MTH 522: Advanced Mathematical Statistics

Predicting Crab Sizes Before and After Molting Using Linear Regression

By: Dhyey Doshi

02/19/2023

Issues

The purpose of this report is to determine a relationship between pre and post molt lengths of crabs. The dataset used to examine the relationship between the pre-molt and post-molt sizes and sum up results graphically and numerically. A linear model, statistical analysis, a quantile plot and histogram were created to help visualize the distribution of the data. To confirm that the model is accurate we investigate the residuals and determine how close they are to a normal distribution telling us about how the model can predict.

Findings

After finding solution to above issues, we can conclude that the model performs great with an r squared of 0.982 or 98 percent which represents that the model performed well. We can also conclude that the residuals follow distribution saying that the predictions are quite accurate with low error percentage amongst the data.

Discussion

Data analysis was performed using library statsmodel in python. Using the data, a linear regression model was constructed. The data analysis contains statistical analysis of pre-molt and post-molt shell sizes for crabs. The Statistical analysis consisted of the mean, median, standard deviation, kurtosis and skewness. Once this was achieved, the residuals was created histogram and a quantile plot to help visualise the distribution of the data. Using the relationship between pre-molt and post-molt, we can create a method for prediction pre-molt size from post-molt.

Appendix A: Methods

The dataset is a mixture of data from several other data sets, containing both laboratory data and capture recapture data. The data variables are pre molt size, post molt size. A linear regression model was built using the data. The statistical analysis of pre- and post-molt crab shell diameters is included in the data analysis. The mean, median, standard deviation, kurtosis, and skewness were the components of the statistical analysis. The minimum and maximum values were included in the statistical analysis along with the prior data. Since the pre-molt sizes cannot be determined because the crabs were

collected from the ocean after they molted, the linear model enables one to predict the pre-molt sizes.

Appendix B: Results

Descriptive Statistics on the both variable.

	Post-molt	Pre-molt
count	416.000000	416.000000
mean	142.722356	128.110337
std	15.832564	16.948336
min	38.800000	31.100000
25%	136.425000	120.875000
50%	145.400000	131.200000
75%	152.900000	139.525000
max	166.500000	155.100000

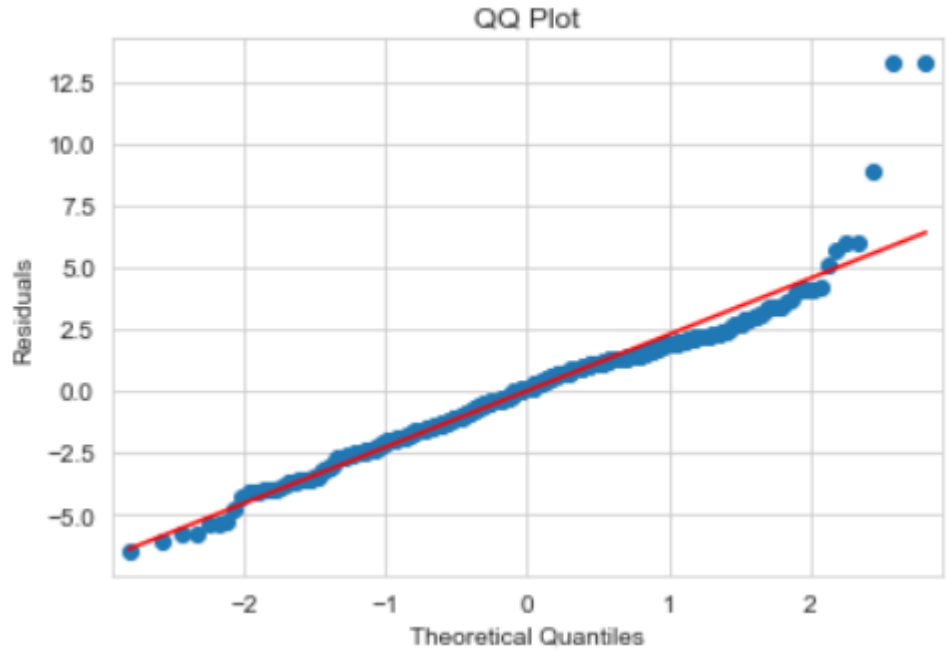
Displaying the summary statistics for linear model.

Dep. Variable:	Pre-molt	**R-squared:**	0.982
Model:	OLS	**Adj. R-squared:**	0.982
Method:	Least Squares	**F-statistic:**	2.263e+04
Date:	Sun, 19 Feb 2023	**Prob (F-statistic):**	0.00
Time:	22:15:35	**Log-Likelihood:**	-931.08
No. Observations:	416	**AIC:**	1866.
Df Residuals:	414	**BIC:**	1874.
Df Model:	1		
Covariance Type:	nonrobust		

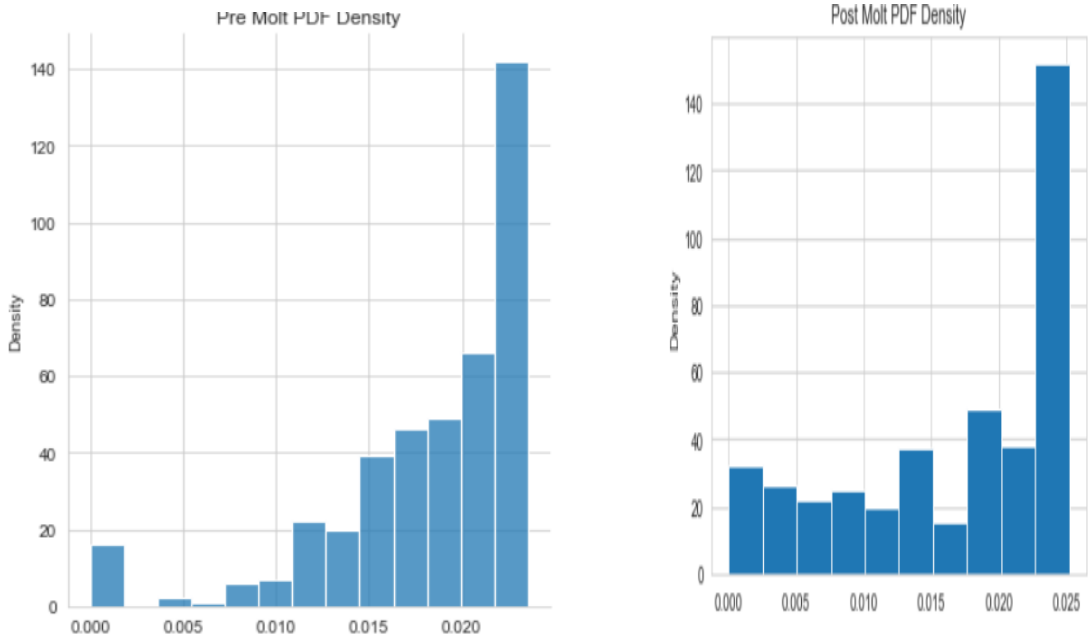
	coef	std err	t	P>|t|	[0.025	0.975]
const	-23.2918	1.013	-23.004	0.000	-25.282	-21.301
Post-molt	1.0608	0.007	150.447	0.000	1.047	1.075

Omnibus:	93.790	**Durbin-Watson:**	1.976
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	526.876
Skew:	0.825	**Prob(JB):**	3.89e-115
Kurtosis:	8.261	**Cond. No.**	1.30e+03

Below is a quantile plot can be found of the regression model residuals, this is testing whether the residuals follow a normal distribution



Probability density function Histogram of both variable



Appendix C: Code

```
import pandas as pd
import scipy.stats as sts
import statsmodels.api as sm
from statsmodels.formula.api import ols
import matplotlib.pyplot as plt
import seaborn as sns


df = pd.read_excel('E:\\Study\\MTH\\crab_molt_data_doshi_dhyey.xlsx')


df.describe()

post_molt = df['Post-molt'] #independent variable
pre_molt = df['Pre-molt'] #dependent variable

# Describe each variable “Post-molt” size and “Pre-molt” size, including: minimum, maximum, median, mean, standard deviation, skewness, and kurtosis.


print("Minimum")
df.min()
print("Maximum")
df.max()
print("Mean")
df.mean()
print("Median")
df.median()
print("Standard Deviation")
df.std()
print("Skewness")
df.skew()
print("Kurtosis")
df.kurt()

# Make a probability density function (PDF) histogram of each variable.


df_mean_pre = df.mean()['Pre-molt']
df_std_pre = df.std()['Pre-molt']
pdf_pre = sts.norm.pdf(df['Pre-molt'], df_mean_pre, df_std_pre)
sns.displot(pdf_pre)
plt.ylabel('Density')
```

```
plt.title('Pre Molt PDF Density')
plt.show()


df_mean_post = df.mean()['Post-molt']
df_std_post = df.std()['Post-molt']
pdf_post = sts.norm.pdf(df['Pre-molt'], df_mean_post, df_std_post)
plt.hist(pdf_post)
plt.ylabel('Density')
plt.title('Post Molt PDF Density')
plt.show()

sts.pearsonr(pre_molt, post_molt)

#Linear Regression

post_molt = sm.add_constant(post_molt)
linear = sm.OLS(pre_molt, post_molt).fit()
linear.resid
linear.summary()

linear_resid = linear.resid
linear_predict = linear.predict(post_molt)
plt.scatter(x=linear_predict, y=linear_resid)
plt.xlabel('Predicted')
plt.ylabel('Residuals')
plt.axhline(y=0)
plt.show()

sns.displot(linear_resid, kde=False, bins=18)
plt.ylabel('Density')
plt.xlabel('Residuals')
plt.title('Histogram')
plt.show()

#Quantile Plot

sm.qqplot(linear_resid, line='s')
plt.title('QQ Plot')
plt.ylabel('Residuals')
plt.show()

#Shapiro-Walks Test
```

```
a, b =sts.shapiro(linear_resid)
print('Statistics', a, 'p-value', b)
```

References

https://bpb-us-w2.wpmucdn.com/sites.umassd.edu/dist/7/886/files/2020/02/report3.pdf