Unlocking the Crab's Mystery: Predicting Pre-Molting Size

Authors: Veda Sahaja Bandi, Sindhuja Baikadi

The Issues:

Driven by the captivating mystery of Crab Molting, a dataset from Washington State sheds light on this fascinating transformation, capturing their sizes before and after molting in both natural habitats and controlled laboratory environments. By dissecting the pre- and post-molt size variations documented within, we aim to unlock the secrets of crab growth and unveil:

- How do crab sizes vary before and after molting, and what factors influence this variation?
- Do the sizes of crabs before and after molting follow a normal distribution, and what causes deviations from this pattern?
- How does the regression model explain the relationship between pre-molt and post-molt sizes, and what does this reveal about crab growth?
- What do the residual patterns in the predictive model suggest about its accuracy and potential nonlinear growth patterns?
- Does the location where crabs molt (field vs. lab) affect their size changes?

Exploring these inquiries sheds light on the nuanced effects of molting on crabs at individual and population levels, while also unraveling the ecological ramifications of their size changes in the wider ecosystem. A thorough investigation of these aspects offers a deeper understanding of crab development post-molt, as well as the broader ecological dynamics at play. Insight into such issues is essential for fostering comprehensive knowledge of crustacean life cycles and their role in marine biodiversity.

Findings:

Upon examining the crab molting data from Washington State, intriguing insights emerge regarding the shell sizes of crabs before and after molting. A notable finding reveals a significant increase in crab size post-molt, with an average difference of 15 mm observed. This underscores the crucial role of the molting process in crab development and adaptation within their habitats.

The results of normality tests indicate deviations from a normal distribution in both pre-molt and post-molt size distributions. This suggests the presence of complex underlying factors influencing crab sizes, warranting further exploration and understanding. Moreover, the linear regression models exhibit a strong correlation between pre-molt and post-molt sizes. With post-molt data accurately predicting pre-molt data at a rate of 98%, these models provide valuable insights into crab growth dynamics.

Furthermore, the analysis reveals significant variations in residual patterns, indicating variance in the data. Differences in regression model outcomes between lab and field data suggest the nuanced effects of environmental conditions on the molting process. Lab crabs tend to adhere more closely to a linear growth pattern compared to field crabs, reflecting environmental influences.

Insights gleaned from this study can inform improved management strategies for crab populations, particularly in aquaculture, thereby contributing to the preservation of ecological balance and biodiversity. In essence, this analysis sheds light on the intricate interplay between molting, crab growth, and environmental factors, offering valuable knowledge for understanding crab development and its ecological implications.

Discussion:

The analysis of crab molting patterns in Washington State has revealed significant insights into the molting process and its impact on crab populations. The observed size differences before and after molting highlight the biological significance of this phenomenon. However, deviations from normality in size distributions suggest underlying factors contributing to growth variability, such as genetics, age, or environmental stresses.

The linear model's ability to predict pre-molt crab sizes using post-molt data underscores its importance in understanding crab growth dynamics. Differences in regression outcomes between lab and field data indicate potential environmental influences on the molting process. This prompts exploration into how factors like water temperature and food availability may affect molting, essential for the crabs' physiological development.

The study emphasizes the need for careful management of natural and aquaculture environments to support successful molting. Insights into environmental factors influencing molting can inform conservation strategies to protect crab populations against climate change and human activities. Overall, the research advances our understanding of crab molting, providing valuable knowledge for conservation efforts and the aquaculture industry.

Appendix A: Method

Data was downloaded as a comma-separated (.csv) file and imported into Jupyter Notebook. Header rows and all data entries with null values were removed. The analysis was done using the pre-molt and post-molt data along with their Location. We have explored the basic statistical properties of the dataset, including maximum, minimum, median, mean, and standard deviation for both pre-molt and post-molt sizes. Skewness and kurtosis were analyzed to understand the symmetry and peaks in the distributions.

To assess normality assumptions, quantile plots were generated for both pre-molt and post-molt sizes, supplemented by quantitative tests including Anderson-Darling, Kolmogorov-Smirnov, Cramér-von Mises, and Shapiro–Wilk tests. Visual Comparative Analysis involved smooth histograms of post-molt and pre-molt variables to visually discern changes, along with a scatterplot illustrating the relationship between pre-molt and post-molt sizes.

We built a Linear Least Squares model to predict pre-molt sizes from post-molt sizes, incorporating Pearson’s R-squared value for model performance evaluation. Residuals of the regression model were scrutinized for normality and heteroskedasticity, with tests including visual inspection, Breusch–Pagan test, and White test. Model Validation and Comparison utilized k-fold (k=5) cross-validation to estimate prediction accuracy and explored potential differences in linear models for lab versus field crabs.

This meticulous methodological approach aimed to offer a comprehensive analysis of crab size changes due to molting, unraveling statistical significance, patterns, and predictive modeling of size changes.

Appendix B: Results

The CSV File contains 472 data points containing both pre and post-molting sizes of crabs along with their Location of Molting. There are no null values in the dataset.

Descriptive Statistics

The analysis began by exploring the **size characteristics** of crabs before and after their molting process, revealing significant growth.

Pre-Molt Crab Sizes:

- The smallest crab measured 31.1mm, and the largest was 155.1mm pre-molt.
- The average size before molting was 129.21mm, with a median slightly higher at 132.80mm, indicating a relatively balanced but slightly right-skewed distribution.
- The standard deviation of 15.86mm suggests variability in sizes, while the skewness (-2.0035) and kurtosis (6.7663) values indicate a distribution that leans towards larger crabs with a heavier tail than a normal distribution.

Post-Molt Crab Sizes:

- Post-molt, crabs exhibited sizes ranging from 38.8mm to 166.8mm.
- The mean size increased to 143.90mm, with the median size at 147.40mm, demonstrating the molting growth spurt.
- The distribution's skewness (-2.3469) and kurtosis (10.1160) heightened post-molt, reflecting even more pronounced non-normality and variability in growth outcomes.

Statistical Analysis for Normality

Quantile plots for the distributions of sizes pre and post-molting suggest that molting affects the size distribution of crabs indicating relatively serious departures from normality (Fig 1):

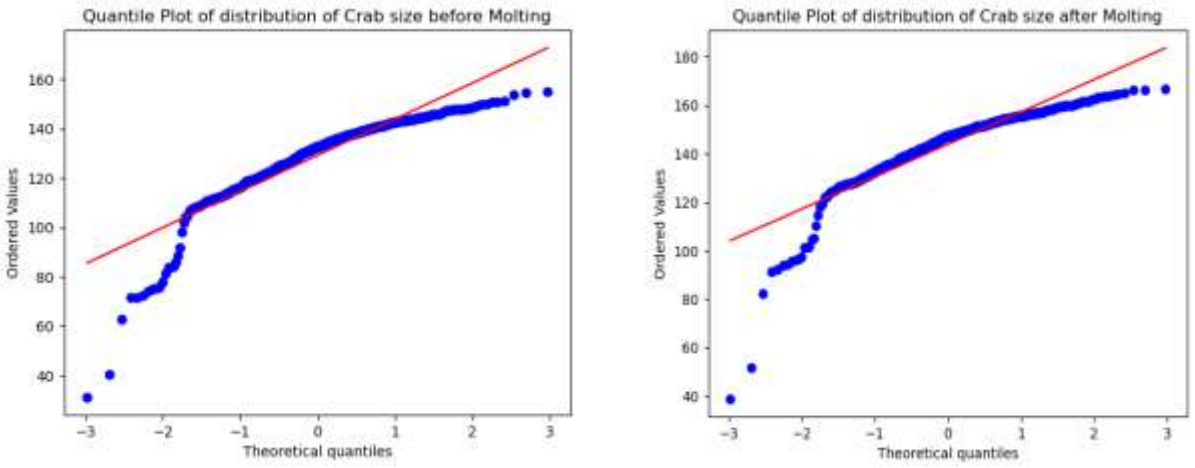

Figure 1: Quantile plots of Crab size distributions before and after molting

Further normality tests were conducted highlighting the non-normal nature of size distributions, both pre-and post-molt, with p-values from the Anderson-Darling (5.84×10^{-31}, 0.0), Kolmogorov-Smirnov (0.0, 0.0), Cramér–von Mises (0.0, 0.0), and Shapiro–Wilk (8.99×10^{-21}, 4.86×10^{-22}) tests less than 0.05, robustly rejecting the normal distribution hypothesis.

Visual Comparisons and Correlation Analysis

A comparison of smooth histogram approximations to the size distribution of crabs before and after molting shows a significant shift to the left and elevation after molting indicating that the size of crabs increased after the molting process (Fig 2).

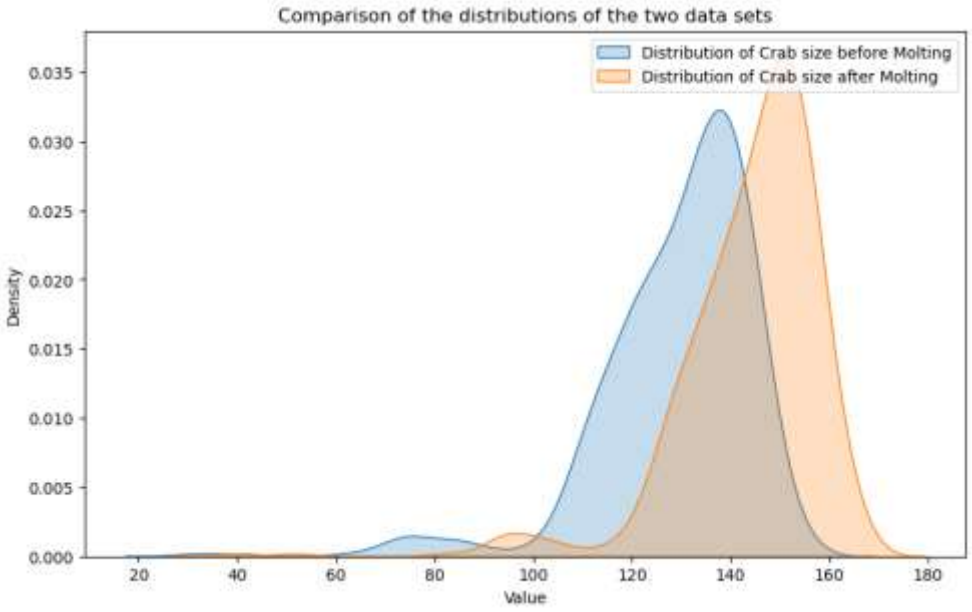

Figure 2: Crab Size distributions before and after molting

The scatterplot of pre-molt versus post-molt sizes confirmed a strong positive relationship, showcasing the universal nature of growth during molting across the sampled population.

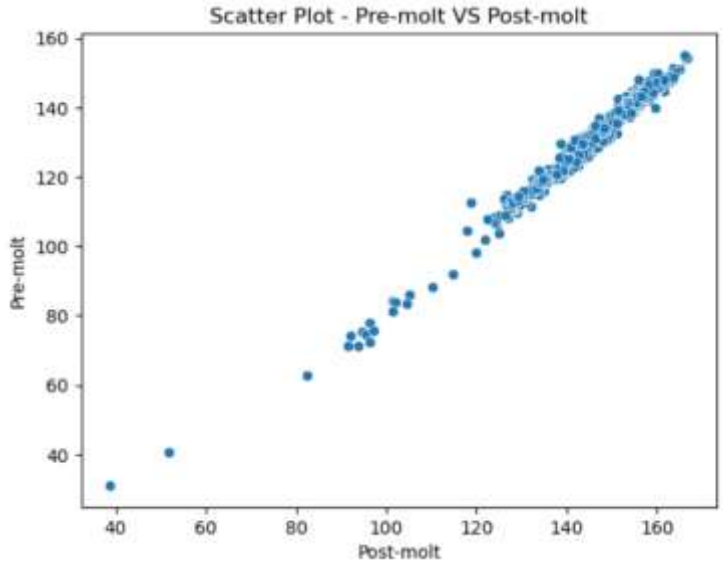

Figure 3: Scatter Plot between Pre-Molt and Post-Molt Data

The scatter plot (Fig 3) illustrates a strong positive linear relationship between the pre-molt and post-molt sizes of crabs. The data points are tightly clustered along a line that ascends from the bottom left to the top right of the plot, which suggests that as the pre-molt size increases, the post-molt size increases correspondingly. There are a few outliers, particularly at the lower end of the scale, but the overall trend indicates that the molting process consistently increases in size. This consistency across the data points indicates that the size before molting is a reliable predictor of the size after molting.

Linear Regression

A linear regression model predicting pre-molt sizes from post-molt sizes yielded a high Pearson's R-squared value of 0.9808, signifying an excellent fit. However, the analysis of residuals revealed skewness and kurtosis, indicating deviations from ideal model assumptions.

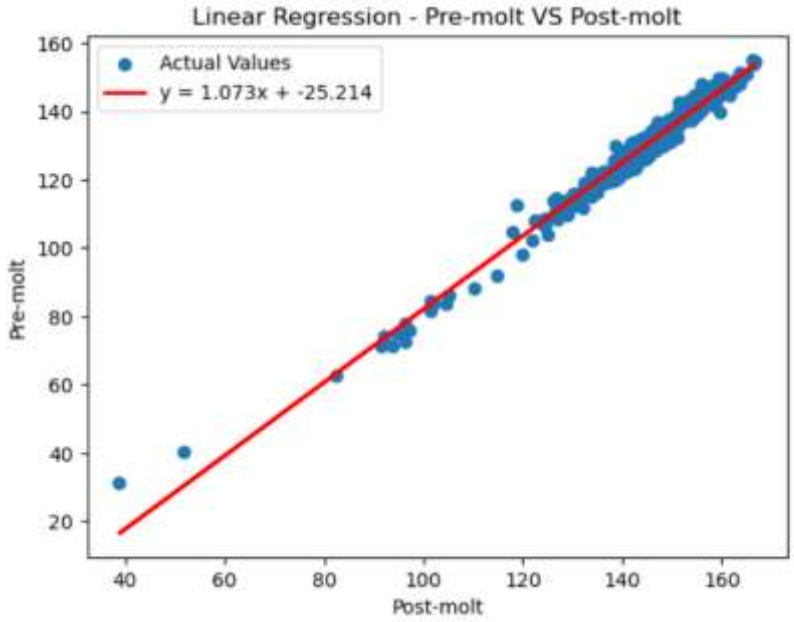

Figure 4: Linear Regression – Pre-molt Vs Post-molt

The scatter plot with the linear regression line (Fig 4) illustrates the relationship between crab sizes pre-molt and post-molt. The linear regression equation, $y = 1.073x - 25.214$, indicates that for every unit increase in post-molt size, the pre-molt size is expected to increase by approximately 1.073 units. The negative intercept suggests that at the point where the regression line intercepts the y-axis (when the post-molt size is zero), the pre-molt size would be -25.214. However, this negative value likely has no practical biological meaning since it falls outside the range of observed data.

The strong linear trend and the clustering of data points around the regression line suggest a strong positive correlation between pre-molt and post-molt sizes. The results support the conclusion that molting significantly influences an increase in crab size and that the relationship between sizes before and after molting can be effectively modeled with linear regression.

Analysis of Residuals

The KDE plot of the residuals (Fig 5) indicates that the regression model's errors are approximately normally distributed with a slight right skew, as the peak is near zero and the tail extends more to the right. This suggests that the model predictions are generally accurate, with a few larger-than-expected values.

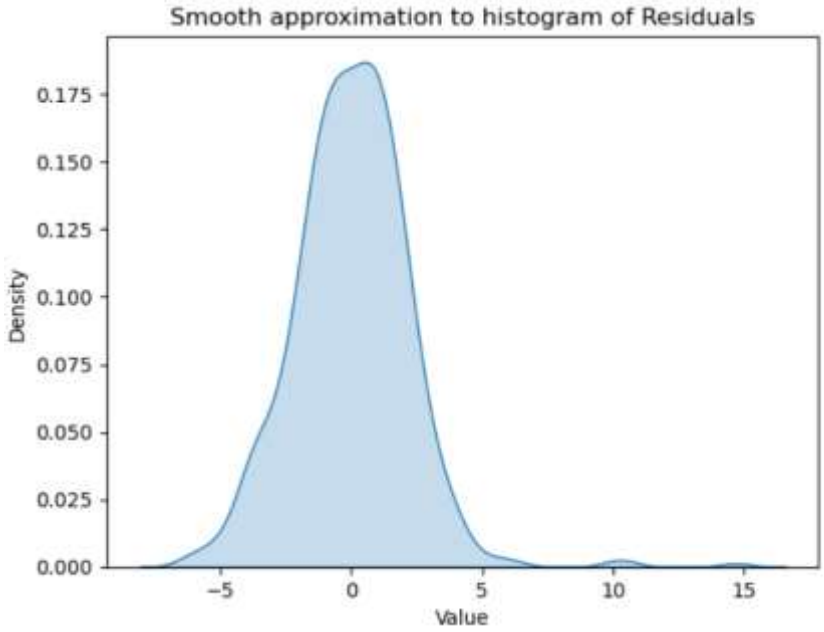

Figure 5: Plot of Residuals

The reported skewness and kurtosis of the residuals are 0.8455 and 5.3787, respectively. This indicates that the residuals from the regression model show a moderate right skew and are leptokurtic, meaning they have a sharper peak and fatter tails compared to a normal distribution.

The Quantile-Quantile (Q-Q) plot (Fig 6) of the regression model's residuals mostly follows the expected line for a normal distribution, suggesting that the residuals are normally distributed. However, there is a noticeable deviation from the line in the upper quantiles, which indicates potential outliers or skew in the distribution of residuals, typically representing heavier tails than a normal distribution would exhibit.

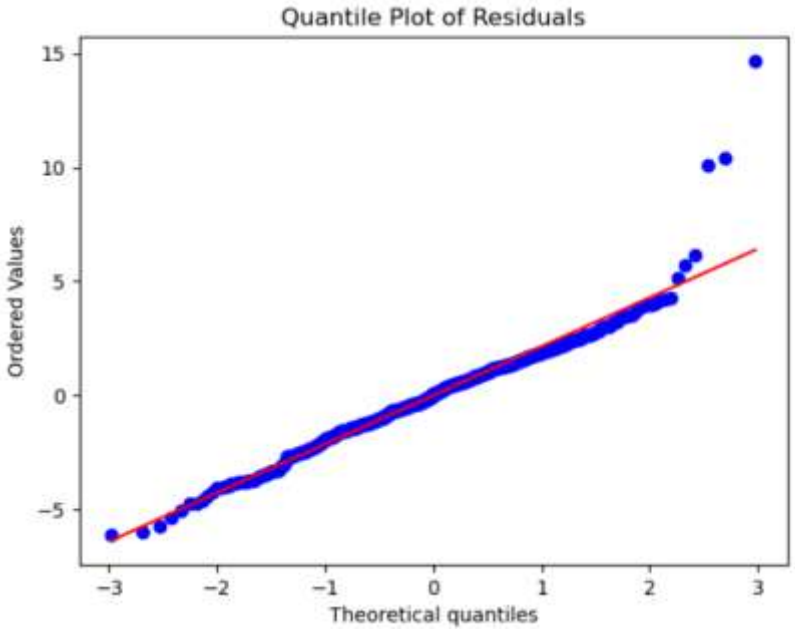

Figure 6: Quantile Plot of Residuals

The results from various tests for normality on the residuals of the regression model are as follows:

- Anderson-Darling Test: The p-value is 0.000115, suggesting that there is evidence against the residuals being normally distributed.
- Shapiro-Wilk Test: The p-value is approximately 6.36×10^{-12}, which is significantly low, again indicating non-normality in the residuals.
- Kolmogorov-Smirnov Test: The p-value is 2.56×10^{-12}, reinforcing the conclusion that the residuals do not follow a normal distribution.
- Cramér–von Mises Test: The p-value, identical to the Kolmogorov-Smirnov test at, 2.56×10^{-12} further confirms the residuals' departure from normality.

All the p-values are less than 0.05, leading to the rejection of the null hypothesis that the residuals are normally distributed. This suggests that the regression model may not meet the normality assumption for the residuals, which could have implications for inference and prediction intervals derived from the model.

Heteroskedasticity:

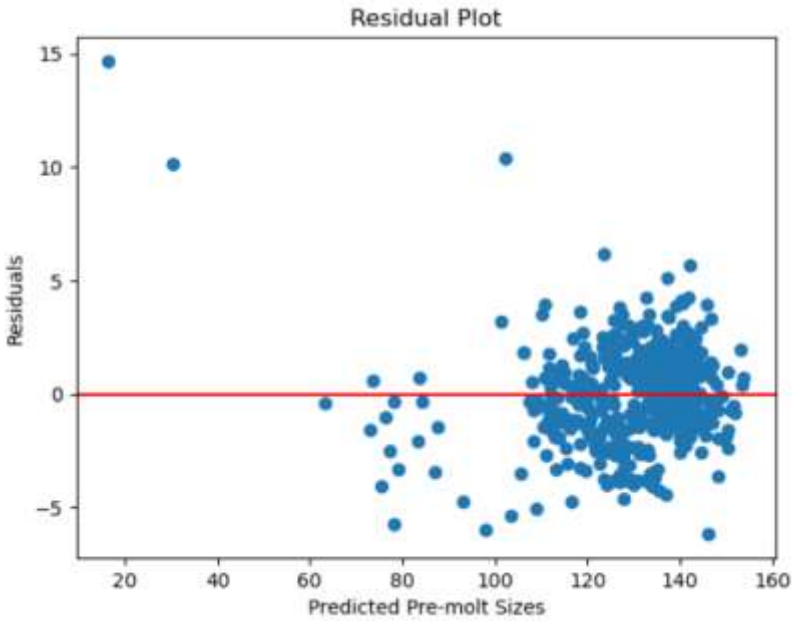

Figure 7: Quantile Plot of Residuals

The Visual estimation of the residuals (Fig 7) illustrates the differences between observed and predicted pre-molt sizes from the regression model. The residuals are fairly scattered around the zero line, suggesting no systematic bias in predictions. However, there are some outliers, particularly for larger predicted sizes, which may indicate potential issues with heteroscedasticity or model fit for these values. The plot shows that there is a variance in the residuals.

The results from the Breusch–Pagan and White tests for heteroskedasticity in the regression model residuals are as follows:

- The Breusch–Pagan test yields a very low p-value 3.25×10^{-21}, indicating significant heteroskedasticity.
- Similarly, the White test produces an even lower p-value 1.86×10^{-52}, reinforcing the presence of heteroskedasticity.

Heteroskedasticity tests, including the Breusch–Pagan and White tests, pointed to varying error variances, prompting further scrutiny of model assumptions.

Analysis of potentially different Linear models for Lab versus Field crabs

The comparative analysis between lab and field data hinted at environmental influences on molting outcomes.

The Linear Regression plot (Fig 8) between pre-molt and post-molt sizes of crabs in the field, with linear equation $y = 1.042x - 20.402$, shows a strong positive correlation between post-molt and pre-molt sizes. The Pearson's R-squared value of 0.9328 indicates that approximately 93.28% of the variability in pre-molt sizes can be explained by the post-molt sizes, suggesting a good fit of the model to the data.

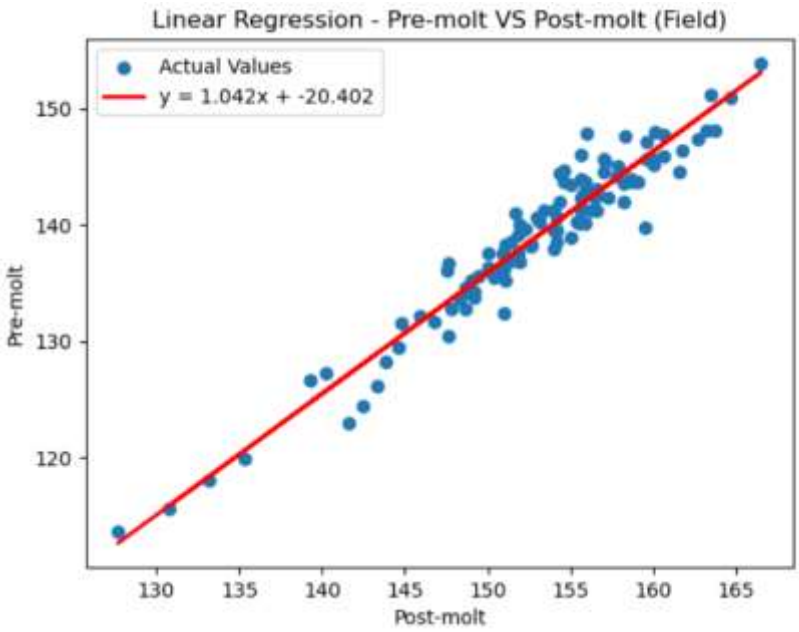

Figure 8: Linear Regression – Field Crabs

The linear Regression plot (Fig 9) for Lab crab data, $y = 1.074x - 25.344$ shows a strong positive relationship between pre-molt and post-molt sizes, with an R-squared value of 0.981, indicating a very good fit compared to field Crabs which has R-Squared value of 0.9328.

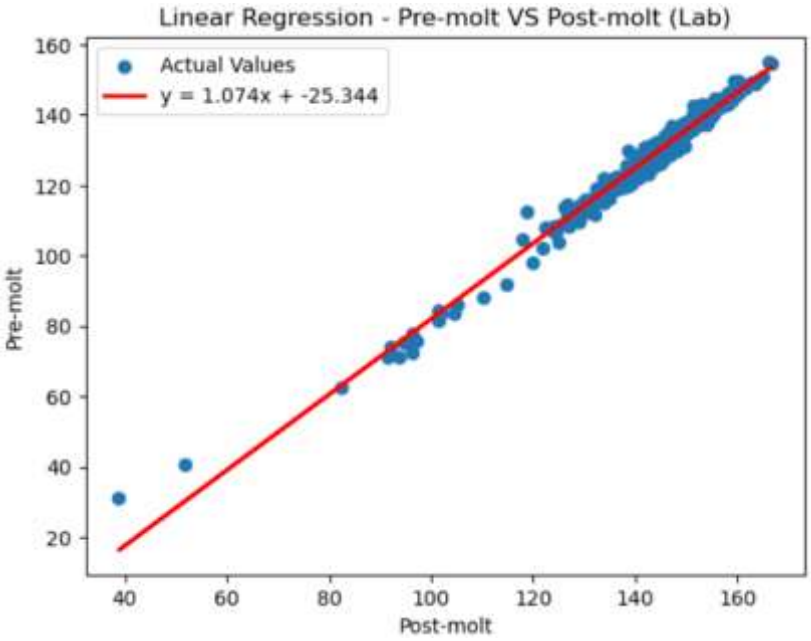

Figure 9: Linear Regression – Lab Crabs

Utilizing K-Fold Cross-Validation to estimate the prediction accuracy of linear models

These results from the k-fold cross-validation where k = 5 indicate that the linear regression model performs consistently well across different subsets of the data, with high predictive accuracy. The mean R-squared score near 0.957 demonstrates the model's reliability in explaining the variance in the response variable.

Appendix C: Code

In this appendix, we document the Python code for building a linear least squares model to predict the sizes of crab shells before molting (pre-molt size) from the sizes of crab shells after molting (post-molt size).

```
# Descriptive Statistics for Pre-Molt Data

print("Minimum size of Crab before Molting:", min(pre_molt))

print("Maximum size of Crab before Molting:", max(pre_molt))

print("Mean size of Crab before Molting: {:.2f}".format(pre_molt.mean()))

print("Median size of Crab before Molting: {:.2f}".format(pre_molt.median()))

print("Standard Deviation of Crab size before Molting: {:.2f}".format(pre_molt.std()))

print("Skewness of distribution of Crab size before Molting: {:.4f}".format(skew(pre_molt, axis=0, bias=True)))

print("Kurtosis of distribution of Crab size before Molting: {:.4f}".format(kurtosis(pre_molt, axis=0, bias=True)))
```

Output

```
Minimum size of Crab before Molting: 31.1
Maximum size of Crab before Molting: 155.1
Mean size of Crab before Molting: 129.21
Median size of Crab before Molting: 132.80
Standard Deviation of Crab size before Molting: 15.86
Skewness of distribution of Crab size before Molting: -2.0035
Kurtosis of distribution of Crab size before Molting: 6.7663
```

```
# Quantile Plot for Pre-Molt Data

plt.figure()

stats.probplot(pre_molt, dist="norm", plot=plt)

plt.title('Quantile Plot of distribution of Crab size before Molting')

plt.show()
```

```
# Various Tests of Normality for Pre-Molt Data

pvalue_pre_molt_ad = normal_ad(pre_molt)[1]

pvalue_pre_molt_sw = stats.shapiro(pre_molt).pvalue

pvalue_pre_molt_ks = stats.kstest(pre_molt, "norm").pvalue

pvalue_pre_molt_cv = stats.cramervonmises(pre_molt, "norm").pvalue

print("P-value for Pre Molt data using Anderson-Darling Test: ", pvalue_pre_molt_ad)
```

```
print("P-value for Pre Molt data using Shapiro-Wilk Test: ", pvalue_pre_molt_sw)

print("P-value for Pre Molt data using Kolmogorov-Smirnov Test: ", pvalue_pre_molt_ks)

print("P-value for Pre Molt data using Cramér–von Mises Test: ", pvalue_pre_molt_ks)
```

Output

```
P-value for Pre Molt data using Anderson-Darling Test:  5.84129257221062e-31
P-value for Pre Molt data using Shapiro-Wilk Test:  8.999501305758897e-21
P-value for Pre Molt data using Kolmogorov-Smirnov Test:  0.0
P-value for Pre Molt data using Cramér–von Mises Test:  0.0
```

```
# Descriptive Statistics for Post-Molt Data

print("Minimum size of Crab after Molting:", min(post_molt))

print("Maximum size of Crab after Molting:", max(post_molt))

print("Mean size of Crab after Molting: {:.2f}".format(post_molt.mean()))

print("Median size of Crab after Molting: {:.2f}".format(post_molt.median()))

print("Standard Deviation of Crab size after Molting: {:.2f}".format(post_molt.std()))

print("Skewness of distribution of Crab size after Molting: {:.4f}".format(skew(post_molt, axis=0, bias=True)))

print("Kurtosis of distribution of Crab size after Molting: {:.4f}".format(kurtosis(post_molt, axis=0, bias=True)))
```

Output

```
Minimum size of Crab after Molting: 38.8
Maximum size of Crab after Molting: 166.8
Mean size of Crab after Molting: 143.90
Median size of Crab after Molting: 147.40
Standard Deviation of Crab size after Molting: 14.64
Skewness of distribution of Crab size after Molting: -2.3469
Kurtosis of distribution of Crab size after Molting: 10.1160
```

```
# Quantile Plot for Post-Molt Data

plt.figure()

stats.probplot(post_molt, dist="norm", plot=plt)

plt.title('Quantile Plot of distribution of Crab size after Molting')

plt.show()
```

```
# Various Tests of Normality for Post-Molt Data

pvalue_post_molt_ad = normal_ad(post_molt)[1]

pvalue_post_molt_sw = stats.shapiro(post_molt).pvalue
```

```
pvalue_post_molt_ks = stats.kstest(post_molt, "norm").pvalue
pvalue_post_molt_cv = stats.cramervonmises(post_molt, "norm").pvalue
print("P-value for Pre Molt data using Anderson-Darling Test: ", pvalue_post_molt_ad)
print("P-value for Pre Molt data using Shapiro-Wilk Test: ", pvalue_post_molt_sw)
print("P-value for Pre Molt data using Kolmogorov-Smirnov Test: ", pvalue_post_molt_ks)
print("P-value for Pre Molt data using Cramér–von Mises Test: ", pvalue_post_molt_ks)
```

Output

```
P-value for Pre Molt data using Anderson-Darling Test:  0.0
P-value for Pre Molt data using Shapiro-Wilk Test:  4.862027690351223e-22
P-value for Pre Molt data using Kolmogorov-Smirnov Test:  0.0
P-value for Pre Molt data using Cramér–von Mises Test:  0.0
```

```
# Plotting the Smooth Approximation of Crab Data (Pre and Post Molt)
plt.figure(figsize=(10, 6))
sns.kdeplot(pre_molt, fill=True, label='Distribution of Crab size before Molting')
sns.kdeplot(post_molt, fill=True, label='Distribution of Crab size after Molting')
plt.xlabel('Value')
plt.ylabel('Density')
plt.title('Comparison of the distributions of the two data sets')
plt.legend()
plt.show()
```

```
# Plotting Scatter Plot of Crab Data (Pre and Post Molt)
sns.scatterplot(x=post_molt, y=pre_molt)
plt.xlabel('Post-molt')
plt.ylabel('Pre-molt')
plt.title('Scatter Plot - Pre-molt VS Post-molt')
plt.show()
```

```
# Linear Regression to predict Pre-Molt data using Post-Molt data
# Reshape the data
X = np.array(post_molt).reshape(-1, 1)
Y = np.array(pre_molt)
```

```
# Fit linear regression model
model = LinearRegression()
model.fit(X, Y)
# Predict pre-molt sizes
predicted_pre_molt = model.predict(X)
r_squared = model.score(X, Y)
r_squared
```

Output

Pearson’s R Squared: 0.9808325947886156

```
# Residual Plot and Normality Tests on Residuals
sns.kdeplot(residuals, fill=True)
plt.xlabel('Value')
plt.ylabel('Density')
plt.title('Smooth approximation to histogram of Residuals')
plt.show()
print("Skewness of Residuals: {:.4f}".format(skew(residuals, axis=0, bias=True)))
print("Kurtosis of Residuals: {:.4f}".format(kurtosis(residuals, axis=0, bias=True)))
plt.figure()
stats.probplot(residuals, dist="norm", plot=plt)
plt.title('Quantile Plot of Residuals')
plt.show()
# Various tests for Normality
pvalue_residuals_ad = normal_ad(residuals)[1]
pvalue_residuals_sw = stats.shapiro(residuals).pvalue
pvalue_residuals_ks = stats.kstest(residuals, "norm").pvalue
pvalue_residuals_cv = stats.cramervonmises(residuals, "norm").pvalue
print("P-value for Residuals data using Anderson-Darling Test: ", pvalue_residuals_ad)
print("P-value for Residuals data using Shapiro-Wilk Test: ", pvalue_residuals_sw)
print("P-value for Residuals data using Kolmogorov-Smirnov Test: ", pvalue_residuals_ks)
print("P-value for Residuals data using Cramér–von Mises Test: ", pvalue_residuals_ks)
```

Output

Skewness of Residuals: 0.8455
Kurtosis of Residuals: 5.3787
P-value for Residuals data using Anderson-Darling Test: 0.00011506617586914685
P-value for Residuals data using Shapiro-Wilk Test: 6.358130168193643e-12
P-value for Residuals data using Kolmogorov-Smirnov Test: 2.5554400301036748e-12
P-value for Residuals data using Cramér–von Mises Test: 2.5554400301036748e-12

```
# Heteroskedasticity
plt.scatter(predicted_pre_molt, residuals)
plt.xlabel('Predicted Pre-molt Sizes')
plt.ylabel('Residuals')
plt.title('Residual Plot')
plt.axhline(y=0, color='r', linestyle='-')
plt.show()
# Add constant term to X if needed
X_ = add_constant(X)  # Make sure X has at least two columns
# Perform Breusch-Pagan test
bp_test = het_breuschpagan(residuals, X_)
labels = ['LM Statistic', 'LM-Test p-value', 'F-Statistic', 'F-Test p-value']
for label, result in zip(labels, bp_test):
    print(f"{label}: {result}")
# Perform White test
white_test = het_white(residuals, X_)
labels = ['LM Statistic', 'LM-Test p-value', 'F-Statistic', 'F-Test p-value']
for label, result in zip(labels, white_test):
    print(f"{label}: {result}")
```

Output

LM Statistic (Breusch-Pagan): 89.38639408257598
LM-Test p-value (Breusch-Pagan): 3.2476245276302648e-21
F-Statistic (Breusch-Pagan): 109.80164993891431
F-Test p-value (Breusch-Pagan): 3.1294545123620833e-23
LM Statistic (White): 238.22876801746258
LM-Test p-value (White): 1.8590039377747353e-52
F-Statistic (White): 238.97143214041003
F-Test p-value (White): 2.767618354610456e-72

```
# Analysis of potentially different linear models for Lab versus Field crabs
```

```
# Field Crabs
data_field = data[data["Location"] == "Field"]
data_field = data_field.drop("Location", axis = 1)
pre_molt_field = data_field['Pre-molt']
post_molt_field = data_field['Post-molt']
# Reshape the data
X_field = np.array(post_molt_field).reshape(-1, 1)
Y_field = np.array(pre_molt_field)
# Fit linear regression model
model_field = LinearRegression()
model_field.fit(X_field, Y_field)
# Predict pre-molt sizes
predicted_pre_molt_field = model_field.predict(X_field)
r_squared = model_field.score(X_field, Y_field)
r_squared
# Lab Crabs
data_lab = data[data["Location"] == "Lab"]
data_lab = data_lab.drop("Location", axis = 1)
pre_molt_lab = data_lab['Pre-molt']
post_molt_lab = data_lab['Post-molt']
# Reshape the data
X_lab = np.array(post_molt_lab).reshape(-1, 1)
Y_lab = np.array(pre_molt_lab)
# Fit linear regression model
model_lab = LinearRegression()
model_lab.fit(X_lab, Y_lab)
# Predict pre-molt sizes
predicted_pre_molt_lab = model_lab.predict(X_lab)
r_squared = model_lab.score(X_lab, Y_lab)
r_squared
```

 # Utilizing cross-validation to estimate prediction accuracy of linear model

```
# Perform k-fold cross-validation
k = 5  # Number of folds
cv_scores = cross_val_score(model, X, Y, cv=k, scoring='r2')
for i, score in enumerate(cv_scores, start=1):
    print(f"R-squared Score (Fold {i}) = {score}")
print("Mean R-squared Score = ", cv_scores.mean())
```

Output

```
R-squared Score (Fold 1) = 0.9347890038239732
R-squared Score (Fold 2) = 0.9881082547655841
R-squared Score (Fold 3) = 0.9605659682730662
R-squared Score (Fold 4) = 0.9505222363907887
R-squared Score (Fold 5) = 0.9493029121354887
Mean R-squared Score =  0.95665767507778
```

Contributions:

Sindhuja Baikadi - 02128756: Worked on the Issues, Findings, Discussion, Method, and Results sections. Also self-plotted the graphs to analyze the data using the various methods discussed in the report.

Veda Sahaja Bandi - 02105111: Outlining key observations and insights, worked on the coding portion of the project to implement necessary functionalities and features.