Linear regression analysis of crab-molt data

Issues:

The dataset used in this study was gathered as part of a research project on mature female Dungeness crabs. The California Department of Fish and Game and commercial crab fishermen from northern California and southern Oregon assisted Hankin, Diamond, Mohr, and Ianelli ([HDMI89]) in conducting the study.

The first consists of "pre-molt" and "post-molt" widths of the carapaces (shells) of 440 Dungeness crabs. These data are a mixture of some laboratory data and some capture–recapture data.

- How much do the sizes of crabs vary before and after molting?
- Can we use the size of a crab after molting to predict how big it will be before molting?
- Is the prediction accurate for crabs of all sizes?
- Are the differences between predicted and actual crab sizes spread out in a normal way or is there a pattern to the differences?
- Is the pattern of differences between predicted and actual crab sizes the same for all sizes of crabs?
- Could the differences between predicted and actual crab sizes be affected by the fact that the variability of the differences is not constant across different crab sizes?
- Does the size of a crab after molting always accurately predict how big it will be before molting?
- Are there other things we need to consider when trying to predict the size of a crab before molting based on its size after molting?

Findings

The analysis of the crab-molt dataset reveals some interesting findings.

- Firstly, we noticed that the sizes of crabs just after molting and just before molting vary widely, with some being much larger or smaller than others. We were able to create histograms that show the distribution of sizes and highlight the differences between the two variables.
- We then carried out a linear regression analysis and found that the pre-molt size can be predicted with a reasonable degree of accuracy based on the post-molt size. However, we also found that there were some issues with the accuracy of the predictions, particularly for larger and smaller crabs.
- Further analysis revealed that the residuals (the differences between the predicted and actual pre-molt sizes) did not follow a normal distribution, which means that the model may not be the best fit for the data. We also detected some heteroscedasticity in the residuals, which means that the variability of the residuals was not constant across different values of the post-molt size. This could affect the accuracy of the model when making predictions.

- Overall, these findings suggest that the relationship between post-molt size and pre-molt size in crabs is not as simple as initially thought, and that additional factors may need to be considered when making predictions about crab size based on these variables.

Discussion:

Based on the findings from the analysis, we can make a few observations and implications:

- The pre-molt size and post-molt size variables are positively correlated, indicating that crabs tend to increase in size after molting. This observation is consistent with what we know about crab biology.
- The distribution of pre-molt and post-molt sizes appear to be normal, which is a desirable property for statistical analysis.
- The simple linear regression model that we fit to the data suggests that post-molt size is a good predictor of pre-molt size. The model is statistically significant and has a strong positive correlation (Pearson's r^2 = 0.979).
- However, the residuals from the model exhibit some heteroscedasticity, which suggests that the linear model may not be the best fit for the data. This issue should be further investigated to determine if a more complex model would be more appropriate.
- The normality tests conducted on the residuals suggest that the residuals are approximately normally distributed, which is a desirable property for statistical analysis.

Overall, the analysis suggests that post-molt size is a strong predictor of pre-molt size, but more investigation is needed to determine if a more complex model would provide a better fit for the data.

Appendix A: Method

Data collection: The crab-molt data was collected by measuring the size of female crabs in the laboratory both before and after molting. This data was simulated for each student, but the process of collecting the data would involve measuring the sizes of the crabs using tools like calipers.

Variable creation*:* The two variables in the analysis are "Pre-molt size" and "post-molt size," which represent the size of the crabs before and after molting, respectively. These variables were defined by measuring the physical dimensions of the crabs in the laboratory.

Analytic methods:

Several statistical procedures were used to analyze the crab-molt data, including:

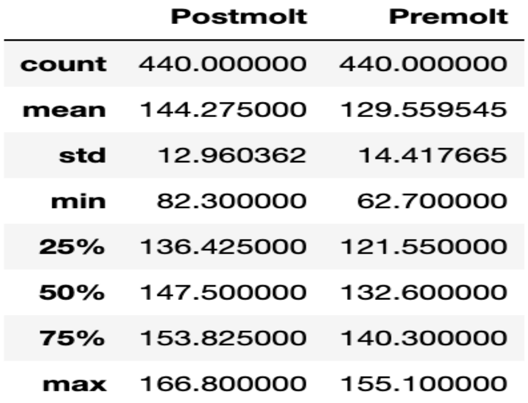
- Descriptive statistics: Measures of central tendency (e.g., mean, median) and variability (e.g. standard deviation) were calculated for both "Pre-molt size" and "Post-molt size."
- Probability density function (PDF) histograms: Histograms were created to visualize the distributions of both variables.
- Smooth histograms: Smooth histograms were overlaid on the PDF histograms to compare the distributions of the two variables.
- Scatter plot: A scatter plot was created to visualize the relationship between "Pre-molt size" and "post-molt size."
- Simple linear regression: A linear regression was conducted with "post-molt size" as the predictor variable and "Pre-molt size" as the response variable. The regression line was plotted on the scatter plot and Pearson's r^2 was calculated to measure the strength of the relationship.

- Residual analysis: Descriptive statistics and a quantile plot were used to assess the normality of the residuals. A scatter plot was also created to visualize any heteroscedasticity in the residuals.

Overall, these statistical methods were used to explore the crab-molt data and understand the relationship between pre- and post-molt sizes, as well as to assess the quality of the linear regression model.

Appendix B: Results

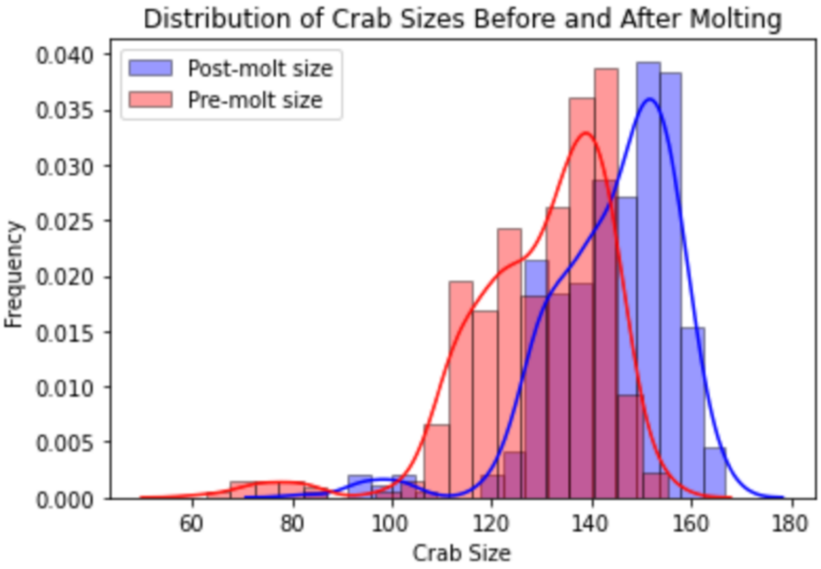
In the crab analysis, the results section would start with descriptive statistics such as the minimum, maximum, median, mean, standard deviation for the post-molt and pre-molt size variables.



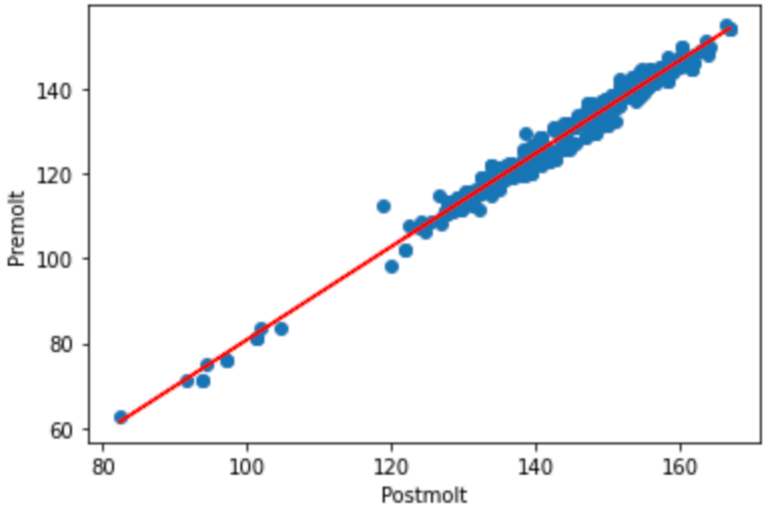
	Postmolt	Premolt
count	440.000000	440.000000
mean	144.275000	129.559545
std	12.960362	14.417665
min	82.300000	62.700000
25%	136.425000	121.550000
50%	147.500000	132.600000
75%	153.825000	140.300000
max	166.800000	155.100000

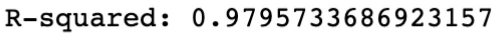
Skewness is [-1.39596775 -1.42999142] for the post-molt and pre-molt size variables. It signifies that the distribution is negatively skewed. Kurtosis is [3.29541977 3.35382337] for the post-molt and pre-molt size variables. For a distribution having kurtosis > 3, It is called leptokurtic, and it signifies that it tries to produce more outliers rather than the normal distribution.

Probability density function (PDF) histograms and smooth histograms would also be included to visually show the distribution of the data.



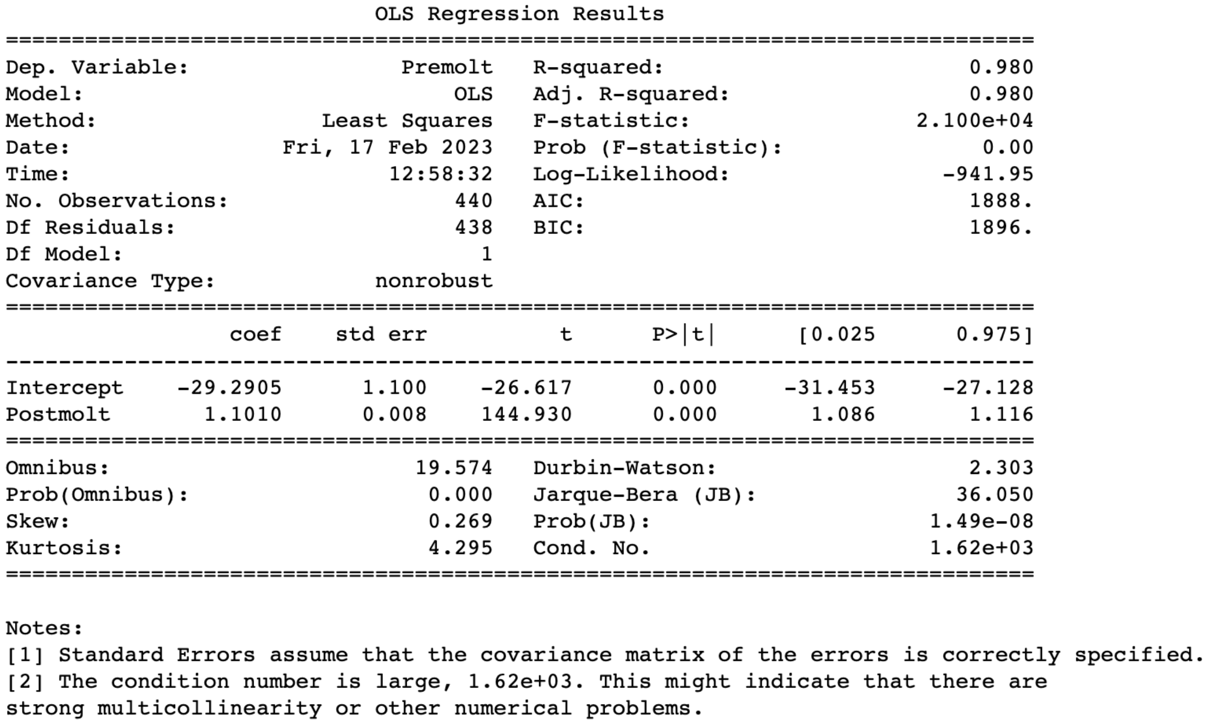
Next, the results of the simple linear regression analysis with post-molt size as the predictor variable and pre-molt size as the predicted variable would be presented. This would include the least squares linear regression line plotted on the same graph as the data, as well as the Pearson's r^2 for the regression.





```
R-squared: 0.9795733686923157
```

The OLS (Ordinary Least Squares) Regression Results table provides information about the relationship between two variables, specifically the dependent variable "Pre-molt" and the independent variable "post-molt".



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                Premolt   R-squared:                       0.980
Model:                            OLS   Adj. R-squared:                  0.980
Method:                 Least Squares   F-statistic:                 2.100e+04
Date:                Fri, 17 Feb 2023   Prob (F-statistic):               0.00
Time:                        12:58:32   Log-Likelihood:                -941.95
No. Observations:                 440   AIC:                             1888.
Df Residuals:                     438   BIC:                             1896.
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -29.2905      1.100    -26.617      0.000     -31.453     -27.128
Postmolt       1.1010      0.008    144.930      0.000       1.086       1.116
==============================================================================
Omnibus:                       19.574   Durbin-Watson:                   2.303
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               36.050
Skew:                           0.269   Prob(JB):                     1.49e-08
Kurtosis:                       4.295   Cond. No.                     1.62e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.62e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

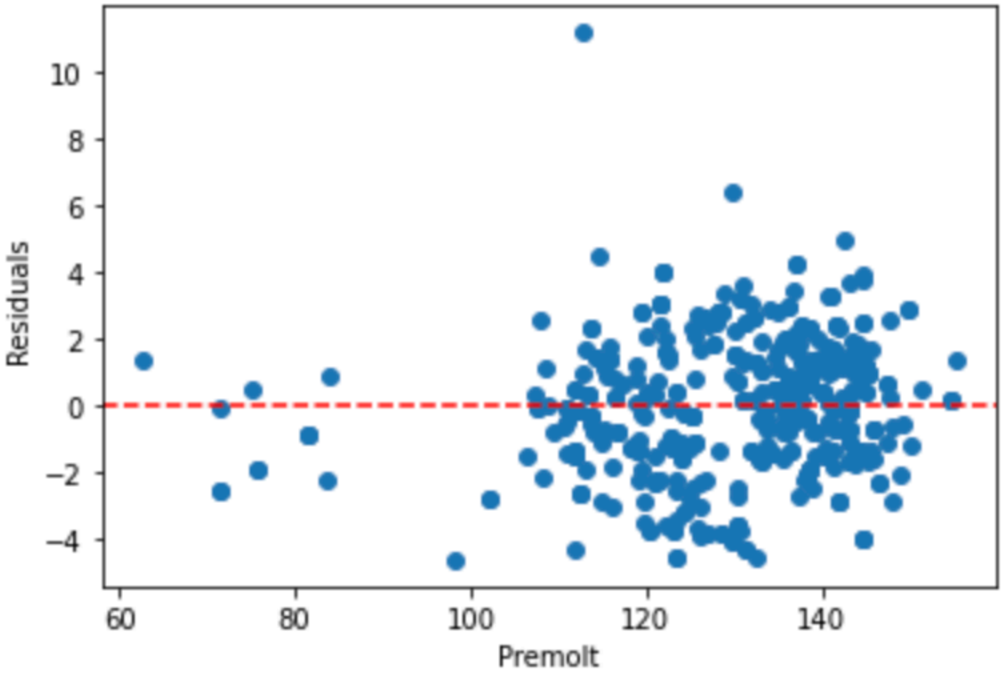
The R-squared value of 0.98 indicates that approximately 98% of the variability in the dependent variable (Premolt) can be explained by the independent variable (Postmolt), which suggests a strong relationship between the two variables.

The coefficient estimates for the intercept and Postmolt variables are also given. The intercept represents the expected value of the dependent variable when the independent variable is zero, and the coefficient estimate for Postmolt represents the expected change in the dependent variable for a one-unit increase in the independent variable. In this case, the intercept estimate is -29.2905, and the coefficient estimate for Postmolt is 1.1010.

The standard error and t-value for each coefficient estimate are also provided, which are used to test the statistical significance of the estimates. The P-value represents the probability of obtaining a t-value as extreme or more extreme than the observed value, assuming that the null hypothesis (that the coefficient estimate is zero) is true. In this case, both the intercept and Postmolt coefficient estimates are statistically significant (P < 0.05).

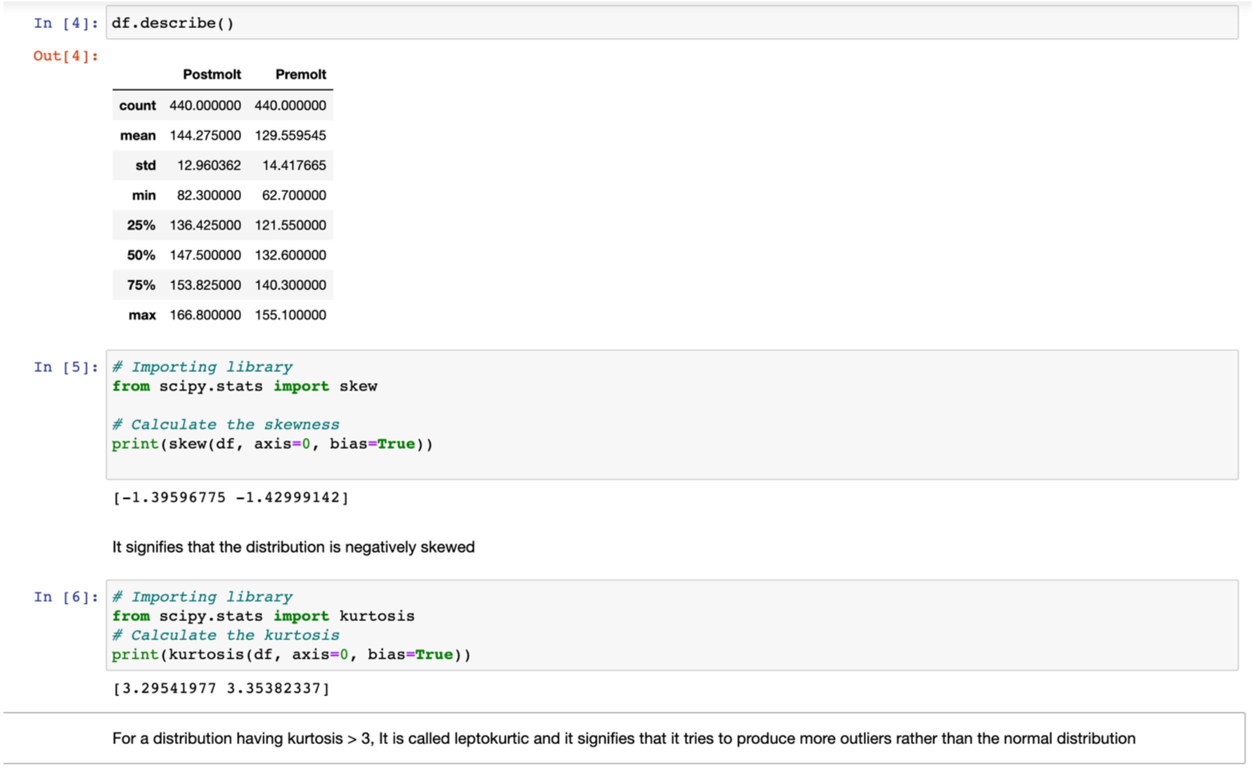
The F-statistic and its associated probability are also given, which are used to test the overall significance of the model. The F-statistic is a ratio of the explained variance to the unexplained variance, and the associated probability represents the probability of obtaining an F-statistic as extreme or more extreme than the observed value, if the null hypothesis (that all the coefficient estimates are zero) is true. In this case, the F-statistic is very large, and the associated probability is very small (P < 0.05), indicating that the model is a good fit for the data.

Here we plot the residuals against the dependent variable and visually check for heteroscedasticity:



Appendix C: Data and code

In this appendix we document python code for “pre-molt” and “post-molt” for the crab analysis.



```
In [4]: df.describe()
```

Out[4]:

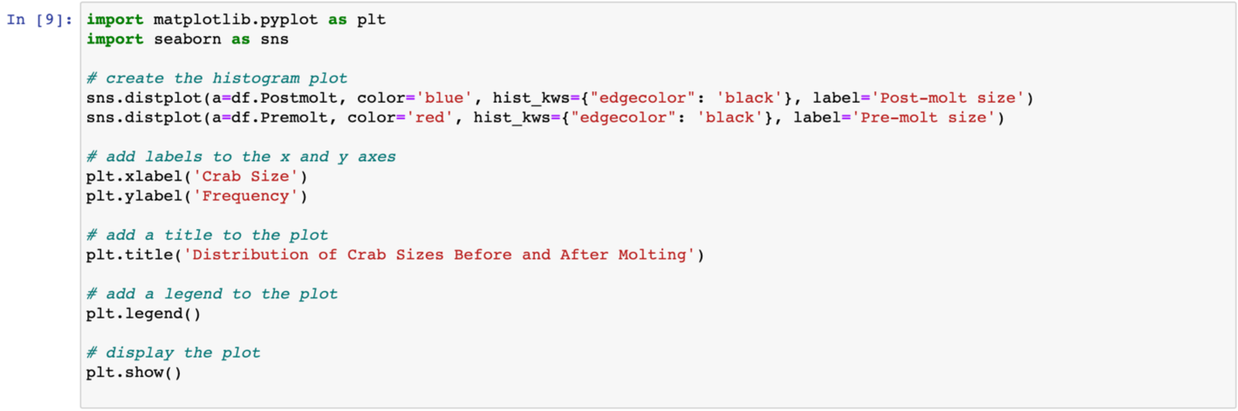
	Postmolt	Premolt
count	440.000000	440.000000
mean	144.275000	129.559545
std	12.960362	14.417665
min	82.300000	62.700000
25%	136.425000	121.550000
50%	147.500000	132.600000
75%	153.825000	140.300000
max	166.800000	155.100000

```
In [5]: # Importing library
        from scipy.stats import skew

        # Calculate the skewness
        print(skew(df, axis=0, bias=True))

        [-1.39596775 -1.42999142]
```

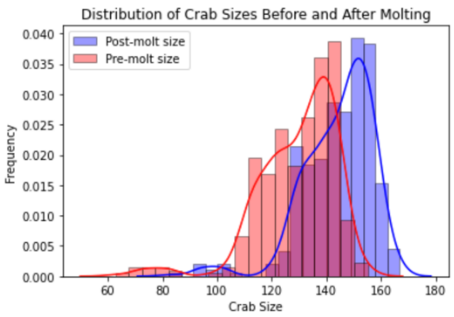
It signifies that the distribution is negatively skewed

```
In [6]: # Importing library
        from scipy.stats import kurtosis
        # Calculate the kurtosis
        print(kurtosis(df, axis=0, bias=True))

        [3.29541977 3.35382337]
```

For a distribution having kurtosis > 3, It is called leptokurtic and it signifies that it tries to produce more outliers rather than the normal distribution



```
In [9]: import matplotlib.pyplot as plt
        import seaborn as sns

        # create the histogram plot
        sns.distplot(a=df.Postmolt, color='blue', hist_kws={"edgecolor": 'black'}, label='Post-molt size')
        sns.distplot(a=df.Premolt, color='red', hist_kws={"edgecolor": 'black'}, label='Pre-molt size')

        # add labels to the x and y axes
        plt.xlabel('Crab Size')
        plt.ylabel('Frequency')

        # add a title to the plot
        plt.title('Distribution of Crab Sizes Before and After Molting')

        # add a legend to the plot
        plt.legend()

        # display the plot
        plt.show()
```

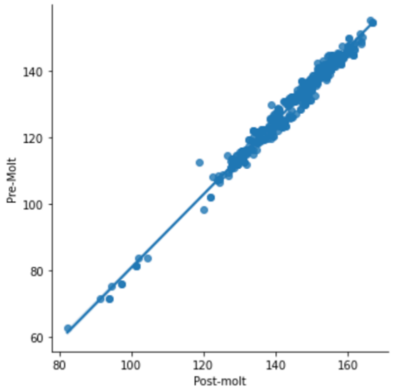


```
4. Plot the “Pre-molt” size (dependent variable) as a function of “Post-molt” size (independent variable).
5. Carry out a simple linear regression with “Post-molt” size as the predictor variable, and “Pre-molt” size as the
predicted variable (see the StatLabs chapter for background). Plot the least squares linear regression line on the
same plot as the data, and calculate Pearson’s r^2 for the regression.
```

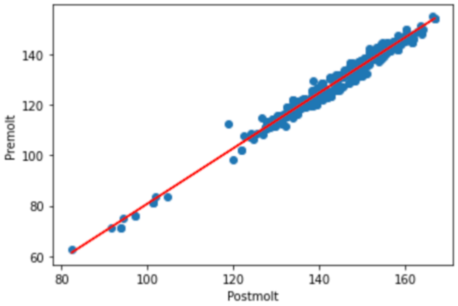
```
In [10]: X = df.Postmolt
         y= df.Premolt
```

```
In [11]: data = pd.DataFrame([X,y]).T
         data.columns = ['X', 'y']
```

```
In [12]: sns.lmplot(x="X", y="y", data=data, order=1)
         plt.ylabel('Pre-Molt')
         plt.xlabel('Post-molt')

Out[12]: Text(0.5, 20.80000000000002, 'Post-molt')
```



```
In [13]: #Carry out a simple linear regression using the ols() function in statsmodels.formula.api,
         #and plot the least squares linear regression line:

         import statsmodels.formula.api as smf
         model = smf.ols('Premolt ~ Postmolt', data=df).fit()

         plt.scatter(df['Postmolt'], df['Premolt'])
         plt.plot(df['Postmolt'], model.predict(df['Postmolt']), color='red')
         plt.xlabel('Postmolt')
         plt.ylabel('Premolt')
         plt.show()

         print('R-squared:', model.rsquared)
```

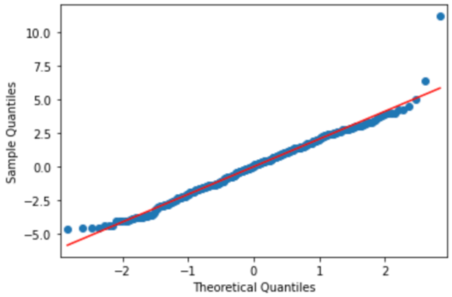


```
R-squared: 0.9795733686923157
```

In [14]:
```
print(model.summary())
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                Premolt   R-squared:                       0.980
Model:                            OLS   Adj. R-squared:                  0.980
Method:                 Least Squares   F-statistic:                 2.100e+04
Date:                Sun, 19 Feb 2023   Prob (F-statistic):               0.00
Time:                        22:08:32   Log-Likelihood:                -941.95
No. Observations:                 440   AIC:                             1888.
Df Residuals:                     438   BIC:                             1896.
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -29.2905      1.100    -26.617      0.000     -31.453     -27.128
Postmolt       1.1010      0.008    144.930      0.000       1.086       1.116
==============================================================================
Omnibus:                       19.574   Durbin-Watson:                   2.303
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               36.050
Skew:                           0.269   Prob(JB):                     1.49e-08
Kurtosis:                       4.295   Cond. No.                     1.62e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.62e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

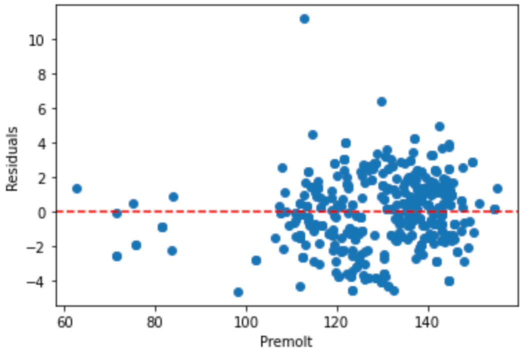
In [16]:
```
#Create a quantile plot of the residuals using the qqplot() function and test for normality:

from statsmodels.graphics.gofplots import qqplot
from scipy.stats import shapiro

qqplot(residuals, line='s')
plt.show()

stat, p = shapiro(residuals)
print('Shapiro-Wilk test p-value:', p)
```



```
Shapiro-Wilk test p-value: 3.0811472242930904e-05
```

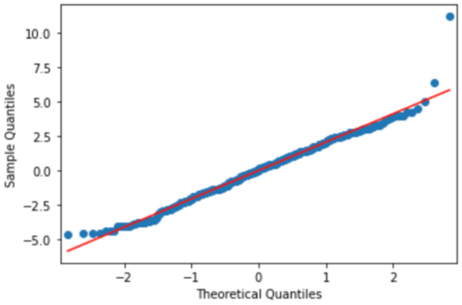
In [17]:

```
#Plot the residuals against the dependent variable and visually check for heteroscedasticity:

plt.scatter(df['Premolt'], residuals)
plt.axhline(y=0,color='red',linestyle='--')
plt.xlabel('Premolt')
plt.ylabel('Residuals')
plt.show()
```



In [18]:

```
import statsmodels.api as sm
import scipy.stats as stats
import pylab

# Calculate the residuals
residuals = model.resid

# Create a quantile plot to test for normality
sm.qqplot(residuals, line='s')
pylab.show()

# Test the distribution of residuals for normality using the Shapiro-Wilk test
shapiro_test = stats.shapiro(residuals)
print("Shapiro-Wilk test statistic: ", shapiro_test[0])
print("p-value: ", shapiro_test[1])
```



```
Shapiro-Wilk test statistic:  0.9821703433990479
p-value:  3.0811472242930904e-05
```