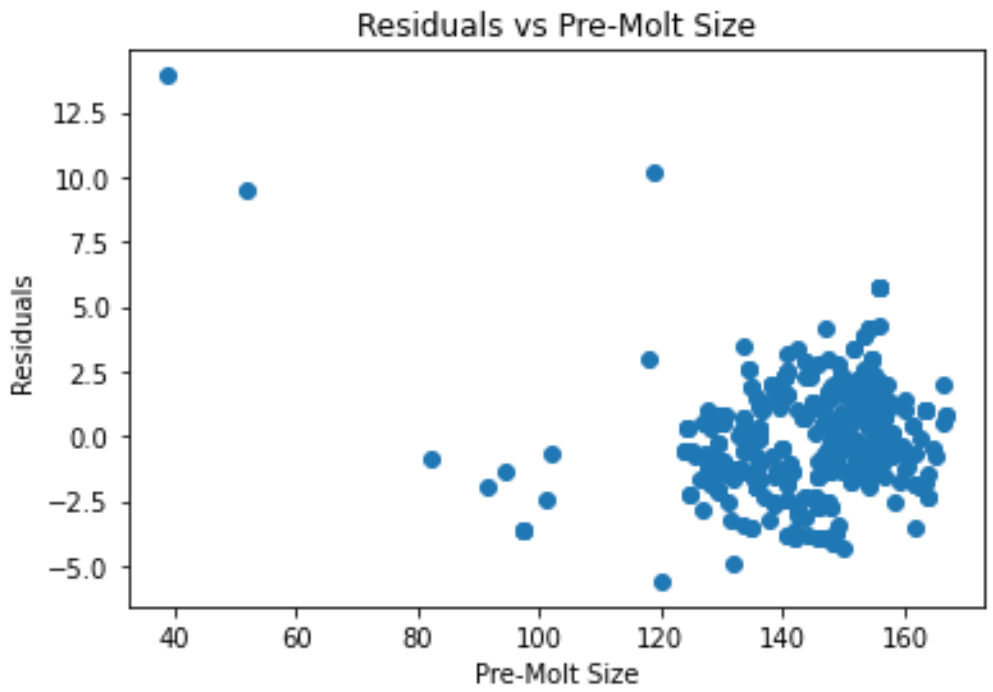


When we look at the summary of the model, we can find out the pearson's r squared value is 0.9803 and a P value of 0. We used the method of OLS which is ordinary least squares. The condition number is large, 1.51e+03. This might indicate that there are strong multicollinearity or other numerical problems.



Next, we create a quantile plot for the residuals. During this process, we can also find out the statistics of these residuals. They have a mean of -1.46 and median of -0.04. We also have the maximum and minimum values which are 13.92 and -5.588 respectively.

The residuals are not normally distributed and its p value is 3.676.



We plot a residual against the dependent variable, pre-molt size. A scatter plot is helpful in presenting the data. As we can see, the majority of the data points are collected in a single area. This means If the plot of residuals against the pre-molt size variable shows that the residuals are all collected in a single space, it suggests that there may be little or no heteroscedasticity in the model. This means that the variance of the residuals is relatively constant across the range of the pre-molt size variable, which is a key assumption of linear regression.
A lack of heteroscedasticity may be good, as it suggests that the model is likely to accurately predict new observations.

Appendix C: Code

```
import pandas as pd
import numpy as np
from scipy.stats import skew, kurtosis
import matplotlib.pyplot as plt
import statsmodels.api as sm
import scipy.stats as stats
import seaborn as sns
```

```
df = pd.read_excel("crab_molt_data_porandla_rithik.xls")
df

df.describe()

#splitting the dataset into two separate, independent variables

df1 = df.iloc[:, :int(df.shape[1]/2)]
df2 = df.iloc[:, int(df.shape[1]/2):]

print(df1, df2)

#to find the skewness

data_skewness = df

skewness = skew(data_skewness)

print("Skewness: ", skewness)

#to find the kurtosis

data_kurt = kurtosis(df)

print("Kurtosis: ", data_kurt)

#histogram for post_molt data

sns.histplot(df1, color='blue', label='Post-Molt data', alpha=0.5)
plt.title('Probability Density Function of post molt data')
plt.xlabel('Data Values')
plt.ylabel('Probability Density')

plt.plot()

#histogram for pre_molt data

sns.histplot(df2, label='Pre-Molt data', alpha=0.5)
plt.title('Probability Density Function of pre molt data')
plt.xlabel('Data Values')
plt.ylabel('Probability Density')

plt.plot()

#creating smooth histograms of each variable, overlooked


fig, ax = plt.subplots(figsize=(8,6))
```

```
sns.distplot(df1, hist=True, kde=False, kde_kws={'linewidth': 2},
label='Post_molt data', ax=ax)
sns.distplot(df2, hist=True, kde=False, kde_kws={'linewidth': 2},
label='Pre_molt data', ax=ax)

plt.xlabel('Data Values')
plt.ylabel('Probability Density')
plt.title('Smooth Histogram of Data')

ax.legend()

plt.show()

#Plotting the “Pre-molt” size as a function of “Post-molt” size

post_molt_size = df1
pre_molt_size = df2

fig, ax = plt.subplots(figsize=(10, 6))

ax.scatter(post_molt_size, pre_molt_size)

ax.set_title('Pre-molt Size as a Function of Post-molt Size')
ax.set_xlabel('Post-molt Size')
ax.set_ylabel('Pre-molt Size')

plt.show()

#Simple Linear Regression

post_molt_size = df1
pre_molt_size = df2

post_molt_size = np.array(post_molt_size)
pre_molt_size = np.array(pre_molt_size)

post_molt_size = sm.add_constant(post_molt_size)
model = sm.OLS(pre_molt_size, post_molt_size).fit()

prediction = model.predict(post_molt_size)

fig, ax = plt.subplots(figsize=(10, 5))

ax.scatter(post_molt_size[:,1], pre_molt_size, label='Data')
ax.plot(post_molt_size[:,1], prediction, 'r', label='Regression Line')
ax.set_title('Pre-molt Size as a Function of Post-molt Size')
ax.set_xlabel('Post-molt Size')
ax.set_ylabel('Pre-molt Size')
ax.legend(loc='best')

plt.show()
```

```
r_squared = model.rsquared
print("Pearson's r^2 for the regression:", r_squared)

model.summary()



#Descriptive statistics of the residuals and quantile plot to test the
normality

post_molt_size = df1
pre_molt_size = df2

post_molt_size = np.array(post_molt_size)
pre_molt_size = np.array(pre_molt_size)

post_molt_size = sm.add_constant(post_molt_size)
model = sm.OLS(pre_molt_size, post_molt_size).fit()

residuals = model.resid

print("Descriptive statistics of the residuals:")
print("\n")
print("Mean:", np.mean(residuals))
print("Standard deviation:", np.std(residuals))
print("Minimum value:", np.min(residuals))
print("Maximum value:", np.max(residuals))
print("25th percentile:", np.percentile(residuals, 25))
print("50th percentile (median):", np.percentile(residuals, 50))
print("75th percentile:", np.percentile(residuals, 75))

fig, ax = plt.subplots(figsize=(10, 5))

stats.probplot(residuals, plot=ax, fit=True)
ax.set_title("Quantile Plot of Residuals")
ax.set_xlabel("Quantities of standard normal")
ax.set_ylabel("Residuals")

plt.show()

#Shapiro wilks test

post_molt_size = df1
pre_molt_size = df2

post_molt_size = np.array(post_molt_size)
pre_molt_size = np.array(pre_molt_size)

post_molt_size = sm.add_constant(post_molt_size)
model = sm.OLS(pre_molt_size, post_molt_size).fit()
```

```
residuals = model.resid

fig, ax = plt.subplots(figsize=(10, 5))

stats.probplot(residuals, plot=ax, fit=True)
ax.set_title("Quantile Plot of Residuals")
ax.set_xlabel("Quantities of standard normal")
ax.set_ylabel("Residuals")

plt.show()

_, p_value = stats.shapiro(residuals)

if p_value < 0.05:
    print("The residuals are not normally distributed, p-value =",
p_value)
else:
    print("The residuals are normally distributed, p-value =", p_value)

#Plotting the residual against the dependant variable and visual
checking for heteroscedasticity

pre_molt_size = df1

residuals = model.resid

plt.scatter(pre_molt_size, residuals)

plt.xlabel('Pre-Molt Size')
plt.ylabel('Residuals')
plt.title('Residuals vs Pre-Molt Size')

plt.show()

"""## Based on the scatter plot of residuals against pre-molt size
variable that we have plotted, it appears that there maybe no
heteroscedasticity in the residuals. This means that the variance of
the residuals is relatively constant across the range of the pre-molt
size variable, which is a key assumption of linear regression.

## Also, if the residuals are all collected in a single space, it may
be difficult to visually check for heteroscedasticity in the first
place, as there may be little variation in the distribution of the
residuals.
"""
```